Lecture Notes in Computer Science 2875

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Emile Aarts   René Collier
Evert van Loenen   Boris de Ruyter (Eds.)

Ambient Intelligence

First European Symposium, EUSAI 2003
Veldhoven, The Netherlands, November 3-4, 2003
Proceedings

Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany
Juris Hartmanis, Cornell University, NY, USA
Jan van Leeuwen, Utrecht University, The Netherlands

Volume Editors

Emile Aarts
René Collier
Evert van Loenen
Boris de Ruyter
Philips Research Laboratories
Professor Holstlaan 4, 5656 AA Eindhoven, The Netherlands
E-mail: {emile.aarts,rene.collier,evert.van.loenen,boris.de.ruyter}@philips.com

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

CR Subject Classification (1998): H.5, H.4, D.2, C.3, C.2, I.2, J.5, K.4, K.8

ISSN 0302-9743
ISBN 3-540-20418-0 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by PTP-Berlin GmbH
Printed on acid-free paper SPIN: 10967492 06/3142 5 4 3 2 1 0

The concept of ambient intelligence refers to a world in which people are surr
with electronic environments that are sensitive and responsive to people
specifically, this means that electronic devices such as laptops, PCs, hi-
television sets, DVD players, and others are bound to disappear into p
backgrounds by being integrated into the environment, and having their funct
extended to provide ubiquitous communication, information, and enterta
through unobtrusive and natural interaction with users. Ambient intelligence o
great expectations for novel improved and enhanced lifestyles, supporting p
expression, healthcare, and productivity, as well as the development of new
and communities.

Since its introduction in the late 1990s the concept has attracted much atte
the scientific and industrial worlds. Substantial progress has been made
development of technologies that can support the realization of the concept, a
prototypes have been realized that allow the study of the user experience
obtained within ambient intelligent environments. So this means that first resu
become available and need to be communicated, which calls for a stage that
people to present the progress they have made. EUSAI intends to provide this s

EUSAI is the first international symposium that is entirely devoted
communication of the progress that has been achieved in the development of a
intelligence. EUSAI aims at presenting recent research results on all aspect
theory, design and application of ambient intelligence. At the same time,
should provide a research forum for the exchange of ideas on recent develo
within the community of ambient intelligence researchers. The symposiu
organized as a two-day event and the program contained a manifold of
speeches, contributed papers, demonstrations, and movies. The progra
structured along the lines of the following four major themes:

- ubiquitous computing,
- context awareness,
- intelligence, and
- natural interaction.

These themes are the main components of ambient intelligence, and their s
integration is what is generally believed will provide a novel enhanced exper
users in ambient intelligent environments.

The EUSAI proceedings contain all contributed papers and some of the t
and they are organized along the lines of four tracks following the major
mentioned above. The contributed papers were selected from the set of su
papers in a carefully conducted review process. We strongly believe that this
process resulted in a final set of contributed papers of high technical qua
innovative value. We are greatly indebted to the members of the program co
for their excellent work in selecting the best papers and compiling the final p
We also greatly appreciate the effort of all authors in submitting papers to EU
preparing the final versions of their manuscripts.

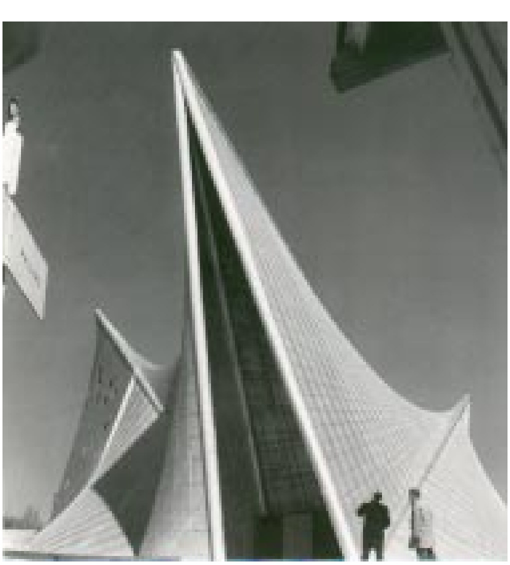
ITEA Ambience project. Many people were involved in this joint effort and greatly indebted to them for their valuable contribution to the organization of Special thanks in this respect go to Ad de Beer for taking care of th arrangements and to Maurice Groten for guaranteeing the financial budget.

EUSAI has succeeded in bringing together a wealth of information on the r progress in ambient intelligence, and we are confident that these proceedin contribute to the realization of the truly great concept that ambient intel provides.

Eindhoven,
August 2003

Emile Aarts
Rene Collie
Evert van L
Boris de Ru

On the occasion of the 1958 World's Fair in Brussels, Le Corbusier designed Philips company a pavilion (see photograph below) that was later referred to neglected building by Le Corbusier, since it was dismantled after the fair. visually compelling book, Treib [1996] brought this object back to li positioned it as an ephemeral structure that exhibited a landmark mul production. The nearly two million visitors to the pavilion were exposed to a show rather than to the typical display of consumer products. The show c *poème électronique* was a dazzling demonstration of ambient color, sound, vo images. It was orchestrated into a cohesive 480-seconds program by Le Corbus his colleagues, including the composer Edgard Varèse, whose distinguished *poème électronique* was composed for this occasion and gave the show its nam



The Philips Pavilion, World Exposition, Brussels, 1958

According to Treib the project has great significance as an early vision on a media, which in his wording can be expressed as follows:

"… the Philips project … can be viewed as a pioneering quest into the pro of modern art, or even as a prototype of virtual reality."

Treib also argues that the project shows how the gap between architecture, and marketing can be bridged, thus providing the means to bring ambient expe to people's homes. Consequently, ambient intelligence can be viewed as *le n poème électronique.*

Treib, M. [1996], *Space Calculated in Seconds: the Philips Pavilion, Le Co Edgard Varèse,* Princeton University Press, Princeton, New Jersey.

Emile Aarts, The Netherlands
Liliana Ardissono, Italy
Niels Ole Bernsen, Denmark
Jan Biemond, The Netherlands
Peter Boda, Finland
Loe Boves, The Netherlands
Jan Burgmeijer, The Netherlands
Louis Chevallier, France
Henrik Christensen, Sweden
Rene Collier, The Netherlands
Yves Demazeau, France
Laila Dubkjaer, Denmark
Rigobert Foka, France
Paul ten Hagen, The Netherlands
Juha Kaario, Finland
David Keyson, The Netherlands
Kristof van Laerhoven, UK
Jan van Leeuwen, The Netherlands
Evert van Loenen, The Netherlands
Paul Lukowicz, Switzerland
Wolfgang Minker, Germany
Bart Nauwelaers, Belgium
Johan Plomp, Finland
Thomas Portele, Germany
Gilles Privat, France
Cliff Randell, UK
Fiorella de Rosis, Italy
Boris de Ruyter, The Netherlands
Egbert Jan Sol, Sweden
Kostas Stathis, UK
Olivero Stock, Italy
Esa Tuulari, Finland
Wim Verhaegh, The Netherlands

Table of Contents

Track 1. Ubiquitous Computing

Track 2. Context Awareness

Track 3. Intelligence

Track 4. Natural Interaction

Find a Meeting

Paul Simons[1], Yacine Ghamri-Doudane[2,] Sidi-Mohammed Senouci[2],
Farid Bachiri[3], Anelise Munaretto[2], Jouko Kaasila[4], Gwendal Le Grand[5], and
Isabelle Demeure[5]

[1] Philips Research Laboratories,
Redhill, United Kingdom
paul.simons@philips.com

[2] LIP6,
Paris, France
{yacine.ghamri,sidi-sohammed.senouci,
anelise.munaretto}@lip6.fr

[3] Thales,
Paris, France
Farid.bachiri@fr.thalesgroup.com

[4] CCC,
Oulu, Finland
Jouko.kaasila@ccc.fi

[5] ENST,
Paris, France
{Gwendal.Legrand, Isabelle.Demeure}@enst.fr

Abstract. This paper presents a highly integrated system for automatically registering, guiding and assisting delegates of a conference. The delegate uses an iPAQ to communicate with a conference server system that delivers a variety of services that use fixed infrastructure to provide an ambient environment. The work was conducted as demo M1 of the Ambience European Project 2003.

1 Introduction

A video is presented which illustrates the integration of a number of new technologies in an ambient environment. This will show for a conference delegate:

- Automatic biometric registration
- Route guidance to a place or person
- Text messaging & file transfer to other delegates
- Robot routing & ad-hoc networking to seamlessly extend network coverage
- Custom applications for meeting chairman

The video presents a scenario that is designed to demonstrate the features of this Ambient system.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 3–14, 2003.

1.1 Scenario

Peter walks into a busy conference venue. His PDA beeps and he is invited to register for the conference by speaking into the PDA's microphone. His voice is recognized and a user interface (UI) relevant to the conference is the automatically displayed on the PDA. This provides a range of useful applications presented in a very user-friendly way. Peter checks his messages. He needs to find his first meeting room, so he requests his PDA to work out a route. A map is downloaded onto the display showing a graphical representation of the building, his current position and visual and textual directions. As he moves, his position is updated on the map and the text window updates with reassuring messages until he reaches his destination.

Peter continues to use his PDA during the meeting to exchange text messages with other participants, follow the agenda and make notes. At the end of the meeting, the participants move to a coffee area. Anticipating that the network capacity will be more useful in the coffee area than the meeting rooms, the mobile robot routers are automatically moved to ensure a seamless communication service.

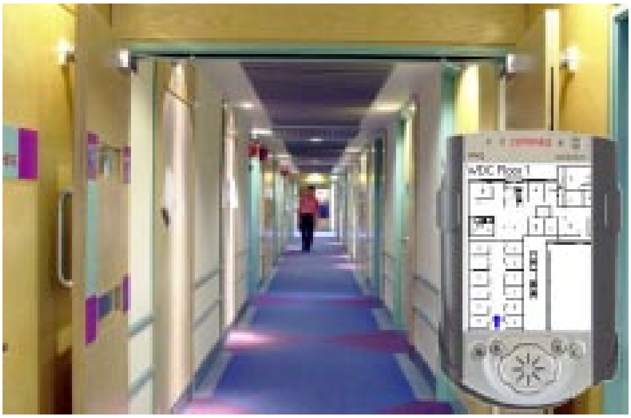


Fig. 1. Interactive route guidance

2 System Overview

Figure 2 shows an overview of the system. At its heart is the server that manages the applications and routes information around the system. The biometric identification is performed by voice recognition that forms the link between the iPAQ client and the delegate.

The iPAQ has a jacket with integrated WLAN and ZigBee radios. A WLAN access point at the server provides the connection to the iPAQs as they roam within the conference centre.

The integrated ZigBee communicates with a fixed ZigBee infrastructure to provide a position estimate derived by integrating Received Signal Strength Information (RSSI) measurements.

Messages are delivered to the iPAQ over WLAN informing the delegate of meetings/schedules. The user may ask to be guided to a room or person within the conference facility. The user's position is continuously monitored causing the server to download an interactive map that updates as the person moves towards the desired location. The messages are defined in XML format [7] to enable efficient data manipulation. The client software drives the messaging through a series of pre-defined message requests to which the server responds. In practice this means that the client applications poll the server regularly for updates.

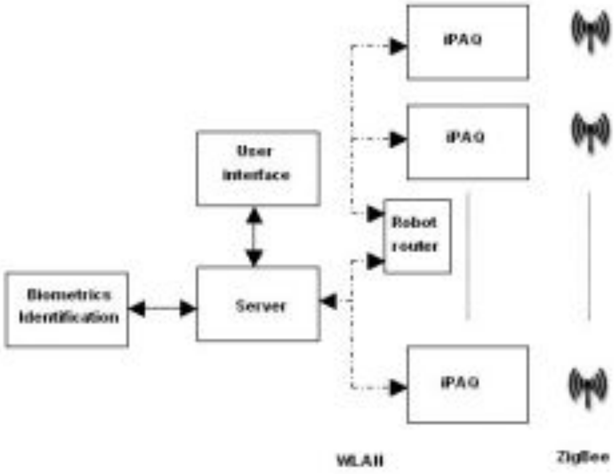


Fig. 2. System overview

Network coverage is extended beyond the fixed infrastructure by the use of ad-hoc networking, using other iPAQs or specially designed robot routers that anticipate network shortage and move to appropriately extend coverage.

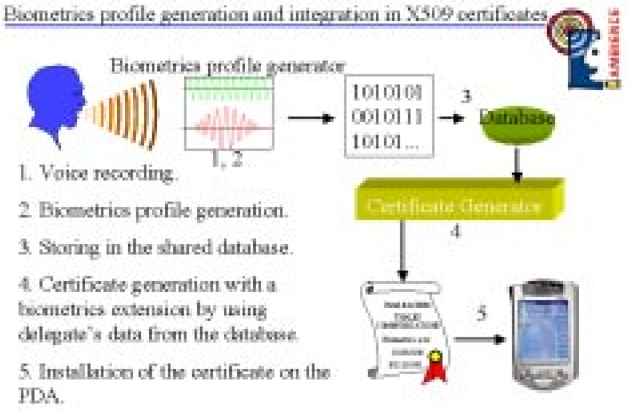
2.1 Biometric Voice Recognition

2.1.1 The Offline Procedure

This phase is designed to generate an "Identity Card" for the delegate. This "Identity card" forms the link between the physical delegate and his digital profile. When entering the conference building, the delegate goes to the reception desk. The receptionist uses the "Welcome application" to launch the authentication procedure in which:

1. The delegate gives proof of identity.
2. The receptionist searches for the delegate in the shared database.
3. When found, the delegate speaks into a microphone to record a voice sample.
4. The "Welcome application" creates a biometric profile from the voice sample and stores it in the database.
5. The "Welcome application" extracts the delegate's data from the database and creates an X509 certificate with a biometric extension. This certificate is the "Identity card".
6. The delegate receives a MultiMedia Card (MMC) with the new "Identity Card".

This process is illustrated in Figure 3.



Biometrics profile generation and integration in X509 certificates

1. Voice recording.
2. Biometrics profile generation.
3. Storing in the shared database.
4. Certificate generation with a biometrics extension by using delegate's data from the database.
5. Installation of the certificate on the PDA.

Fig. 3. Biometric profile generation

2.1.2 Registration for Services

When the PDA carried by the delegate is first detected by the network, it will have no secure user identity. The delegate's identity card is used to establish an SSL connection [4]. The user is prompted to authenticate for the conference services by speaking into the PDA's microphone. A few seconds of natural speech is recorded. The voice sample is sent to the server through the SSL connection. The server searches for the delegate's biometric profile in the shared database and the biometric features are compared to grant or deny access to the conference services.

2.2 ZigBee RSSI Positioning

The positioning technology used for the demonstrator is based on a prototype ZigBee radio system. A number of ZigBee beacons are installed at fixed known positions over the entire length of the demonstration site, so that at least two beacons can be detected at any one point. A ZigBee radio is also built into each iPAQ jacket providing a serial interface to the PDA.

The ZigBee radio in the iPAQ Jacket detects the beacons in range and measures their signal strength. This information is compiled into a beacon report that is received by the PDA and forwarded to the server. The server uses a table to convert signal strength into the range from each fixed beacon. This provides sufficient information to derive user position, using a triangulation algorithm. This process updates the position of each iPAQ every 2-3 seconds. The position is then filtered with previous position estimates to produce a more reliable smoothed position for the use in the demonstration.

2.3 iPaq Client Device

Client devices are equipped with Familiar Linux operating system and Blackdown Java Runtime Environment. Actual client application software is Java 1.3.1 Standard Edition compatible. All messaging between clients and server is based on XML, which can be easily extended to future needs. Choosing open standards, such as Java and XML , gives the benefit of flexibility and the support for wide variety of different client devices and environments.

2.3.1 iPaq Graphical User Interface (GUI)

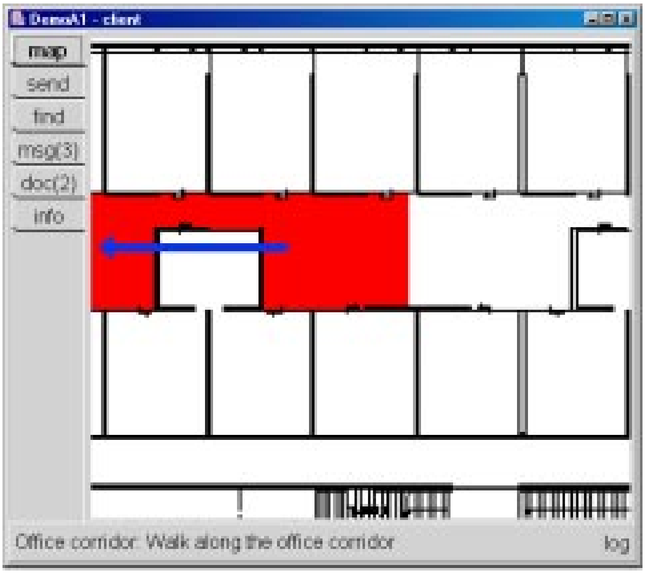
The iPaq runs a Java based application GUI to enable the user to select and control the available applications. It provides a simple user interface that allows the user to switch between a number of views that control each application for:

- Map – display map of local area
- Send – send messages to delegates present
- Find – locate places or people
- Msg – read received messages
- Doc – documents viewer
- Info – display status information of the system

Examples of two such GUI screens are shown in Figure 4 and Figure 5.

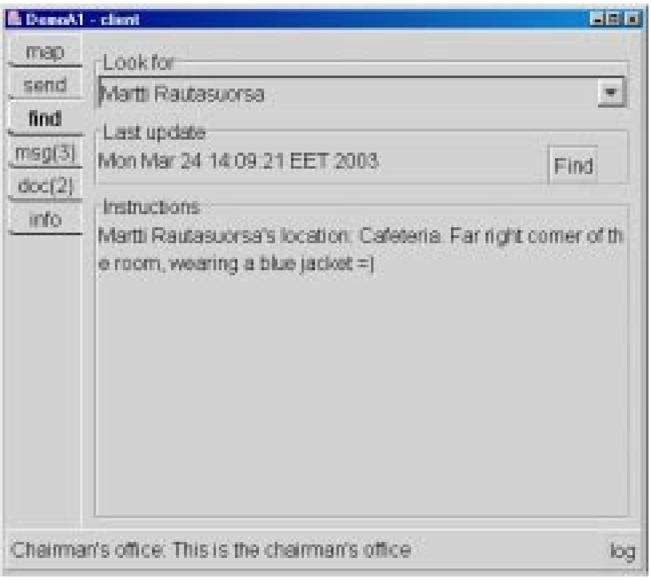
In general the interfaces provide:

- Graphical map of building and visual and textual instructions on how to proceed
- Ability to select recipients from the list of attendees
- Send messages with subject and message fields
- View received documents



Office corridor: Walk along the office corridor

Fig. 4. Map application



Look for: Martti Rautasuorsa

Last update: Mon Mar 24 14:09:21 EET 2003

Instructions: Martti Rautasuorsa's location: Cafeteria. Far right corner of th e room, wearing a blue jacket =)

Chairman's office: This is the chairman's office

Fig. 5. Find application

2.4 Integrated Server System

The applications are made accessible to the users following a client-server architecture. The Tomcat servlet container from the Apache Software Foundation [5] is used, implementing the applications as servlets .

Following a scheme that resembles the popular web services scheme [6], the communications between the GUI (client) and the server rely on the http protocol and the data exchanged between the clients and the server are encoded using XML.

The communications are secured using the SSL protocol. The use of X-509 certificates provide for mutual authentication of clients and server. These security mechanisms are supported by the Apache Tomcat server.

Finally, the data manipulated by the server and the applications are managed by an SQL relational database that supports concurrent accesses, synchronisation and permanent storage. “MySQL” , an open source database management system [3], was chosen for this task.

The architecture of the server is depicted in Figure 5. It shows that TOMCAT contains two servlets: the chairman servlet that initialises the database, and the ambience server servlet that encapsulates the applications currently made accessible to the users. The messages received by this last servlet are parsed and dispatched to the appropriate service.

3 Mobile Robots

The robots [2] have been designed and built at ENST's Computer Science department with standard, low cost hardware components. Their architecture is open and modular to meet the needs of sensor and actuator customisation. The board offers standard communications: Ethernet, 802.11 and Bluetooth. It also provides a wide range of input output interfaces: USB, I2C, PCMCIA, 115 Kbits IRDA, ISO 7816-2 Smartcard, 8 bit audio CODEC, graphical LCD touchscreen.

The battery is controlled by a dedicated circuit that keeps the processor aware of its load level. When the battery's level reaches a low limit, the robot will complete its critical tasks, if any, and move towards a refueling intelligent docking station. The station emits a laser beam to guide the robot back.

3.1 Ad Hoc Networking

In the scenario described when the participants move to coffee area for a coffee break, an ad hoc network is established to guarantee an ambient communication service between them. Ad-hoc networks are multi-hop wireless networks where all nodes cooperatively maintain network connectivity. Such networks provide mobile users with ubiquitous communication capability and information access regardless of location.

The main challenge in forming ad-hoc networks is the routing. Routing protocols in such networks must be adaptive to face frequent topology changes because of node mobility. Minimizing the number of hops is the most common criteria adopted by the routing protocols proposed within the IETF MANET working group. However, the number of hops criteria is not necessarily the most suitable metric to build routing decisions. Route selection must take into account the current link conditions, to offer a certain quality of service (QoS). The basic function of QoS routing is to find a network path that satisfies some given constraints.

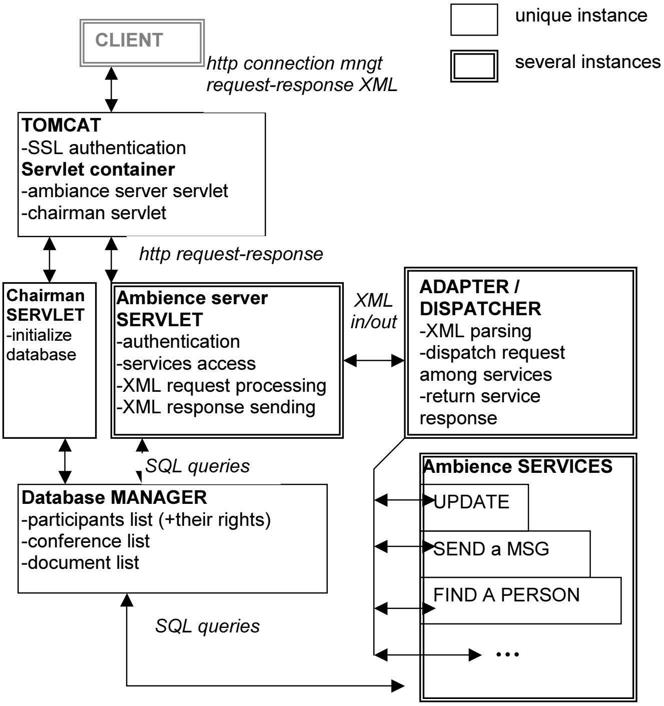


Fig. 6. Server architecture

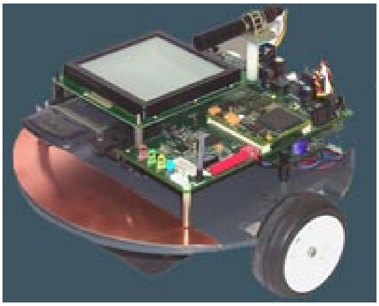


Fig. 7. Mobile robot

For the purpose of this demo, we develop a Quality-of-Service (QoS) routing protocol for mobile ad-hoc Networks. This QoS routing protocol aims to enhance the perceived communication quality within the coffee area [8]. We perform the proposed QoS-enhanced routing based on the Optimized Link State Routing (OLSR) protocol [9], introducing a more appropriate metric than the number of hops. The proposed QoS routing protocol, called QOLSR, produces better performance comparing with the legacy OLSR protocol, as depicted in Figure 8.

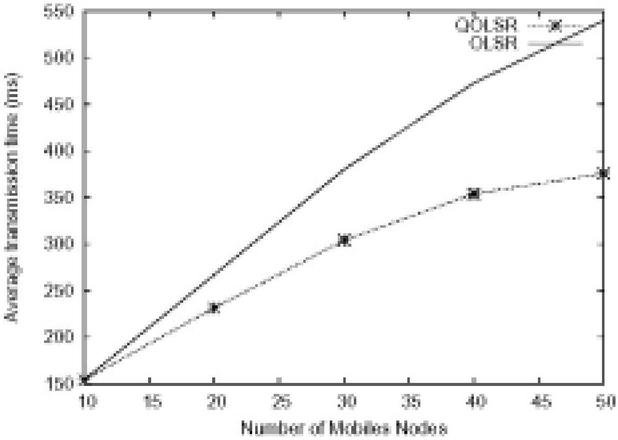


Fig. 8. Performance evaluation of QOLSR

3.2 Robot Routing

The robot implements an ad hoc routing algorithm in order to support multi-hop routing (in an IPv6 environment). In other words, the robot can either be used to connect several wireless devices that are not within the range of one another or as an extension of an existing infrastructure – in which case a gateway between the fixed infrastructure and the ad hoc network is used, as depicted Figure 6. Contrary to many classical ad hoc networks, routing is designed hierarchically such that communication relays are exclusively robots (and not low capacity nodes -- battery, CPU, etc.). The use of a wireless network core made up of mobile robots adds some control in the ad hoc network. This feature is essential in order to provide an acceptable link quality, increase network performance, and thus support real time services in the network, as shown in [1].

3.3 Monitoring Ad Hoc Network

A monitoring station was developed to show the improvements obtained when using the QOLSR or the robot-based routing compared to the OLSR protocol. The monitoring station is based on monitoring agents installed on all wireless nodes (including robots in the robot-based routing). This tool allows the visualisation of the entire network state including the state of each node, links as well as the entire topology. It will also visualise the routing tables of each node (Figure 10). This tool was developed independently of the underlying ad-hoc routing protocol and can be used for either proactive or reactive ad-hoc routing protocols.

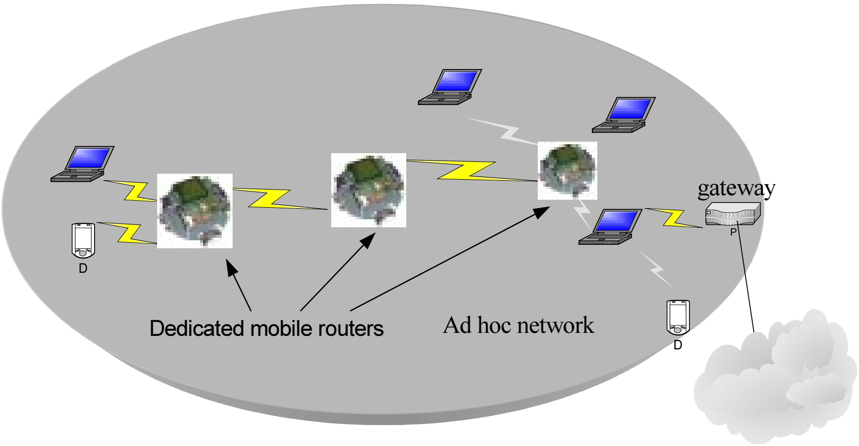


Fig. 9. Ad hoc networking with the mobile robots

Monitoring agents gives the following set of information to the monitoring station (as shown in Figure 10):

- Timestamp (date:jjmmyyyy/hour:hhmmss)
- Node's addressing information (IP address, subnet mask, gateway address, MAC address, SSID)
- Routing table with neighbour's IP addresses
- Monitored wireless link characteristics (radio channel ID, noise, signal strength, packet error ratio)

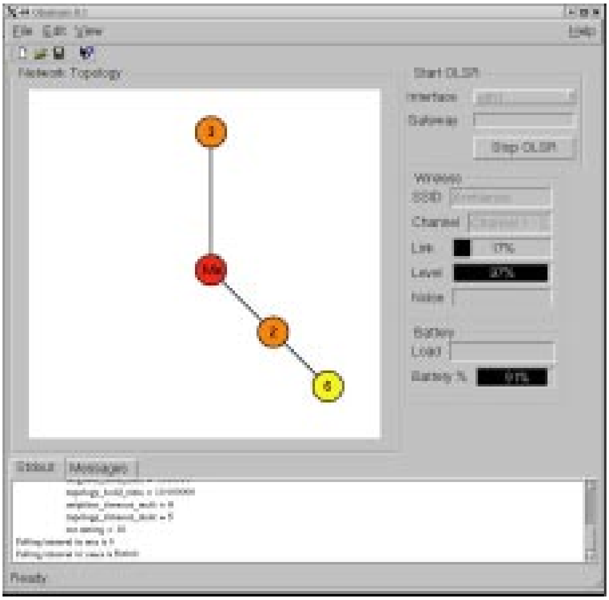


Fig. 10. Monitoring station snapshots

4 Summary of Technology

The key technologies demonstrated are:

- Natural speech recognition – Thales, France
- Java based iPAQ GUI – CCC, Finland
- XML message system – CCC, Finland
- ZigBee positioning – Philips, UK
- Integrated server – ENST, France
- Robot routing/ad-hoc networking – ENST, France
- Network monitoring and analysis – LIP6, France

5 Conclusions

The video demonstrates an ambient environment designed to improve the experience of attending a conference. From the moment the delegate enters the conference venue, registration is simple and automatic which gives rise to a wealth of useful services available through the user's PDA. The system is designed to help make finding people and places easy and allow information to be exchanged.

The real work of this ambient system goes on in the background as the server authenticates people, tracks their movements, routes messages, exchanges documents and handles user requests. The server also successfully implements a scheme of ad-hoc networking and robot routing that fills holes in the network.

Acknowledgments. This work would not have been possible without the support of the European commission. The team would also like to thanks the overall project coordinator Evert van Leonen for his help and guidance. Finally thanks to all the efforts of the demo M1 team represented by the companies CCC, ENST, LIP6, Philips & Thales and for their detailed contributions to this abstract made by Yacine Ghamri Doudane Sidi-Mohammed Senouci, Farid Bachiri, Anelise Munaretto, Jouko Kaasila, Gwendal Le Grand & Isabelle Demeure.

References

1. Le Grand, G., Meraihi, R.: Cross-layer QoS and terminal differentiation in ad hoc networks for real-time service support, proceedings of MedHOC NET 2003, (IFIP-TC6-WG6.8), Mahdia, Tunisia, June 25–27 (2003)
2. ENST, SPIF robots, http://www.infres.enst.fr/~spif/ambience.html
3. MySQL wen page. http://www.mysql.com/
4. SSL 3.0 SPECIFICATION web page. http://wp.netscape.com/eng/ssl3/
5. Apache Jakarta Project web page. http://jakarta.apache.org/tomcat/

6. W3C Web Services Activity web page. http://www.w3.org/2002/ws/
7. Extensible Markup Language (XML) web page. http://www.w3.org/XML/
8. Munaretto, F.A., Badis, H., Al Agha, K., Pujolle, G.: A Link-state QoS Routing Protocol for Ad Hoc Networks, IEEE Conference on Mobile and Wireless Communications Networks, MWCN'02, Stockholm, Sweden, September (2002)
9. Clausen, T., Jacquet, P. (Eds.): Optimized Link State Routing Protocol, IETF Internet Draft, draft-ietf-manet-olsr-09.txt

Some Issues on Presentations in Intelligent Environments

Christian Kray, Antonio Krüger, and Christoph Endres

Saarland University
P.O. Box 15 11 50, 66041 Saarbrücken, Germany
{kray, Krueger, endres}@cs.uni-sb.de

Abstract. Intelligent environments frequently embed a varying number of output means. In this paper, we present an analysis of the task of generating a coherent presentation across multiple displays. We review problems and requirements, and propose a multi-layer approach aimed at addressing some of the previously identified issues. We introduce a low-level architecture for the handling of devices as well as mechanism for distributing coherent presentations to multiple displays. We then discuss what further steps are needed on the way from an abstract presentation goal to a concrete realization.

1 Introduction

From a technical point of view the trend of Ubiquitous Computing is tightly bound to the notion of the disappearing computer, where most of the processing power and communication abilities will be embedded behind the scenes. However, from the user's perspective Ubiquitous Computing will lead to ubiquitous input and output facilities in a given environment. New display technologies (e.g. organic displays) will increase the amount of displays in an instrumented environment significantly, allowing intelligent multimedia presentation systems to plan and render presentations for multiple users on multiple displays. Modern airports are already good examples of environments equipped with several different types of displays: Information on arrivals and departures is presented on huge public boards, at gates plasma screens display information on the actual flights, throughout the building small touch screens are used to provide single users with information on the airport facilities. Finally wireless LAN hotspots allow the usage of private PDA screens to access information on the web at several locations. Until now those displays cannot be used together to present information, but several technical suggestions were recently made to move towards a solution for this problem. In [1] a framework is presented that incorporates different devices to render audio and graphics files. An approach that allows users to access publicly available displays (e.g. from an ATM) for personal use is presented in [2]. Some research has been also carried out on the combined usage of a PDA and large screens [2],[4],[7].

However, only little work was done on issues regarding the planning and rendering of multimedia presentations on multiple displays. In this paper we reflect both on the different technical prerequisites and the concepts that will allow for distributed multimedia presentations in instrumented environments. This involves the handling of

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 15–26, 2003.

devices, e.g. their registration and access, the rendering of the presentations, e.g. synchronizing presentations in time and space and the planning of presentations, e.g. how to guide the users attention from device to device. While the processes underlying a presentation using multiple devices are rather complex and hence call for a sophisticated approach that takes into account a number of factors ranging from device control to guiding the attention of the user, the benefits are worth the effort: On the one hand, the users will enjoy consistent and non-interfering presentations that are adapted to the current situation. On the other hand, these presentations will use all the available output means instead of just small limited set. Hence, the system will deal with the complexity while the users benefit from a simplified interface.

We will first review some key issues in the context of intelligent environments in general, and more specifically point out problems related to handling presentations in such a setting. In order to address some of the issues raised, we will then present a low-level infrastructure capable of dynamically handling various devices. Based on this foundation, we will introduce a mechanism for rendering presentations that are distributed across multiple devices in a synchronous way. These presentations are the result of a planning process that can only partially rely on traditional presentation planning. Consequently, we will point out some areas where further reasoning is necessary. After sketching out a possible solution for this problem, we will shortly summarize the main points of this paper, and provide an outlook on future research.

2 Issues

During the process of planning, generating, and rendering a presentation in intelligent environments several issues arise that go beyond those encountered in traditional multimedia presentation [9]. The issues are mainly caused by the dynamicity of the ubiquitous setup and the diversity of the devices used for a presentation. A further factor that complicates the process is the potential presence of multiple users.

A fundamental issue that does not only impact presentations but all services and tasks running ‘on’ an intelligent environment is the discovery, handling, and control of devices. In the case of presentations mainly output devices such as displays or loudspeakers are of interest in this area. The infrastructure has to provide means to determine what devices are present and available, to control these devices, and to dynamically add and remove devices.

The latter point is also closely related to another key issue: the constant change an intelligent environment may be faced with. On the one hand, new devices and users entering or leaving a room may require the system to re-plan all presentations that are currently active. On the other hand, presentations running in parallel may interfere with each other, e.g. two unrelated but simultaneous audio messages. Therefore, a suitable presentation engine has to constantly reassess the current device assignment, and potentially re-plan frequently. In doing so, it must also guarantee a smooth transition from one assignment to another in order to not confuse the users. The requirement for constant re-planning also calls for a representation format that supports this process.

The potential presence of multiple users causes further problems in addition to interference. On the one hand, planning several presentations instead of a single one results in increased complexity. On the other hand, the sharing of devices (such as a large display) adds a new dimension to the planning as well as to the rendering process. An additional issue concerns the actual rendering or output of a presentation. In order to realize a coherent presentation, it is vitally important to properly synchronize the various output devices. In the synchronization process, device-specific properties such as bandwidth, memory, and speed have to be taken into account. A suitable mechanism for synchronized rendering of presentations would therefore have to provide means for prefetching and pre-timed events.

In the following, we present a basic infrastructure for device handling, a rendering mechanism for distributed presentations, and some ideas on the planning process that are aimed at addressing several of the issues raised above.

3 Device Handling

In a dynamic, instrumented environment, there are several requirements for managing the devices. One key factor to take into consideration is the fluctuation of people and their personal devices (e.g. PDAs or mobile phones), the concurrency of several applications/presentations and the differing features of similar devices (e.g. PDAs with color or monochrome displays, or different display sizes).

The task of classifying devices turns out to be rather complex. One possible approach, which was realized in the FLUIDUM project [8], is to assume a different point of view: Instead of classifying devices, we define device feature objects (e.g. video in, audio out, tag reader, etc) and then describe a device using a list of feature objects it possesses.

As underlying infrastructure, we use a device manager with two remotely accessible interfaces. On the one hand, devices can register and announce their availability; on the other hand, services can register and check for devices with certain features. The structure of the device and the architecture of the device manager are shown in Figure 1.

A device consists of a table containing parameter value pairs (for instance "name=camera01") and a collection of property objects (or "features"). The advantage of realizing these objects as remotely accessible APIs is, that we do not only know which features a device possesses, but we can also access it directly. (We decided to use Java remote method invocation (RMI) since it is an advanced technology for object marshalling and sending, but does not have the communication overhead of more sophisticated approaches such as CORBA or JINI.)

The device registers with the device manager over a "device plug adapter" (an RMI interface). The device manager keeps an up-to-date list of registered devices, and provides information to connected services that are interested in knowing about available devices. The device manager thus serves as a matchmaker between services and devices, and also passes events concerning the deregistering of devices to the services. It hence provides a lookup or yellow page service for a given environment such as a room.

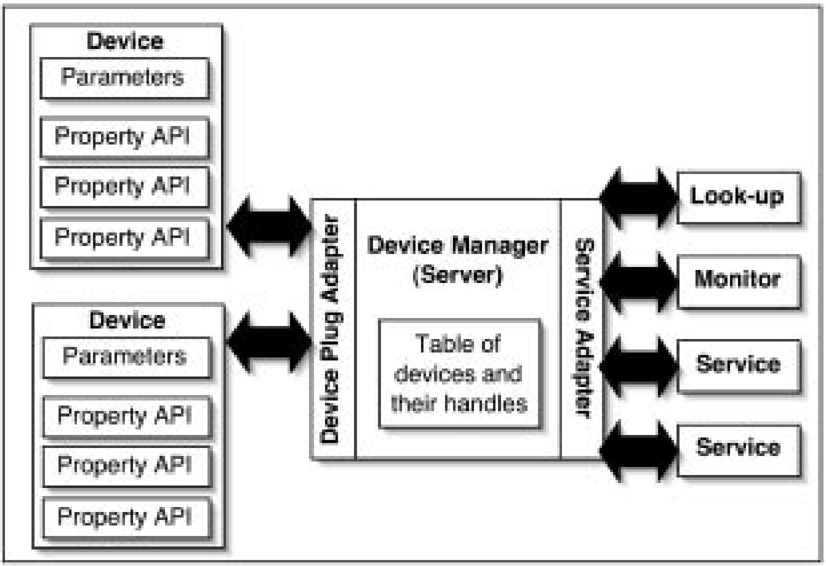


Fig. 1. Architecture of the Device Manager

4 Presentation Rendering

In the previous section, we described an infrastructure aimed at handling the (de-) registration and low-level control of various devices including output means such as displays. While this certainly provides a basis to build upon, further services are needed in order to render presentations in a coherent way. As we have seen previously, the synchronization of a presentation that is spread across multiple displays or output devices is a key concern that still needs to be addressed. Furthermore, the spatial constellation of the devices used for the presentation plays an important role.

For example, we want to avoid to generate a sophisticated presentation and to display it on a screen that is behind the user. In addition, we have to cope with the removal of some devices while a presentation is running, and thus with the incremental and continuous re-planning of the presentation. Therefore, there is a need for a representation format that not only incorporates space and time but that is also modular and can be rendered by a variety of different output devices.

We would like to propose the Synchronized Multi-Media Integration Language (SMIL) [16] to address these issues for several reasons. Firstly, the language incorporates a sophisticated concept of time allowing for the sequential and parallel rendering of various types of media. In addition, it is possible to specify the duration, the beginning and the end of the actual rendering of a part of a presentation both in absolute and relative time. Secondly, there is a similar concept of space, which supports both absolute and relative coordinates. Furthermore, it incorporates a region concept (providing local containers) as well as a simple layer concept that allows for the stacking of regions. Unfortunately, the underlying space is only two-dimensional, which is of limited use in an inherently three-dimensional scenario such as intelligent environments. However, we will point out ways to overcome this limitation later in this section.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
 <smil>
   <head>
     <meta name="title" content="example"/>
     <layout>
       <root-layout title="demo" id="root" width="240" height="300"/>
       <region title="main" height="320" id="main" z-index="1"
         width="240" top="0" left="0"/>
     </layout>
   </head>
   <body>
     <par id="par1">
       <audio begin="0s" region="main" id="audio1" src="song.wav"
         dur="10s"/>
       <seq id="seq1">
         <img begin="0s" region="main" id="img1" src="one.png"
           dur="5s"/>
         <img begin="0s" region="main" id="img2" src="two.png"
           dur="5s"/>
       </seq>
     </par>
   </body>
  </smil>
```

Fig. 2. Example SMIL presentation

A third reason supporting the adoption of SMIL for ambient presentations lies in its inclusion of various media such as text, images, sound and videos at the language level. The short example shown in Figure 2 illustrates the simplicity of including various media. The example describes a presentation that displays a slide show of two pictures (‘‘one.png” and ‘‘two.png”) that are each shown for five seconds while a sound file is playing (‘‘song.wav”). This brings us to a fifth reason for using SMIL, which is the conciseness of the format. Since it is text-based, it can even be compressed and adds very little overhead to the media-data itself. This is especially important in an intelligent environment where many devices will not have a high-speed wired connection but rather an unreliable wireless link with a lower bandwidth.

A final reason favoring the use of SMIL in ambient presentations consists of the availability of the corresponding player software on a wide range of devices. SMIL is a subset of the format used by the REAL player [15], which runs on desktop computers, PDAs and even some mobile phones. Similarly, a slightly outdated version of SMIL is part of the MPEG-4 standard [13] that is at the heart of a variety of services and media players such as QuickTime [11]. The Multimedia Message Service (MMS) [11] -- a recent addition to many mobile phone networks -- is also based on SMIL, and consequently, the current generation of mobile phones supporting MMS provide some basic SMIL rendering capabilities. Furthermore, there are several free players such as S2M2 [14] and X-Smiles [17]. For our purpose, especially the latter one is interesting as it is continuously updated and its source code is available. In addition, it has been specifically designed to run on various devices ranging from desktop computers to mobile phones.

However, before we can actually design a service for controlling and rendering SMIL on multiple output devices, we have to address the problem of SMIL only supporting two dimensions. Fortunately, this is a straightforward task. Assuming that we are dealing with bounded spaces – e.g. we do not consider interstellar space -- we can develop a function that maps the surface of a bounded space, e.g. a room, to a two-dimensional plane. Figure 3 shows an example for such a function. If we just want to deal with rectangular rooms and wall-mounted displays the mapping is very straightforward and consists basically of an `unfolding' of the corresponding three-dimensional box.

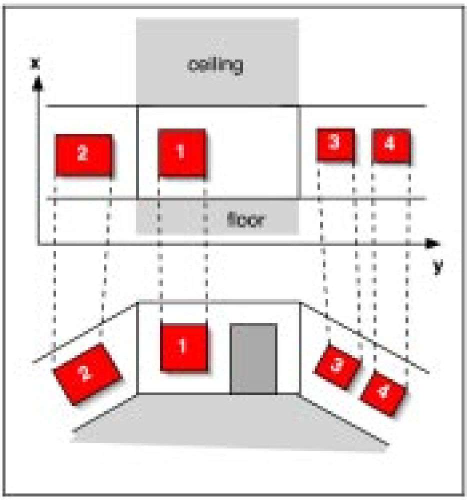


Fig. 3. Mapping of 3D surfaces to a 2D plane

Obviously, the more complex a room is, the harder it is to find an `intuitive' mapping that preserves some of the spatial properties. This is even more so, if we take into account mobile devices such as PDAs or tablet computers. However, as long as the mapping function is bijective, we can always determine which part of a SMIL presentation corresponds to a specific location in 3D space, and vice versa. Even if the mapping does not preserve spatial properties such as neighborhood relations, we can perform spatial reasoning in 3D space by re-projecting 2D parts using the mapping function. Therefore, it is safe to assume the existence of a table-based bijective mapping function based on the basic infrastructure described in the previous section. Since a service maintaining this function, i.e. the presentation-planning component, receives constant updates on the availability and location of various output devices, it can ascertain the integrity of the mapping function. Using such a mapping function, a presentation planner can generate a complete SMIL presentation covering multiple devices in a room. At this stage, we need a piece of software that takes this presentation and sends the various parts of it to the corresponding devices, and that assures that the overall presentation is delivered in a coherent way. We have implemented such a service - the SMIL Manager shown in Figure 4 - that takes care of this part of the presentation pipeline. By introducing a SMIL property object for all devices capable of rendering (parts of) a SMIL presentation, we can rely on the standard service interface for device access and control. The API of the SMIL property object implements the actions of the protocol listed in Table 1.

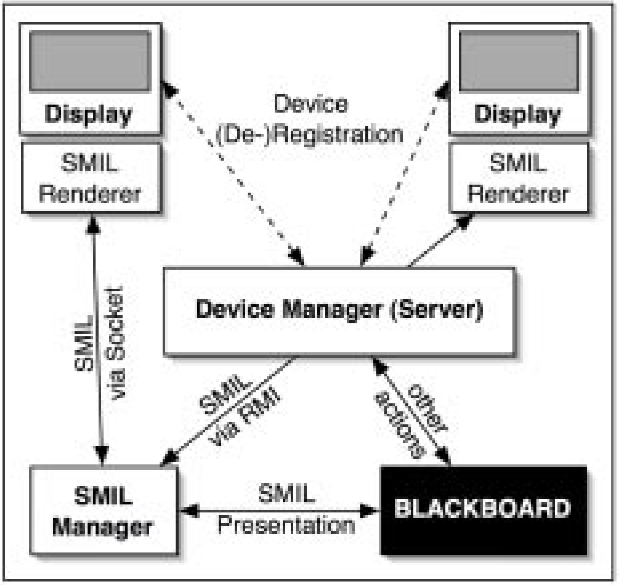


Fig. 4. Architecture of the SMIL Manager

Using the Device Manager (presented in the previous section), the SMIL Manager can hence keep track of all devices capable of rendering SMIL presentations. It relies on the mapping function to identify, which part of the overall presentation is to be rendered on which device. It then sends the various parts of the presentation to the corresponding devices for prefetching. Once all devices have reported back after completing the prefetching of media included in the presentation, the SMIL Manager triggers the simultaneous start of the presentation on all devices.[1]

However, there are a few details in the process that deserve proper attention. In order to assure a synchronized presentation, the SMIL Manager sends out synchronization messages to all attached output devices on a regular basis. Otherwise, the different clocks and time-servers that the various devices rely on would result in unwanted behaviors such as a part of the presentation starting too early or too late. Furthermore, the prefetching of media is vitally important. Especially on a shared wireless network, bandwidth is limited and can result in media not being available when they are needed in the presentation if we do not perform prefetching. But even in case prefetching is included in the process, it is important to actually check whether it has completed on *all* devices since there are great differences in terms of the time that is required to download all media. For example, a screen attached to a wired workstation will probably fetch the required media for a presentation much faster than a PDA connected through WLAN.

Table 1. Protocol for interaction between the SMIL Manager and output devices

Action	Explanation
<LOAD>	Load a presentation specified by an included URL.
<START>	Start a presentation specified by an included URL, optionally a specified time.
<STOP>	Immediately stop a presentation specified by an included URL.
<SYNCHRONIZE>	Synchronize internal clock with time stamp of SMIL Manager.
<REPORT>	Send a message whenever the user activates a link.
<LOADED>	Presentation at included URL has been fully prefetched.
<FINISHED>	Presentation at included URL has finished playing.
<LINK>	User has activated a link pointing to included URL.

Table 1 lists the various actions that the SMIL Manager can perform on the connected devices. These are either transmitted using remote method invocation (RMI) or through a socket connection using plain text messages such as the ones listed in the table. Obviously, there has to be a means to instruct a device to load a presentation, and to start it (at a given time). Additionally, it should be possible to stop a running presentation, and to synchronize the devices with the clock of the SMIL Manager. Finally, there is rudimentary support for interaction (through the <REPORT> action). When enabled, the device reports back to the Manager in case the user clicks on a link embedded in the SMIL presentation. On touch-sensitive screens, for example, this allows for simple interactions. We will discuss some implications of this feature in the

[1] Note that the simultaneous start of the presentation on all devices does not necessarily imply that all devices will actually produce output right away. More often, some will render something directly while others will wait for a predefined time before actually outputting something.

last section of this paper. Furthermore, the devices report back once the current presentation has finished playing.

The SMIL Manager as well as the client application are written in Java and have successfully been run on various desktop computers and PDAs. Since they only require version 1.1 of the Java virtual machine, we are confident that they will run on further devices in the future (such as mobile phones). SMIL rendering relies on the X-Smiles engine [26], and is realized as a dedicated thread that is separate from the communication (either through RMI or a socket connection). Therefore, both the server and the client can communicate while performing other tasks (such as playing a presentation or registering further clients). This is an important feature especially in the context of stopping a currently running application (e.g. when the device has to be used by another service, or when the presentation has to be re-planned due to the removal of some device).

5 Presentation Planning

Now that we have discussed how to control the different devices and how to handle distributed SMIL presentations, there is a final stage still missing in the presentation pipeline: the planning and generation of the presentation. As we have noted in the introduction, there is a lot of research going on in the context of single-user single-device presentations, i.e. for a person sitting in front of a desktop computer. However, when faced with a multi-device (multi-user) presentation, there are some key properties that distinguish the corresponding planning process from traditional setups. Firstly, time and space play a much more prominent role since not only are the various output devices distributed across space but also do they have to be properly synchronized. The latter point refers to the issue discussed in the previous section as well as to the planning stage since the presentation planner has to take into account when to present what on which display. Secondly, it may frequently occur that multiple users are present in a room, which entails several problems not present in single-user scenarios. For example, a number of users may share certain devices, which may interfere with one another, e.g. when one person is making a phone call while a second person uses a voice-based approach to access a database. In addition, security and privacy come into play if there is more than one person involved.

A third important differentiating factor consists of the dynamicity of the environment. While traditionally, it is assumed that the devices used for a presentation do not abruptly disappear, this assumption is not realistic in an intelligent environment. Even more so, it is possible that further output devices actually appear, e.g. a person carrying a PDA enters an intelligent room. All these factors call for an approach that facilitates re-planning on various levels, which in turn is a further difference to more static and traditional presentation planning. In order to take these special requirements into account, we distinguish three different planning problems that are interleaved with each other: (1) the planning of the *content, (2)* the planning of the *media distribution in space and time* and (3) the planning of the *attentional shift* (*focus control)* that has to be supported when users have to focus on different displays in their environment over time.

The first problem is similar to content selection in classical intelligent multimedia presentation systems designed for single display scenarios (e.g. [9]). Hence, we will not discuss it here, as it can be addressed using a classical plan operator based planning technique such as STRIPS [10]. For the planning of the spatio-temporal distribution we plan to adopt a constraint-based mechanism described in [6], where the rendering abilities of each display (e.g. resolution, audio quality, interaction requirements) is modeled and scored with a certain value. The rendering of a video on a PDA display would for example achieve a lower score than on a big plasma screen. Although most of the constraints are soft (e.g. it is possible to render a video on a PDA), some are not: if the presentation contains interactive elements (e.g. buttons) it is not possible to render them on a screen without input capabilities. In this case a plasma screen without a touch screen overlay for example could not be used. One possible solution would consist of rendering the interactive parts of the presentation on a device with a touch-screen (e.g. a PDA) and the rest of the presentation on the plasma screen.

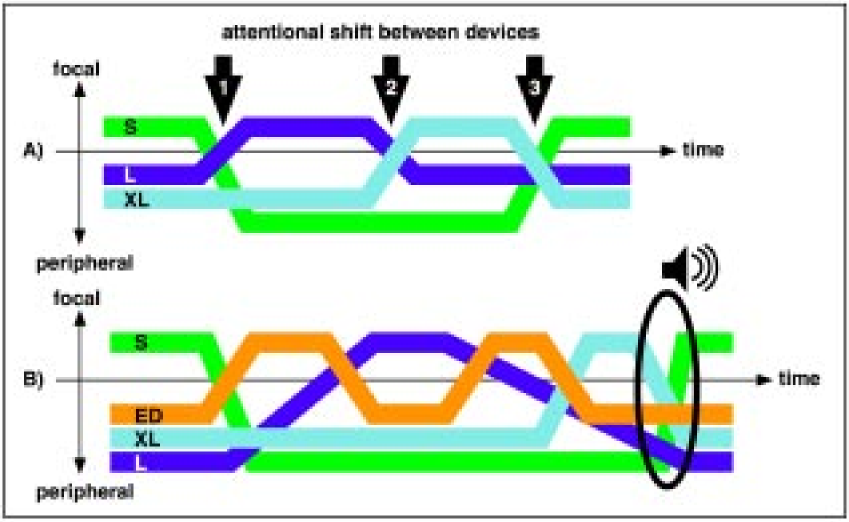


Fig. 5. A distributed presentation: A) without focus control, and B) with focus control using an everywhere display and audio meta-comments

From the users perspective such a distributed presentation can cause problems, if they do not know where to focus their attention at a certain moment of time. One way to overcome this problem is illustrated in Figure 5. The diagram in Figure 5 A) shows an example of a presentation distributed over time on a small ('S' - e.g. a PDA), a large ('L' - e.g. a 17-inch touch-screen) and an extra large ('XL' - e.g. a plasma screen) device. From the user's perspective devices in the environment can be either in focus or peripheral. The three arrows in Figure 5 A) indicate that the user has to perform an attentional shift, whenever a device switches from the focus role to a peripheral role and vice versa. The user's ability to perform this shift may be hindered by several factors, e.g. the spatial distance between the devices and distracting cues in the environment (e.g. noise). If a user is not able to shift the attention to the right device at the right moment in time, the probability increases of the presentation becomes confusing or, even worse, that it is misinterpreted.

Therefore it makes sense to support this shift of attention in the instrumented environment. In Figure 5 B) two additional means are used to facilitate the attentional shift. The shifts 1 and 2 are supported by an everywhere display [18] (ED). An ED is a device that is able to project images on arbitrary surfaces in the environment, i.e. a ceiling-mounted projector attached to a movable stand that can be controlled remotely. Therefore it can be used to guide the attention of the user from one place in the instrumented environment to another place in the environment. If the environment knows the user's position, an ED can even be used to guide the attention from a PDA that is carried by the user towards a fixed display in the environment. Other means to shift the user's attention can be meta-comments that explicitly tell the user where to look next (e.g.: "To stop the video please press the corresponding button on your PDA"). In Figure 5 B) such an audio meta-comment is used to close the third "attentional gap".

We can extend the constraint-based planning approach by introducing another score that represents the difficulty of the users to direct their attention from one display to another in a given situation. Of course these additional elements of the presentation increase the overall presentation time leading to other problems that may result into a backtracking process and a modified distribution of the presentation.

6 Conclusion and Outlook

In this paper, we discussed several issues and possible solutions in the context of generating coherent presentations in an intelligent environment. We presented an analysis of the task of generating a coherent presentation across multiple displays, namely the planning, the rendering and the control of the output devices. We reviewed several problems and requirements such as synchronization and three-dimensionality, and introduced a multi-layer approach aimed at addressing some of the previously identified issues. We described a low-level architecture for the handling of devices as well as mechanism for distributing coherent presentations to multiple displays using the Synchronized Multimedia Interface Language (SMIL). We then discussed the issues related to presentation planning, and provided some ideas for implementing the corresponding process.

Consequently, a major part of future work consists in actually designing a working system for the planning process, and to test it with various services such as navigation, localized information or smart door displays. A second field of major interest is the capture of user interaction. We have already started on this point by incorporating a simple link following action in the protocol used to connect the SMIL Manager and its clients. However, the underlying Device Manager allows for a much more fine-grained interaction, which we intend to include in future versions of the system.

Acknowledgements. The authors would like to acknowledge their funding though the Deutsche Forschungsgemeinschaft (DFG) within the Special Research Center 378: Resource-Adaptive Cognitive Processes and the Fluidum research group.

References

1. T. L. Pham, G. Schneider, S. Goose: A Situated Computing Framework for Mobile and Ubiquitous Multimedia Access Using Small Screen and Composite Devices, in Proc. of the ACM International Conference on Multimedia, ACM Press, 2000
2. Roy Want, Trevor Pering, Gunner Danneels, Muthu Kumar, Murali Sundar, and John Light: The Personal Server: Changing the Way We Think about Ubiquitous Computing, Proc. of Ubicomp 2002, LNCS 2498, Springer, 2002, pp. 192
3. Scott Robertson, Cathleen Wharton, Catherine Ashworth, Marita Franzke: Dual device user interface design: PDAs and interactive television, Proc. of CHI'96, ACM Press, 1996,pp. 79
4. Yasuyuki Sumi , Kenji Mase: AgentSalon: facilitating face-to-face knowledge exchange through conversations among personal agents, Proc. of Autonomous agents 2001, ACM Press, 2001.
5. Brad A. Myers: The pebbles project: using PCs and hand-held computers together, Proc. of CHI 2000, ACM Press, 2000, pp. 14
6. Antonio Krüger, Michael Kruppa, Christian Müller, Rainer Wasinger: Readapting Multimodal Presentations to Heterogeneous User Groups, Notes of the AAAI-Workshop on Intelligent and Situation-Aware Media and Presentations, Technical Report WS-02-08, AAAI Press, 2002
7. Michael Kruppa and Antonio Krüger: Concepts for a combined use of Personal Digital Assistants and large remote displays, Proceedings of SimVis 2003, SCS Verlag, 2003
8. The Fluidum project. Flexible User Interfaces for Distributed Ubiquitous Machinery. Webpage at http://www.fluidum.org/
9. E. Andre, W. Finkler, W. Graf, T. Rist, A. Schauder and W. Wahlster: "WIP: The Automatic Synthesis of Multimodal Presentations" In: M. Maybury (ed.), Intelligent Multimedia Interfaces, pp. 75–93, AAAI Press, 1993
10. R. E. Fikes and N. J. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. Artificial Intelligence, 2: 198–208, 1971
11. The 3rd Generation Partnership Project (3GPP). The Multimedia Message Service (MMS). Available at http://www.3gpp.org/ftp/Specs/html-info/22140.htm
12. Apple Computer Inc. The QuickTime player. Available at http://www.apple.com/quicktime
13. MPEG-4 Industrial Forum. The MPEG-4 standard. Available at http://www.m4if.org
14. The National Institute of Standards and Technology. The S2M2 SMIL Player. Available at http://smil.nist.gov/player/
15. RealNetworks. The REAL player. Available at http://www.real.com
16. The World Wide Web Consortium. The synchronized multimedia integration language (SMIL). Available at http://www.w3.org/AudioVideo/
17. X-Smiles. An open XML-browser for exotic devices. Available at http://www.xsmiles.org
18. Claudio Pinhanez: The Everywhere Displays Projector: A Device to Create Ubiquitous Graphical Interfaces, Proc. of Ubicomp2001, Volume 2201, Lecture Notes in Computer Science, 2001, pp 315

A Reactive QoS Routing Protocol for Ad Hoc Networks

Stéphane Lohier, Sidi-Mohammed Senouci, Yacine Ghamri-Doudane, and Guy Pujolle

LIP6 - Université de Paris VI
8, rue du Capitaine Scott 75015 Paris – France
{Stephane.Lohier, Sidi-Mohammed.Senouci, Yacine.Ghamri, Guy.Pujolle}@lip6.fr

Abstract. Due to bandwidth constraint and dynamic topology of mobile ad hoc networks, supporting Quality of Service (QoS) is a challenging task. In this paper we present a complete solution for QoS routing based on an extension of the AODV (Ad hoc On Demand Vector Distance) routing protocol that deals with delay and bandwidth measurements. The solution is based on lower layer specifics. Simulation results shows that, with the proposed QoS routing protocol, end to end delay and bandwidth on a route can bee improved in most of cases.

1 Introduction

The increasing progress of wireless local area networks (WLAN) has opened new horizons in the field of telecommunications. Ad Hoc networks are multi-hop wireless networks where all nodes cooperatively maintain network connectivity without fixed infrastructures and centralized administration. Due to their capability of handling node failures and fast topology changes, such networks are usually needed in any situation where temporary network connectivity is needed, such as in battlefields, disaster areas, and meetings. These networks provide mobile users with *ubiquitous* communication capability and information access regardless of location.

Throughputs reached today by Mobile Ad hoc NETworks (MANET) [1] enable the execution of complex applications such as multimedia applications (video conference, visiophony, etc.). However, these applications consume significant amounts of resources and can suffer from an inefficient and an unfair use of the wireless channel when they coexist with bursty data services. A lot of work has been done to support QoS on the Internet. However, none of these works can be directly used in MANET due to their specifics. Therefore, new specific QoS solutions need to be developed taking into account the dynamic nature of ad hoc networks. Since ad hoc networks should deal with the limited radiorange and mobility of their nodes, we believe that the best way to offer QoS is to integrate it in routing protocols. Such protocols will have to take into consideration QoS requirements, such as delay or bandwidth constraints, in order to select the adequate routes.

In this paper, we present a complete solution to the QoS routing problem based on an extension of the AODV (Ad hoc On Demand Vector Distance) routing protocol

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 27–41, 2003.

[2]. This solution consists of tracing routes in a reactive way by taking into account the QoS requirements (in terms of bandwidth, delay or both) associated with each flow. This work is inspired from the proposal of QoS extensions made in [3] in which we add QoS loss notifications, and delay/bandwidth measurements. The delay (resp. available bandwidth) are measured based on MAC and PHY layer specifics. Based on these measurements on each node/link, an end-to-end cumulative delay or available bandwidth can be estimated, which will enable route selection. This selection uses only the QoS extensions proposed for AODV. No additional signaling is required.

The rest of the paper is organized as follows. In section 2, we introduce the ad hoc routing issues. Section 3 abstracts the AODV protocol. Section 4 describes our solution and section 5 summarizes simulation results. Section 6 concludes the paper and presents some perspectives.

2 Qos in Ad Hoc Networks

The quality of service in ad hoc networks can be introduced in several interdependent levels [4]:

- At the *medium access protocols (MAC) level*, by adding QoS functionalities to the MAC layer in order to offer guarantees [5];
- At the *routing protocols level*, by looking for more performing routes according to various criteria (in this study we are interested more particularly by this approach);
- At the *signaling level* with routing protocol-independent resource reservation mechanisms. The QoS at the signaling level is responsible of the coordination between other QoS levels as well as other components, such as scheduling or admission control (cf. Figure 1).

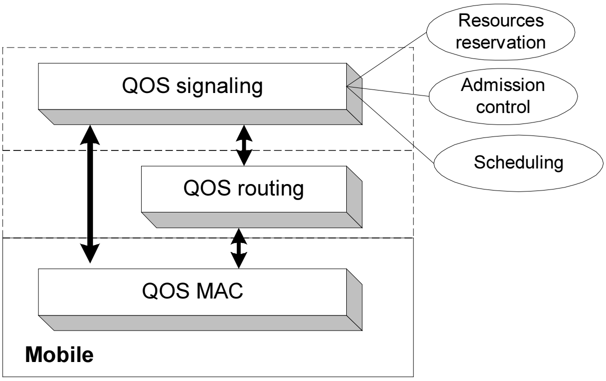


Fig. 1. QoS Model.

The QoS routing objective is to find a route with enough available resources to satisfy a QoS request. Resource reservation on the optimum route, evaluated by the routing protocol, is generally carried out by the signaling layer. The QoS routing in ad hoc networks can be introduced from existing ad hoc routing protocols like AODV, by extending it with the help of mechanisms that allow differentiating end-to-end paths according to chosen metrics (*delay*, *throughput* or *cost* [1]). The advantage of such a solution is to avoid a systematic overhead when QoS is not required.

Among the proposed QoS models, we distinguish a class of solutions called "*soft QoS*" [6]. The basic idea is that if the QoS is guaranteed as long as the path remains valid, it is possible to tolerate, depending on application requirements, transition periods that correspond to route reorganizations. During these periods, the service is only "best effort". This class of solutions seems to be the most suitable for ad hoc networks allowing to offer QoS with a reduced complexity and overhead.

3 AODV Protocol

AODV is a reactive ad hoc routing protocol which uses a broadcast route discovery mechanism. When a route is established, the nodes which are not concerned with the active path do not have to maintain routing tables or to take part into the route-update process.

3.1 Route Discovery

Each node maintains a temporary routing table with an entry for each active route that contains:

- destination IP address;
- destination sequence number;
- hop count (number of hop to the destination);
- next hop;
- list of precursors;
- lifetime of the route.

The route discovery process is initiated whenever a source node needs to communicate with another node for which it has no routing information in its table.

The source node initiates path discovery by broadcasting a route request (RREQ) packet to its neighbours. The RREQ packet contains the following fields: < source addr; source sequence number; broadcast id; dest addr; dest sequence number; hop count >

The pair < source addr; broadcast id > uniquely identifies a RREQ (the source broadcast id is incremented each time it issues a new RREQ). If a node has already received a RREQ, it drops the redundant RREQ and does not rebroadcast it.

When a node receives a new RREQ, it looks in its route table for the destination. If it does not know any route or a fresh enough one (the dest sequence number received

[1] Number of hops, resources requested for each node, utilization ratio of the links, etc.

in the RREQ is greater than the destination sequence number stored in the table), the node rebroadcasts the RREQ to its own neighbours after increasing the hop count.

If it knows a fresh enough route or if the node is the destination, the node stores the new information transported by the RREQ and sends a route reply (RREP) back to the source.

Insofar as the destination node replies to the first received RREQ, only one end-to-end route will be established.

A RREP packet contains the following information:

< source addr; dest addr; dest sequence number; hop cnt; lifetime >.

When an intermediate node receives back a RREP, it updates his table and forward the packet to the source which begins to send data after the first received RREP. The source node will change the route if a new RREP teaches him a better one (greater destination sequence number or lower hop count).

To set up a reverse path and then to be able to forward a RREP, a node records the address of the neighbour from which it received the first copy of the RREQ. These reverse route entries are maintained for at least enough time for the RREQ to traverse the network and produce a reply to the sender.

In the same way, nodes have to store the direct route. As the RREP travels back to the source, each node along the path sets up a forward pointer to the node from which the RREP came.

Figure 2 shows a route discovery with a RREQ broadcast when no intermediate node has a valid route. Figure 3 recall the reply to the first RREQ received by the destination. Note that nodes, B, C…, which are not involved on the initiated route do not have to maintain a routing table.

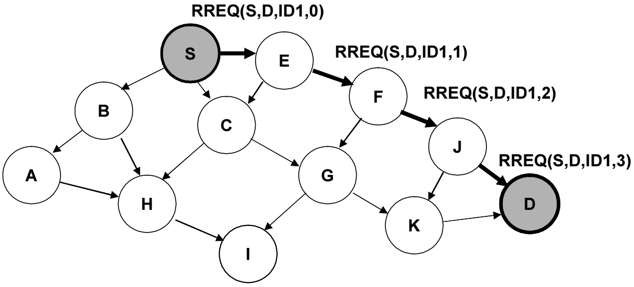


Fig. 2. Route discovery broadcast initiated by the source

3.2 Route Maintenance

If the initiated route breaks, due to node movement or failure, during an active session, the source has to reinitiate the route discovery procedure to establish a new route towards the destination. Periodic hello messages are used to detect link failures.

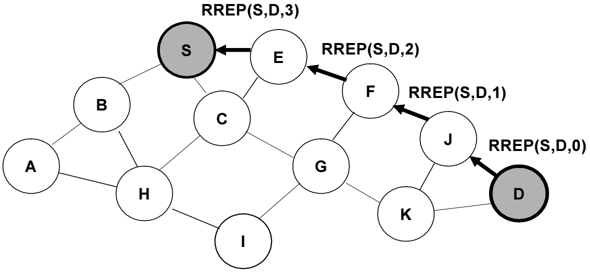


Fig. 3. Route reply unicasted from the destination

When a node detects a link break for the next hop of an active route (or receives a data packet destined to a node for which it does not have an active route), it sends a Route Error packet (RERR) back to all precursors. The RERR packet contains the following fields:

< unreachable dest ; unreachable dest sequence number >

When a node receives a RERR from a neighbour for one or more active route, it must forward the packet to the precursors stored in its table. Routes are erased by the RERR along its way. When a traffic source receives a RERR, it initiates a new route discovery if the route is still needed.

4 QoS Extension for AODV

The QoS routing solution we propose uses two metrics: the delay and the available bandwidth. The QoS route is traced node by node using AODV QoS extensions [3]. For each crossed node, an estimate is made to know whether the maximum delay or minimum bandwidth requirements could be satisfied. If not (i.e. in the case where the delay estimate remains too long at an intermediate node or the available bandwidth too weak on a selected link), the route search will be interrupted. Thus, the QoS routing remains reactive, using only extensions on the AODV request (RREQ) and reply packets (RREP).

4.1 Delay Estimation

The delay estimation uses one of the existing AODV parameters: the NTT (NODE_TRAVERSAL_TIME), initially considered as a constant [2]. In our proposal, the NTT becomes an estimate of the average one hop traversal time for a packet. It includes the transmission delay over the link and the processing time in the node (delays in queues, processes interruption time, etc).

As shown in Figure 4, the NTT parameter for node B is divided on 2 parts:

$$NTT_B = d_{AB} + t_{T\,B} \tag{1}$$

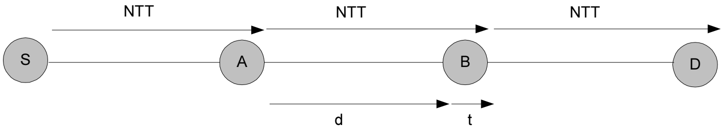


Fig. 4. NTT estimation.

d_{AB} corresponds to the transmission delay between two adjacent nodes introduced by MAC and PHY level operations. For example, on an IEEE 802.11 [7] network, the transmission delay (d_{AB}) is due to the durations of frame transmission (RTS, CTS, data, ACK); to the inter-frame spacing (DIFS, SIFS), to propagation delays and to contention resolution (including possible retransmissions due to collisions).

As numerous MAC level protocols for ad hoc networks uses frame acknowledgments to ensure that no collision occurs during a frame transmission, we can define d_{AB} as the time difference between the time the packet is handled by the MAC layer in the source node and the time its acknowledgment is transmitted back by the destination node.

$$d_{AB} = T_{ACK} - T_{transmission} \tag{2}$$

In order to keep only one time reference for the source node [8], we can take into account the propagation delay, between two nodes, for the acknowledgement. This parameter is a constant and its value depends on PHY layer specifics.

$$d_{AB} = T_{ACK\_reception} - T_{transmission} - T_{propagation} \tag{3}$$

For the NTT calculation at the destination node, d_{AB} can be sent with another AODV extension.

The choice of doing the delay measurement using only RREQ and RREP packets rather than all data and routing packets is motivated by the processing overhead which is reduced when using passive measurement. Note that, the obtained delay d_{AB} depends closely on the packet size. A correction should therefore be made in order to take into account an average size instead of the RREQ or RREP packet lengths used for such measurements. For example, with a control-packet length of 32 bytes and with an average length of 100 bytes for data packets sent at 11Mbit/s, the correction could bee:

$$d'_{AB} = d_{AB} + \frac{(100-32)\times 8}{11.10} \tag{4}$$

Insofar as route delays depend on unpredictable events (node movements, arrivals, extinctions, variations of streams and traffic, etc.), the variance of node-to-node delays can be significant. Two methods exist in order to take these delay variations into account [9]. The first one calculates an average according to a fixed size window. The second method consists of calculating an average, weighted by a forgetting factor (*exponential forgetting*). As our aim is to minimize the overhead, the second method is naturally more suitable. The delay between nodes *A* and *B* is then given as follows:

$$d_{AB}(t)=(1-\lambda)\sum_{k=0}^{\infty}\lambda^{k}.d_{AB}(t-k) \tag{5}$$

where $\lambda\in$ [0,1] is the forgetting factor.

The processing time in the node (t_{TB}) includes a node-specific constant (corresponding to the processing capability of the packet at the different levels) and a variable delay, function of the packet number in the queue. A first estimation is done by computing the average number, over a sliding window, of the queued packets.

The length of the window is based on another specific AODV parameter: ACTIVE_ROUTE_TIMEOUT. This first estimation gives satisfactory results and has to be compared to other more complex queuing delay estimators. This comparison is outside the scope of this paper and is a subject of a future work.

Note that to estimate the end to end delay, we take into account the processing time in the source node. For the destination node, as there is no forwarding, the queuing delay is not considered.

4.2 Bandwidth Estimation

An estimate for the available bandwidth on a link can be formulated as follows [10]:

$$BW_{available} = (1\text{- } u) \times Throughput_{on\ the\ link} \tag{6}$$

where u represents the link utilization.

To calculate the available bandwidth for a node, the link throughput must first be evaluated. An initial evaluation can be done simply by emitting packets and measuring the corresponding delays:

$$Throughput_{on\ the\ link} = \frac{S}{d_{AB}} \tag{7}$$

S being the packet size and d_{AB} the transmission delay between two adjacent nodes defined above. As for the delay estimation, it is necessary to limit the random aspect of the measurement. *Exponential forgetting* can also be used to calculate the average available bandwidth.
The link availability ($1\text{-}u$) is evaluated by the following formula:

$$1-u=\frac{idle\ times\ in\ window}{window\ duration} \tag{8}$$

where '*idle times in window*' is the sum of all transmission idle times measured during a time sliding window of width '*window duration*'. The '*window duration*' is set to ACTIVE_ROUTE_TIMEOUT. Note that, this computation is done only if there are active routes stored in the node's routing table. Otherwise, $u=0$.

4.3 QoS Routing

For each route entry corresponding to each destination, the following fields are added to the routing tables:

- Maximum delay;
- Minimum available bandwidth;
- An extension is foreseen by AODV for its main packets RREP and RREQ (cf. Figure 5).

- 8 bits	- 8 bits	- n bits
- Type	- Length	- Type-specific data…

Fig. 5. AODV Extension format.

- Depending on the packet type, a "delay" extension can have two meanings:
- For an RREQ packet, it means the delay allowed for a transmission between the source (or an intermediate node forwarding the RREQ) and the destination;
- For an RREP packet, it means an estimate of the cumulative delay between an intermediate node forwarding the RREP and the destination.

Thus, a source requiring maximum delay constraint transmits a RREQ packet with a QoS delay extension. Before forwarding a RREQ packet, an intermediate node compares its NODE_TRAVERSAL_TIME with the remaining delay bound indicated in the extension. If the delay bound is inferior, the packet is discarded and the process stops. Otherwise, the node subtracts its NTT from the delay bound provided in the extension, updates the QoS delay extension, and propagate the RREQ as specified by AODV (cf. section 3.1).

In the example of Figure 6, each node in the route satisfies the comparison and the requested delay at the destination (50ms-10ms) remains greater than zero.

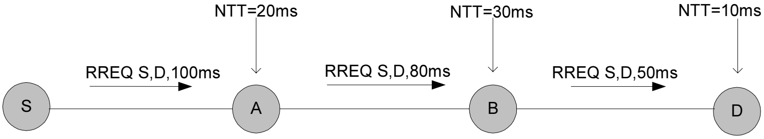


Fig. 6. Example of QoS delay request.

In response to a QoS request (RREQ), the destination sends an RREP packet (cf. Figure 7) with an initial delay corresponding to its NTT. Each intermediate node adds its own NTT to the delay field and records this value in the routing table for the concerned destination before forwarding the RREP. This entry update allows an intermediate node to answer the next RREQ simply by comparing the maximum delay fields of the table with the value of the transmitted extension. The answer of the intermediate node is always valid in time because the old routes are deleted from the table according to the ACTIVE_ROUTE_TIMEOUT parameter.

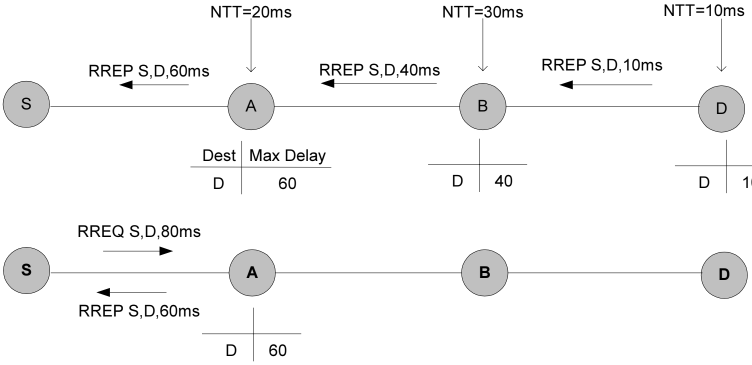


Fig. 7. Examples of QoS delay responses.

For a "bandwidth" extension, the principle remains the same. A source requiring a bandwidth constraint transmits a RREQ packet with QoS bandwidth extension. This extension indicates the minimum bandwidth having to be available on the whole path between the source and the destination. Before forwarding the RREQ packet, an intermediate node compares its available bandwidth to the bandwidth field indicated in the QoS extension. If the bandwidth required is not available, the packet is discarded and the process stops.

In response to a QoS request, the destination sends a RREP packet with its measured available bandwidth. Each intermediate node, forwarding the RREP, compares the bandwidth field of the extension with its own available bandwidth on the selected route and keeps the minimum between these two values to propagate the RREP. This value is also recorded in the routing table for the concerned destination. It indicates the minimum available bandwidth for the destination (see example on Figure 8). This information remains valid as long as the route is valid (lifetime < ACTIVE_ROUTE_TIMEOUT).

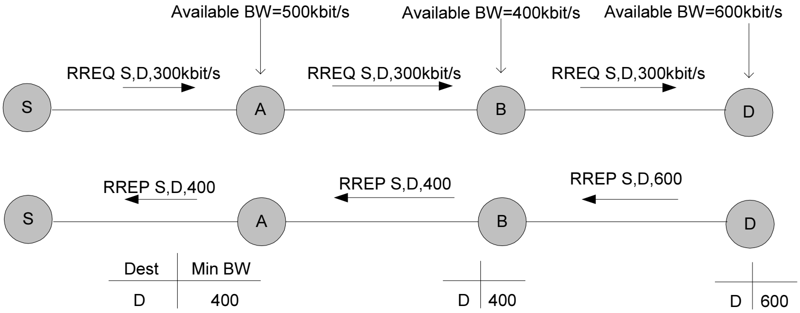


Fig. 8. Example of QoS bandwidth request and response.

If the QoS request concerns both delay and bandwidth, the two extensions can be appended to the same request and reply packets. In this case, both maximum delay and available bandwidth verifications of request (RREQ) and reply (RREP) will be applied simultaneously. RREQ packets are discarded if one of the constraints cannot be satisfied.

To prevent an eventual variation of the NTT on a node and a possible lost of QoS, a predefined QoS Delay Margin (says QDM) can be introduced. A route error packet (RERR) is generated when an intermediate node detects an increase in its NTT that is greater than QDM. The RERR packet is also generated if the node detects a decrease in its available bandwidth that is greater than a QoS Bandwidth Margin (says QBM).

Note that, if the margin is chosen too large, the source node will never be informed of QoS loss. Conversely, if the margin is too small, useless RERR packets can be generated causing new RREQ broadcast. This undesirable control packet transmission induces an undesired overhead, slowing down data packet exchanges, even if the QoS constraint is initially respected. So, an accurate dimensioning of these margins is very important.

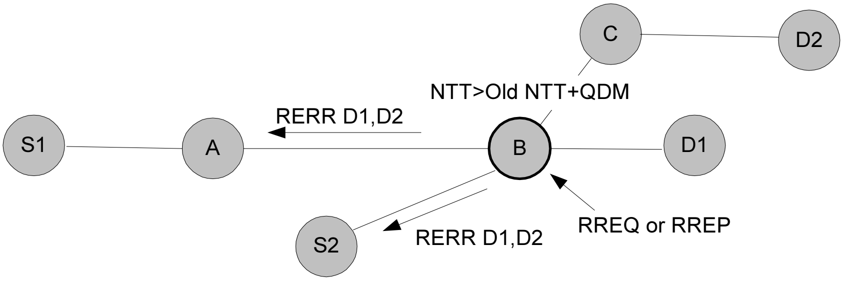


Fig. 9. Example of QoS delay lost.

As for standard AODV route error mechanism, the RERR packets are sent to all the precursors stored for all the routes (cf. Figure 9). Note that the NTT or/and the available bandwidth are measured each time a RREQ or a RREP packet is received by a node, which generally corresponds to a change of the network and traffic load (new source, node mobility…) producing a possible loss of QoS.

5 Performance Evaluation

In order to evaluate the performances of our QoS routing protocol, we simulate the proposed mechanisms using NS-2 [11] extended by a complete implementation of IEEE 802.11 [12].

The radio model allows a bit rate of 2 Mbit/s and a transmission range of 250 m. The number of mobile nodes is set to 20 or 50 nodes giving two simulation sets. These nodes are spread randomly in a 670×670m area network and they move to a random destination every 30 s with a speed randomly chosen between 0 and 10 m/s. Simulations run for 300 s. Traffic sources are constant bit rate (CBR) sending 8 packets of 512 bytes per second.

Several simulations are realized by varying the source node percentage from 10% to 100%. The QoS constraint is set to 100ms for delay and 100kbit/s for bandwidth. For the QoS loss detection mechanism, the optimal margins are 30ms for QDM and 30kbit/s for QBM. First simulation results shows that these values are optimal and give a good compromise between overhead due to RERR generation and QoS loss detection.

The delay estimation uses a constant processing time in the node equal to 3 ms, a first order (*k*=1 in (1)) variance correction is applied and the optimum forgetting factor λ is set to 0.2. ACTIVE_ROUTE_TIMEOUT is a constant set to 10s. The performance of our algorithm has been evaluated by measuring the average end to end delay, the average throughput, and the overhead induced.

Figures 10 and 11[2] present the average end to end delay for data packets on all QoS routes when the number of QoS sources varies. On a 20-nodes network, the delay remains lower than 100ms with QoS routing whatever the number of sources. On a high density (50 nodes), the QoS constraint of 100ms is respected with QoS routing if the number of sources does not exceed 70%. Without QoS, the delay can reach several seconds (Figure 11).

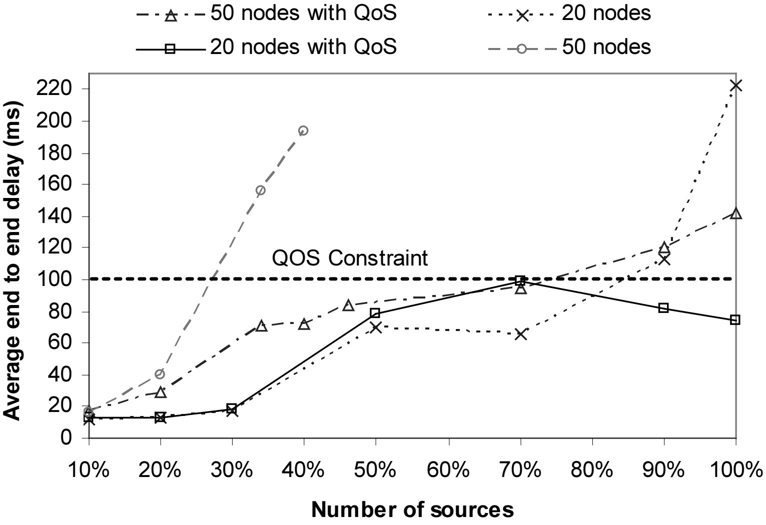


Fig. 10. Average end to end delay / Number of sources.

The overhead [3] due to the AODV control messages is slightly higher when using QoS extensions (cf. Figure 12). The increase of this overhead when using QoS extensions

[2] In Figure 11 the scale is given in *s*, rather than *ms* (Figure 10), for better visibility.

[3] The overhead is computed as the bandwidth percentage consumed by the control packets (RREQ, RREP, and RERR).

remains lower than 6% whatever the density of the network and the number of sources.

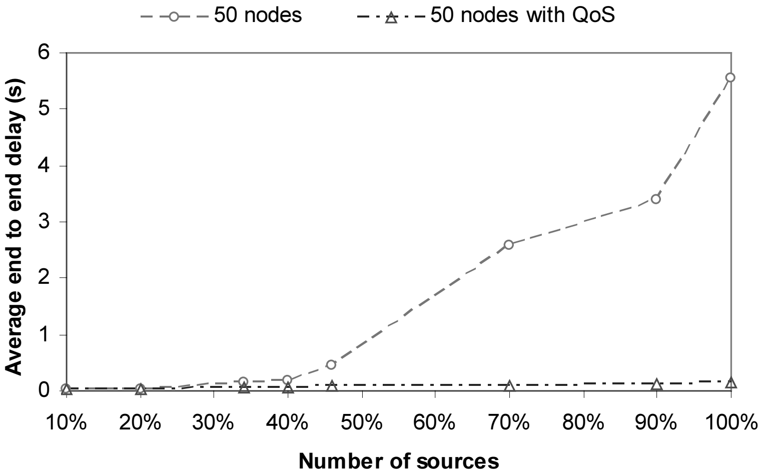


Fig. 11. Average end to end delay / Number of sources (50 nodes).

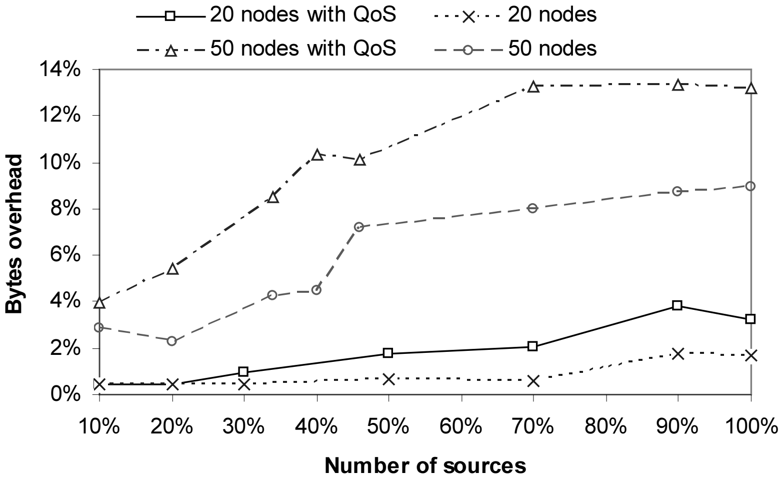


Fig. 12. Overhead / Number of sources (delay constaint).

Figure 13 presents the average throughput on all QoS routes when data packets are sent from a source to a destination. For a 20-nodes network, the bandwidth constraint is respected whatever the number of sources. On a high density network (50 nodes), the QoS routing becomes more efficient when the number of sources, and then the traffic, increases. Note that without QoS extensions, the throughput becomes quickly very weak when the number of sources increases (under the 100kbit/s requested when there are more than 20% of sources).

The overhead is higher when using the QoS extension especially for a small number of QoS sources (cf. Figure 14). The traffic generated by AODV control packets, and particularly RERR packets, is relatively important in this case. For a dense network with high load, the overhead becomes equivalent. This is a logical result since all new demands are quickly rejected, even by the first encountered node: the overhead is then considerably reduced.

The last results show that the QoS routing algorithm with bandwidth extensions is more suitable for a high density network with an important traffic.

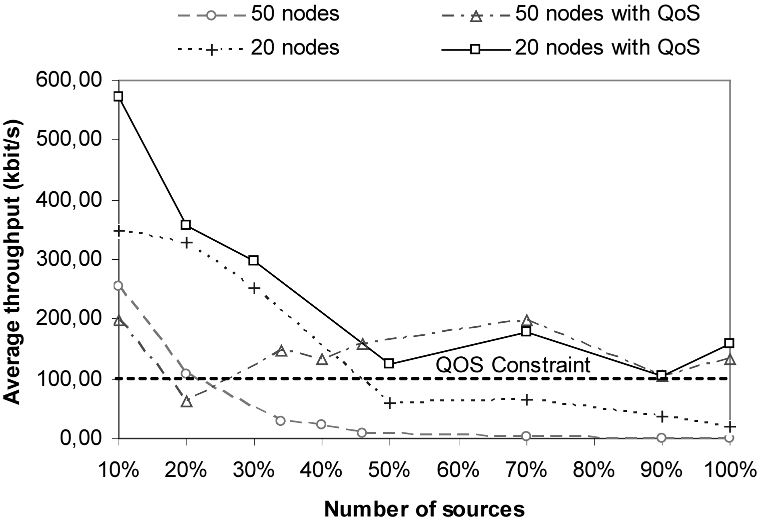


Fig. 13. Average throughput / Number of sources

6 Conclusion

In this paper, we have proposed and evaluated a QoS routing solution based on AODV. This solution uses delay and bandwidth measurement and preserves the reactive nature of AODV.

[4] The overhead is computed as the bandwidth percentage consumed by the control packets (RREQ, RREP, and RERR).

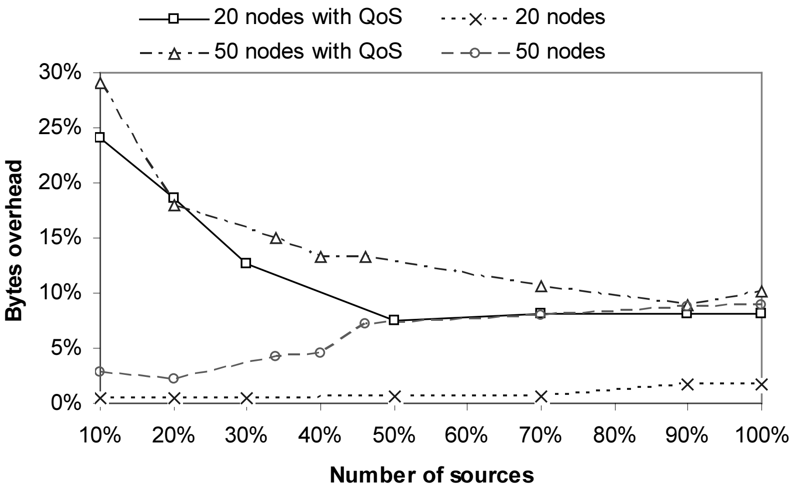


Fig. 14. Overhead / Number of sources (bandwidth constaint).

The QoS routes are traced node by node and the proposed routing algorithm uses extensions of the AODV request (RREQ) and reply (RREP) packets. The delay and bandwidth measurements are initiated only on RREQ or RREP arrivals in a node (these times correspond to a network state change: arrival of a new flow). Note that measurements on each data or routing packet would increase the overhead unnecessarily. Corrections are however made in order to take into account variations due to the dynamic nature of ad hoc network and network traffic.

The proposed QoS routing with QoS loss notification gives satisfying results, especially for the delay extension. For the bandwidth extension, good performances are obtained for high density networks with an important load. New algorithms for bandwidth measurement and queuing-delay estimation are currently under study. This study will allow choosing the best estimators to run with the proposed QoS routing protocol.

References

1. IETF MANET WG (Mobile Ad hoc NETwork) www.ietf.ora/html.charters/manet-charter.html
2. C. E. Perkins, E. M. Royer, and S. R. Das, “Ad hoc on-demand distance vector routing,” Internet Draft, 2002
3. C. E. Perkins, E. M. Royer "Quality of service for ad hoc on-demand distance vector routing". IETF Internet Draft
4. Kui Wu and Janelle Harms “QoS Support in Mobile Ad Hoc Networks” Computing Science Department University of Alberta

5. A. Veres, Campbell, A. T, Barry, M and L-H. Sun, "Supporting service differentiation in wireless packet using distributed control", IEEE Journal of Selected Areas in Communications (JSAC) Vol. 19, No. 10, pp. 2094–2104, October 2001
6. H. Xiao, K.G. Seah, A. Lo and K.C. Chua, "A flexible quality of service model for mobile ad-hoc networks", IEEE Vehicular Technology Conference (VTC Spring 2000), Tokyo, Japan, May 2000, pp. 445–449
7. LAN MAN Standards of the IEEE Computer Society. “Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification”. IEEE Standard 802.11, 1997
8. Kay Römer, "Time Synchronization in Ad Hoc Networks." In Proceedings of MobiHoc 2001, ACM, Oct. 2001
9. P. Gupta and P. R. Kumar "A system and traffic dependent adaptive routing algorithm for ad hoc networks" , Proceedings of the 36th IEEE Conference on Decision and Control, pp. 2375–2380, San Diego, Dec. 1997
10. M Kazantzidis, “End-to-end versus explicit feedback measurement in 802.11 networks,” Technical Report N° 010034 UCLA Computer Science WAM Lab,. 2001
11. The Network Simulator - NS-2: http://www.isi.edu/nsnam/ns/
12. The Monarch Project : http://www.monarch.cs.rice.edu/

User Mobility Model in an Active Office

Teddy Mantoro and Chris Johnson

Department of Computer Science, Australian National University,
North Road, ACT 0200, Australia
{teddy,cwj}@cs.anu.edu.au

Abstract. User mobility in an Active Office represents human activity in a context awareness and ambient intelligent environment. This paper describes user mobility by detecting their changing locations. We have explored precise, proximate and predicted user location using a variety of sensors (e.g. WiFi and Bluetooth) and investigated how the sensors fit in an Active Office to provide interoperability to detect them. We developed a model to predict and proximate user location using wireless sensors in the Merino layering architecture, i.e. the architecture for scalable context processing in an Intelligent Environment.

1 Introduction

The Intelligent Environment (IE) is an environment with rapid and rich computing processing. Ubiquitous and ambient intelligent computing by embedded sensing devices within the environment will be the key point in IE because of the significant processing done by sensors. As a consequence, the environment provides information on anywhere-anytime devices so that a user may access it on an anywhere-anytime basis too.

Unfortunately, computers and communications systems today are underused because the range of control mechanisms and application interfaces are too diverse. Context awareness mechanisms could be the best way to implement computer applications in IE. It is necessary to consider the mechanism that might allow users to manipulate systems in simple and ubiquitous ways and to make computers more aware of the facilities in their surroundings [4],[15].

This paper discusses an important problem in ubiquitous and ambient intelligent computing, namely how to handle users' mobilities in an Active Office. This work contributes to a model is given to predict and proximate users' locations to understand user mobility in an Active Office using wireless sensor and scalable context processing.

There are significant challenges. These are making use of already existing devices which enable the design of systems and software, ease of deployment, network and sensor scalability, through sensor fusion rather than in precision, and enabling wider acceptance through better design for user needs, by human factored interfaces and increased human trust in the system's care for user privacy and security.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 42–55, 2003.

In a context awareness mechanism, the question of Who, What, Where and When in IE is responded to by Identity, Activity, Location and Timestamp. In this system the scope of 'who' will be user identity, persona, profile, personalisation/ internationalisation and user model. 'What' is responded to by users' activity that can be represented by user mobility, i.e. the changing of a user's location to another location. Thus, a context awareness mechanism, based on user's locations and time stamps, could deduce a user's activity.

These awareness-mechanisms bring the computer into the user's daily activity. We explore computer capabilities to recognise user location, activity and social context as defined by the presence of other people and to assist people with the variety of activities at work.

This paper describes an Active Office as an implementation model of an IE. An Active office is defined as a normal office, which consists of several normal rooms with minimal additional decorations (minimal intrusive detectors and sensors), without badging people. An Active Office uses wireless communication i.e. Bluetooth and WiFi to enable user mobility. In this paper we look at the use of Wifi rather than the use of Bluetooth.

The Active Office uses a scalable distribution context processing architecture (Merino service layers architecture) to manage and respond to rapidly changing aggregation of sensor data. The Merino architecture has four tightly coupled layers to manage the interaction between users and the environment in the Active Office by detecting the movements and transformations between the raw physical devices/sensors within the environment and the application programs. The infrastructure supports interoperability of context sensors/widgets and applications on heterogeneous platforms [2], [5]. In order for an Active Office to provide services to users, an Active Office must be able to detect its current state/context and determine what actions to take based on the context [2], [5].

Location awareness is the most important part of context awareness in an Active Office. The off-the-shelf availability and everyday use of a number of moderate-cost mobile devices (e.g. handheld, laptop computers), installed wireless and wired networking, and associated location information, lead us to focus our attention on context awareness computing that rests lightly on our everyday environment, and to ask in particular: "What effective location awareness computing can be achieved with minimal, unobtrusive, commodity hardware and software?"

In an Active Office, the terms of changing user location need to be clearly defined, since we are not using coordinate mapping for the Active Office. The user is considered to be not significantly moving when he is typing on his computer, or opening the drawer of his desk. The user needs to changed location significantly to be considered as his having change location. "Significant" means moving from room to room or moving from one side of the room to the centre or to the other side. We explore changing user location to find the user moving from one location to another location to explain user mobility. In an Active Office we also consider two important variables: speed and location resolution. We will discuss this problem in detail in section Location Characteristics in an Active Office.

This paper describes user mobility in an Active Office as a representation of human activity in the context awareness environment. First, we will present an overview of Merino service layers architecture, followed by a description of a distributed system architecture in an Active Office, mobility in an Active Office, location user model, and precise, proximate and predicted user location. Then, discussion of user mobility will be followed by a conclusion and further research topics arising from this work will be suggested.

2 Merino Service Layers Architecture

Merino Architecture has two important parts. The first part is the managing of the interoperability between physical devices/sensors within the IE environment and the program application which manages the interoperability between user and environment. This part contains its abstraction layers i.e. the core Sensors and Device Layers, Context and Device Abstraction Layers, and the highest abstraction level i.e. the Smart Environment Agent Layers [5].

The Sensor Layer is the innermost service layer that represents the range of sensors, which detect the physical environment. The next layer is the Context Layer, which performs core tasks of filtering and aggregating the raw data from the Sensor Layer and ensures maintenance of the interoperability between sensors.

The second part of Merino Architecture consists of a Context Repository and a User Model. A Context Repository is a key element that unifies and manages the whole object in the environment. The Context Layer interacts with the Context Repository. This handles an object's global name structure and it manages the naming/subnaming authority.

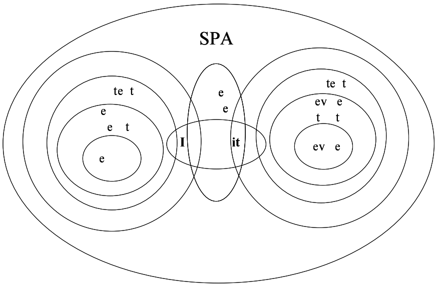


Fig. 1. Merino service layers architecture for an Intelligent Environment

A User Model overlaps with a Context Repository. The Merino Architecture treats them almost similarly, the main difference is that the User Model is about people only. The User Model contains personal identity, security and privacy, to confirm the emerging environment in a local and a global space, whereas the Context Repository holds context information from the Context Layer. Data from the Context Repository, which is associated with the user, is stored in the User Model. For example, data from

a movement detector, a WiFi or a Bluetooth detector, and a keyboard activity monitor are held in the Context Repository. However, once it is associated with an individual user, it will be kept in the User Model.
The devices and Device Abstraction Layer are motivated by the need to send data to low level devices in the IE including attributes of the device e.g. a phone number, or a Bluetooth MAC address in a mobile phone. A Smart Environment Agent may communicate with the phone to change its state, e.g. to request switching to a silent mode during a meeting.

3 Distributed System Architecture in an Active Office

The key role of a distributed context processing in IE is an IE domain. An IE domain is an administrative domain, which at least contains an IE repository, a resources manager, knowledge based and various sensors (above the level of dumb sensor that communicates using standard protocol).

A resources manager uses sensors/widget agents to have direct communication between sensors/devices and the IE repository. A resources manager also distributes sensor data or aggregate sensor data to another IE repository in the other IE domain.

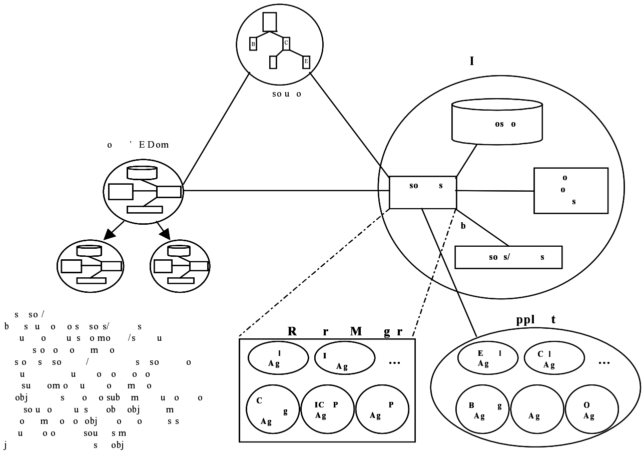


Fig. 2. Distributed Architecture for Active Office

The structure of an IE domain can be hierarchical (it might have a single higher level IE domain or have several lower level IE domains) or scattered without level (Figure 2). The communication between IE domains is based on a request from each resources manager (peer-to-peer basis). The resources manager will send a request to the resolution server to require information about the object. The resolution server accepts the object's global name and uses persistent mapping from the object's global name to send the persistent location (persistent URL) of the object to the resources manager. Finally, the resources manager sends a request to a relevant resources manager in another IE domain to obtain the object including the characteristics of the object.

Every object in the IE domain is defined to be self-describing object data-structures. This means that the structure, as well as the value, is always included in the object's content. This approach has two advantages: firstly, it avoids the complexity of separate schemes and maintains an object across multiple distributed servers. Secondly, it permits meta-tools e.g. browsers to manipulate the object without knowledge of the specific contents [11].

Communication within the IE domain uses logical multicast. This means that communication between the IE repository and the sensors could be implemented using IP multicast on a well-known group 'content filtering' in the receivers or message server distribute content routing/filtering (such as Elvin or Spread) for an automatic update of the content object purpose [13]. A particular object has a unique identifier and is located using a search that starts in the local IE domain, and expands to adjacent and higher lever IE domains such as a resolution server (using persistent URL or Handle system).

For example, Anna works for the CS department. When she goes to the department of Engineering the sensor at the CS department would aggregate that Anna is not present and another sensor at the department of Engineering would detect an unknown object and ask its local resources manager for the address of the server for this object. The local resources manager multicasts the request to the resolution server and discovers the correct home server for the object. The object's location information in the home server is then updated appropriately.

An IE repository keeps all object information including the characteristic of the object. It holds data about any rapid changing of the object. It contains all relevant data within the IE domain such as the sensor data, the aggregate sensor data, the context data and the rich context data. The IE repository uses a distributed object database for Context Repository and User Model purposes.

A resources manager is a network management map application that provides information about the status of objects (devices/sensors) in the IE network. The resources manager detects, controls, manages and concludes the functionality of all objects in the IE domain.

The resources manager will show any failures in low level devices i.e. sensors, access points, hubs, bridges and routers. Furthermore, the resources manager will detect any traffic problems and will identify the what, the where, the cause and the time-length of the problems. The resources manager has the capability to control and manage the functionality of an IP network using the Internet Control Message Protocol (ICMP) and 'trap' mechanism using the Simple Network Management Protocol (SNMP), so that the fail-safe mechanism can be implemented.

To set up the resources manager a set policy (knowledge-based) needs to be established that determines the acceptable levels of traffic, broadcasts and errors on any devices/sensors at any segment.

The knowledge-based context contains context rules of the interaction between the user and the Active Office environment. Rules are used to represent heuristic, or "rules of thumb" which specify a set of actions that need to be performed in a given situation. A generic rule is composed of an *if* portion and a *then* portion. The *if* portion of a rule is a series of patterns which specify the facts where the rule would be applicable in the IE domain. The process of matching the facts to the patterns (pattern-matching) is done by a specific agent in the resources manager called an inference engine agent. The inference engine agent will automatically match the facts against the patterns and determine which rules are applicable to the context. The *then* portion is a set of actions that needs to be executed when the rule is applicable to the situation (context). The actions of applicable rules are executed when the inference engine agent is being instructed to begin the execution. The inference engine agent selects the rules, and then the actions of the selected rule are executed. The process continues until there are no more applicable rules.

The knowledge-based context also has a dynamic knowledge cooperation aspect [3] to allow the rules and the sensor/devices to dynamically affect the actions of the service selection process within the IE domain or the inter-IE domain.

In the IE domain, all objects can be distributed to other physical locations or logical locations. Every object has a global resolution name and the global resolution name needs to be registered in the resolution name server. IE resolution is a global resolution server, which can be implemented using URN, Persistent URL (PURL), or Handle Server for persistent global mapping/resolution purposes. The resolution mechanism is identical to the DNS for the Naming Authority Pointer (NAPTR) Resources Record (RR) to delegate the lookup's name [1],[10]. We use the Unique Resolution Name (URN) as a global unique name and the Unique Resolution Locator (URL) as a locator to locate any object, anywhere, anytime.

4 User Mobility in an Active Office

A location is the most important aspect to provide a context for mobile users, e.g. finding the nearest resources, navigation, locating objects and people. A location in the context awareness application needs a model of the physical environment and a representation of the location [4], [7], [12].

This paper explores user mobility in an Active Office. We began from understanding user location, then changing location from a current location to another location. By analysing the history data we can get the pattern of the user mobility. We strongly believe that by understanding user mobility we can better understand user activity.

In the following part we will explore the user location model and location characteristics in an Active Office to match the location model in an Active Office.

4.1 User Location Model

Numerous location models have been proposed in different domains, and can be categorized into two classes:

- Hierarchical (topological, descriptive or symbolic).
- Cartesian (coordinate, metric or geometric).

The other issue is the location of representation. Most context awareness applications adopt a distributed collaborative service framework that stores location modelling data in a centralized data repository. Location related queries issued by end-users and other services are handled by a dedicated location service. This distributed service paradigm is attractive because of its scalability and modularity. However, we need an effective and efficient location representation method to make this work [6].

We use hierarchical models for representing locations and enabling rapid changes of location information between distributed context services within space.

The hierarchical location model has a self-descriptive location representation. It decomposes the physical environment to different levels of precision. We use a tree structure to handle location structure and we store it as an object/entity in a relational database model.

Cartesian location imposes a grid on a physical environment and provides a coordinate system to represent locations. The GPS coordinate system that is defined by longitude, latitude, altitude or relative location in space using the Cartesian system can be used as a Cartesian location (var x, var y, var z), e.g. the user X is located at (18, 5, 15). In this paper, we are not going to explore the shape and other extension locations of the object.

Any simple mobile device has a physical location in space. In the spatial dimension, there are some devices (e.g., GPS-based map systems) where the exact Cartesian position in 2D or 3D space is important in defining a sense of absolute physical location. If the location information is sufficient in understanding the position, the location is considered in relation to other existing objects or sensors.

GPS does not work in an indoor Active Office environment because it has approximately 5-15 meters accuracy and the signal is too weak. It works only for an outdoor environment.

In the Wireless LAN environment, the location of mobile devices can be determined with spatial precision for a group of three or four individual offices, by measuring the signal strengths of a few of the most visible access points [14]. This accuracy is sufficient to support everyday tasks in the Active Office. Both IEEE 802.11b and Bluetooth can be used to find the proximity of physical user locations.

4.2 Location Characteristics in an Active Office

Since an Active Office is also a ubiquitous computing environment, we assume that sensor and actuators, simple push button and sliders, and computer access will be available in every area. The users can be identified by mobile computing devices (PDA/handheld), vision image recognition, or by active/passive badge. Users also can be identified by activity when accessing available registered resources at static locations.

In an Active Office we assume that the user has a regular work schedule, has some routine activity that can be used to predict his location in a specific timestamp.

User Location in an Active Office implies the ability of IE to understand a user changing location on a ‘significant’ scale. The user is considered to be moving if his location is changing. The term of “change location on a ‘significant’ scale” means that the change in user location affects the user’s access to available resources. For example, picking up the phone or picking up a glass from the table while the user is typing at the computer is not a significant location change. By ‘significant’ we mean when the user moves from room to room or moves from one side of the room to the centre, or to the other side, or he moves from one building to another.

The approximate significant location scale could be affected by two important aspects: speed and location resolution. A sensor continuously senses on a real time base, i.e. the speed of a user changing location in the office between 0-20 Km per hour and a location resolution of about 1-3 meters in the office. The slower the user changes location, the more the sensor detects and delivers service information about the location the user passes through.

The Active Office can be designed to understand ‘significant’ change in user location by using sensors that can measure proximate location. When the user moves, it means that the user’s access to the resources also changes. The availability of resources depends on the user location. For example, when the user moves from his room to a meeting room, the proximity or the availability of resources, e.g. a printer, may also change.

We believe that accessible resources for users in an Active Office can be detected based on their proximate location. We also believe that a hierarchical location model will be more relevant than a Cartesian location model because a hierarchical location model could scale room and building, while the technology for gridding/mapping the office using a Cartesian model is not available yet at the time of writing.

5 Sensor Aggregation for User Location

User locations in an Active Office can be categorised as follows:

1. Precise Location
2. Proximate Location
3. Predicted Location

The above category is based on the sensor’s capability in covering an area. The problem is to combine these known location data to determine the user’s actual location for office activity purposes which is a different precision matter and sometimes user location is not available at all at arbitrary times. How to characterise the pattern of user location based on aggregate current sensor data and history sensor data is also a problem.

5.1 Precise User Location

Precise location is based on sensors that cover less than a meter range, e.g. swipe card, keyboard activity, biometric sensor/finger-print, iButton, etc.

By registering the location of desktop computers, swipe cards, iButtons, fingerprints, as sensors, an Active Office will have more precise information about user locations than from proximate sensors, e.g. WiFi or Bluetooth.

5.2 Proximate User Location

Proximate location is based on a sensor that covers more than a meter range, e.g. WiFi, Bluetooth, WiMedia, ZigBee, active/passive badge (depending on the range), voice recognition (microphone), face recognition (digicam), smart floor, etc.

Proximate location is detected by Wireless LAN is an interesting proximate sensor in an Active Office because it can be used to access the network and also it can be used to sense a user location within the scale of a room or an office. Bluetooth, as a wireless personal area network, favours low cost and low power consumption, over a range and peak speed. On the other hand, WiFi as a wireless local area network, favours higher speed and greater range but has higher cost and power consumption. The range of the Bluetooth to sense another Bluetooth in a closed space, such as an Active Office, is about 3 meters for class 2 and 25 meters for class 1. The range of WiFi to sense a user with a WiFi device is about 25 meters.

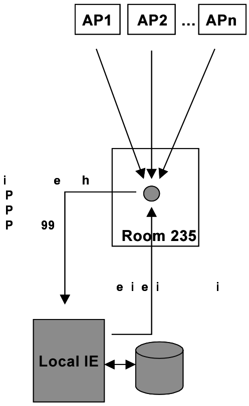


Fig. 3. Measure by devices of WiFi APs' signal strengths

Bluetooth permits scanning between devices: when Bluetooth capable devices come within range of one another, the location of one Bluetooth will be in the range of the other Bluetooth. We can use Bluetooth capable devices as sensors or an access point to sense a user with Bluetooth capable devices. From our experiment, unfortunately a Bluetooth signal strength is not useful enough to sense a user's location. We used the Bluetooth access point as a sensor for several rooms within the range without measuring any signal strength. For example, when a user is close to a certain access point, the user's location will be proximately close to the access point and it could represent user location from several rooms.

WiFi does not only have a higher speed and longer range than Bluetooth but the signal strength of Wi-Fi also can be used to detect user location. We have two scenarios to determine user location using WiFi. Firstly, by determining the signal strength from the WiFi capable device which stores data in a local IE repository with the server sending the current user location. Secondly by determining the signal strength from the WiFi access points and storing the signal strength data in the local IE repository with the server sending the user location. The difference between these scenarios is that in the first scenario the process of sensing is in the devices, so we need a good user's mobile device, whereas in the second scenario the process of sensing is in the access point, so we do not require a user's mobile device with a high capability.

We use a self organizing map (Kohonen map) approach of artificial neural network to cluster the signal strength data by giving random weight, doing normalisation, calculating Euclidian distance, finding the winner for clustering. The training is continued until the learning rate is zero and the final weight is obtained, which leads to output activation and cluster winner allocation [8]. A self organizing map using a Kohonen map is suitable to cluster locations based on signal strength data measure in the local IE. Using this method we can directly determine current user location.

In our experiment using WiFi, we used 11 access points to measure signal strength on two buildings, 5 in the Department of Engineering and 6 in the Department of Computer Science. The result was good enough to predict current user location. On the 2nd level of the Department of Computer Science, we found that most places had a good signal from more than two access points and we could predict accurately (96%) in rooms of 3 meters width. On the 3rd level, where not all rooms/locations were covered by more than one access point, we had only a reasonable degree of accuracy (75%) in predicting a user's current location.

5.3 Predicted User Location

Since IE is also a ubiquitous and ambient computing environment, we assume that sensors and actuators, simple push button and sliders, and computer access will be embedded and available in every area. People can be identified by the activity of accessing available resources at static locations or by sensing the user's mobile computing devices (PDA/handheld).

Table 1. HistoryDB entity

UserID	LocationID	Date	Time	Device
TM	125	13/8/02	04.02	Ibutton1
TM	323	20/8/02	05.01	VR10
…	…	…	…	…
TM	125	26/8/02	03.02	Sun15
TM	125	26/8/02	03.58	FR4
TM	125	27/8/02	04.05	PC5
…	…	…	…	…

We can identify the user's location by recording a history database of events, whenever a user accesses to identify himself (such as when using iButton, typing at a desktop computer or logging into the network) or whenever the receptor/sensor /actuator (such as webcam, handheld, active/passive badge) captures the user's identity in a particular location.

We develop a history data from precise users' locations (see a section on Precise User Location). Table 1 is an example of a locations history data in an IE Repository. The history data above can be used to predict user location. It is also possible to develop a probabilistic model to find the most probable location of a user in the log of history data based on a particular policy. We have implemented the policies below for a user's location checkpoints i.e. [7]:

1. The same day of the week (assuming regular work schedules, to find a user's location based on the history data of his location at almost the same time and on the same day of the week).
2. All the days in a one week range (to find a user's location based on the history data of his location at almost the same time within a week).

We use simple extended SQL query to implement the above policies to find user location. In addition, we also develop SpeechCA (speech context agent) using speech synthesis and speech recognition by cross platform Java Speech API (JSAPI). It makes the Active Office recognise and understand instructions from the user when he questions the Active Office, then at different levels, he queries the history data and feeds the answer to the speech synthesizer [7] as the Active Office responds to the user's query.

6 Discussion

In this part we discuss how the Active Office determines user location based on three location categories: precise location, proximate location, predicted location.

In Figure 4, we show how an Active Office processes the information to determine a user's location by aggregating the relationship between user data and location data.

Aggregate precise location has first priority and is followed by proximate and predicted locations respectively. This means that when the Active Office receives information from aggregate precise location, then the current user's location is determined. If not, then we check using aggregate proximate location data.

In the case of there being no user location at all in aggregate, precise, or proximate location then predicted location will be used, as in the case, for example, when user X works in room Y which is not covered by WiFi or Bluetooth access points, and there is only a workstation present. If at the time of query user X is not accessing the workstation, history data is used to find the most probable user location based on data for the same day at the certain time of the week or using all the days at the certain time in a one week range.

Our experiment using Wireless LAN i.e. WiFi and Bluetooth as proximate location, gave a significant result to sensing a user's proximate location. The result can be improved by getting interoperability between sensors to aggregate sensor data. From our experiment, compared with Bluetooth, WiFi seems the better way to obtain precise location data using proximate sensors.

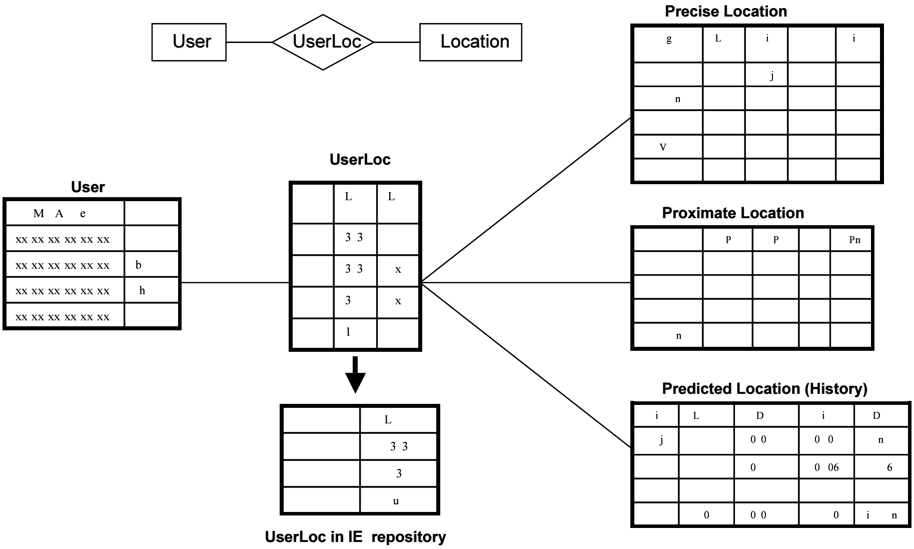


Fig. 4. Aggregate user's locations in an Active Office

7 Conclusion

This paper proposed an implementation model of an Active Office environment which is simple, efficient, scalable, fault tolerant and applicable to a range of heterogeneity computing platforms.

This paper has described the requirement of recognizing user locations and activities in the Merino service layer architecture, i.e. the architecture for scalable context processing in a ubiquitous computing environment and ambient intelligent (embedded sensing devices) environment.

In an Active Office, a user has a regular work schedule. A user has a routine activity that can be used to predict his location in a specific timestamp. A user's activity can be represented by user mobility, and user mobility can be seen from the user's changing location on a significant scale. So, in an Active Office once we can capture a user's location then we can map a pattern of user mobility.

The users' locations in an Active Office can be categorised into three, i.e. precise location, proximate location, predicted location. The category is based on the sensor capability to sense the area and the use of history data. The priority order in deciding a current user location is first precise location followed by proximate location, and then predicted location.

Our experiment using WiFi and Bluetooth in determining proximate location in an Active Office has shown good results in sensing user location. At present, WiFi rather than Bluetooth seems to be the better way to fix precise location using proximate sensors . The result can be improved by developing interoperability between sensors to yield aggregate sensor data.

Further research topics that can be considered arising from this work are:

- Aggregation of smart sensors (more interoperability between sensors) using Elvin notification system to notify the difference between current location and previous notification.

- Managing location information in Merino service layer architecture:
  - format representation,
  - conflict resolution,
  - privacy of location information.

References

1. Daniel, R. and Mealling. M. Resolution of Uniform Resource Identifiers using the Domain Name System. RFC 2168 (1997)
2. Dey, A. K., Abowd G. D., et al.: A Context-Based Infrastructure for Smart Environments. 1st International Workshop on Managing Interactions in Smart Environments. MANSE'99. (1999)
3. Gajos, K.: A Knowledge-Based Resource Management System For The Intelligent Room. Master Thesis. Massachusetts Institute of Technology. Massachusetts. (2000): 58
4. Harter, A. and Hopper A.: A Distributed Location System for the Active Office. IEEE Network Vol 8, No 1, (1994)
5. Kummerfeld B. and Quigley A. et al.: Merino: Towards an intelligent environment architecture for multi-granularity context description. Workshop on User Modelling for Ubiquitous Computing. Pittsburgh, PA, USA. June 2003
6. Jiang, C. Steenkiste P.: A Hybrid Location Model with a Computable Location Identifier for Ubiquitous Computing. The Fourth International Conference on Ubiquitous Computing (UBICOMP 2002). Goteborg, Sweden. (2002)
7. Mantoro, T. Johnson C. W.: Location History in a Low-cost Context awareness Environment. Workshop on 'Wearable, Invisible, Context awareness, Ambient, Pervasive and Ubiquitous Computing'. Australian Computer Science Communications. Volume 25, Number 6, Adelaide, Australia. February 2003
8. Mantoro, T.: Self-Organizing Map and Digraph Analysis of Electronic Mail Traffic to Support Interpersonal Resources Discovery. Master Thesis. Department of Computer Science. Asian Institute of Technology. Bangkok, Thailand. December 1994: 96 (CS-94-44)
9. Manzoni, P. and Cano J.: Providing interoperability between IEEE 802.11 and Bluetooth protocols for Home Area Networks. Computer Networks 42 (1).2003: 23–37
10. Mealling, M. and Daniel R.: The Naming Authority Pointer (NAPTR) DNS Record. Updated RFC 2168. (2000)
11. Schilit, W. N.: A System Architecture for Context awareness Mobile Computing. PhD Thesis. The Graduate School of Arts and Sciences. Colombia University. Colombia. (1995) 144

12. Schmidt, A. Beigl M. et al.: There is more to context that location. Computer & Graphics 23 (1999): 893–901
13. Segall, B. Arnold, D. et al.: Content Based Routing with Elvin4. In Proc. AUUG2K (June 2000)
14. Small, J. Smailagic A. et al.: Determining User Location For Context awareness Computing Through the Use of a Wireless LAN Infrastructure. Institute for Complex Engineered Systems. Pittsburgh, USA. (2000)
15. Weiser, M.: Some Computer Science Issues in Ubiquitous Computing. Communications of the ACM. 6(7) (1993).: 75–84

Multimodal Mobile Robot Control Using Speech Application Language Tags

Michael Pucher and Marián Képesi

Telecommunications Research Center Vienna,
Donau-City-Str. 1, 1220 Vienna, Austria
{pucher, kepesi}@ftw.at

Abstract. This paper describes the design and architecture of a multimodal interface for controlling a mobile robot. The architecture is build up from standardized components and uses Speech Application Language Tags. We show how these components can be used to build complex multimodal interfaces. Basic design patterns for such interfaces are presented and discussed.

1 Introduction

At our institute we develop and investigate multimodal interfaces for mobile devices for next generation telecommunication networks. Multimodal interfaces provide a promising paradigm for next generation user interface design. These interfaces combine inputs from different modalities like voice, vison and gesture and produce output with different modalities. In a mobile context multimodal input can overcome several restrictions of the mobile input bottleneck. In mobile usage situations the user has to deal with small displays in terms of size and resolution, small and limited keyboards and no access to the visual display. A typical mobile usage situation is driving a bicycle. All these restrictions can be overcome by multimodal interfaces, which can also be used to build more robust speech recognition interfaces.

1.1 Mobile Robotics and Ambient Intelligence

As we see it, mobile robotics is relevant for ambient intelligence in at least two different points:

– Robots can be seen as devices having sensors (inputs) and actuators (outputs) which can be combined to behaviours [5]. The so conceived robots provide an analogy to certain forms of ambient intelligence. The intelligent home for example, can be seen as a device (the apartment) which has a set of sensors (temperature, videocam) and a set of actuators (door, window), which are combined into intelligent behaviours.

– Furthermore a mobile robot can turn a normal apartment into an intelligent home without the need to attach sensors and actuators to the “home” directly. If I move to a new apartment, I can place my personal robot there and turn my

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 56–64, 2003.

- new apartment to an intelligent one. All functions, like measuring the temperature, checking if the lights are off, that are not apartment-specific can be used. The intelligence of the robot is transferred to the apartment in this way. This ambient intelligent solution is also mobile.

1.2 Multimodality and Ambient Intelligence

Ambient intelligence should allow the user to interact naturally with an intelligent interface. Because human face-to-face communication is multimodal, multimodal communication is perceived as natural and intelligent. Therefore well designed multimodal interfaces can have these required properties of ambient intelligence.

2 The Application

Our web interface allows a user to control a robot using a multimodal user interface which shows the map of that room where the robot is located including the robots location. It is possible to send the robot around and let it fulfill different tasks like measuring the temperature or checking if the windows are closed. For example the user could click on the room-icon and ask “What’s the temperature in this room?”.

At the moment we have only implemented dummy sensors, which produce a random sensor output, because our main interest was to investigate the relation between the multimodal interface and the (partially simulated) robot capabilities. The user can also ask the robot about it’s functionality or use voice commands to control the robot remotely. The robot “answers” the question posed by the user, by using a Text to Speech (TTS) system. (A Text to Speech system converts a given text to synthesized speech)

Our application implements third order multimodality [9], which is the most advanced form of multimodality and allows coordinated, simultaneous multimodal input. This means that the input from different modalities produces multimodal events which drive the application.

We see the main advantage of our architecture in the use of standardized components and protocols which make the architecture flexible and maintainable.

3 Standardized Components

3.1 SALT

Speech Application Language Tags [8] (SALT) are a standardized set of EXtensible Markup Language [10] (XML) tags, which can be used inside of HyperText Markup Language [11] (HTML) pages. The standardization of SALT is driven by major companies in the field. At the moment SALT is available for desktop browsers and Pocket PC browsers [6].

The Microsoft Speech SDK (Software Development Kit) Version 1.0 Beta 3 already supports the Pocket PC platform which runs on mobile clients. Our application was implemented using an older version of the SDK which supported only desktop browsers. Several implementations of SALT browsers can be found on [8].

We will show how one can deal with the complexity of multimodal interface development using SALT. Any SALT enabled client can control the robot via the HyperText Transfer Protocol [12] (HTTP). To interpret the SALT pages the browser needs to have access to a local or remote Text to Speech (TTS) and Automatic Speech Recognition (ASR) systems. (Automatic Speech Recognition means the "correct" conversion of speech into text)

3.2 Java

Java is a programming language and platform developed by Sun Microsystems [2], which provides amongst other things, standardized interfaces for building web applications.

Java Server Pages (JSP) are programs which run on a server. They can be called by a client via HTTP, keeping track of the clients state and producing dynamic output for the client. The advantage of Java Server Pages for our application is, that we can include the dynamic and static HTML, and the control of the robot in one JSP.

4 Architecture

A complete voice platform consists of at least three components. The TTS and ASR, and the dialog management. The dialog management manages the applications dialog flow.

In Figure 1 the building blocks of our application are shown. The logic of the application is distributed between the SALT client, the Webserver and the Robot.

The dialog management is done by the SALT client. The SALT tags in combination with JavaScript functions allows us to implement the dialog. (JavaScript is a scripting language which can be used inside of webpages and is interpreted by the clients Webbrowser) The dynamic SALT pages are requested from the Webserver and are sent to the client via HTTP. They include a response message from the robot.

The Graphical User Interface of the SALT client presents a map including the location of the robot. It can be used to send the robot around, and track it's movement.

The Webserver establishes the communication between the client and the robot. The robot's default interface is an infrared transmitter/receiver. Because infrared receivers only work in line of sight, this restricts the robot's mobility. A radio interface to the robot was implemented so that it could be controlled also when out of sight. The robot communicates through frequency-shift-keyed (FSK) data using an 868 MHz transceiver. (transmitter/receiver) The FSK is used to modulate the binary data in order to achieve errorless transmission of robot control commands. We call these two components radio/infrared proxy and infrared/radio proxy in analogy to the term's use in IP networks.

The robot interprets the commands it gets from the client. The commands are divided into simple commands which translate into simple robot actions, such as "Turn right", "Stop" etc. and complex commands, like "Is the light in this room on?" which trigger complex robot programs. There are some commands, like "What can you do?" which are processed directly by the Webserver. In case of this command a string with a description of the robots functionality is returned.

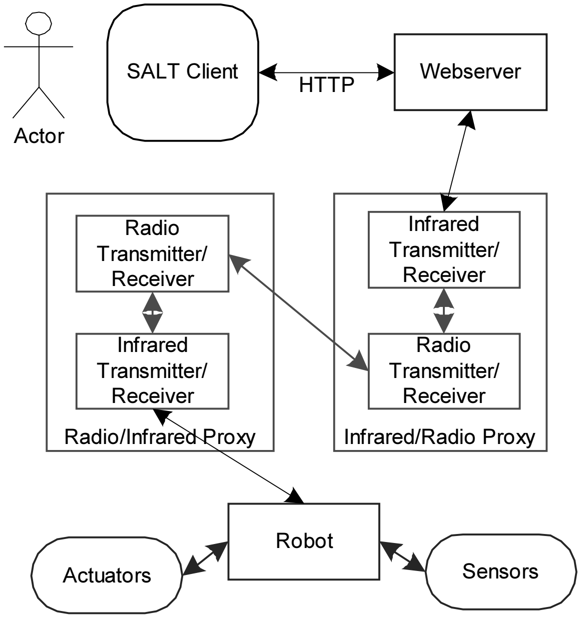

Actor
SALT Client
HTTP
Webserver
Radio Transmitter/ Receiver
Infrared Transmitter/ Receiver
Radio/Infrared Proxy
Infrared Transmitter/ Receiver
Radio Transmitter/ Receiver
Infrared/Radio Proxy
Robot
Actuators
Sensors

Fig. 1. Logical Architecture of a multimodally controlled robot

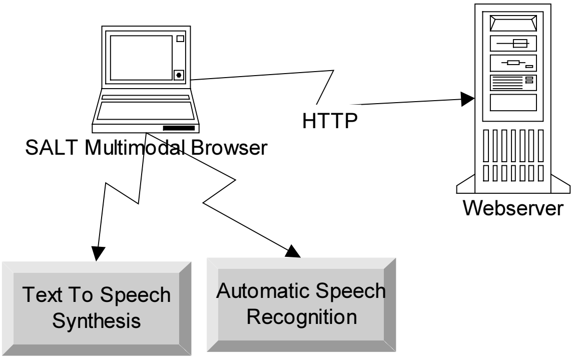


Fig. 2. System Architecture

Figure 2 shows the system diagram of the SALT client and the Webserver. In our case a laptop computer is used as a client running a SALT multimodal browser, but a palmtop computer, a smartphone or any other mobile SALT enabled client could be used. The multimodal browser accesses the TTS and ASR, which runs on the same machine as the browser in our case, but could also be remote.

The Webserver returns SALT pages in the users language. Localization of applications, e.g. translation into other languages is easy from the application providers point of view, because the SALT client chooses the ASR and TTS to use.

5 Multimodal Dialog Management Using SALT

5.1 Dialog Managment

Block schemes of finite-state machines describing the two different standard SALT tags are shown on Figure 3. The first tag is called SALT:prompt. It plays a specified prompt to the user, using the systems TTS.

The second tag is SALT:listen which starts a recognition process for a specified amount of time and goes into one of the end states afterwards. Both objects "start" methods can be called inside of JavaScript functions.

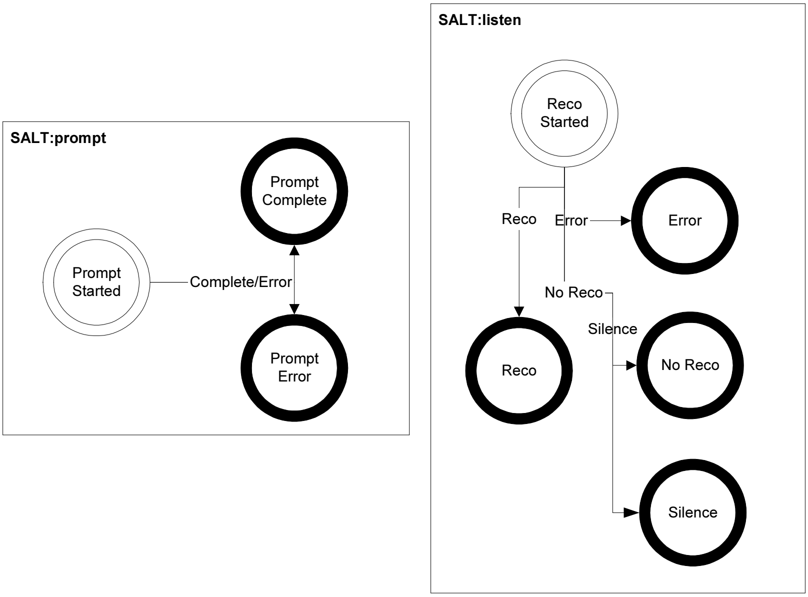


Fig. 3. Recognition and Synthesis Blocks

The description of the SALT tags as finite-state machines is convenient for the definition of the systems's application logic. Finite-state machines are a common method for the design of dialog systems. These blocks can be combined to implement the dialog management of our application. Figure 4 shows a design pattern for an application that uses two standard SALT tags. This pattern could be called the "endless prompt and recognition dialog", where the recognition determines the next prompt. The application plays a prompt, tries to recognize the user's speech command (specified by a grammar) and loads a new Webpage. At this point the application starts again.

As one can see from Figure 4, the application runs endlessly unless the user closes the browser. The starting point is the starting of the prompt where the synthesized string depends on the dynamic SALT content generated by the JSP. If, for example, the last recognized command was the question "What can you do", which was sent to the robot, then the synthesis string will be "I can walk around, measure the temparature,…". When the SALT page is loaded for the first time the user is welcomed.

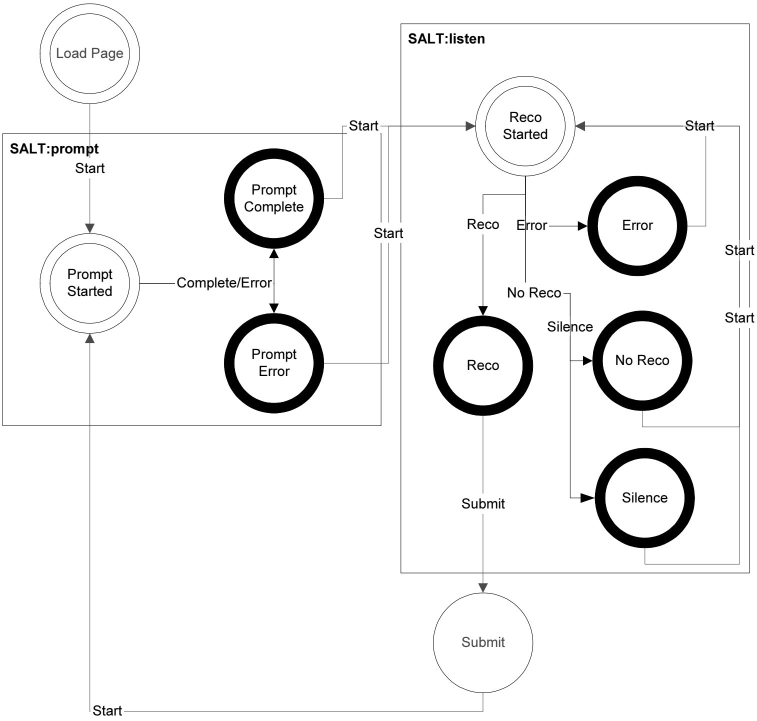


Fig. 4. Dialog Management of the application

The prompt has to terminate before the recognition is started, because otherwise it can happen that the robot gives itself commands. If the synthesis string is "I am going forward" and "forward" is also present in the list of commands, the robot will tell itself to walk forward, if the synthesis and recognition process overlap and the synthesis is played using the loudspeakers.

When the prompt ends the recognition process is started. While the recognition is running the user can see a progress bar showing the voice energy on the screen. If the recognition ends with an error it is started again, otherwise a JSP is called that generates a new SALT page.

In case of multimodal commands the JSP is only called if certain visual (mouseclicks) and speech commands occur at the same time.

This dialog can be seen as a design pattern for multimodal remote control dialogs. One important concept which is not mentioned in the above diagram, but is needed for the SALT:listen tags are grammars.

They can be specified in the HTML page or in a separate file.

```
<SALT:listen id="MainReco" onreco="HandleOnReco()"…>

<SALT:grammar …>

<grammar version="1.0" >

<rule id="numbers">

         <one-of>

         <item>forward</item>

         <item>back</item>

         <item>left</item>

         <item>right</item>

         <item>stop</item>

         <item>bye</item>

         …

         </one-of>

</rule>

</grammar>

</SALT:grammar>

</SALT:listen>


function StartRecognition()
{
         try{
         MainReco.Start();
         }catch(e){
         alert("Recognition error"); }

}
```

The above code shows the definition of a SALT:listen tag with an associated grammar and a JavaScript function which is called to start the recognition. In the first line of the SALT:listen tag the function is defined which is called when the object comes into the “reco” state, meaning that something was recognized.

5.2 Multimodal Integration

Under multimodal integration we understand the process of combining the input from the different modalities to a merged multimodal input and the generation of a multimodal output from this input. The first step is done by the SALT client in our case, while the second processing step is done at the server.

The dialog management and the multimodal integration are nicely separated in our application. The integration happens at one point in the applications dialog flow, namely before the submission of the webpage.

This separation is specific for our application, because SALT allows us to merge the dialog and the integration at any point in the above state diagram.

There are many different formalisms for multimodal integration like typed feature structures [3] or rule-based integration. We implemented a simple rule-based system using JavaScript functions. This method is sufficient to implement multimodal systems that support coordinated, simultaneous multimodality which is the most advanced form of multimodality.

The ontology of the application can be derived from the applications rules. "Ontology" in this context means the basic things or events that are represented in different modalities. The object/event "Hello World" could be represented as a text string or as a voice command. To know which representations refer to a common object one has to look at the rules of the system. W. Wahlster recently argued that a multimodal system should have an underlying ontology for the system to understand its own multimodal output [13]. To implement this form of "symmetric multimodality" using SALT one would have to add an ontology module at the server side.

One of our rules is, that if a room on the map is choosen with a mouseclick and the speech command "What's the temperature here?" is uttered at the same time, then these inputs will be combined to a multimodal command.

```
if(submitForm.x1.value != "0" && submitForm.sp.value ==
"temperature")

submitForm.submit();
```

The code above shows the part of a function which implements this rule. Through the combination of multimodal rules complex interfaces can be created.

Although the multimodal integration can be triggered at any state of the dialog, the integration and dialog can be logically seperated using this rule-based integration design.

6 The Robot

Our robot was built with the Lego Mindstorms kit [7]. It can drive around, has a touch sensor and a simulated temperature and vision sensor. To enable the communication with the Webserver we installed the Lejos Java Virtual Machine on the robot, so that we can run small Java programs on the robot [4].

The behaviour of the robot, including the navigation was implemented according to the behaviour control theory from [1].

7 Conclusion

We showed how to build a sophisticated multimodal interface using standardized components. We also discussed the relevance of these kinds of interfaces for ambient intelligence. Such standardized components are of special importance for future interfaces for mobile devices. There are and will be many different types of such devices, leading to a crucial need of having standardized basic building blocks.

The work presented here is part of the project Speech&More carried out at the Telecommunications Research Center Vienna (ftw.) ftw. is supported by the Kplus program of the Austrian Federal Government.

References

1. Bagnall B., Core Lego Mindstorms Progamming, Prentice Hall 2002, pp 196
2. Java 2 Platform, Enterprise Edition (J2EE), http://java.sun.com/j2ee
3. Johnston M., Unification-based Multimodal Parsing, COLING-ACL, Montreal, Canada, 1998, pp 624–630
4. Lejos – Java for the RCX, http://www.lejos.org/
5. Maes, P. and R. A. Brooks, Learning to Coordinate Behaviors, AAAI, Boston, MA, August 1990, pp. 796–8
6. Microsoft .NET Speech SDK Version 1.0 Beta, http://www.microsoft.com/speech/getsdk
7. Mindstorms - Robotics Invention System 2.0, http://www.mindstorms.com/
8. SALTFORUM: Speech Application Language Tags, http://www.saltforum.org
9. W3C: Multimodal req. for Voice Markup Lang. Working draft 10. July, 2000, http://www.w3.org/TR/multimodal-reqs
10. W3C: http://www.w3.org/XML/
11. W3C: http://www.w3.org/MarkUp/
12. W3C: http://www.w3.org/Protocols/
13. Wahlster, W., SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell In: Krahl, R., Günther, D. (eds): Proceedings of the Human Computer Interaction Status Conference 2003, June 2003, Berlin: DLR, p. 47–62

Energy Consumption Routing for Mobile Ad Hoc Networks

Hasnaa Moustafa and Houda Labiod

GET/ ENST / INFRES Paris, France
46 rue Barrault, 75634 Paris Cedex 13, France
{moustafa,labiod}@enst.fr

Abstract. The specific features of mobile ad hoc networks (MANETs) impose new requirements for routing protocols. The aim of this paper is to present an approach that enhances routing performance by integrating ad hoc related characteristics. We propose a loop-free adaptive path energy conserving scheme which attempts to minimize both routing and storage overhead in order to provide efficiently robustness to host mobility, adaptability to wireless channel fluctuations and optimization of network resources use in large-scale networks. The strength of our scheme, based on source routing approach, named Energy Conserving Routing (ECR), is to handle network state related constraints such as node's energy consumption, link availability and path quality. The performance of the proposed scheme is evaluated via simulation and is compared to DSR protocol. Better results are obtained with ECR in terms of path lifetime, forwarding efficiency, delay, and total overhead.

1 Introduction and Background

The advent of ubiquitous computing and the proliferation of portable computing devices have raised the importance of mobile and wireless networking [1]. A Mobile Ad hoc NETwork (MANET) is an autonomous collection of mobile nodes forming a dynamic network and communicating over wireless links. Ad hoc communication concept allows users to communicate with each other in a temporary manner with no centralized administration and in a dynamic topology that changes frequently.

MANETs do not rely on any pre-established infrastructure and can therefore be deployed in places with no infrastructure. This is useful in disaster recovery situations and places with non-existing or damaged communication infrastructure where rapid deployment of a communication network is needed [2]. Ad hoc networks can also be useful in conferences where people participating in the conference can form a temporary network without engaging the services of any pre-existing network [3].

Due to the limited propagation range of wireless environment, routes in ad hoc networks are multihop, and mobile nodes in these networks dynamically establish routing among themselves to form their own network "on the fly"[4]. Each participating node acts both as a host and a router and must therefore be willing to

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 65–75, 2003.

forward packets for other nodes. Nodes in such a network move arbitrarily causing frequent and unpredictable changes in the network topology, and allowing unidirectional links as well as bi-directional links to occur.

Moreover, these networks suffer from wireless channel bandwidth limitation. The scarce bandwidth decreases even further due to the effects of signal interference, and channel fading. Network hosts of ad hoc networks such as laptops and personal digital assistants operate on constrained battery power, which will eventually be exhausted, and limited CPU and storage capacity.

Building ad hoc networks implies a significant technical challenge because of many constraints related to the environment, such as unreliable wireless links, limited energy consumption, and dynamic network topology. This introduces a trade-off between link maintenance in a highly unreliable networks and power conservation for users with little battery power.

Routing in mobile ad hoc networks is a significant research topic where various mechanisms have been already proposed concerning this issue. In fact, conventional routing protocols are not well suited in ad hoc environment for several reasons. Firstly, they are designed for static topology, which is not the case in ad hoc network. Secondly, they are highly dependent on periodic control messages; this is in contradiction with resource-limited ad hoc environment. Moreover, classical protocols try to maintain routes to all reachable destinations, which wastes resources and consumes energy. Another limitation comes from the use of bi-directional links, which is not always the case in ad hoc environment. Actually, there is a need for new routing protocols, adapting to the dynamic topology and the wireless links' limitations. Routing protocols in such networks should provide a set of features including [3]: distributed operation, loop freedom, on-demand based operation, unidirectional link support, energy conservation, multiple routes, efficiency, scalability, security and quality of service support. However, none of the proposed protocols have all the above desired properties, but these protocols are still under development and are being probably enhanced and extended with more functionality. Until now, no standard has been adopted and many critical issues remain to be solved. In this paper, our work focuses on one critical issue in MANETs that is energy conserving routing. Indeed, the advantages mainly expected are providing efficient saving in bandwidth and network resources, reducing communication cost, supplying efficient data delivery with highly unpredictable nodes' mobility, and supporting dynamic topology with unreliable wireless links. Until now, only a few energy conserving routing protocols have been proposed.

Consequently, we propose a novel energy conserving routing protocol. Our scheme named Energy Conserving Routing (ECR) protocol, operates in a loop-free manner and attempts to minimize both routing and storage overhead in order to provide efficiently robustness to host mobility, adaptability to wireless channel fluctuations, and optimization of network resources use. It applies the source routing mechanism defined by the Dynamic Source Routing (DSR) unicast protocol [5] to avoid channel overhead and improve scalability. ECR outperforms other routing protocols by providing available stable paths based on future prediction for links' states. These paths also guarantee nodes stability with respect to their neighbors, strong connectivity between nodes, and higher battery life. In fact, ECR selects routes according to four criteria: association stability between nodes, strong connectivity between neighbors, link availability between nodes, and minimum energy consumption at each node. Performance analysis has shown robustness against mobility.

This paper is organized as follows. Section II, provides a brief survey on routing protocols in MANETs, pointing out the advantages of energy conserving routing in the context of multihop wireless communications, and stating the most recent protocols proposed by the Internet Engineering Task Force (IETF) MANET working group. Section III, gives a detailed description of our proposed protocol ECR. Section IV, analyzes our performance results with a comparison to DSR protocol. Finally, section V provides concluding remarks and highlights our future work.

2 Routing: Brief Survey

Ad hoc networks specific features impose new requirements on routing protocols. The most important features are the dynamic topology due to nodes' mobility, and the nodes' limited resources in term of battery power and bandwidth. In this section, we briefly describe routing protocols in ad hoc networks. Then, we discuss the energy-conserving routing approach since it is the focal point of the study undertaken in this work.

Many classifications are found in the literature. The first one reflects the existence of three main categories based on the routing strategy. Firstly, there are protocols, which use a proactive approach to find routes between all source-destination pairs regardless of the need of such routes. The main feature of this class consists of keeping continuous up-to-date routing information from each node to each other node in the network. Destination-Sequence-Distance-Vector (DSDV) protocol [6] is an example of this approach; it comes as an improvement of Distributed Bellman Ford (DBF) protocol [7]. This approach also includes Wireless Routing Protocol (WRP) [8], Global State Routing (GSR) and Fishey State Routing (FSR) protocols [9], which are based on link state algorithms. Landmark Ad hoc Routing (LANMAR) [2], and Optimized Link State (OLSR) [10] protocols recently proposed by the MANET group also fall into this category.

Secondly, there are the reactive (on-demand) routing protocols suggested with the key motivation of reducing routing load. Contrarily to proactive mechanisms, these protocols do not maintain routes to each destination continuously. Instead, they initiate routing procedures on an "on-demand" basis. DSR, Signal Stability Routing (SSR) [11], Associativity Based Routing (ABR) [8], and Temporally Order Routing Algorithm (TORA) [8] are typical on-demand routing protocols.

In addition to the above-mentioned protocols, hybrid protocols combine reactive and proactive characteristics, which enable them to adapt efficiently to the environment evolution. This approach comprises Zone Routing Protocol (ZRP) [12]. Table 1, states the advantages and disadvantages of the main approaches cited above. Routing protocols can be also classified in terms of an architectural view. Most traditional classification contains hierarchical protocols and flat protocols [8].

Table 1. Proactive vs. reactive vs. hybrid approaches

	Advantages	Disadvantages
Proactive	Up-to-date routing information Quick establishment of routes Small delay	Slow convergence Tendency of creating loops Large amount of resources needed Routing information not fully used
Reactive	Reduction of routing load Saving Resources Loop-free	Not always up-to-date routes Large delay Control traffic and overhead cost
Hybrid	Scalability Limited search cost Up-to-date routing information within zones	Arbitrary proactive scheme within zones Inter-zone routing latencies More resources for large size zones

A third classification identifies two types of mechanisms according to the location characteristic: Physical Location Information (PLI)-based protocols giving approximate location for mobile nodes, and PLI-less protocols.

Various metrics are used to determine optimal paths. The most common metric is the shortest path applied in DSR (Dynamic Source Routing), DSDV, TORA, and WRP. Nevertheless, not necessarily optimal routes are obtained for different possible network configurations. Some protocols, however, can use the shortest delay as the metric, although longer paths have higher probability to reach destination in an ad hoc environment. SSA has been advocated to improve routing performance, based on using link quality metrics. In this case, a shortest hop route may not be always used. However, this method could not avoid frequent rerouting due to node mobility.

In this paper, we address the problem of energy conserving routing in mobile ad hoc networks, in terms of both battery power and consumed resources. Our approach aims at extracting from dynamic, irregular topology of a MANET to obtain a topology with minimum energy conserving, available and more stable links, and higher path quality.

In fact, in the case of medium and large size MANETs (greater number of nodes), the size of the routing tables increases significantly and the topology changing events will grow proportionally. In addition, large number of nodes introduces higher interference and weak link quality. Therefore, we should consider the tradeoff between the increasing interference, and the higher connectivity of the network. Furthermore, we need to balance the tradeoff between the power consumption and the link maintenance.

Obviously, energy conserving routing is a young research domain, no standard has been adopted yet and many issues have to be addressed and more studies are needed. Actually, most existing protocols face several problems in routes maintenance and frequent reconfiguration when link failures occur. Other protocols depend on upstream and downstream nodes requiring storage overhead and consuming resources. Moreover, some protocols consider the shortest path as a criterion for path selection, which is not usually suitable to the high and unpredictable variation of the topology and does not assure the minimum energy consumption.

As nodes in an ad hoc network can enter and leave the network randomly, reliability of the network nodes is low. At the same time, wireless links suffer from fading and multipath causing low reliability of the network links. This introduces a tradeoff between link maintenance in a highly unreliable network and power conservation for users with a little battery power. Users in an ad hoc network have to strike a balance between two objectives: conserving power and mitigating interference to other users on the one hand, and increasing power to maintain links on the other hand [13].

In this paper, we address the problem of energy efficient routing in MANET. This is a challenging environment as every node operates on limited battery resources and multihop routes are used over a changing network environment due to node mobility. The battery energy of a transmitting node can be depleted due to: (a) processing at the node, (b) transmission attenuation due to path loss, (c) the need to maintain the transmission above a certain threshold due to signal interference [14].

The goal of our work in this paper is to propose new characteristics in ad hoc routing to provide efficient saving in bandwidth and network resources, and to insure minimum energy consumption along the different used routes. In this context, we propose our Energy Conserving Routing (ECR) protocol. ECR is an on-demand protocol, routes are obtained on-demand to use efficiently network resources, avoiding channel overhead and improving scalability. It conserves energy and provides robust paths.

3 The Proposed Protocol

This section gives an overview of our new adaptive path energy conserving routing protocol ECR, describing its operation and route's selection approach. We introduce four different metrics based on battery power at each node, node's stability with respect to neighbors, quality of the link between nodes in terms of the received signal level, and availability of the link using future prediction for the link's state. Our scheme operates in a loop-free manner and attempts to minimize both routing and storage overhead providing robustness to host mobility, adaptability to wireless channel fluctuations and optimisation of network resources use. To allow efficient use of network resources, route discovery process is initiated on demand avoiding channel overhead and improving scalability.

ECR applies a selection criterion using the defined four metrics to provide stable paths based on: links availability according to future prediction of links state, higher battery life paths tending to power conserving, link quality information to select one among many different routes, and location stability of nodes biasing the route selection towards routes with relatively stationary nodes. This scheme can be viewed as "limited scope" flooding within a properly selected set of nodes. The mechanism of source routing proposed in DSR unicast protocol is applied.

ECR uses an energy level metric. This metric periodically calculates the current battery power, which is a decreasing function of time and processed packets. We introduce this metric for power conservation of nodes. Paths with higher battery life, indicating less power consumption, are only selected.

MAC layer beacons are used to provide each node with neighbors' existence information. When a node receives a neighbor's beacon, it updates or creates the corresponding entry of this neighbor in its *Neighbor_Stability_Table*, Table 2. Entry update takes place through incrementing the *node stability* field, and setting the *signal strength* field according to the level of strength the beacon is received. In addition, the node performs continuous prediction for link's availability towards the neighbor and updates its *link availability* field. If no beacons are received by a node from a certain neighbor up to a certain period of time, the node indicates neighbor's movement and updates its stability table fields towards this neighbor.

Table 2. Neighbor_stability_table

Neighbor	Type	Node Stability	Signal Strength	Link Availability

Operation starts when a source node has data to transmit to a certain destination and it has no route to that destination. A route discovery procedure is then invoked through broadcasting a *request* packet to a selected set of neighbors searching for the destination. Neighbors' selection takes place following our selection criteria among the neighbors, through using the four proposed selection metrics. Each neighbor node receiving the *request* packet repeats the process of selection and *request* transmission; meanwhile the source route is accumulating in the packet header, until reaching the destination node. A destination receiving a *request* packet will take the accumulated route in its header and transmit a *reply* packet in the reverse route direction. The *reply* packet continues to be forwarded along the source route that is stored in the packet header, constituting of a set of selected nodes, until reaching the source. The source receives the packet and stores in its cache the route that is found in the packet header. This constructs a route between the source and the destination consisting of selected nodes to forward data. Finally, the source selects the shortest path route from its routing cache to transmit its data.

In fact, minimizing the broadcasting scope in the route discovery process attempts to minimize the message overhead of computing routes. A benefit of these types of selected routes also is that they are energy efficient and there will be little need to modify them frequently. This protocol investigates the routing problem in MANETs through considering a distinctive approach. Basically, it addresses three issues in this problem: energy conserving, path availability, and nodes' strong connectivity.

4 Performance Analysis

Network Simulator*2* (ns*2*) is used in our performance analysis, it is a discrete event simulator developed at Berkeley University targeted at networking research [15]. The aim of our performance analysis is to evaluate the behavior of our proposed ECR protocol and to compare its performance to traditional DSR protocol, as an on demand routing protocol applying the source routing concept.

4.1 Simulation Model and Scenarios

The overall goal of our simulation study is to analyze the behavior of our protocol under a range of various mobility scenarios. Our simulations have been run using a MANET composed of 20 nodes moving over a rectangular 1200 m x 300 m space, and operating over 600 seconds of simulation time. The radio and MAC models used are described in [15]. Nodes in our simulation move according to the Random WayPoint mobility model [16]. The movement scenario files used in each simulation are characterized by pause times; we studied 6 different pause times (0, 30, 60, 120, 300, 600). A pause time of 600 represents a stationary network, while a pause time of 0 represents a network of very high mobility in which all nodes move continuously. Table 3 and Table 4, show respectively the network configuration and the simulation parameters used in our simulations run.

Table 3. Network configuration parameters

Configuration	Parameters Values
Topography	1200m X 300m
# of nodes	[20]
Max. speed	20 m/s
Traffic Type	CBR
Traffic sources	3
Pkt Size	512 bytes
Emission rate	4 pkts/sec

Our performance evaluation is a result of 60 different simulations, using 10 different simulations for each pause time. At each pause time, we study runs with a max nodes movements' speed of 20 m/s. The traffic sources in our simulation are constant bit rate (CBR) traffic. Each traffic source originates 512 bytes data packets, using a rate of 4 packets/second.

4.2 Results and Analysis

We used the following main performance metrics: average end-to-end delay, delivery ratio, dropped packets, control packets overhead, and control bytes overhead. The following section shows the analysis for our obtained results.

4.2.1 Average End-to-End Delay

Figure 1 shows the average end-to-end delay as a function of the mobility scenario. In this figure, results for the protocol ECR are compared with those of DSR. This delay is calculated only for the data packets that have been successfully received. We can see that the delay has nearly the same behavior for both protocols at intermediate and low mobility. At high mobility, DSR causes an increase in delay over ECR thanks to the nodes' selection mechanism applied in ECR. In fact, using the selection mechanism of ECR minimizes the broadcast scope and constructs more minimum hops stable paths with longer route lifetime consuming less energy, thus achieving better impact on the delay. When the network becomes stable, the delay difference is reduced. In this case the two protocols show nearly the same behavior since routes become stable.

Table 4. Simulation parameters

Parameters	**Values**
Simulation time	600 sec
Tmax between requests	10.0 sec
Twait for requests	0.5 sec
Twait for reply	2 sec
Neighbors update interval	3.0 sec
Nodes' pause time	(0to 600) sec
Total simulations	60(10/pausetime)

4.2.2 Average Delivery Ratio

Results comparison for the delivery ratio is depicted in Figure 2. Delivery ratio is determined by the ratio of the number of data packets actually delivered to the destination versus the number of data packets supposed to be received. This number presents the effectiveness of the protocol. We can see that the two protocols point up the same behavior in all mobility cases. ECR outperforms DSR showing incremental delivery ratio for all mobility cases; this refers to its rigid long-lived routes by means of selecting stable high-energy paths. The obtained result confirms the expected behavior of ECR, which tends to choose links' quality paths reacting better towards frequent distortion and interference.

4.2.3 Average Packets Drop

Figure 3 shows the effect of ECR and DSR on the number of dropped packets. It is obvious that ECR has a weaker impact on packets drop for all mobility cases, while DSR has more impact on packets drop in these cases. When link failure occurs, queue congestion is increased in DSR causing more packets drop. In contrast, ECR avoids frequent link failures. We can explain this by the fact that the route lifetime of ECR paths is longer, at the same time the stability features allows ECR to outperform DSR.

4.2.4 Average Control Overhead

The control overhead comparison is subsequently illustrated in Figure 4 and Figure 5. We observe a significant difference between ECR and DSR in terms of control packets generated during simulation for all cases of mobility. ECR shows less control overhead providing better results in terms of both packets and bytes overhead. The effective node selection mechanism in ECR causes packets generation only to certain nodes, and the fact of selecting high-energy paths decreases the probability of link failure and the need to send more routing packets to recover this failure. On the contrary, DSR relies on broadcast flooding in its route discovery process.

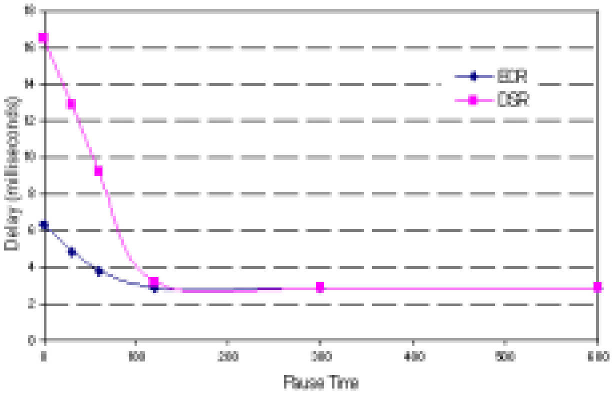


Fig. 1. Average end-to-end delay

ECR outperforms DSR in bytes overhead as the source route accumulation takes place for the selected nodes only, thus saving lot of bytes caused by source route headers. On the contrary DSR accumulates the source route during route request broadcast, thus consuming more overhead bytes.

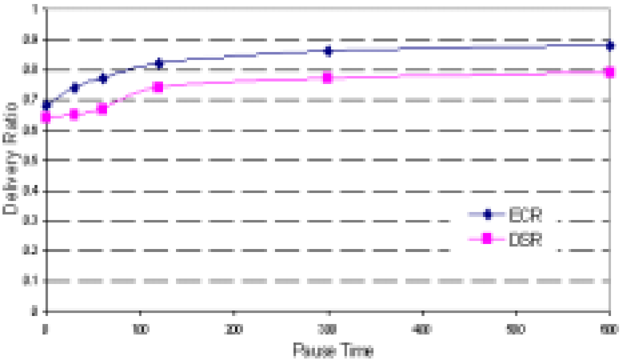


Fig. 2. Average delivery ratio

5 Conclusion and Future Work

In this paper, we propose the Energy Conserving Routing (ECR) protocol. ECR uses no periodic network flood of control packets. Thanks to its selection criteria, stable paths with future links availability and higher battery life are provided. Simulation results have shown that by using our four proposed selection metrics we reduced incredibly the routing overhead. Furthermore, our performance results indicate the robustness of ECR against mobility. This is fulfilled in terms of the significant increase of the route lifetime due to the lower frequency of link failure. An interesting property of using quality adaptive energy conserving routing is that the packet delay is not greatly affected and the packet delivery is enormously increased. ECR has good impact on the delay compared to DSR and shows better delivery ratio at different mobility's. It outperforms DSR in the amount of dropped packets and has significantly lower overhead

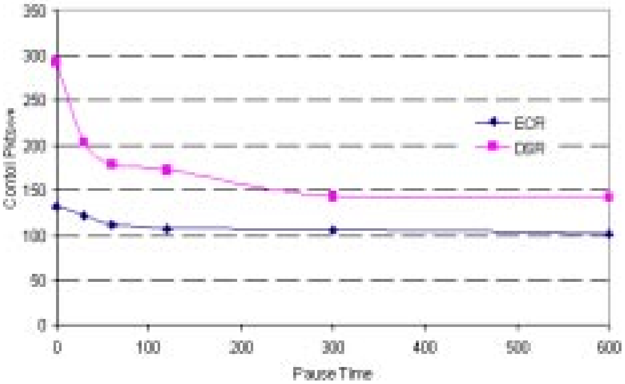


Fig. 3. Average packets drops

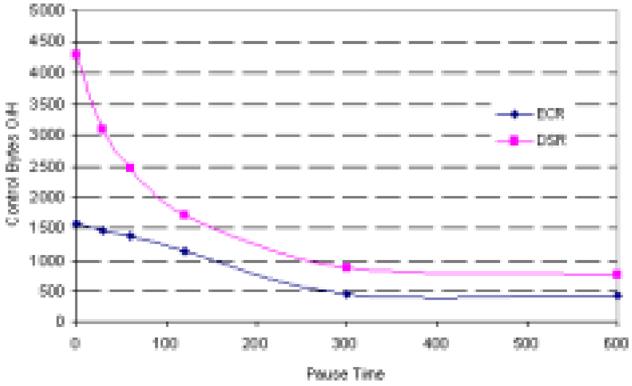


Fig. 4. Packets control overhead

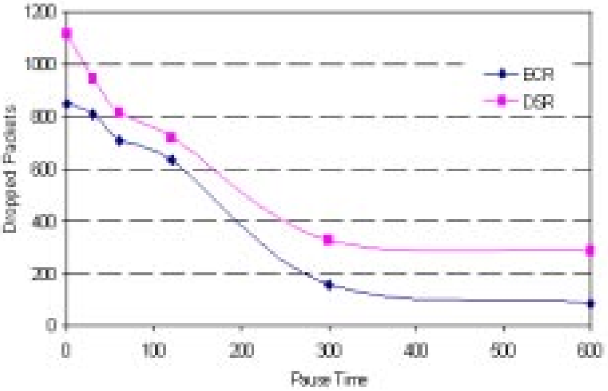


Fig. 5. Bytes control overhead

For future work, we intend to compare ECR with other energy conserving routing protocols, considering new performance metrics such as energy-based mobility and link stability metrics. We also intend to consider large scalable networks in our performance analysis.

References

1. Mobile Ad Hoc NETwork (MANET). (1999). URL: http://www.ietf.org/html.charters/manet-charter.html
2. Pei, G., Gerla, M., Hong, X.: LANMAR: landmark routing for large scale wireless ad hoc networks with group mobility. Proceeding of IEEE/ACM MobiHoc 2000, Boston, MA, August 2000
3. Broch, J., Maltz, D., Johnson, D., Hu, Y., Jetcheva, J.: A performance comparison of multi hop wireless ad hoc network routing protocols. MOBICOM 98
4. Larsson, T., Hedman, N.: Routing protocols in wireless ad hoc networks – A simulation study. Master Thesis, Stockholm ericsson switched lab. 1998
5. Johnson, D., Maltz, D.: Dynamic source routing in ad hoc wireless networks. In Mobile Computing, Imielinski, T. and korth, H. Eds. Norwell, MA: Kluwer,1996
6. Perkins, C., Bhagwat, P.: Highly dynamic destination-sequenced-distance-vector routing (DSDV) for mobile computers. ACM SIGCOMM, vol. 24, no. 4, October 1994, PP. 234–244
7. Murthy, S., Garcia, J.: An efficient routing protocol for wireless networks. MONET, vol. 1, October 1996, PP. 183–197
8. Royer, E., Toh, C.: A review of current routing protocols for ad hoc mobile wireless networks. IEEE personal communication, April 1999
9. Moustafa, H.: Multicast Routing in Mobile Ad hoc Networks. Master Thesis, ENST 2001
10. Jacquet, P., Mahlethaler, P., & al.: Optimized link state routing protocol. Internet draft, IETF, March 2001
11. Dube, R., & al.: Signal Stability based Adaptive Routing (SSR) for Ad Hoc Mobile Networks. IEEE personal communication, February 1997, pp. 36–45
12. Pearlman, M., Hass, Z.: Determining the optimal configuration for the zone routing protocol. IEEE selected area in communication, 1999
13. Chiang, M. and Carlsson, G.: Admission Control, Power Control, and QoS Analysis for Ad Hoc Wireless Networks. IEEE International Conference on Communication ICC, 2001, volume:1, 11–14 June 2001
14. Kang, I. and Poovendran, R.: On the Lifetime Extension of Energy-Efficient Multihop Broadcast Networks. IEEE International Joint Conference on Neural Networks 2002, IJCNN'02
15. Fall, K. and Varadhan, K.: NS Notes and Documentation. The VINT project, UC Berkeley, LBL, USC/ISI, and Xerox PARC, May 1998. Work in progress
16. Bettstetter, C., Hartenstein, H., and Pérez-Costa, X.: Stochastic Properties of the Random Waypoint Mobility Model: Epoch Length, Direction Distribution, and Cell Change Rate. ACM MSWiM'02, 2002

Ubiquitous Attentiveness – Enabling Context-Aware Mobile Applications and Services

Herma van Kranenburg, Alfons Salden, Henk Eertink, Ronald van Eijk, and Johan de Heer

Telematica Instituut, Drienerlolaan 5, Enschede, The Netherlands
{Herma.vanKranenburg, Alfons.Salden, Henk.Eertink, Ronald.vanEijk, Johan.deHeer}@telin.nl

Abstract. We present a concept called ‘ubiquitous attentiveness’: Context information concerning the user and his environment is aggregated, exchanged and constitutes triggers that allow mobile applications and services to react on them and adapt accordingly. Ubiquitous attentiveness is particularly relevant for mobile applications due to the use of positional user context information, such as location and movement. Key aspects foreseen in the realization of ubiquitously attentive (wearable) systems are acquiring, interpreting, managing, retaining and exchanging contextual information. Because various players own this contextual information, we claim in this paper that a federated service control architecture is needed to facilitate ubiquitous attentive services. Such a control architecture must support the necessary intelligent sharing of resources and information, and ensure trust.

1 Introduction

Nowadays, our everyday world offers us a heterogeneous environment with multiple network providers, several mobile and wireless technologies, numerous terminals, manifold administrative domains and various application providers. On the other hand millions of users co-exist that all have different interests and priorities. Providers may want to tailor their services to the end-user preferences and available resources. In addition, knowledge about situational contexts, such as location, nearby facilities, available resources, and preferences of (other) parties, can be quite beneficial in tailoring as well. Hence, there is a clear need for open service architectures that enable these ubiquitous attentive services.

Context can refer to any relevant fact in the environment. Examples are users’ location and related parameters (coordinates, being indoors or outside, velocity, temperature, humidity etc). Also the facilities users have at their disposal (cellular phone, PDA, Operation System and software running on mobile device, graphical capabilities, etc) form bits and pieces of the context. Further context information includes users’ preferences. What do we like (e.g. spoken or written messages: while driving a car possibly spoken messages, while at the same time the car stereo volume should be

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 76–87, 2003.

lowered) and who may interrupt us (e.g. while being in a meeting, or giving a presentation: in these cases a disturbance normally is not preferred for a social chat, in contrast to the case that a serious accident has happened to one of your family members). Also nearby persons and services have a relevancy to mobile users as well as to mobile application providers. All these (and more) information pieces form integral parts of the (user) context. Context information can thus be defined as any information that can be used to characterize the situation of an entity.

Context awareness functionality provides the means to, starting from the user, deliver optimal tailored applications, services and communication. This requires functionality to sense, retain and exchange the context in which a person is in at a certain moment in time. The facilitating technologies and supporting architectures thus need to supply context-awareness functionality, and moreover ability to take appropriate action accordingly. The latter implies an exchange between systems (platforms, domains, terminals, parties, etc) and a pro-active response of the environment are essential too. These three paradigms together constitute ubiquitous attentiveness:

- Ubiquitous computing (loads of small sensors, actuators and terminals).
- Information processing and exchange between different systems in different domains.
- Pro-active responsiveness of the ubiquitous environment.

Our current paper addresses key aspects of ubiquitously attentive (wearable) systems that need to be solved to facilitate a mobile user in a global and heterogeneous environment. The organization of the paper is as follows: after addressing key issues and requirements of ubiquitous attentiveness we illustrate them by a near futuristic mobile medical service scenario. In the following section we point out how to retain the envisioned ubiquitous attentive mobile services using commonly available technologies.

2 Key Issues and Requirements

There are many issues in the mobile and wireless settings that ask for sophisticated technological and architectural solutions:

- Information sources are heterogeneous in nature and are distributed across heterogeneous environments with unpredictable and dynamically varying network resources and varying terminal processing capabilities. On the other hand users like to access (in a transparent way, not being bothered about underlying technologies) multimedia applications that in turn typically are distributed. The mobility of users typically requires query-functionalities regarding the context (like where is…).

- The current architectures and platforms offer isolated, and sometimes also partial information. Interconnection between service infrastructures and domains is not available now. There is a strong requirement for inter-working between platforms using standardized means. One of such is by federated coupling between service platforms as investigated in the 4PGLUS project [1], see also figure 1. In that service control solution existing mobile, wireless and information access architectures are federated using connections between the service platform functionalities in each of the domains. This results in seamless service-access for end-users. Such service control layer hide technical, syntactic, schematic, and semantic heterogeneities from the user.

- User expectations need to be considered as well. Seamless roaming is expected to be possible by mobile users, with a maximum degree of freedom and flexibility (and minimal costs) while not being bothered about technological issues. Users consider their mobile device as a personal assistant that offers unique opportunities such as real-time adaptation of services to a dynamic user environment. In addition, context awareness is expected to assist both in reducing the overload of available content and services and in increasing the efficiency of content-consumption, e.g. by filtering the offered services according to the user's situational preferences and needs.

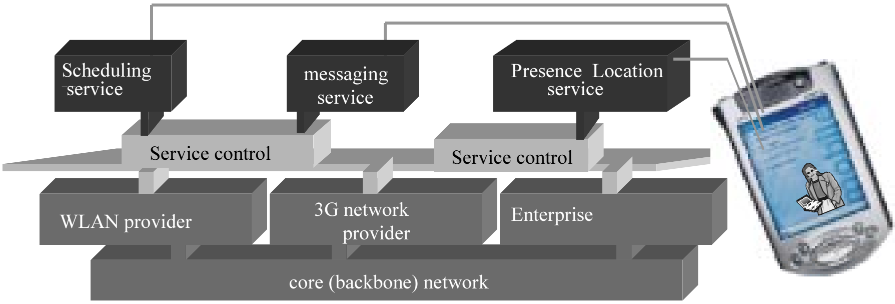


Fig. 1. Service platform based on a federated service control architecture

A ubiquitous attentive solution that enables true context-aware services and facilitates mobile users anytime and anyplace should take into account an integration of all relevant contextual aspects in relation to the communication and computational capabilities of the environment. As the mobile user is roaming across different administrative domains and different technologies services have to deal with a dynamic environment. The adaptation to this dynamic context and service brokerage in turn needed for this should be completely transparent for the user and offer mobile service providers the highest customer retention values. Vice versa, the environment of the user can be influenced by the availability and activities of the user and adapt itself accordingly.

3 Scenario: Bandaging and Salving Alice

The following scenario illustrates how several context parameters (location, presence, agenda) in a heterogeneous environment are collected, distributed, accessed, interpreted and used to pro-actively plan actions of a mobile medical assistant keeping in mind for example presence or availability information.

3.1 Scenario

Alice a highly qualified nurse pays a visit to several patients in need for bandage replacement and salving their skins. This medical treatment in the past required consultation of and treatment by a dermatologist at a hospital. Nowadays, however, Alice carries out such a treatment at the patients' premises using a remote consultation service provided by an available dermatologist.

Alice is always on and so is her consultation service. On the one hand, her mobile consultation service consists of a service providing remote videophone connection with one or more dermatologists distributed over various hospitals. On the other hand this videophone service is integrated with an agent-based scheduling service for making consultations. Either Alice or her software agent 'Consul' retrieves the availability information of the doctors by simply inspecting their agenda's and planning video meeting between Alice and one of the available dermatologists at a suitable time for them both. The dermatologist gets a notification about the arranged video meeting at an appropriate time, being at a time when he is not busy treating other patients nor while he is a meeting room. During the video meeting, the quality of the video stream is adapted to the bandwidth of the (wireless) networks available at the hospital or the patient's home.

Alice likes to spend all her time on and with her patients. She hates being bothered by the crippled MS browser that still comes with her PDA user interface – "A relict", she thinks, "that humanity better had done away with straight away". "Consul", she reflects, "takes all the hustle of (re)arranging meetings from my shoulders. Consul is my man: I can really count on him!"

4 Categories of Context

Due to the nature of mobile services and use, contextual information (often stored in profiles) is gathered by multiple parties, stored in multiple places (e.g. in the terminal, the access networks, the Service Platform and at service providers) and used and managed by multiple stakeholders. The distributed nature of context information (and profiles) must be supported by the service architecture. Exchange of part of the information between and across boundaries of administrative domains can be supported in an federated service control architecture [1]. Depending on the privacy policies, entities in different domains can share and exchange relevant profile data. Privacy policies are crucial in this respect. Moreover trust is vital between the parties involved in sharing contextual information. The user should have trust in all system-components and parties that collect and store his – privacy sensitive context and have the ability to distribute it to other parties, e.g. other users and services. Both globalization and virtualization of society have contributed to a greater privacy risk. User acceptance and his perceived privacy is a mobile business enabler [2]. Protection of profile data and trust must be addressed - to the satisfaction of the users - in the federated service architecture. Without access to user-related context data many mobile services and ubiquitous attentive systems will surely not exist.

With respect to mobile services and applications delivered over heterogeneous networks in a distributed environment we distinguish the following relevant categories of contextual information:

- User context [3], typical including the person's environment (entities surrounding the user, humidity, light), and describing the personal (physiological and mental context), task, social and spatiotemporal (location, in- or outdoors, time) domain. The user context includes user preferences about the use of surrounding (wireless) access networks, e.g. UMTS and WLAN. These preferences for using specific networks are for example based on the availability, price, quality and domain of the network. The user may also have a preference about the level of control for switching to other networks while using applications and services on its terminal.
- Service/Application context [1]: the applications currently available or used at the users terminal and the services the user is currently subscribed to. Both have requirements about the terminal (display size) and the network (bandwidth).
- Session context [1]. Description of sessions currently used by the user. With whom is he communicating (which service, which other users). What kind of data is communicated (voice or video). Is it a high bandwidth session (multimedia streaming) or a low bandwidth session (browsing, emailing, voice call).
- Access Network context [1]. The properties of available (wireless) networks in the neighborhood of the user's terminal, such as availability, domain, price and bandwidth.
- Terminal context [1]. All properties of the user's terminal that are relevant for running applications: display size, memory, communication interfaces (Bluetooth, GPRS, WLAN, UMTS)
- Service platform context [1]. The Service Platform (SP) has information about the subscriber, e.g. about its access rights and privacy aspects with respect to identity, call handling settings and location of the end-user. Furthermore, the SP has roaming agreements with other SP's to enable seamless user authentication in all visited networks, irrespective the radio technology or domain of such networks.

A relevant contextual parameter-set differs for different applications. Consider a mobile user with a terminal with multiple applications being simultaneously active. While leaving a network domain eg when leaving his office premises, a video application needs contextual information about the Access Network (bandwidth dropping from WLAN to GPRS), while another application (e.g. a restaurant "push" service) might be more interested in the activity of the user (e.g. changing from business to private role).

5 Context Representation and Modeling

To cope with incomplete context-parameter-sets and therefore changing context structure, ontologies instead of fixed (profile) schemes have to be used. An ontology gives meanings to symbols and expressions within a given domain language. It allows for a mapping of a given constant to some well-understood meaning and vice versa. For a given domain, the ontology may be an explicit construct or implicitly encoded with the implementation of the software [4]. Thus ontologies enable all participating components to ascribe the same meaning to values and structures stored in a profile.

In general dealing with dynamical changing and incomplete, distributed contextual parameters requires a set of context knowledge bases that define various typical real world situations. With these context knowledge bases user service and system adaptation preferences can be associated and thus the problem resolves to determining the appropriate user context for the environmental variables and clues provided by the real world situation, and then applying the referenced service options per user situation.

In the field of artificial intelligence, modeling of ontologies is done for quite some time. It is e.g. being applied to knowledge representation languages and DAI such as Open Knowledge Based Connectivity (OKBC). Interesting examples of context theories – enabling reasoning within and across contexts - are e.g. proposition logic [5, 6], local model semantics [7, 8] and multi-context systems [9].

In W3C and the Semantic Web a number of standard initiatives and tool developments exist that specialise in modeling of ontology services for the web, such as, RDF Schema [10], DARPA Agent Markup Language [11], Ontology Interchange Language OIL [12], DAML-S [13]. In [14] these are elaborated on. DAML+OIL [15] is the basis for the ontology web language OWL [16] that facilitates greater machine readability of Web content by providing additional vocabulary along with a formal semantics.

In the next section a very closely related topic is described, that adds processing to the modeled context.

6 Crunching

Context information is useless, unless an evaluation takes place, which can take advantage of the additional situational information, like e.g. is done in an ambient intelligent home environment [17]. The crunching functionality includes algorithms for robust and reliable interpretation of multiple contextual information streams, ontologies, context modeling, context theories, meta-profiles (metadata on processing order/sequence and algorithms to be used). Thus crunching comes very close to the previous section on context representation and modeling. Example references where contextual reasoning is tackled with regard to integration of heterogeneous knowledge and databases are [18, 19, 7].

Different parts of the actor's situation and of the context of the information may have different weights in determining the ambient of the actor. In addition, the ordering of the processing may affect the resulting context awareness.

The crunching functionality should also be able to compensate for limited and incomplete information.

This could be dealt with in several ways [20]:

- Use cached information stored during older sessions.
- Using only the parameters that are available (Fuzzy logic, estimate value or thresholds for unknown parameters).
- Let the user make decision and do the inference. I.e. just present the context that is available to the user and let him decide.
- Supervised learning the above [21, 22, 23].

7 Acquiring Context Data

In order to acquire context information, sensors or human-machine-interfaces can gather information about users and their current situation. Users can also provide this information themselves. A special category of context is position and location information. This part of the user context is mostly exploited in today's mobile services and sensory input plays a major role in providing this information. Sensor technologies will be more and more embedded in the mobile equipment and networks while services will sense better and better who the user is, where he is, what he is doing, what he is up to, what the environmental conditions are, etc. The environment itself will also be equipped with sensors that perceive users and communicate with their devices.

Advances in sensor technology and extension of sensor types are needed to reach further capabilities for adaptation of services to - and co-operation with - the environment of users. Aggregation of the sensed values in relation to a known reference is needed for a rightful, reliable and robust interpretation. As sensors can also interact, crunching of all contextual parameters is complicated even more. Many devices in our current world already adapt to a very crude and limited extend to their situational context. For example, some types of television sets adjust their image contrast to the ambient lighting level.

Many context aware applications make use of multiple sensor information inputs in order to steer one or other modality. Achieving in this respect e.g. cross-modality mapping by means of ubiquitous attentive mechanisms will be an inviting research challenge. The presence of multi-modality and the availability of relevant contextual information will require the introduction of even more intelligent human-machine-interface-management, enabling intelligent I/O behavior adaptation. This means, that the input/output behavior of a user end-system - for example the modality - can change dynamically based on the actual context. For example the user interface to an e-mail-system is typically text-based, whereas it would become speech-based if the user drives a car.

8 Access and Exchange

The access and exchange of context information must ensure that applications (executed on an end-user device) or services (executed somewhere in the universally connected world) get proper access to the information that describe the context of the device, the network, the end-user, etc. Hence, we will need in the future even more transparent access to - and exchange functionalities between - distributed, autonomous, heterogeneous information sources. Thereto, we need shared supervised learned scalable and extensible indexing, query and retrieval schemes that can stand time despite the increase in types of acquisition, processing and categorization technologies arising for new mobile service usage contexts.

The massive number of actors and objects in the wireless world make it simply impossible to have centralized solutions for access to information or exchange of contextual information. Luckily, ubiquitous network connectivity allows for building global software infrastructures for distributed computing and storage. The provisioning of context-information can profit a lot from intelligent and secure sharing of resources. We need at least the following functions in the information-provisioning environment:

- Scalable synchronization and update mechanisms. For service-oriented information provisioning (e.g. personal location services) information from several domains can be collected, and analyzed. For these types of applications, synchronization and update mechanisms for contextual information are needed that are completely different from traditional database synchronization solutions. Domain-owners (e.g. end-users for their PANs; 3G operators for their mobile networks; enterprises for their wireless LAN infrastructures) will have their own policies for this type of synchronization and update mechanisms. It is necessary that an open infrastructure will be created, that allows one to exchange information on-demand; comparable to current infrastructures for presence information (as e.g. are currently being standardized in the IETF SIMPLE group [24].
- Federated authentication and authorization mechanisms, in order to provide trust. A key issue is to ensure that the end-user is still in control: he has to be confident that only the people (or providers) that he trusts are able to use his context information. Identity protection and privacy are key requirements. We need federated solutions for this (e.g. building on architectures coming from the Liberty Alliance [25]) in order to allow for global access and distribution of private information across several domains.
- Service discovery solutions. Single-domain service discovery is already in place in the video and audio entertainment sector: HAVi (Home Audio Video Interoperability) and UPnP (Universal Plug and Play) are existing standards that are widely implemented. UPnP is one of the architectures for pervasive peer-to-peer network connectivity of PCs of all form factors, intelligent appliances, and wireless devices. However, in multi-domain situations there is a lot of work to be done. There are efforts that head for global service discovery (e.g. the UDDI directory-mechanisms used in the web-service community), but these directory mechanisms are not scalable enough for the amount of devices and services that exist in a universally connected world. New mechanisms are needed. This is of course strongly related to the synchronization and update mechanisms discussed earlier.

- Well-defined extensible data-formats for contextual information (as described in earlier sections). This calls for standardization.
- Open interfaces for context information. There is a growing need for standardization of not only the data formats of contexts, but also for the access functions towards this information, and the control-mechanisms that allow both a device or service-owner (interface provider) and the interface-user to control the behavior of the information exchange mechanisms.

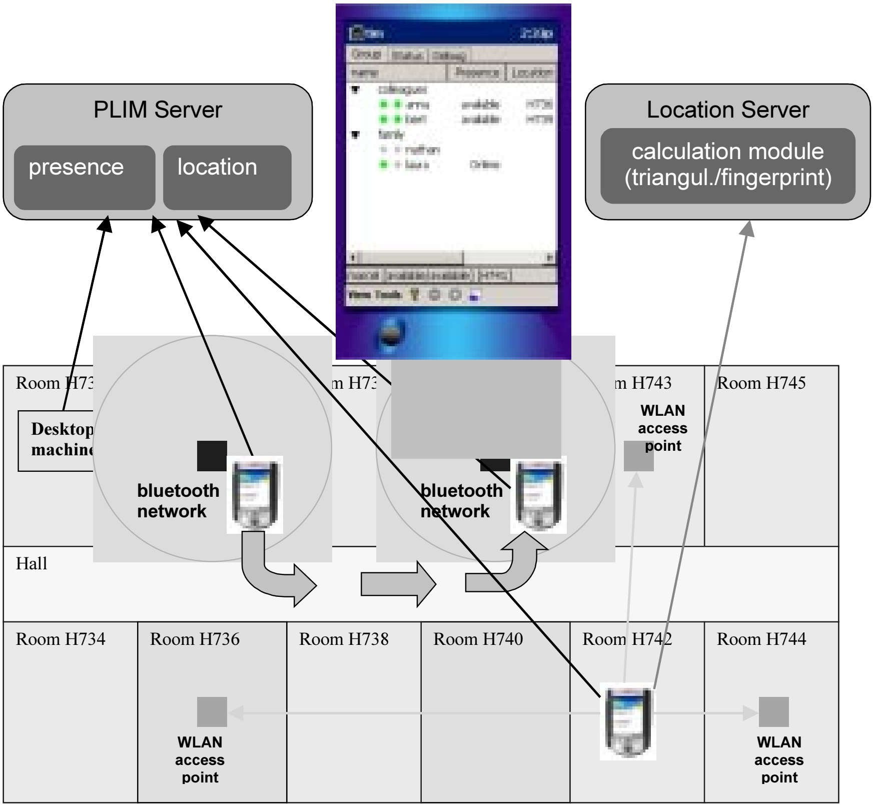
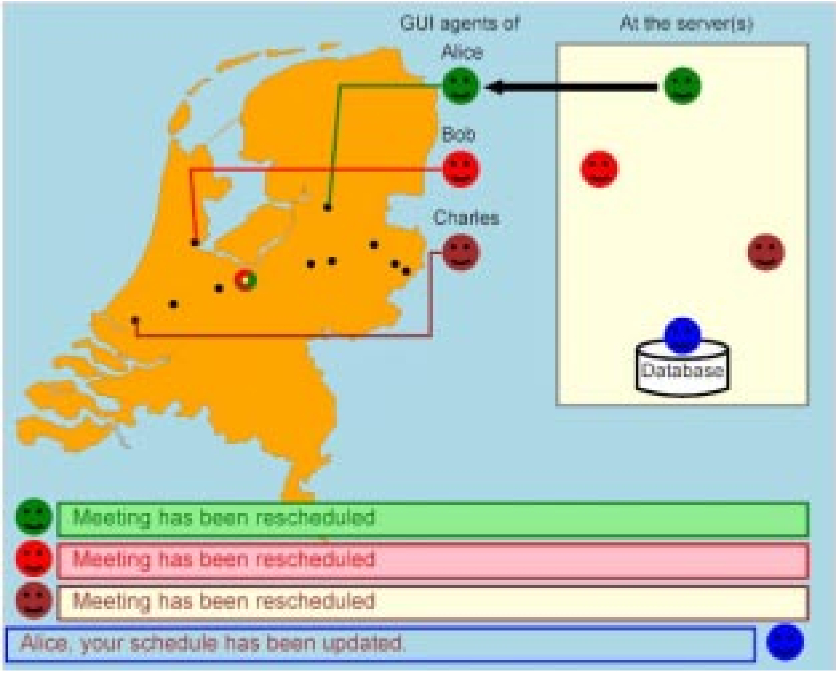


Fig. 2. The user interface of PLIM

9 Example of Implementations (Consul Says to Alice: "At Your Service")

We developed and implemented several platforms with ubiquitous attentive functionality. In the context of our scenario (described in section 3.1) we briefly describe two of them: Presence Location Instant Messaging (PLIM) [26] platform and Scheduler Agent System (SAS) platform [27].

The PLIM server (modified Jabber Instant Messaging server) stores the dermatologists' locations in its database. The location (based on Bluetooth and WLAN) is represented in technology independent XML formats. The nurse logs onto the PLIM server and has access to location/presence information of all dermatologists in the all hospitals she is subscribed to. Distribution of the location information to other users makes it possible to view (see Figure 2) each others context (situational awareness) on the same application i.e., an Instant Messaging application that wants to know if a dermatologist is available for a videoconference consult.



Meeting has been rescheduled

Meeting has been rescheduled

Meeting has been rescheduled

Alice, your schedule has been updated.

Fig. 3. The visualization of negotiations between scheduler agents that can represent the dermatologist (Bob and Charles) and the nurse (Alice)

Additionally, we build a context-aware personalized scheduling service for a mobile business-to-employee (B2E) setting, where software agents collectively arrange new meetings at different locations (in this case cities see Figure 3) and times keeping in mind the upcoming meetings of the employees (e.g. dermatologists, nurses). The software agents also simultaneously look after privacy or security policies of the dermatologists and nurses or their hospitals they work for, e.g. with respect to location information, time schedules, personal preferences or patient sensitive information. We developed and deployed our scheduling service on the JADE-agent platform using PDA's connected to a server using WLAN and GPRS networks.

10 Summary

A (wearable) system is ubiquitously attentive when it is cognizant and alert to meaningfully interpret the possibilities of the contemporary user environment and intentionally induces some (behavioral/system) action(s). Context awareness is as part of ubiquitous attentiveness, as exchanging information across and between (heterogeneous) domains and the pro-active responsiveness of the ubiquitous environment is. Key aspects of ubiquitously attentive (wearable) systems to us appear to be acquiring, interpreting, managing, retaining and exchanging contextual information.

A ubiquitous attentive solution that integrates all contextual aspects in relation to the communication and computational capabilities of the environment is needed for true context-aware services offered to mobile users anytime and anyplace. Dynamic adaptation of services to a dynamic changing user context (and vice versa) with unpredictable resources and incomplete context information should be solved in a completely transparent manner for the user. Seamless roaming of users in a global and heterogeneous setting must be supported by a open service architecture (loosely coupled), assisting in exchange of a variety of context parameters across network boundaries, such as resources, user identification, knowledge of optimal settings, session context, etc. Key aspects in facilitating connectivity and applications in a distributed heterogeneous environment are a federated service control architecture, intelligent sharing of resources, authentication and trust.

References

1. 4GPLUS project, http://4GPLUS.freeband.nl
2. Lankhorst, M.M., van Kranenburg, H., Salden, A., Peddemors, A.: Enabling technology for personalizing mobile services. In Proceedings of 35th Annual Hawai'i International Conference on System Sciences, HICSS (2002)
3. IST Project 2001-34244-AmbieSense, Deliverable D1 – Requirements Report
4. FIPA-ACL. The Foundation for Intelligent Physical Agents (FIPA), http://www.fipa.org
5. Buvac, S., Mason, I.A.: Propositional logic of context. In Proceedings of the 11th National Conference on Artificial Intelligence (1993)
6. Buvac, S.: Quantificational Logic of Context. In Proceedings of the 13th National Conference on Artificial Intelligence (1996)
7. Ghidini, C., Serafini, L.: Model Theoretic Semantics for Information Integration. In F. Giunchiglia (ed.), Proceedings AIMSA'98, 8th International Conference on Artificial Intelligence, Methodology, Systems, and Applications, Vol. 1480 of LNAI, Springer-Verlag. (1998)
8. Ghidini, C., Giunchiglia, F.: Local Model Semantics, or Contextual Reasoning = Locality + Compatibility, Artificial Intelligence, 127(2) (2001) 221–259
9. Sabater, J., Sierra, C., Parsons, S., Jennigs, N.: Using multi-context systems to engineer executable agents. In Proceedings of the sixth international workshop on agent theories, architectures, and languages (ATAL'99) (1999)

10. RDF Vocabulary Description Language: RDF Schema http://www/w3.org/TR/rdf-schema
11. DARPA Agent Markup Language (DAML) http://www.daml.org/
12. Ontology Inference Layer (OIL) http://www.ontoknowledge.org/oil/
13. DAML-S Coalition, DAML Services, http://www.daml.org/services/
14. Pokraev, S., Kolwaaij, J., Wibbels, M.: Extending UDDI with context-aware features based on semantic service descriptions. In Proceedings of The First International Conference on Web Services (ICWS'03), parallel-conference of The 2003 International Multiconference in Computer Science and Computer Engineering, Las Vegas, Nevada, USA (June 23–26, 2003)
15. DAML+OIL. Reference Description, http://www.w3.org/TR/daml+oil-reference (March 2001)
16. Web Ontology Language (OWL) http://www.w3.org/TR/owl-ref/
17. Demonstrated by Aarts, E.H.L. at the WWRF meeting in Eindhoven and the ICT Kenniscongres in Den Haag (2002). Also in Snijders, W.A.M.: Ubiquitous communication: lubricant for the ambient intelligence society, to be published in:`Ambient Intelligence' (book publication by Philips Design; ed. E.H.L. Aarts), http://www.extra.research.philips.com/cgibin3/reprints.pl?zkauthor=ambient&zkkey1=&zkkey2=&zksrt=kort
18. Farquhar, A., Dappert, A., Fikes, R., Pratt, W.: Integrating information sources using context logic. In Proceedings of AAAI Spring Symposium on Information gathering from distributed heterogeneous environments (1995)
19. Mylopoulos, J., Motschnig-Pitrip, R.: Partitioning information bases with contexts. In Proceedings of the Third International Conference on Cooperative Information Systems, (1995).
20. Dey, A.K.: Understanding and Using Context. Personal and Ubiquitous Computing Journal, 5(1) (2001) 4–7
21. Song, X., Fan, G.: A Study of Supervised, Semi-Supervised and Unsupervised Multiscale Bayesian Image Segmentation. In Proceedings of the 45th IEEE International Midwest Symposium on Circuits and Systems Tulsa, Oklahoma, USA (2002)
22. Aldershoff, F., Salden, A.H., Iacob, S., Kempen, M.: Supervised Multimedia Categorisation. IS&T/SPIE's 15-th Annual Symposium Electyronic Imaging Science and Technology, Storage and Retrieval for Media Databases 2003 (EI24), Santa Clara, California, USA (January 20–14 2003)
23. Salden, A.H., Kempen, M.: Enabling Business Information and Knowledge Sharing. IASTED International Conference, Information and Knowledge Sharing (IKS 2002), St. Thomas, Virgin Islands, USA (November 18–20, 2002)
24. IETF SIMPLE group http://www.ietf.org/html.charters/simple-charter.html
25. Liberty Alliance www.projectliberty.org
26. Peddemors, A., Lankhorst, M., de Heer, J.: Presence, location and instant messaging in a context-aware application framework. 4th International Conference on Mobile Data Management (MDM2003), Melbourne, Australia (21–24 January, 2003)
27. Bargh, M.S., van Eijk, R., Ebben P., Salden A.H.: Agent-based Privacy Enforcement of Mobile Services. In International Conference on Advances in Infrastructure for Electronic Business, Education, Science and Medicine and Mobile Technologies on the Internet (SSGRR2003w: http://www.ssgrr.it/en/ssgrr2003w/index.htm), L'Aquila, Italy (January 6 –12, 2003)

Real Time Application Support in an Ambient Network Using Intelligent Mobile Robots

Rabah Meraihi, Gwendal Le Grand, Samir Tohmé, and Michel Riguidel

Get Telecom Paris,
46, rue Barrault - 75634 Paris Cedex 13, France
{rabah.meraihi, gwendal.legrand, samir.tohme,
michel.riguidel}@enst.fr

Abstract. In this paper, we present an intelligent and controllable ad hoc network using mobile robot routers in heterogeneous mobile environments. The goal of the mobile robots is to ensure a global wireless connectivity, and interconnection to a wired infrastructure. The support of real time multimedia services are also allowed by a cross-layer quality of service management combined with a terminal based routing protocol, providing good wireless link quality.

1 Introduction

This work is part of the ambient network architecture of the Ambience Work Package 1 platform. In this context, an ubiquitous access to an ambient intelligent mobile network is allowed with real time communication support. In such an environment, the network must adapt to mobile environment variations and responsive to application and user's needs. The goal of the Ambience Work Package 1 is to provide seamless and transparent services to different users moving in a building or in public environments across a range of access networks and terminal capabilities within a completely heterogeneous environment.

In the Ambience Demo M1 scenario, as a new member enters the conference hotel (where several meetings take place) he is identified by voice recognition and gets some meeting details on his PDA. A member can read a message, find a person, a meeting or read a document. A member may also use his PDA during the meeting to send text messages to other participants, make notes and follow the agenda. In a more general case, real time multimedia communications can be established between users. In this context, the support of real time services (voice or multimedia applications) is absolutely necessary. However, due to the bandwidth constraint, dynamic topology and shared link of mobile networks, providing quality of service (QoS) for real time applications (requiring a fixed bandwidth, a low delay and little jitter) is a challenging task.

The network architecture of the Ambience Work Package 1 consists of a combination of a wired and wireless network, where the wireless network connectivity is ensured by an auto configured ad hoc network. An ad hoc network is a

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 88–98, 2003.

set of mobile stations dynamically forming a temporary network, interconnected by wireless links, and that does not require the use of any existing network infrastructure or centralized administration. In our study, focused on ambient environment, two types of constraints should be considered to support real time flows: application constraints (required bandwidth, end-to-end delay...) and network constraints (route stability, path availability, congestion and loss probability, load balancing ...).

Three aspects are mainly treated in this paper: wireless mobile network connectivity, interconnection with a fixed infrastructure and real time multimedia applications support. We designed an ad hoc routing protocol providing network connectivity, and performing terminal differentiation (CPU, routers, batteries) using intelligent mobile routers. The routing is performed preferentially by controllable robots that do not have resources or mobility issues. We also investigate QoS concerns: IP (Layer 3) quality of service combined with a Medium Access Control (MAC IEEE 802.11 [21]) service differentiation, in order to support real time traffics.

This document is organized as follows: section 2 gives a survey of the state of the art in ad hoc network interconnection with fixed networks, and QoS works in MANETs (Mobile Ad hoc NETworks). Section 3 details the principles of our solution that combines a routing protocol scheme using dedicated intelligent mobile routers (robots) and cross-layer QoS management in ad hoc environment. Section 4 summarizes the simulation results of our scheme. Finally, the main lessons learned are presented in section 5.

2 Related Work

Usually, an ad hoc network is auto-configurable and easy to deploy, so it has no central control device to configure it and manage the classical problems associated with creating and maintaining a network. This creates a whole set of issues that need to be resolved in order to allow MANETs to be formed and function under all eventualities in a reliable fashion. They are in general, low capacity networks as the link layer is over a radio interface which is by nature a narrowband channel, and also loss (large packet loss ratio). They are usually small - fewer than 100 nodes, and are often used at the edge of larger permanent networks as temporary extensions for meetings or conferences or in situations where wired networks are not feasible.

2.1 Interconnection of Ad Hoc Network with Fixed Networks

In the following, we provide a means to connect heterogeneous ad hoc networks to an infrastructure. This is very useful when extending an existing infrastructure without deploying any new access points and also when an existing wireless node density increases. The ad hoc network is made up of heterogeneous wireless devices (phones, laptops, PDAs …) that use heterogeneous wireless technologies (IEEE 802.11b, Bluetooth, GPRS …). An important functionality that must be integrated in ad hoc networks is their ability to connect to wireless gateways to fixed networks. In order to provide Internet services via a gateway, a MANET must be capable of attaching itself to the fixed network that is behind it. The interconnection with fixed networks consists in giving nodes within a MANET the possibility to operate as if they were part of a fixed Ethernet bus. In order for any solutions to the problem of IP connectivity to be useful, the MANET must be able to function reliably as an autonomous Mobile Ad Hoc Network.

When using the ad hoc network as an extension of the infrastructure, all nodes need to be connected to a gateway to the Internet. When this gateway is unreachable, several solutions might be envisaged:

- the network may tell some of its nodes that they should move in order to recover connectivity;
- a number of nodes dedicated to routing in the ad hoc network may move in order to provide connectivity to disconnected nodes.

For example, in Ambience demo M1 scenario, it is unlikely that people having a meeting together and therefore wearing several devices might decide to conclude their meeting in order to provide connectivity to other users. Many wireless nodes may belong and be worn by the same user. Again, it seems unreasonable to separate these nodes from the user. Moreover, parameters like wireless link quality can hardly be maintained to an acceptable level when node roaming is totally uncontrolled. A user may not wish that his devices be used as wireless routers if he gains no benefit from it. Finally, many devices may not be connected to the infrastructure if they do not use interfaces that are compatible with the other nodes.

2.2 Quality of Service Works in Mobile Ad Hoc Networks

Recently, numerous researches on quality of service have been carried out on certain aspects of MANETs [11]: QoS routing, QoS MAC, resource reservation protocol, and QoS model for ad hoc networks.

In this section we describe briefly ‘FQMM’ (a quality of service model for Manet), followed by the IEEE 802.11 protocol and describe some mechanisms for service differentiation at MAC layer. Other QoS researches in Mobile Ad Hoc networks are also presented.

2.2.1 FQMM (Flexible Quality of Service Model for Manets)

FQMM is a QoS model designed for small or medium size Ad-Hoc networks of not more than 50 nodes and using a flat non-hierarchical topology [6]. Nodes can be classified under three main types: ingress nodes, internal nodes and egress nodes.

The main purpose of FQMM was to propose a QoS model for MANETs that takes into consideration the dynamic topology of Ad-Hoc networks and the air interface that lies between the nodes. Their proposition is inspired from the performances of IntServ and DiffServ; FQMM is actually a combination of the two with an attempt to undermine the weaknesses of each when applied to Ad-Hoc networks. The next sub section presents how QoS can be provided at the 802.11 MAC layer.

2.2.2 IEEE 802.11 MAC Quality of Service

The default MAC access scheme used in IEEE 802.11 [21] is the Distributed Coordination Function (DCF) which provides a fair and stochastic access to the wireless link. It uses CSMA/CA (Carrier Sense Multiple Access/ Collision Avoidance) that is very similar to Ethernet's CSMA/CD.

EDCF (Enhanced DCF) [8] has been developed by the IEEE 802.11e group. It defines new QoS mechanisms for wireless LANs. EDCF enhances the access mechanisms and proposes a distributed access scheme that provides service differentiation. Priorities are managed by the terminals by modifying the basic DCF access scheme.

Other works on 802.11 MAC service differentiation can be found in [7][9].

2.2.3 Other QoS Approaches

Other attempts to introduce QoS in wireless and ad hoc networks can be identified. QoS routing [14] [3] [16] consists in selecting a route according to QoS constraints. Power adaptation [13] maximizes transmission power for high priority traffic. Load management may be used as in DLAR [15] to distribute load amongst several nodes of the network. Link stability may increase network performance; several schemes optimizing this point have been proposed. For example, ABR [16] is a reactive protocol that selects routes having the longest lifetime; SSA (Signal Stability-based Adaptive Algorithm) [17] selects a route according to the signal strength and its stability in time; [18] combines both link stability and multipath routing. In mobile environments, battery saving is a critical issue. [20] proposes several algorithms based on DSR that take into account the remaining energy of each node. [3] describes several metrics that can be taken into account by a shortest path algorithm in order to increase battery lifetime. [4] describes four algorithms (MTPR, MBCR, MMBCR, and CMMBCR) that minimize the overall transmission power.

3 Our Approach

In this section we will describe our solution based on the combination of a routing protocol scheme using dedicated intelligent routers and a cross-layer QoS management. The ad hoc network connectivity and interconnection with fixed infrastructure are provided by mobile robot routers which ensure a seamless communication service with a better network coverage.

3.1 Intelligent Mobile Router

The mobile robot [19] was developed at ENST. It is based on an open and modular architecture, which offers standard wireless communications: Ethernet, 802.11 and Bluetooth.

For power management needs, the battery is controlled by a dedicated circuit which keeps the processor aware of its load level. When the battery's level reaches a low limit, the robot will complete its critical tasks, if any, and move towards a refueling intelligent dock. The robot will find its way to the dock by following a laser beam delivered by this dock. The three stands dock hardware is built on the same basic board as the robots.

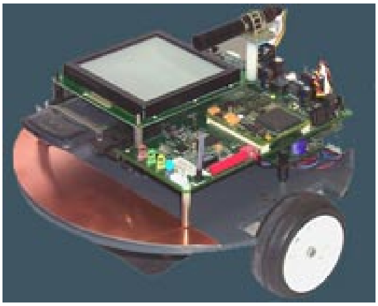


Fig. 1. Intelligent mobile robot

We propose to design an ad hoc network corresponding to (Figure 2). This allows the use of slow motion mobile routers with respect to call duration. The mobile robot-routers are a means to build a network core over which the infrastructure has some control, in order to provide some added value to the network. In this context, the robots will move to provide wireless connectivity (no user mobility required). In addition, battery consumption is not a concern since the robots may have a high autonomy and charge when they need it (on a dedicated docking station).

Having control over the network is essential in order to provide an acceptable link quality and add QoS mechanisms in the network (as we will show in the following). Hence, the amount of traffic and wireless hops may be managed and thus good quality links may be maintained.

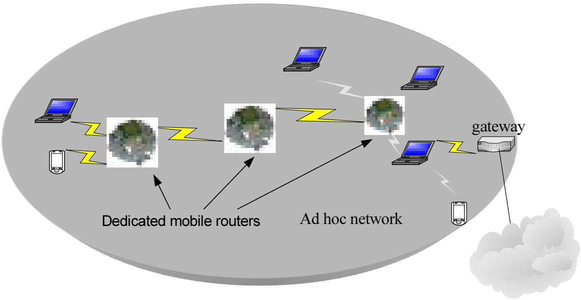


Fig. 2. Interconnecting an ad hoc network to an infrastructure with dedicated robots

The ad hoc network is considered as the extension of an existing infrastructure, that is, it constitutes a means to access the fixed network even when the terminals are not within the wireless range of an access point. In this context, mobile robots will move, function to network state, to provide wireless connectivity, wireless quality link adaptation. In addition, the robots since they are dedicated to routing may support several wireless interfaces in order to provide a bridge between different wireless technologies.

3.2 Improving Overall Performance with Mobile Robots and Terminal Differentiation

We will now aim at describing a routing protocol for a hierarchical ad-hoc network running on the IPv6 protocol stack. This stack has been chosen since it is better fitted to mobile environments than IPv4. Generally speaking, there will be two groups of terminals in the network – routers and nodes that usually do not forward packets. Moreover, there will be a possibility of distinguishing several kinds of non forwarding nodes – laptops, PDAs and Bluetooth devices – in order to provide real-time traffic support and a better aggregated bandwidth.

Most of the ad-hoc networking protocols proposed today suggests an optimization of the number of hops from a source to a destination. However, in a wireless environment using IEEE 802.11, the throughput of a link decreases with the signal quality. This is due to the fact that a host far away from an Access Point is subject to important signal fading and interference. To cope with this problem, hosts change their modulation type, which degrades the bit rate to some lower value. When an optimal number of hops scheme is selected, several problems are thus likely to occur:

- the throughput of the wireless link will drop from 11 to 5.5, 2 or even 1 Mb/s, when repeated unsuccessful frame transmissions are detected, causing a performance degradation. This performance degradation is not only detrimental to the nodes that use a lower throughput, but also to all the nodes within cell associated to these nodes, as shown in [12]. This is a consequence of the CSMA/CA channel access method that guarantees an equal long term channel access probability to all hosts. When one host captures the channel for a long time because its bit rate is low, it penalizes other hosts that use higher rates,
- a next hop that is already far from a source is likely to get out of the range of the source within a short time, thus triggering route discovery mechanisms that use some network resources.

Our scheme provides a solution to these problems by routing the packets preferentially on a wireless network core made of intelligent mobile routers that move in the environment in order to maintain a good link quality (and thus a high throughput) within the core. As a consequence, the link quality remains stable and QoS can be offered, as we will show in the following. Moreover, since the intelligent routers are high capacity autonomous entities, they have no particular power consumption issues and are not reluctant to perform routing for other terminals. We can easily imagine that the routing function can be implemented in dedicated terminals, mobile robots in our case, that roam intelligently in the environment, or even in cars in the case of a large scale metropolitan network. We, in the Ambience project, use a beacon based location management using Zigbee interfaces (Philips technology) for the localization system. For example, in an outdoor environment, the GPS system may be used. If the routing functionality is performed by routers implemented in vehicles, no behavior can be enforced in the set of vehicles. Instead, the routing algorithm has to adapt to maintain a high quality and reliable network core.

3.3 Cross-Layer QoS Management

In this section, we present a combined approach to combine IP and MAC QoS mechanisms in order to support real time traffic in a stable core ad hoc network using terminal differentiation

We showed in [5] that at layer 3, QoS functionalities differentiate packet treatment within a node but do not guarantee access to the link. At layer 2, a differentiated access to the link is provided, but packets are not classified and scheduled within a node. Therefore, they may be delayed by best effort packets within the same node that should be served before (and that may not get access to the link if higher priority packets are waiting on other nodes). We proposed a cross-layer QoS mechanism that takes the best of each QoS management methods.

Obviously, the IP service differentiation in wired or wireless networks provides a better quality for high priority applications. However, in wireless mobile networks, the medium is fairly shared between mobile stations being in the same range of transmission, and hence, all the packets will have the same probability to access the channel. On the one hand, the aim of IP prioritization is to differentiate packets in each node (locally); on the other hand, MAC QoS controls the medium access between mobile stations sharing the same wireless link. For example, in a dynamic environment, a service differentiation provided by IP layer can be limited, where packets can be delayed or dropped due to medium contentions and interference, and consequently IP mechanisms not sufficient. The IP service differentiation consists of a simple prioritization principle in node queues using a strict priority queue, giving thus a differential treatment to packets of each class. Priority is based on the DSCP field (in IP packet). It can be noted that other IP quality of service mechanisms can be used, such as admission control, in order to monitor the network load, and/or reservation protocol to ensure a quantitative quality of service for real-time flows. In addition, a modified 802.11 MAC protocol is used. It allows two types of services, where prioritization is rather simple, based on the DIFS parameter of the Distributed Control Function. A short DIFS is attributed to high priority packets, contrary to low priority packets.

Our goal is to maintain the packet priority already assigned in the IP layer to access the wireless shared medium. We define two service classes, a high priority class (real-time application: delay and packets loss sensitive), and a best effort class with no quality of service requirements. QoS is thus performed both at layer 2 and at layer 3 and a QoS mapping is done.

In the following section, we will demonstrate using simulations that QoS support is clearly improved when used in conjunction with an ad hoc routing protocol performing terminal differentiation.

4 Simulation and Results

The goal of our simulations is to compare the performance of the network under different conditions and its impact on the high priority flows (ex: real time applications).

To ensure real time applications quality of service, end to end packet delay and jitter (delay variation) must be bounded and reduced (delay less than 150 ms), and throughput guaranteed. Without QoS mechanisms, these requirements can not be met, where packets can be delayed or dropped at IP layer (node queue), or MAC layer (access to the shared medium).

Our simulation model is built using the Network Simulator tool (NS-2) [10]. It consists of an ad hoc network, connected to a wired infrastructure, with few nodes (10 nodes) moving in an area of 1000x1000 meters. Figure 3 and 4 show the throughput and end to end delay experienced by two flows (with 130 Kb/s rate) with and without terminal differentiation. We notice that, as others flows start (at t=100s), congestion occurs because of the limited link capacity. The two traffics suffer from a high increase in their end-to-end (490 ms) delay and a lower throughput.

When no differentiation mechanisms, all packets have the same priority in mobile node queues, and the same probability to access the wireless shared medium. The wireless channel is fairly shared between different flows, and packets are treated in a FIFO principle. In the terminal differentiation case, we observe a better throughput and end-to-end delay. However, flow 1 throughput still not guaranteed. In this case, we ensure good quality of wireless links, increasing thus the ad hoc network capacity, but no differentiation between different applications requirements. Thus, other traffics can degrade delay and throughput sensitive applications.

In Figure 5 and 6, we show real time traffic performance when combining two approaches, with a link quality sensitive routing algorithm and service differentiation at both IP and MAC layer.

We observe bounded end-to-end delays (32 ms) and jitter (17 ms) for real-time flows, and guaranteed throughput which allow a good quality of real-time communications.
For more details on simulation model and results see [5].

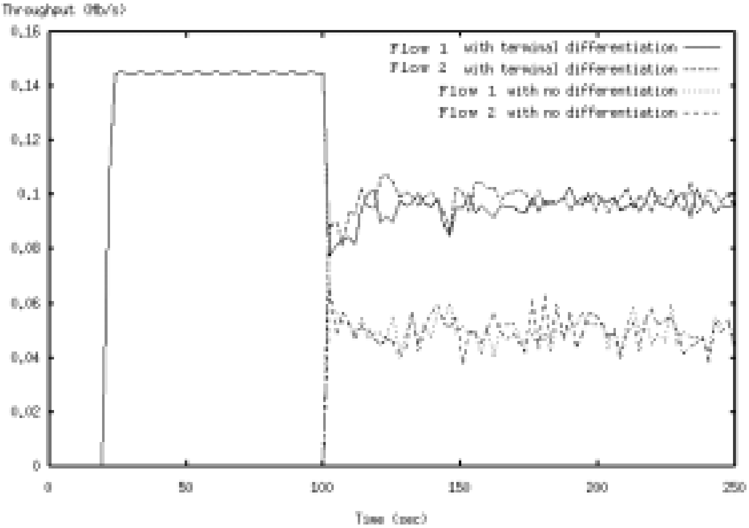


Fig. 3. Throughputs with/without terminal differentiation

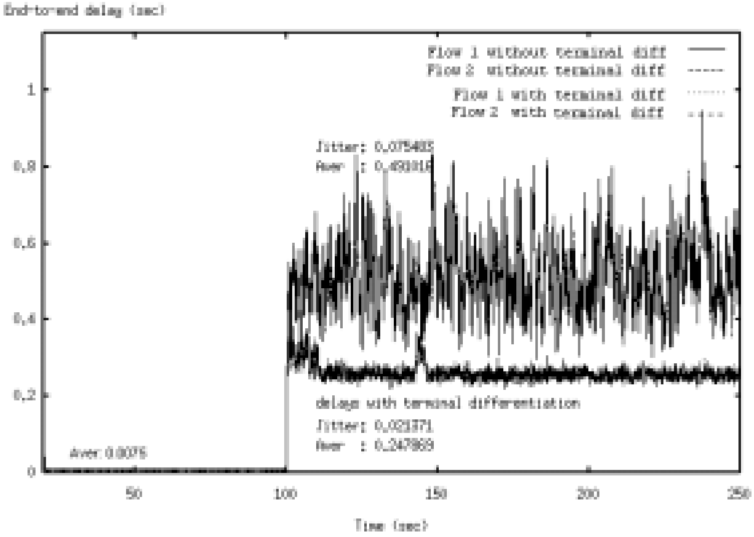


Fig. 4. Delays with/without terminal differentiation

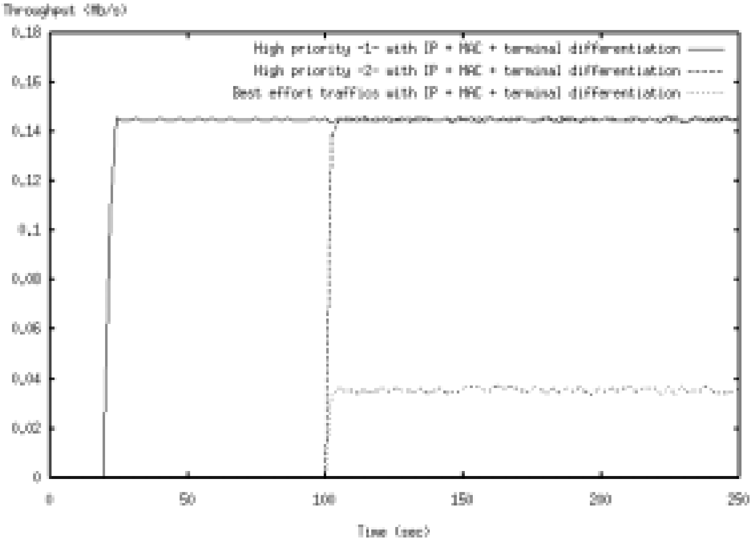


Fig. 5. Throughputs with cross-layer QoS and terminal differentiation

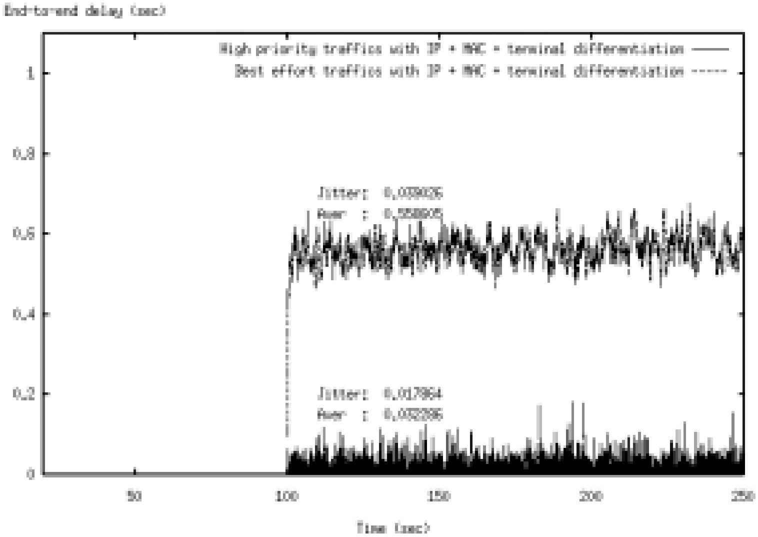


Fig. 6. Delays with cross-layer QoS and terminal differentiation

The combined mechanisms thus provide a good performance in the ad hoc network environment. Terminal differentiation provides stable, high quality and high data rate wireless links; cross-layer QoS reduces the end-to-end delay and the drop probability of real-time traffic.

5 Conclusion

In this paper, we presented an efficient approach to support real time multimedia applications in an intelligent mobile network using mobile robot routers. These are used to ensure a seamless communication service with a better network coverage and interconnection to fixed infrastructure. A cross-layer quality of service management IP and MAC is also performed to meet real time flows requirements.

The next steps of this work concern:

- The definition of a distributed admission control mechanism that limits the network load, taking into account the variable quality of wireless links,
- Robots movement management in order to reach a better ad hoc network capacity and performance.

References

1. Ambience project (ITEA), http://www.extra.research.philips.com/euprojects/ambience/
2. C. E. Perkins, ed., Ad Hoc Networking, Addison-Wesley, Boston, 2001, pp.2–3

3. C.-K. Toh, “A novel distributed routing protocol to support ad-hoc mobile computing”, Wireless Personal Communication, Jan. 1997
4. C.K. Toh, H. Cobb, D.A. Scott, «Performance Evaluation Of battery Life Aware routing Schemes for Wireless Ad hoc Networks» 2001. 0-7803-7097-1 IEEE
5. G. Le Grand, R. Meraihi, Cross-layer QoS and terminal differentiation in ad hoc networks for real-time service support, to appear in the proceedings of MedHOC NET 2003, (IFIP-TC6-WG6.8), Mahdia, Tunisia, June 25–27 2003
6. H. Xiao, K.C. Chua and K.G. Seah, ‘Quality of Service Models for Ad-Hoc Wireless Network’, ECE-ICR, Wireless Communications Laboratory, Appeared in "Handbook of Ad hoc Wireless Networks" which was published in late 2002 by CRC Press, FL, USA http://www.ece-icr.nus.edu.sg/journal1/fqmmhandbook02.pdf
7. I. Aad, C. Castelluccia, ‘Differentiation mechanisms for IEEE 802.11’, IEEE Infocom 2001, Anchorage - Alaska, April 22–26^{h}, 2001
8. IEEE 802.11 WG, Draft Supplement to STANDARD FOR Telecommunications and Information Exchange Between Systems-LAN/MAN Specific Requirements – Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) specifications: Medium Access Control (MAC) Enhancement for Quality of Service (QoS), IEEE 802.11e/Draft 3.0, May 2002
9. J. L. Sobrinho, A. S. Krishnakunar, “Quality-Of-Service in Ad hoc Carrier sense multiple access Wireless networks”, IEEE Journal on Selected Preas in Communications, pp. 1353–1368, August 1999
10. K. Fall, K. Varadhan, The ns manual, http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf
11. K. Wu and J. Harms, "QoS Support in Mobile Ad Hoc Networks," Crossing Boundaries-the GSA Journal of University of Alberta*,* Vol. 1, No. 1, Nov. 2001, pp.92–106
12. M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, “Performance anomaly of 802.11b”. To appear in Proceedings of IEEE INFOCOM 2003, San Francisco, USA, March 30-April 3, 2003
13. M. Pearlman, Z. Hass, “Determining the optimal configuration for the zone routing protocol”, IEEE selected area in communication, August, 1999
14. N. Roux. “Cost Adaptive Mechanism (CAM) for Mobile Ad Hoc Reactive Routing Protocols”, Master Thesis, ENST 2000
15. S.-J. Lee, M. Gerla, «Dynamic Load-Aware Routing in Ad hoc Networks». Proceedings of ICC 2001, Helsinki, Finland, June 2001
16. S.-J. Lee, M. Gerla, and C.-K. Toh, “A Simulation Study of Table-Driven and On-Demand Routing Protocols for Mobile Ad-Hoc Networks”, IEEE Network, vol. 13, no. 4, Jul. 1999, pp. 48–54
17. S. Jiang, D. He, and J. Rao , “A Link Availability Prediction Model for Wireless Ad Hoc Networks”, proceedings International Workshop on Wireless Networks and Mobile Computing Taipei, Taiwan, April 10–13, 2000
18. S. K. Das, A. Mukherjee, Bandypadhyay, Krishna Paul, D.Saha, «Improving Quality-of-service in Ad hoc Wireless NetWorks with Adaptive Multi-path Routing»
19. SPIF, **h**ttp://www.enst.fr/~spif
20. S. Sheng, “Routing Support for Providing guaranteed end-to-end Quality of Service”,PH.D Thesis, University of IL at Urbana Champaign, http://cairo.cs.uiuc.edu/papers/SCthesis.ps, 1999
21. The Institute of Electrical and Electronics Engineers, Inc. IEEE Std 802.11 – Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications. The Institute of Electrical and Electronics Engineers, Inc., 1999 edition

Context Driven Observation of Human Activity

James L. Crowley

Laboratoire GRAVIR, INRIA Rhône Alpes,
655 Ave de l'Europe, F-38330 Montbonnot, France
Crowley@inrialpes.fr
http://www-prima.imag.fr

Abstract. Human activity is extremely complex. Current technology allows us to handcraft real-time perception systems for a specific perceptual task. However, such an approach is inadequate for building systems that accommodate the variety that is typical of human environments. In this paper we define a framework for context aware observation of human activity. A context in this framework is defined as a network of situations. A situation network is interpreted as a specification for a federation of processes to observe humans and their actions. We present a process-based software architecture for building systems for observing activity. We discuss methods for building systems using this framework. The framework and methods are illustrated with examples from observation of human activity in an "Augmented Meeting Environment".

1 Introduction

This paper presents a framework for context aware observation of human activity. Within this framework, contexts are modeled as a network of situations. Situation models are interpreted to dynamically configure a federation of processes for observing the entities and relations that define a situation. We propose a software architecture based on dynamically assembled process federations [1], [2]. Our model builds on previous work on process-based architectures for machine perception and computer vision [3], [4], as well as on data flow models for software architecture [5].

Within this framework, a situation is described by a configuration of relations for observed entities. Changes in relations correspond to events that signal a change in situation. Such events can be used to trigger actions by the system. Situation models are used to specify an architecture in which reflexive processes are dynamically composed to form federations for observing and predicting situations. We believe that this architecture provides a foundation for the design of systems that act as a silent partner to assist humans in their activities in order to provide appropriate services without explicit commands and configuration. In the following section we review the use of the term "context aware" in different domains. This leads us to a situation-based approach to modeling context.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 101–118, 2003.

2 A Brief History of Context

The word "context" has come to have many uses. Winograd [6] points out that the word “Context” has been adapted from linguistics. Composed of “con” (with) and “text”, context refers to the meaning that must be inferred from the adjacent text. Such meaning ranges from the references intended for indefinite articles such as “it” and “that” to the shared reference frame of ideas and objects that are suggested by a text. Context goes beyond immediate binding of articles to the establishment of a framework for communication based on shared experience. Such a shared framework provides a collection of roles and relations with which to organize meaning for a phrase.

Early researchers in both artificial intelligence and computer vision recognized the importance of a symbolic structure for understanding. The “Scripts” representation [7] sought to provide just such information for understanding stories. Minsky’s Frames [8] sought to provide the default information for transforming an image of a scene into a linguistic description. Semantic Networks [9] sought to provide a similar foundation for natural language understanding. All of these were examples of what might be called “schema” [10]. Schema provided context for understanding, whether from images, sound, speech, or written text. Recognizing such context was referred to as the “Frame Problem” and became known as one of the hard unsolved problems in AI.

In computer vision, the tradition of using context to provide a framework for meaning paralleled and drew from theories in artificial intelligence. The “Visions System” [11] expressed and synthesized the ideas that were common among leading researchers in computer vision in the early 70’s. A central component of the “Visions System” was the notion of a hierarchical pyramid structure for providing context. Such pyramids successively transformed highly abstract symbols for global context into successively finer and more local context terminating in local image neighborhood descriptions that labeled uniform regions. Reasoning in this system worked by integrating top-down hypotheses with bottom-up recognition. Building a general computing structure for such a system became a grand challenge for computer vision. Successive generations of such systems, such as the “Schema System”[12] and “Condor” [13] floundered on problems of unreliable image description and computational complexity. Interest in the 1990’s turned to achieving real time systems using “active vision” [14], [15]. Many of these ideas were developed and integrated into a context driven interpretation within a process architecture using the approach “Vision as Process” [16]. The methods for sensing and perceiving context for interaction described below draws from this approach.

Context awareness has become very important to mobile computing where the term was first introduced by Schilit and Theimer [17]. In their definition, context is defined as “the location and identities of nearby people and objects and changes to those objects”. While this definition is useful for mobile computing, it defines context by example, and thus is difficult to generalize and apply to other domains. Other authors, such as [18] [19] and [20] have defined context in terms of the environment or situation. Such definitions are essentially synonyms for context, and are also difficult to apply operationally. Cheverest [21] describes context in anecdotal form using scenarios from a context aware tourist guide. His system is considered one of the early models for a context aware application.

Pascoe [22] defines context to be a subset of physical and conceptual states of interest to a particular entity. This definition has sufficient generality to apply to a recognition system. Dey [23] reviews definitions of context, and provides a definition of context as "any information that can be used to characterize situation". This is the sense in which we use the term context. Situation refers to the current state of the environment. Context specifies the elements that must be observed to model situation. However, to apply context in the composition of perceptual processes, we need to complete a clear definition with an operational theory. Such a foundation is provided by a process-based software architecture.

3 Perceptual Components for Context Awareness

In this section we describe a process-based software architecture for real time observation of human activity. The basic component of this architecture is a perceptual process. Perceptual processes are composed from a set of modules controlled by a supervisory controller. We describe several common classes of modules, and describe the operation of the controller. We also present several classes of perceptual processes and discuss how they can be combined into process federations according to a network of expected situations.

3.1 Perceptual Processes

A system's view of the external world is driven by a collection of sensors. These sensors generate observations that may have the form of numeric or symbolic values. Observations may be produced in a synchronous stream or as asynchronous events. In order to determine meaning from observations, a system must transform observations into some form of action. Such transformations may be provided by perceptual processes.

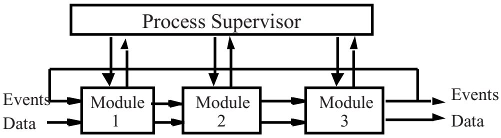


Fig. 1. A perceptual process integrates a set of modules to transform data streams or events into data streams or events.

Perceptual processes are composed from a collection of modules controlled by a process supervisor, as shown in figure 1. Processes operate in a synchronous manner within a shared address space. In our experimental system, the process supervisor is implemented as a multi-language interpreter [24] equipped with a dynamic loader for precompiled libraries. This interpreter allows a processes to receive and interpret messages containing scripts, to add new functions to a process during execution.

The modules that compose a process are formally defined as transformations applied to a certain class of data or event. Modules are executed in cyclic manner by the supervisor according to a process schedule. We impose that transformations return an auto-critical report that describes the results of their execution. Examples of information contained in an auto-critical report include elapsed execution time, confidence in the result, and any exceptions that were encountered. The auto-critical report enables a supervisory controller to adapt parameters for the next call in order to maintain a execution cycle time, or other quality of service.

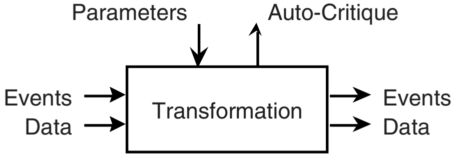


Fig. 2. Modules apply a transformation to an input data stream or events and return an auto-critical report.

3.2 Examples: Modules for Observing, Grouping, and Tracking

A typical example of a module is a transformation that uses table look-up to convert a color pixel into a probability of skin, as illustrated in figure 3. Such a table can easily be defined using the ratio of a histograms of skin colored pixels in a training image, divided by the histogram of all pixels in the same image [25]. Skin pixels for an individual in a scene will all exhibit the same chrominance vector independent of surface orientation and thus can be used to detect the hands or face of that individual [26]. This technique has provided the basis for a very fast (video rate) process that converts an RGB color image into image of the probability of detection based on color using a look-up table.

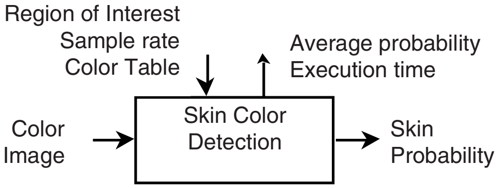


Fig. 3. A module for detecting skin colored pixels with a region of interest

A color observation module applies a specified look-up table to a rectangular “Region of Interest” or ROI using a specified sample rate. The sample rate, S, can be adapted to trade computation time for precision. The output from the module is an image in which pixels inside the ROI have been marked with the probability of detection. The auto-critical report returns the average value of the probabilities (for use as a confidence factor) as well as the number of microseconds required for execution. The average probability can be used to determine whether a target was detected within the ROI. The execution time can be used by the process supervisor to assure that the overall execution time meets a constraint. This module can be used either for initial detection or for tracking, according to the configuration specified by the supervisor.

The color observation module is one example of a pixel level observation module. Pixel level observation modules provide the basis for an inexpensive and controllable perceptual processes. In our systems, we use pixel level observation modules based on color, motion, background subtraction [27], and receptive field histograms [28]. Each of these modules applies a specified transformation to a specified ROI at a specified sample rate and returns an average detection probability and an execution time.

Interpretation requires that detected regions be grouped into "blobs". Grouping is provided by a grouping module, defined using on moments, as shown in figure 4.

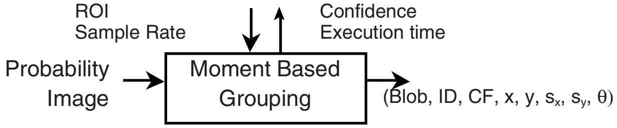


Fig. 4. A module for grouping detected pixels using moments

Let w(i,j) represent an image of detection probabilities provided by a pixel level observation process. The detection mass, M, is the sum of the probabilities within the ROI. The ratio of the sum of probability pixels to the number of pixels, N, in the ROI provides a measure of the confidence that a skin colored region has been observed.

$$M = \sum_{i,j \in ROI} w(i,j) \qquad CF = \frac{M}{N}$$

The first moment of the detected probabilities is the center of gravity in the row and column directions (x, y). This is a robust indicator of the position of the skin colored blob.

$$x = \frac{1}{M} \sum_{i,j \in ROI} w(i,j) \cdot i \qquad y = \frac{1}{M} \sum_{i,j \in ROI} w(i,j) \cdot j$$

The second moment of w(i, j) is a covariance matrix. Principal components analysis of the covariance matrix formed from σ_{ii}^2, σ_{jj}^2,and σ_{jj}^2 yield the length and breadth of (s_x, s_y), as well as its orientation θ, of the blob of detected pixels.

$$\sigma_{ii}^2 = \frac{1}{M} \sum_{i,j \in ROI} w(i,j) \cdot (i-x)^2 \qquad \sigma_{jj}^2 = \frac{1}{M} \sum_{i,j \in ROI} w(i,j) \cdot (j-y)^2$$

$$\sigma_{ij}^2 = \frac{1}{M} \sum_{i,j \in ROI} w(i,j) \cdot (i-x) \cdot (j-y)$$

Tracking is a cyclic process of recursive estimation applied to a data stream. The Kalman filter provides a framework for designing tracking processes [29]. A general discussion of the use of the Kalman filter for sensor fusion is given in [30]. The use of the Kalman filter for tracking faces is described in [31].

Tracking provides a number of fundamentally important functions for a perception system. Tracking aids interpretation by integrating information over time. Tracking makes it possible to conserve information, assuring that a label applied to an entity at time T_1 remains associated with the entity at time T_2. Tracking provides a means to focus attention, by predicting the region or interest and the observation module that should be applied to a specific region of an image. Tracking processes can be designed to provide information about position speed and acceleration that can be useful in describing situations.

In perception systems, a tracking process is generally composed of three phases: predict, observe and estimate, as illustrated in figure 4. Tracking maintains a list of entities, known as "targets". Each target is described by a unique ID, a target type, a confidence (or probability of existence), a vector of properties and a matrix of uncertainties (or precisions) for the properties.

The prediction phase uses a temporal model (called a "process model" in the tracking literature) to predict the properties that should be observed at a specified time for each target. For many applications of tracking, a simply linear model is adequate for such prediction. A linear model maintains estimates of the temporal derivatives for each target property and uses these to predict the observed property values. For example, a first order temporal model estimates the value for a property, X_{T1}, at time T_1 from the value X_{T0} at a time T_0 plus the temporal rate of change multiplied by the time step, $\Delta T = T_1 - T_0$.

$$X_{T1} = X_{T0} + \Delta T(dX/dt)_{T0}$$

Higher order linear models may also be used provided that the observation sample rate is sufficiently fast compared to the derivatives to be estimated. Non-linear process models are also possible. For example, articulated models for human motion can provide important constraints on the temporal evolution of targets.

The prediction phase also updates the uncertainty (or precision model) of properties. Uncertainty is generally represented as a covariance matrix for errors between estimated and observed properties. These uncertainties are assumed to arise from imperfections in the process model as well as errors in the observation process.

Restricting processing to a region of interest (ROI) can greatly reduce the computational load for image analysis. The predicted position of a target determines the position of the ROI at which the target should be found. The predicted size of the target, combined with the uncertainties of the size and position, can be used to estimate the appropriate size for the ROI. In the tracking literature, this ROI is part of the "validation gate", and is used to determine the acceptable values for properties.

Observation is provided by the observation and grouping modules described above. Processing is specific for each target. A call to a module applies a specified observation procedure for a target at a specified ROI in order to verify the presence of the target and to update its properties. When the detection confidence is large, grouping the resulting pixels provides the information to update the target properties.

The estimation process combines (or fuses) the observed properties with the previously estimated properties for each target. If the average detection confidence is low, the confidence in the existence of a target is reduced, and the predicted values are taken as the estimates for the next cycle. If the confidence of existence falls below a threshold, the target is removed from the target list.

The detection phase is used to trigger creation of new targets. In this phase, specified observation modules are executed within a specified list of "trigger" regions. Trigger regions can be specified dynamically, or recalled from a specified list. Target detection is inhibited whenever a target has been predicted to be present within a trigger region.

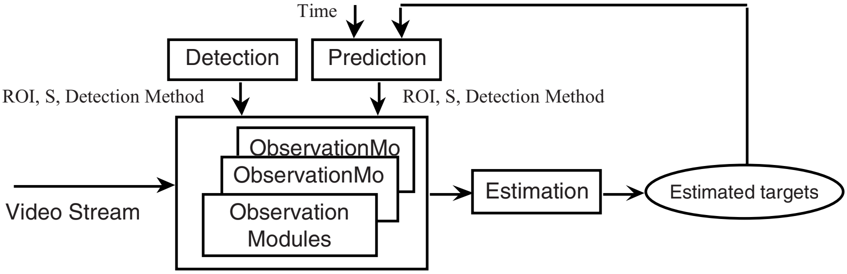


Fig. 5. Tracking is a cyclic process of four phases: Predict, Observe, Detect and Estimate. Observation is provided by the observation and grouping modules described above.

A simple zeroth order Kalman filter may be used to track bodies, faces and hands in video sequences. In this model, targets properties are represented by a "state vector" composed of position, spatial extent and orientation (x, y, s_x, s_y, θ). A 5x5 covariance matrix is associated with this vector to represent correlations in errors between parameters. Although prediction does not change the estimated position, it does enlarge the uncertainties of the position and size of the expected target. The expected size provides bounds on the sample rate, as we limit the sample rate so that there are at least 8 pixels across an expected target.

3.3 A Supervisory Controller for Perceptual Processes

The supervisory component of a process provides four fundamental functions: command interpretation, execution scheduling, parameter regulation, and reflexive description. The supervisor acts as a programmable interpreter, receiving snippets of code script that determine the composition and nature of the process execution cycle and the manner in which the process reacts to events. The supervisor acts as a scheduler, invoking execution of modules in a synchronous manner. The supervisor regulates module parameters based on the execution results. Auto-critical reports from modules permit the supervisor to dynamically adapt processing. Finally, the supervisor responds to external queries with a description of the current state and capabilities. We formalize these abilities as the autonomic properties of auto-regulation, auto-description and auto-criticism.

A process is auto-regulated when processing is monitored and controlled so as to maintain a certain quality of service. For example, processing time and precision are two important state variables for a tracking process. These two may be traded off against each other. The process controllers may be instructed to give priority to either the processing rate or precision. The choice of priority is dictated by a more abstract supervisory controller.

An auto-critical process maintains an estimate of the confidence for its outputs. Such a confidence factor is an important feature for the control of processing. Associating a confidence factor to every observation allows a higher-level controller to detect and adapt to changing circumstances. When supervisor controllers are programmed to offer "services" to higher-level controllers, it can be very useful to include an estimate of the confidence of their ability to "play the role" required for the service. A higher-level controller can compare responses from several processes and determine the assignment of roles to processes.

An auto-descriptive controller can provide a symbolic description of its capabilities and state. The description of the capabilities includes both the basic command set of the controller and a set of services that the controller may provide to a more abstract supervisor. Such descriptions are useful for the dynamic composition of federations of controllers.

3.4 Classes of Perceptual Processes

We have identified several classes of perceptual processes. The most basic class is composed of processes that detect and track entities. Entities may generally be understood as spatially correlated sets of properties, corresponding to parts of physical objects. However, correlation may also be based on temporal location or other, more abstract, relations. From the perspective of the system, an entity is any association of correlated observable variables.

Formally, an entity is a predicate function of one or more observable variables.

$$\text{Entity-process}(v_1, v_2, \ldots, v_m) \Rightarrow \text{Entity}(\text{Entity-Class, ID, CF}, p_1, p_2, \ldots, p_n)$$

Entities may be composed by an entity detection an tracking processes, as shown in figure 6.

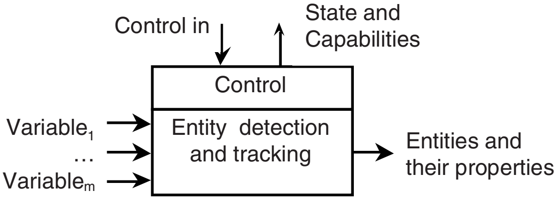


Fig. 6. Entities and their properties are detected and described by a special class of perceptual processes.

The input to an entity detection process is typically a stream of numeric or symbolic data. The output of the transformation is a stream including a symbolic token to identify the class of the entity, accompanied by a set of numerical or symbolic properties. These properties allow the system to define relations between entities. The detection or disappearance of an entity may, in some cases, also generate asynchronous symbolic signals that are used as events by other processes.

A fundamental aspect of interpreting sensory observations is determining relations between entities. Relations can be formally defined as a predicate function of the properties of entities. Relations may be unary, binary, or N-ary. For example, Visible(Entity1), On(Entity1, Entity2), and Aligned(Entity1, Entity2, Entity3) are examples of relations.

Relations that are important for describing situations include 2D and 3D spatial relations, as well as temporal relations [32]. Other sorts of relations, such as acoustic relations (e.g. louder, sharper), photometric relations (e.g. brighter, greener), or even abstract geometric relations may also be defined. As with entity detection tracking, we propose to observe relations using Perceptual processes.

Relation-observation processes are defined to transform a list of entities into a list of relations based on their properties, as shown in figure 7. Relation observation processes read in a list of entities tracked by an entity detection and tracking process and produce a list of relations that are true along with the entity or entities that render them true. Relation observation uses tracking to predict and verify relations. Thus they can generate an asynchronous event when a new relation is detected or when a previously detected relations becomes false.

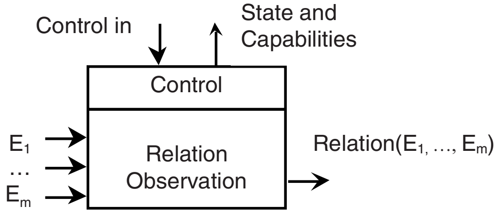


Fig. 7. Relations are predicates defined over one or more entities. Relation observation processes generate events when relations become true or false.

Composition processes assemble sets of entities into composite entities. Composition processes are similar to relation observation entities, in that they operate on a list of entities provided by an entity observation process. However, entity observation processes produce a list of composite objects satisfying a set of relations. They can also measure properties of the composite object. As with relation observation processes, composition processes can generate asynchronous events when a composite object is detected or lost.

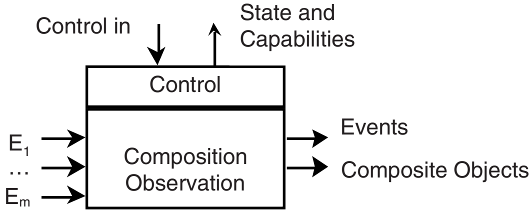


Fig. 8. Composition processes observe and track compositions of entities.

3.5 Process Federations

Perceptual processes may be organized into software federations [2]. A federation is a collection of independent processes that cooperate to perform a task. We have designed a middle ware environment that allows us to dynamically launch and connect process on different machines. In our system, processes are launched and configured by a "meta-supervisor". The meta-supervisor configures a process by sending snippets of control script to be interpreted by the controller. Each control script defines a command that can be executed by a message from the meta-supervisor. Processes may be interrogated by the meta-supervisor to determine their current state and the current set of commands.

Meta-supervisors can also launch and configure other meta-supervisors so that federations can be built up hierarchically. Each meta-supervisor invokes and controls lower level supervisors that perform the required transformation. At the lowest level are Perceptual processes that observe and track entities and observe the relations between entities. These are grouped into federations as required for to observe the situations in a context.

As a simple example of a federation of perceptual processes, consider a system that detects when a human is in the field of view of a camera and tracks his hands and face. We say that observed regions can be selected to "play the role" of torso, hands and faces. We call this a FaceAndHand observer. The system uses an entity and detection tracking process that can use background difference subtraction and color modeling to detect and track blobs in an image stream. The set of tracked entities are sent to a composition process that labels likely blobs as a torso, face or a left or right hand.

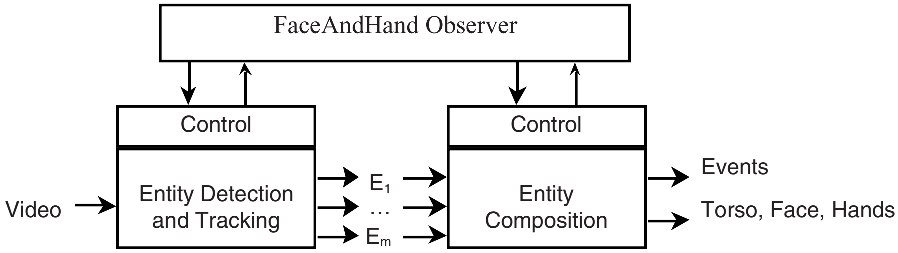


Fig. 9. A simple process federation composed of an entity detection process, a composition process and a meta-supervisor.

The control for this process federation works as follows. The meta-supervisor begins by configuring the entity detection and tracking processes to detect a candidate for torso by looking for a blob of a certain size using background subtraction in a pre-configured "detection region". The acceptance test for torso requires a blob detected by background subtraction in the center region of the image, with a size within a certain range. Thus the system requires an entity detection process that includes an adaptive background subtraction detection.

When a region passes the torso test, the composition process notifies the meta-supervisor. The meta-supervisor then configures new trigger regions using color detection modules in the likely positions of the hands and face relative to the torso. The acceptance test for face requires a skin colored region of a certain range of sizes in the upper region of the torso. Hands are also detected by skin color blob detection over a regions relative to the torso. Sets of skin colored regions are passed to the composition process so that the most likely regions can be assigned to each role. We say that the selected skin-colored regions are assigned the "roles" of face, left hand and right hand. The assignments are tracked so that a change in the entity playing the role of hand or face signals an event. Such role assignment is a simple example of a more general principle developed in the next section.

4 Context and Situation

Perceptual processes provide a means to detect and track compositions of entities and to verify relations between entities. The design problem is to determine which entities to detect and track and which relations to verify. For most human environments, there is a potentially infinite number of entities that could be detected and an infinite number of possible relations for any set of entities. The appropriate entities and relations must be determined with respect to a task or service to be provided.

In this section we discuss the methods for specifying context models for human activity, We define the concept of a "role" and explain how roles can help simplify context models. We define three classes of events in such systems, and describe the system reaction to each class. We then present two examples of simple context models. An early version of the concepts presented in this section was presented in [33] . This paper refines and clarifies many aspects of this framework in the light of experience with implementing systems based on this model.

4.1 Specifying a Context Model

A system exists in order to provide some set of services. Providing services requires the system to perform actions. The results of actions are formalized by defining the output "state" of the system. Simple examples of actions for interactive environments include adapting the ambient illumination and temperature in a room, or displaying a users "availability for interruption". More sophisticated examples of tasks include configuring an information display at a specific location and orientation, or providing information or communications services to a group of people working on a common task.

The "state" of an environment is defined as a conjunction of predicates. The environment must act so as to render and maintain each of these predicates to be true. Environmental predicates may be functions of information observed in the environment, including the position, orientation and activity of people in the environment, as well as position, information and state of other equipment. The information required to maintain the environment state determines the requirements of the perception system.

The first step in building a context model is to specify the desired system behavior. For an interactive environment, this corresponds to the environmental states, defined in terms of the variables to be controlled by the environment, and predicates that should be maintained as true. For each state, the designer then lists a set of possible situations, where each situation is a configuration of entities and relations to be observed. Although a system state may correspond to many situations, each situation must uniquely belong to one state. Situations form a network, where the arcs correspond to changes in the relations between the entities that define the situation. Arcs define events that must be detected to observe the environment.

In real examples, we have noticed that there is a natural tendency for designers to include entities and relations that are not really relevant to the system task. Thus it is important to define the situations in terms of a minimal set of relations to prevent an explosion in the complexity of the system. This is best obtained by first specifying the system output state, then for each state specifying the situations, and for each situation specifying the entities and relations. Finally for each entity and relation, we determine the configuration of perceptual processes that may be used.

4.2 Simplifying Context Models with Roles

The concept of role is an important (but subtle) tool for simplifying the network of situations. It is common to discover a collection of situations for an output state that have the same configuration of relations, but where the identity of one or more entities is varied. A role serves as a "variable" for the entities to which the relations are applied, thus allowing an equivalent set of situations to have the same representation. A role is played by an entity that can pass an acceptance test for the role. In that case, it is said that the entity can play or adopt the role for that situation. In our framework, the relations that define a situation are defined with respect to roles, and applied to entities that pass the test for the relevant roles.

For example, in a group discussion, at any instant, one person plays the "role" of the speaker while the other persons play the role of "listeners". Dynamically assigning a person to the role of "speaker" allows a video communication system to transmit the image of the current speaker at each instant. Detecting a change in roles allows the system to reconfigure the transmitted image.

Entities and roles are not bijective sets. One or more entities may play a role. A role may be played by one or several entities. The assignment of entities to roles may (and often will) change dynamically. Such changes provide the basis for an important class of events : role-events. Role events signal a change in assignment of an entity to a role, rather than a change in situation.

Roles and relations allow us to specify a context model as a kind of "script" for activity in an environment. However, unlike theater, the script for a context is not necessarily linear. Context scripts are networks of situations where a change in situations is determined based on relations between roles.

4.3 Context and Situation

To summarize, a context is a composition of situations that concerns a set of roles and relations. A context determines the configuration of processes necessary to detect and observe the entities that can play the roles and the relations between roles that must be observed. The roles and relations should be limited to the minimal set necessary for recognizing the situations necessary for the environmental task. All of the situations in a context are observed by the same federation.

$$\text{Context} \Rightarrow \{\text{Role}_1, \text{Role}_2,\ldots,\text{Role}_n; \text{Relation}_1,\ldots,\text{Relation}_m\}$$

A situation is a kind of state, defined by a conjunction of relations. Relations are predicate functions evaluated over the properties of the entities that have been assigned to roles. A change in the assignment of an entity to a role does not change the situation, unless a relation changes in value.

Entities are assigned to roles by role assignment processes. The context model specifies which roles are to be assigned and launches the necessary role assignment processes. A meta-supervisor determines what kind of entities can play each role, and launches processes to detect and observe these entities. A description of each detected entity is returned to the role assignment process where it is subjected to the acceptance test to determine its suitability for the role based on type, properties and confidence factor. The most suitable entity (or entities) is (are) assigned to the roles. Relations are then evaluated, and the set of relations determines the situation.

The situation is a set of relations computed on the entities assigned to roles. Situation changes when the relations between entities change. If the assignment of entities to situations changes, the situation remains the same. However, the system may need to act in response to a change in role assignment. For example, if the person playing the role of speaker changes, then a video communication system may need to change the camera view to center on the new speaker.

4.4 Classes of Events

From the above, we can distinguish three classes of events: Role Events, Relation events and Context events.

Role events signal a change in the assignment of entities to roles. Such a may result in a change in the system output state and thus require that the system act so as to bring the state back to the desired state. For example, the change in speaker (above) renders a predicate Camera-Aimed -At(Speaker) false, requiring the system to selected the appropriate camera and orient it to the new speaker. Situation events or (relation events) signal changes in relations that cause a change in situation. If the person playing the role of speaker stops talking and begins writing on a blackboard, then the situation has changed. Context events signal changes in context, and usually require a reconfiguration of the perceptual processes.

Role events and Situation-Events are data driven. The system is able to interpret and respond to them using the context model. They do not require a change in the federation of Perceptual processes. Context events may be driven by data, or by some external system command .

4.5 A Simple Example: An Interuptibility Meter

As first simple example, consider a system whose task is to display the level of "interruptibility" of a user in his office environment. Such a system may be used to automatically illuminate a colored light at the door of the office, or it may be used in the context of a collaborative tool such as a media-space [34]. The set of output actions are very simple. The environment should display one of a set of interruptibility states. For example, states could be "Not in Office", "Ok for interruptions", "urgent interrupts only" and "Do not Disturb".

Suppose that the user has decided that his four interruptibility states depend on the following eight situations, labeled s1 to s8: (S1) The user is not in office when the user is not present. He is interruptible when (S2) alone in his office or (S3) not working on the computer or (S4) talking on the phone. He may receive urgent interruptions when (S5) working at his computer, or when (S6) visitors are standing in the office. The user should not be interrupted (S7) when on the phone, or (S8) when the visitors are sitting in his office.

The roles for these situations are <User>, <Visitor>, <Computer>, <Phone>. The <User> role may be played by a human who meets an acceptance test. This test could be based on an RFID badge, a face recognition system, a spoken password, or a password typed into a computer. A <Visitor> is any person in the office who has not met the test for <User>. A person is a class of entity that is detected and tracked by a person observation process. For example, this can be a simple visual tracker based on subtraction from an adaptive background.

The predicate "Present(User)" would be true whenever a person observed to be in the office has been identified as the <User>. The fact that entities are tracked means that the person need only be identified once. Evaluating the current situation requires applying a logical test for each person. These tests can be applied when persons enter the office, rather then at each cycle.

Situation S1 would be true if the no person being tracked in the office passes the test for user. Situation S2 also requires a predicate to know if a person is playing the role of visitor. States S4 and S7 require assigning an entity to the role <Phone>. This can be done in naïve manner by assigning regions of a certain color or texture at a certain location. However, if we wish to include cellular phones we would need more sophisticated vision processes for the role assignment.

For assigning an object to the role of <Computer> a simple method would be to consider a computer as a box of a certain color at a fixed location. The <User> could then be considered to be using the computer if a face belonging to his torso is in a certain position and orientation. Facing would normally require estimating the position and orientation of a person's face. The test would be true if the orientation of the position of the face was within a certain distance of the computer and the orientation of the face were opposite the compute screen. Again, such tests could be arbitrarily sophisticated, arbitrarily discriminant and arbitrarily costly to develop. Situations S6 and S8 require tests for persons to be sitting and standing. These can be simple and naïve or sophisticated and expensive depending on how the system is to be used.

4.6 Second Example: A Video Based Collaborative Work Environment

As a second example, consider a video based collaborative work environment. Two or more users are connected via high bandwidth video and audio channels. Each user is seated at a desk and equipped with a microphone, a video communications monitor and an augmented work surface. Each user's face and eyes are observed by a steerable pan-tilt-zoom camera. A second steerable camera is mounted on the video display and maintains a well-framed image of the user's face. The augmented workspace is a white surface, observed by a third video camera mounted overhead.

The system task is to transmit the most relevant image of the user. If the user is facing the display screen, then the system will transmit a centered image of the users face. If the user faces his drawing surface, the system will transmit an image of the drawing surface. If the user is facing neither the screen nor the drawing surface then the system will transmit a wide-angle image of the user within the office. This can be formalized as controlling two functions: transmit(video-stream) and center(camera, target). For each function, there is a predicate that is true when the actual value corresponds to the specified value.

The roles that compose the context are 1) the user 2) the display screen, 3) a writing surface. The user is a composite entity composed of a torso, face and hands. The system's task is to determine one of three possible views of the user: a well-centered image of the user's face, the user's workspace and an image of the user and his environment. Input data include the microphone signal strength, and a coarse resolution estimation of the user's face orientation. The system context includes the roles “speaker” and “listener”. At each instant, all users are evaluated by a meta-supervisor to determine assignment to one of the roles "speaker" and “listener”. The meta-supervisor assigns one of the users to the role speaker based on recent energy level of his microphone. Other users are assigned the role of listener. All listeners receive the output image of the speaker. The speaker receives the mosaic of output images of the listeners.

The user may place his attention on the video display, or the drawing surface or “off into space”. This attention is manifested by the orientation of his face, as measured by positions of his eyes relative to the center of gravity of his face (eye-gaze direction is not required). When the user focuses attention on the video display, his output image is the well-framed image of his face. When a user focuses attention on the work surface, his output image is his work-surface. When the user looks off “into space”, the output image is a wide-angle view of the user’s environment. This system uses a simple model of the user’s context completed by the system’s context to provide the users with the appropriate video display. Because the system adapts its display based on the situation of the group of users, the system, itself, fades from the user’s awareness.

5 Conclusions

A context is a network of situations concerning a set of roles and relations. Roles are services or functions relative to a task. Roles may be “played” by one or more entities. A relation is a predicate defined over the properties of entities. A situation is a configuration of relations between the entities.

This ontology provides the basis for a software architecture for the perceptual components of context aware systems. Observations are provided by perceptual processes defined by a tracking process or transformation controlled by reflexive supervisor. Perceptual processes are invoked and organized into hierarchical federations by reflexive meta-supervisors. A model of the user’s context makes it possible for a system to provide services with little or no intervention from the user.

Acknowledgment.
This work has been partly supported by the EC project IST FAME project (IST-2000-28323) and IST CAVIAR (IST 2001-37540) as well as French national project RNTL/ProAct CONTACT. This work has been performed in active collaboration with Joelle Coutaz, Gaetan Rey, Patrick Reignier, Dave Snowdon, Jean-Luc Meunier and Alban Caporossi.

References

[1] Software Process Modeling and Technology, edited by A. Finkelstein, J. Kramer and B. Nuseibeh, Research Studies Press, John Wiley and Sons Inc, 1994.

[2] J. Estublier, P. Y. Cunin, N. Belkhatir, "Architectures for Process Support Ineroperability", ICSP5,Chicago, 15-17 juin, 1997.

[3] J. L. Crowley, "Integration and Control of Reactive Visual Processes", Robotics and Autonomous Systems, Vol 15, No. 1, décembre 1995.

[4] J. Rasure and S. Kubica, “The Khoros application development environment “, in Experimental Environments for computer vision and image processing, H. Christensen and J. L. Crowley, Eds, World Scientific Press, pp 1-32, 1994.

[5] M. Shaw and D. Garlan, Software Architecture: Perspectives on an Emerging Disciplines, Prentice Hall, 1996.

[6] T. Winograd, “Architecture for Context”, Human Computer Interaction, Vol. 16, pp401-419.

[7] R. C. Schank and R. P. Abelson, Scripts, Plans, Goals and Understanding, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

[8] M. Minsky, "A Framework for Representing Knowledge", in: The Psychology of Computer Vision, P. Winston, Ed., McGraw Hill, New York, 1975.

[9] M. R. Quillian, "Semantic Memory", in Semantic Information Processing, Ed: M. Minsky, MIT Press, Cambridge, May, 1968.

[10] D. Bobrow: "An Overview of KRL", Cognitive Science 1(1), 1977.

[11] A. R. Hanson, and E. M. Riseman, , VISIONS: A Computer Vision System for Interpreting Scenes, in Computer Vision Systems, A.R. Hanson & E.M. Riseman, Academic Press, New York, N.Y., pp. 303-334, 1978.

[12] B. A.Draper, R. T. Collins, J. Brolio, A. R. Hansen, and E. M. Riseman, "The Schema System", International Journal of Computer Vision, Kluwer, 2(3), Jan 1989.

[13] M.A. Fischler & T.A. Strat. Recognising objects in a Natural Environment; A Contextual Vision System (CVS). DARPA Image Understanding Workshop, Morgan Kauffman, Los Angeles, CA. pp. 774-797, 1989.

[14] R. Bajcsy, Active perception, Proceedings of the IEEE , Vol. 76, No 8, pp. 996-1006, August 1988.

[15] J. Y. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active Vision", International Journal of Computer Vision, Vol. 1, No. 4, Jan. 1988.

[16] J. L. Crowley and H. I Christensen, Vision as Process, Springer Verlag, Heidelberg, 1993.

[17] B. Schilit, and M. Theimer, “Disseminating active map information to mobile hosts”, IEEE Network, Vol 8 pp 22-32, 1994.

[18] P. J. Brown, “The Stick-e document: a framework for creating context aware applications”, in Proceedings of Electronic Publishing, ’96, pp 259-272.

[19] T. Rodden, K.Cheverest, K. Davies and A. Dix, “Exploiting context in HCI design for mobile systems”, Workshop on Human Computer Interaction with Mobile Devices 1998.

[20] A. Ward, A. Jones and A. Hopper, “A new location technique for the active office”, IEEE Personal Comunications 1997. Vol 4. pp 42-47.

[21] K. Cheverest, N. Davies and K. Mitchel, “Developing a context aware electronic tourist guide: Some issues and experiences”, in Proceedings of ACM CHI ’00, pp 17-24, ACM Press, New York, 2000.

[22] J. Pascoe “Adding generic contextual capabilities to wearable computers”, in Proceedings of the 2nd International Symposium on Wearable Computers, pp 92-99, 1998.

[23] Dey, A. K. “Understanding and using context”, Personal and Ubiquitous Computing, Vol 5, No. 1, pp 4-7, 2001.

[24] A. Lux, "The Imalab Method for Vision Systems", International Conference on Vision Systems, ICVS-03, Graz, april 2003.

[25] K. Schwerdt and J. L. Crowley, "Robust Face Tracking using Color", 4th IEEE International Conference on Automatic Face and Gesture Recognition", Grenoble, France, March 2000.

[26] M. Storring, H. J. Andersen and E. Granum, "Skin color detection under changing lighting conditions", Journal of Autonomous Systems, June 2000.

[27] J. Piater and J. Crowley, "Event-based Activity Analysis in Live Video using a Generic Object Tracker", Performance Evaluation for Tracking and Surveillance, PETS-2002, Copenhagen, June 2002.

[28] D. Hall, V. Colin de Verdiere and J. L. Crowley, "Object Recognition using Coloured Receptive Field", 6th European Conference on Computer Vision, Springer Verlag, Dublin, June 2000.

[29] R. Kalman, "A new approach to Linear Filtering and Prediction Problems", Transactions of the ASME, Series D. J. Basic Eng., Vol 82, 1960.

[30] J. L. Crowley and Y. Demazeau, “Principles and Techniques for Sensor Data Fusion“, Signal Processing, Vol 32 Nos 1-2, p5-27, May 1993.

[31] J. L. Crowley and F. Berard, "Multi-Modal Tracking of Faces for Video Communications", IEEE Conference on Computer Vision and Pattern Recognition, CVPR '97, St. Juan, Puerto Rico, June 1997.

[32] J. Allen, "Maintaining Knowledge about Temporal Intervals", Journal of the ACM, 26 (11) 1983.

[33] J. L. Crowley, J. Coutaz, G. Rey and P. Reignier, "Perceptual Components for Context Aware Computing", UBICOMP 2002, International Conference on Ubiquitous Computing, Goteborg, Sweden, September 2002.

[34] J. L. Crowley, J. Coutaz and F. Berard, "Things that See: Machine Perception for Human Computer Interaction", Communications of the A.C.M., Vol 43, No. 3, pp 54-64, March 2000.

[35] Schilit, B, N. Adams and R. Want, “Context aware computing applications”, in First international workshop on mobile computing systems and applications, pp 85 - 90, 1994.

[36] Dey, A. K. “Understanding and using context”, Personal and Ubiquitous Computing, Vol 5, No. 1, pp 4-7, 2001.

Interacting in Desktop and Mobile Context: Emotion, Trust, and Task Performance

Mark Neerincx and Jan Willem Streefkerk

TNO Human Factors, P.O. Box 23,
3769 ZG Soesterberg, The Netherlands
{neerincx, streefkerk}@tm.tno.nl

Abstract. The Personal Assistant for onLine Services (PALS) project aims at attuning the interaction with mobile services to the momentary usage context. Among other thing, PALS should adequately address emotional states of the user and support users building up an adequate trust level during service interactions. In an experiment, participants performed interaction tasks with mobile services, on a small handheld device or a laptop. Before each task session, they watched film clips with different emotional impact (i.e. valence and arousal). In addition to performance measures, we measured trust and heart rate. In the handheld condition, task *performance* was substantially worse and showed a more extensive navigation path (i.e. more 'wrong' hyperlinks) to find the target information. Furthermore, during the experiment *trust* in the web services hardly increased in the handheld condition whereas trust substantially improved in the laptop condition. Device and service proved to be the major factors that determine the user experience. There is a clear need to improve the mobile interaction with web services in order to establish an adequate performance *and* trust level (e.g. by attentive interactive displays).

1 Introduction

Due to the development of (wireless) networks and mobile devices, an increasing number and variety of users can access electronic services in a continuous changing usage context. The PALS project aims at a "Personal Assistant for onLine Services", which attunes the user interface to the momentary, individual user interests, capacities, usage history, device and environment. PALS will be developed using a cognitive engineering development approach that provides theoretically and empirically founded concepts for adaptation [8]. A scenario analysis and literature research provided high-level user requirements [12]. Human-Computer Interaction (HCI) knowledge and enabling technologies are lacking to fully realise the identified user requirements, in particular with respect to (1) how to address user's momentary attention, emotion and trust, and (2) how to dynamically structure navigation and derive a user model.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 119–132, 2003.

Nagata [11] provides the first results of the study on *attention*, showing that disruptions in a multitasking setting have a critical impact on a person's attention, limiting task performance, in particular for mobile devices compared to desktop. The disruption is larger when it appears on the device (e.g. instant messaging) than when it comes from an "external object" (e.g. a phone call). In particular for the mobile device, the expectation by the user of receiving an interruption decreased the disruption effect. These results provided an empirical foundation for design concepts to support users' attention by mediating interruptions and an attentive interactive display for better web task performance.

Herder and Van Dijk [3] show the first developments on the technology for adapting the *navigation* structure and deriving *user models*. This adaptation is based on the notion that user navigation paths indicate user interests and navigation strategies. From navigation paths one can predict "lostness in hyperspace" and "navigation efficiency and effectiveness". Navigation structures are modelled as graphs and user navigation is viewed as an overlay of the site graph.

This paper presents a study on *emotion* and *trust* as important elements of the user experience with mobile services. PALS should adequately address emotional states of the user and support users building up an adequate trust level during service interactions. In a similar way as for attention, we first need insight in the effects of device and emotional state on user behaviour and subjective trust in order to create and improve design concepts.

2 User Experience

2.1 Emotion

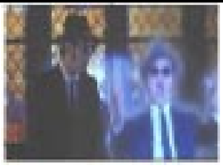
Human behaviour comprises physical, cognitive and affective processes. From the 90's, researchers started to study affection in more detail providing new insights in user experience (e.g., Picard, 1997). For example, Klein, Moon, and Picard [7] showed that "emotional support", consisting of sympathy and compassion expression, leads to less frustration and more prolonged interaction with a computer game. Norman [13] stated that a positive emotional state leads to divergent and creative thinking and problem solving. On the other hand, a negative emotional state causes tunnel vision and higher concentration on the task at hand. PALS should take into account the user's emotional state and possible effects of the human-computer interaction on these states. Based on the Pleasure-Arousal-Dominance (PAD) model of Mehrabian [1], we distinguish two dimensions to define the emotional state: the arousal level—low versus high—and the valence—positive versus negative. We do not distinguish a separate dominance dimension like the original PAD-model, because the dominance scale proved to explain the least variance and had the highest variability in terms of its inferred meaning in previous research [14].

The first emotion experiment will investigate the effects of device type (iPAQ versus laptop) and emotional state on user's behaviour and trust.

2.2 Trust

In addition to the importance of emotion, trust has being received more attention due to the development of e-commerce services and privacy issues. Trust is being viewed as an important constraint for establishing successful financial transactions and sharing of information in (mobile) network environments [6],[9],[2]. Therefore, PALS should support the realisation of an adequate trust level. User interface appearances, personalisation elements and interaction aspects prove to affect trust [5] and we assume that emotion plays a role in these effects. Trust depends on persistent, competent behaviour of a system that has a purpose of serving the user. It develops as a function of experience, and is influenced by errors from the system, interface characteristics and individual differences on part of the user [10].

Table 1. The target emotional states and description of the film clips

		Valence	
		NEGATIVE	POSITIVE
Arousal	HIGH	 “**Koyaanisqatsi**” Several high-speed video fragments depict commuters in a station, highway scenes and crowds of people. The music is choir music with a repetitive nature. The duration of the clip is 2 min. 31 sec.	 “**Blues Brothers**” The two main characters attend a church service where the reverend and the choir perform “The Old Landmark”. The duration of the clip is 2 min. 55 sec.
	LOW	 “**Stalker**” Two persons, filmed from behind, sitting on a moving train. The clip is black and white and no music was played here, only the sound of the train is heard. The duration of the clip is 2 min. 6 sec.	 “**Easy Rider**” The three main characters ride their motorcycles on a desert highway. The soundtrack is the song “The Weight” played by “The Band”. The duration of this clip is 2 min. 38 sec.

3 Experiment

This first experiment should provide insight in the relations between emotion, device, trust and performance, in order to develop a personalisation mechanism that establishes adequate user behaviour and an appropriate sense of trust for different interaction devices and emotional states. To induce specific emotional states (high and low arousal, positive and negative valence), the participants view film clips and listen to sound tracks with different emotional impacts (see Table 1). Subsequently, they perform tasks with four web services on two different devices, and fill in trust questionnaires. The tasks consist of searching for information and executing financial transactions.

We expect that the mobile device will lead to less efficient performance on these services, compared to a desktop computer. In particular, the experiment tests if emotional state affects performance and trust, and if the effect (size) is related to usage of a specific type of device. In this way, the experiment investigates how trust builds up for a mobile and a desktop device, and whether there are differences that should be addressed by PALS. In this experiment, we investigate the effects for users that already have experience with computers and Internet.

3.1 Method

3.1.1 Experimental Design

A 2 (device) x 2 (valence) x 2 (arousal) mixed model design was adopted, with device as a between subjects variable, and valence and arousal as within subjects variables. Each subject viewed all the film clips and worked with all the services.

3.1.2 Participants

Twenty-four (12 male, 12 female) students participated in this experiment. Mean age was 21 years (min. 18 and max. 26 years). The participants used a PC or laptop and the Internet on a daily basis and used various web services at least once a week. Most of them had never before used a handheld computer.

3.1.3 Stimuli

Five film clips were used in this experiment. Four of them were used previously in a study by Wensveen, Overbeeke, and Djajadiningrat [16] and were validated to induce the target emotional states. Table 1 summarizes the film clips that were used. The presentation order of these film clips was balanced across participants to rule out possible order effects.

The fifth clip was an emotionally neutral film, which showed a busy shopping street with corresponding sound. During the tasks, the soundtrack of the clip played on at a lower volume.

3.1.4 Tasks

The web tasks consisted of searching for information and executing financial transactions with four web services. Two of these four web services were current "commercial" sites, the site of the Rabobank, which is a Dutch bank, and the site of Gouden Gids, which is the Dutch Yellow Pages. Two other sites originated from the Microsoft .Net software development environment: a book store site and a stockbroker service. Each task for a specific service had five alternatives, designed in such a way that they could be compared on execution time and number of steps, but with different parameters. The presentation order for the web services and alternatives was balanced, to minimise possible order effects. The description of the web sites and tasks can be found in Table 2

Table 2. The description of the web services used in the experiment

Web site	Task
www.goudengids.nl	**Looking up information on restaurants, swimming pools, and cinemas.** - Type the name or branch in the search field and hit "Find". - Choose the appropriate item from the search results. - Click it for more detailed information. - Find the specific information item (e.g. telephone number).
Www.rabobank.nl	**Looking up information on insurance, bankcards, and accounts.** - Click on the appropriate link in the main menu on the homepage. - Click on the appropriate link in the submenus on the next 2 pages. - Read and scroll to the specific information item.
Book site	**Compare book prices.** - Type in the name of the first book or author in the search field. - Remember the price. - Type in the name of the second book or author in the search field. - Compare the prices.
Stocks site	**Buying stocks.** - Log on to the web site with user name and password. - Navigate to the search field and type in the name of the company. - Remember the code and navigate to the "Buy" field. - Enter the code and number of stocks, and click "Buy". - Navigate to the portfolio, and remember the amount of money.

3.1.5 Device

The participants used one of two different devices for interacting with the web service. The first was a Compaq iPAQ 3800 Handheld Computer. This device uses touch screen and pen-based navigation and input. The screen resolution was 320 x 240 pixels. The second device was a Sony VAIO Notebook with standard keyboard and mouse for user input. The screen resolution of the notebook was 1024 x 768 pixels. Both devices used Wireless LAN to connect to the Internet. During the whole experiment, the connection was excellent. Browsing was done using Microsoft Internet Explorer. All the instructions and questionnaires were shown on another notebook, which was placed on the table in front of the participant. The answers to the questionnaires had to be typed in on this notebook.

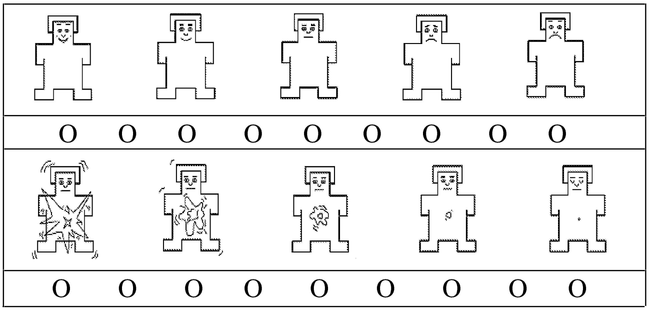


Fig. 1. The valence (top) and arousal (bottom) scales of the Self Assessment Manikin [1]

3.1.6 Measures and Questionnaires

As each subject had to complete five sessions with four services each, and the questionnaires had to be filled out after every service, the questionnaires had to be brief.

Measurement of emotion is done using the Self-Assessment Manikin (SAM). This subjective scale, developed by Lang [1] is based on the PAD model of emotion and measures emotion on the three dimensions of valence, arousal and dominance. The scale consists of three rows of cartoons, on which the participant has to characterize the experienced emotion. When used in studies, the dominance scale proved to explain the least variance, and had the highest variability in terms of its inferred meaning. Therefore, for the present experiment, the dominance scale was omitted. Figure 1 shows the SAM scale.

The trust questionnaire was based on a scale developed by Jian et al. [4]. As the original was too cumbersome to fill out after each task (i.e. interaction with a web service), an adapted version was developed. It consists of three questions that measure the elements of the definition of trust proposed here. Question 1 concerns the service-oriented attitude and question 2 the quality of persistent and competent behaviour. The third question asks directly to what extent the user trusts the service. Trust was measured on a 7-point scale; a higher score corresponded with a higher level of trust. To see whether participants discriminated between trusting the web service and trusting the device, trust in device was measured with a similar questionnaire as described above. This measure was only taken prior to and after the experiment. Reliability of the questionnaire is assessed in the results section.

For every service, effectiveness and efficiency were measured. Effectiveness was defined by correct completion of the task. Efficiency was defined by time on task in seconds, number of typing errors (that were corrected) during the task, and number of extra links that were clicked. By comparing the navigation path of the user with the optimal path, in terms of number of links, this last measure is obtained. The optimal path of navigation for every service was established prior to the experiment. In addition, subjective mental load was measured using the one-dimensional Rating Scale Mental Effort [17]. In this experiment, a digital version of the RSME was used. Participants could move a pointer over the scale using the mouse, and click the appropriate score, ranging from 0 to 150.

Recording heart rate (HR) was done using a Vitaport recorder system and software. Three electrodes were attached to the participants' chest for the duration of the experiment. Afterwards, the duration between two beats in the raw data was sampled with a frequency of 4 Hz. The mean HR for any given period is obtained by averaging over these samples. HR is measured as number of beats per minute (bpm). In addition, heart Rate Variability (HRV) was measured as an indication of the level of arousal.

3.1.7 Procedure

In short, the procedure consisted of:

- Briefing and instruction
- Trust questionnaire
- Training on the device
- Neutral session
    - Film clip
    - SAM questionnaire
    - Four tasks
    - After each task: RSME & trust questionnaire
- Four experimental sessions
    - Film clip
    - SAM questionnaire
    - Four tasks
    - After each task: RSME & trust questionnaire
- Trust questionnaire & debriefing

Participants were told that they participated in a usability experiment and were asked to perform the tasks as quickly and as accurately as possible. They rated their trust in the device that they would be using, either the laptop or the iPAQ. Then the baseline HR measure was taken. The subject was connected to the Vitaport and asked to remain calmly seated for a period of 5 to 7 minutes. After this, a short training session on the device took place. With the experimenter present, subjects completed two tasks to familiarize themselves with the device and questionnaires. After receiving feedback on performance, there was an opportunity to ask questions.

Subjects were left alone in the darkened room, with enough illumination to distinguish the keys on the keyboard of the notebook. In the first session, the neutral film clip was shown. After the film, subjects filled out the SAM questionnaire, and proceeded with the four web services. The soundtrack of the film played on at a lower volume. After each service, the RSME and trust questionnaire were filled out. Upon completion of the four services, the sound of the film was turned off and the SAM questionnaire was presented again.

Similar to the first session, four additional sessions were completed, in which the subjects viewed all the film clips. Between each session, a moment of rest was provided. Once all the sessions were completed, subjects again rated their trust in the device and were debriefed on the true nature of the experiment.

3.2 Results

3.2.1 Measurement Assessment

As the *HRV* measure did not discriminate between conditions, this measure is omitted from further analysis.

A reliability analysis for the *trust* questionnaire showed that questions 1 and 2 correlated for .823 and both 1 and 2 correlated .576 with question 3. Cronbach's alpha for the questionnaire was .85.

3.2.2 Emotion Manipulation

A repeated measures ANOVA was performed on the averaged arousal and valence scores for the SAM questionnaire as well as HR during film viewing, and compared to the neutral first session. Both arousal and valence showed significant main effects (F (4, 92) = 19.10; p = 0.000 and F (4, 92) = 17.84; p = 0.000, respectively). Post-hoc analysis revealed that only the high arousal, positive valence and low arousal, negative valence conditions did significantly differ from baseline (p = 0.000 for both conditions). Figure 2 shows the target emotional states, as described in Table 1, and the actual SAM scores after the film clip.

A main effect of HR was found between conditions (F (4, 88) = 6.63; p = 0.000), i.e. all conditions showed a lower HR than baseline (average 73.9 for conditions and 77.8 for baseline). Post-hoc analysis revealed no significant differences between either high or low arousal (p = 0.366) and positive and negative valence (p = 0.974). Thus, the emotional induction procedure was successful for two distinct emotional states, although no physiological evidence (e.g. differences in HR) for this induction can be found in the results.

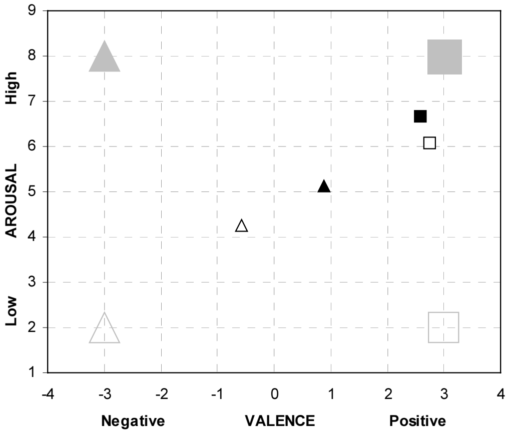


Fig. 2. The target emotional states (big, gray) and the actual SAM scores after the film clip (cf. Table 1)

3.2.3 Effects of Device on Performance and Trust

A significant main effect of device is found for time on task (F (1, 17) = 58.27; p = 0.000), number of typing errors (F (1, 17) = 22.94; p = 0.000) and number of links (F (1,17) = 12.42; p = 0.001). No main effect of device was found for mental effort. Averaged over services, subjects took longer to perform the services with the iPAQ (127 sec.) than with the laptop (55 sec.). In addition more errors and wrong links occurred with the iPAQ (1.0 and 1.0 respectively) than with the laptop (0.5 and 0.4 respectively).

Wilcoxon Matched Pairs test revealed an overall increase of trust in the laptop as a result of experience with this device. The scores after the experiment on the servitude (Z (12) = 2.67; p < 0.01), persistence (Z (12) = 2.37; p < 0.05) and trust (Z (12) = 2.67; p < 0.01) questions all differed significantly from the pre-measurement. Figure 3 shows these results. Scores for the iPAQ did not differ between pre and post measurement.

We also measured trust in the service, before and after the experiment. Figure 4 shows that, averaged over services, scores on the trust questionnaire for web services were higher after using the laptop than after using the iPAQ. The scores after the experiment for laptop users on the servitude (Z (12) = 2.97; p < 0.01), persistence (Z (12) = 2.63; p < 0.01) and trust (Z (12) = 2.85; p < 0.01) questions all differed significantly from the pre-measurement.

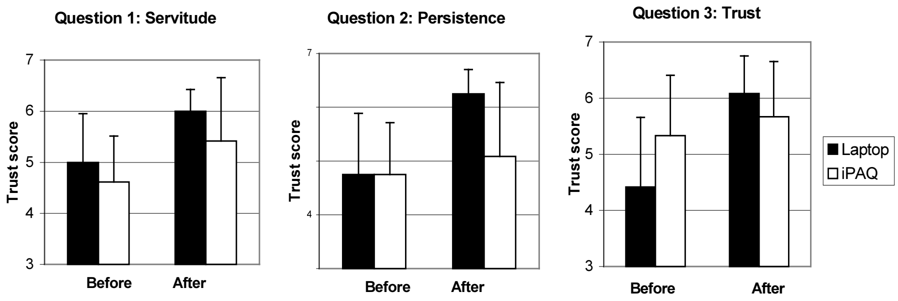


Fig. 3. Scores for trust in DEVICE, both before and after the experiment, for both devices separately. Error bars denote the standard deviation (SD)

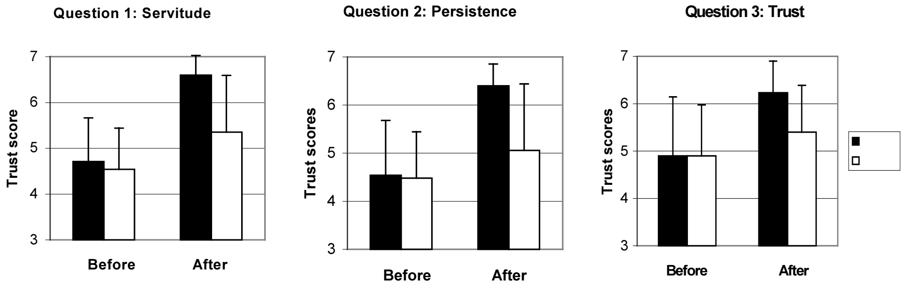


Fig. 4. Scores for trust in WEB SERVICES, both before and after the experiment for both devices separately. Error bars denote the standard deviation (SD)

3.2.4 Effects of Emotional State on Performance and Trust

No main effects of emotional state were found significant for any of the performance measures or any of the trust scores. The two way “emotional state x web service” interaction effect was found significant for time on task (F (9, 153) = 3.64; p = 0.000) and number of links (F (9, 153) = 3.24; p = 0.001). It appeared that performance was worse in the high arousal, positive valence condition for the Yellow Pages service. For trust in web services, the “emotional state x web service” interaction effect approached significance: (F (9, 198) = 1.72; p = 0.086). Again, on the Yellow Pages service, lower trust scores were obtained in the high arousal, positive valence condition. This corresponds nicely with the analysis of the performance measures.

In addition, a three way interaction of “device x emotional state x web service” was only significant for time on task (F (9, 153) = 3.84; p = 0.000). Time on the Yellow Pages service using the iPAQ was significantly higher in the high arousal, positive valence condition (258 sec.) than in the rest of the conditions (average 172 sec.).

There were no significant main effects on HR during work with the web services. However, a “device x web service” interaction effect was significant (F (3, 63) = 10.30; p = 0.000). It turns out that when using the book service and stocks service, HR is higher in the laptop condition (76.1 and 76.3 bpm respectively, compared to 73.4 averaged). HR is not significantly different between tasks in the iPAQ condition.

3.2.5 Multiple Regression

In order to gain more insight into the relations between emotion, trust, performance and device, the following questions were assessed using standard stepwise multiple regression:

- Is performance predicted by emotion or by device?
- Is trust predicted by performance, emotion or device?

In answering the first question, the predictors included in the regression analysis are the SAM scores after the film, heart rate during tasks and device. This analysis was done for all performance measures separately. For time on task and answer on task, device explained 27% and 26% of the variance respectively. The variance explained for number of links and typing errors is rather low (6%). The analysis showed that lower performance accompanied the use of the iPAQ. Arousal after the film did significantly explain some variance in time on task, however the added value is low (1%).

To assess which factors predict trust, all performance variables, device and SAM scores after the film as well as heart rate during tasks were included in the analysis. It appeared that 39% of the variance was explained by two performance measures alone, time on task and answer on task. Long time on task and more wrong answers accompanied lower trust scores. In comparison, device explained only 2% of the variance in the trust scores. Table 3 lists the results of the multiple regression analysis.

Table 3. Results of the multiple regression analysis. Codes in the regression equation are between brackets in the first two columns

Dependent variable	Predictor	Cum. R^2	Regression equation
Time on task (TOT)	Device (DEV) Arousal (AR)	27% 27%	TOT = -25.83 + 0.52*DEV + 0.08*AR
No. of links (LIN)	Device	6%	LIN = -0.32 + 0.24*DEV
No. of typing errors	Device	6%	TYP = -0.65 + 0.25*DEV
Answer wrong (ANS)	Device	26%	ANS = - 0.78 + 0.50*DEV
Trust (averaged scores on the 3 trust questions)	TOT ANS Heart Rate (HR) LIN Device	30% 39% 42% 44% 46%	Trust = 8.95 - 0.55*TOT - 0.32*ANS - 0.19*HR - 0.20*LIN - 0.15*DEV

3.2.6 Emotional state after Task Performance

In order to investigate the effects of device used and task performance on emotional state, the arousal and valence scores after the film and after the tasks were compared. A repeated measures ANOVA with factors device, film, pre/post measurement was performed, and the neutral condition was excluded from the analysis. For arousal scores, only a main effect for film (F (3, 60) = 3,84; p = 0,01) is observed. No significant effects or interactions were found in this analysis. The same analysis was run for valence scores. Main effects are found for device (F (1, 20) = 6,16; p < 0,05), film (F (3, 60) = 11,78; p = 0,000) and pre/post measurement (F (1, 20) = 8,79; p < 0,01). Valence is more positive after using the laptop (1,5) than after using the iPaq (0,8).

4 Conclusions

In general the type of device proves to have a substantial effect on performance and trust (see Figure 5). As expected, task *performance* was worse for the iPAQ. Furthermore, an interesting effect appeared: users of the small, mobile device used more (‘wrong’) links to find the target information. Probably, users need an overview of possible links (navigation paths) in order to assess which link is appropriate. For a small device, part of the current navigation space is out of vision so that assessments are more difficult. In conclusion, there proves to be a real need to diminish the navigation space for such devices, which the PALS project is developing [3].

The experiment shows that the service itself *and* the type of device influence user’s *trust* in the service. Corresponding to the theory, trust builds up during the interaction with services in the laptop condition. However, during the interaction with the mobile device trust hardly increases (note that the user tasks and service content are completely similar for the mobile and laptop device). The regression analysis shows that interaction performance affects trust, causing a lower level of trust for the small device. Based on these results, we expect that improving the performance, for example with an attentive interactive display [11] and decreasing the navigation space [3], will also result in increased trust. It is interesting to note that our results fit with the distinction of two levels of trust in design [9]. At the first level, the user forms a sense of trust based upon appearance only. Subsequently, as the user begins to interact with a product's explicit as well as implicit functionality, a sense of experiential trust is established.

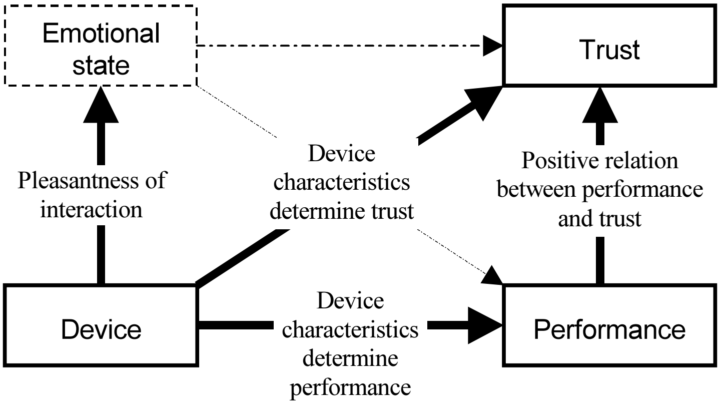

Emotional state
Trust
Device characteristics determine trust
Pleasantness of interaction
Positive relation between performance and trust
Device
Device characteristics determine performance
Performance

Fig. 5. The relations between trust, performance, device and emotion that were demonstrated in this study. Dotted lines indicate relations that could not be demonstrated unequivocally. These relations need further investigation

In particular when accessing the information search service (Yellow Pages) with the mobile device, there were distinct decreases in performance in the high arousal, positive valence condition, compared to the other conditions. Subjects took up to 300% longer to complete work on the service then in the laptop condition. In addition, subjects reported a 70% increase in mental effort, compared to the laptop. In this condition, emotional state, device and information search interfere. Further research should investigate to what extent an aroused, positive emotional state influences searching for information.

Task performance with the mobile device resulted in a lower valence score than with the laptop. In other words, users experienced a more negative *emotion* with the mobile device. Similar to trust, this effect might be attributed to the decreased performance.

In our experiment, heart rate proved not to discriminate between device, but it did discriminate between services for the laptop (i.e. heart rate was higher for the Books and Stocks services in the laptop condition). Probably, a substantial increase in effort for the mobile device would hardly help the users to improve their performance level. Consequently, the users do not feel inclined to increase their effort as with data-driven task performance compared to resource-driven task performance [15]. The task demands for Books and Stocks services are higher and extra effort helps to realise an adequate performance level. In this way, PALS may profit from a 'smart' heart rate sensor in order to attune the interaction to user's state changes that are caused by task demand fluctuations.

In summary, device and service proved to be the major factors that determine the user experience. The experiment shows a clear need to improve the mobile interaction with web services in order to establish an adequate performance *and* trust level. The results provide an empirical foundation for working out the PALS concept (among other things by attentive interactive displays and personalised navigation spaces [8]).

Acknowledgements. The PALS project is partly funded by the IOP MMI programme of Senter. Hans Veltman provided important support for the heart rate analysis and Bert Bierman developed the software for this study. PALS partners are TNO Human Factors, University of Utrecht, Technical University Twente, LogicaCMG, Rabobank and KPN Mobile.

References

1. Bradley, M., Lang, P.: Measuring emotion: The Self-Assessment Manikin and the Semantic Differential. Journal of Behavioral Therapy & Experimental Psychiatry 25 (1994) 49–59
2. Corritore, C.L., Kracher, B., Wiedenbeck, S.: On-line trust: concepts, evolving themes, a model. Int. J. Human-Computer Studies 58 (2003) 737–758
3. Herder, E. & Dijk, B. van: From browsing behavior to usability issues. Proc. 10th Int. Conference on Human-Computer Interaction, Vol. 3. Lawrence Erlbaum, Hillsdale, NJ (2003) 452–456
4. Jian, J., Bisantz, A., Drury, C.: Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics 4 (2000) 53–71
5. Karvonen, K. The beauty of simplicity. Proceedings of the 2000 ACM Conference on Universal Usability. Arlington, VA (2000) 85–90
6. Kim, J., Moon, J.: Designing towards emotional usability in customer interfaces - trustworthiness of cyber-banking system interfaces. Interacting with Computers 10 (1998) 1–29
7. Klein, J., Moon, Y., Picard, R.: This computer responds to user frustration: Theory, design, and results. Interacting with Computers 14 (2002) 119–140
8. Lindenberg, J., Nagata, S.F., Neerincx, M.A.: Personal assistant for online services: Addressing human factors. Proc. 10th Int. Conference on Human-Computer Interaction, Vol. 3. Lawrence Erlbaum, Hillsdale, NJ (2003) 497–501
9. Marsh, S., Meech, J.: Trust in design. Proceedings of the 2000 Conference on Computer Human Interaction, The Hague (2000) 45–46.
10. Muir, B.: Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37 (1994) 1905–1922
11. Nagata, S.F.: Multitasking and Interruptions During Mobile Web Tasks. Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2003)
12. Nagata, S.F., Neerincx, M.A., Oostendorp, H.: Scenario Based Design: Concepts for a Mobile Personal Service Environment. Adj. Proc. 10th Int. Conference on Human-Computer Interaction. Crete University Press, Heraklion, Crete (2003)
13. Norman, D.: Emotion & Design. Attractive things work better. Interactions 9 (2002) 36–42
14. Scheirer, J., Fernandez, R., Klein, J., Picard, R.: Frustrating the user on purpose: a step toward building an affective computer. Interacting with Computers 14 (2002) 93–118

15. Veltman, J.A., Jansen, C.: Differentiation of mental effort measures: consequences for adaptive automation. In: G.R.J. Hockey, A.W.K. Gaillard, A. Burov (eds.): Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks. (2003) Chapter 20
16. Wensveen, S., Overbeeke, K., Djajadiningrat, T.: Push me, shove me and I show you how you feel. Proceedings of the 2002 ACM Conference on Designing Interactive Systems. London (2002) 335–340
17. Zijlstra, F.R.H.: Efficiency in work behavior. A design approach for modern tools. (PhD thesis, Delft University of Technology). Delft University Press, Delft, The Netherlands (1993)

Ultrasonic 3D Position Estimation Using a Single Base Station

Esko Dijk[1,2], Kees van Berkel[1,2], Ronald Aarts[2], and Evert van Loenen[2]

[1] Eindhoven University of Technology,
Postbus 513, 5600 MB Eindhoven, The Netherlands
esko@ieee.org
[2] Philips Research Laboratories Eindhoven,
Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands
evert.van.loenen@philips.com

Abstract. In indoor context awareness applications the location of people, devices or objects is often required. Ultrasound technology enables high resolution indoor position measurements. A disadvantage of state-of-the-art ultrasonic systems is that several base stations are required to estimate 3D position. Since fewer base stations leads to lower cost and easier setup, a novel method is presented that requires just one base station. The method uses information from acoustic reflections in a room, and estimates 3D positions using an acoustic room-model. The method has been implemented, and verified within an empty room. It can be concluded that ultrasonic reflection data provides useful clues about the 3D position of a device.

1 Introduction

In many context aware computing applications, the locations of people, devices or objects is part of the required context information. For some applications, also the orientation of devices [1], people or objects plays a role. Typically, indoor applications require higher resolution than 'on the move' and outdoor applications, anywhere between room-size and sub-centimeter. Within the Phenom project [2] we focus on the *in-home* application area. Context awareness in the home can make technology more enjoyable and easy-to-use or enable new experiences that users may find attractive. Several application scenarios were developed in our project that require location information. For example, one scenario enables users to intuitively 'move' content from any handheld or fixed display device, to any other nearby display. Proximity between displays can be determined by using the devices' mutual positions and orientations. A second application is the control of light, music and atmosphere settings in the living room, based on the positions of small devices or 'smart objects' that users can move around. For these applications, we need a resolution of at least one meter.

The required resolution is much higher than wide-area systems –like GPS– can deliver indoors at the moment. Therefore many research and commercial

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 133–148, 2003.

systems exist that can provide indoor location information at a finer scale. Systems based on either radio (RF) or ultrasonic technologies are most popular and promising at the moment. RF systems typically cover wider areas like buildings, but have difficulties reaching high accuracy [3,4]. Ultrasonic systems can routinely offer very high (centimeter) resolution as shown in the Bat [5], Cricket [1], Dolphin [6], and InterSense [7] systems. But they have a limited operating range which makes deployment throughout a building difficult. Because of the potential high resolution, we are investigating ultrasound technology for home applications.

State-of-the-art ultrasonic systems are based on calculating distances (or distance-differences) by measuring ultrasound time-of-flight. A position estimate can be obtained using triangulation algorithms with these distances as inputs. A disadvantage of this approach is that several units of infrastructure are required, to generate sufficient input for a triangulation algorithm. Generally, four base stations (BS) are needed in a non-collinear setup to estimate 3D position. For special cases we can do with three, for example three ceiling-mounted BSs. However, more than three or four BSs are often employed for increased robustness [5,1], or the ability to estimate speed-of-sound from measurements [1]. For RF systems the same problem exists. The need for high resolution leads to large numbers of units within an infrastructure (e.g. [3]), increasing installation and maintenance (e.g. battery replacement) effort and cost.

An important goal of indoor positioning systems research is to realise ubiquitous deployment in an economically viable way. This means that a technology is preferably cheap, easy to set up, and low-maintenance. We argue that these requirements can be better met if the infrastructure is minimal. Fewer BSs could mean lower cost and easier setup. The resulting research question is whether a 3D positioning system can work with fewer BSs, and if in the extreme case a single BS (of small size) would be sufficient. Two concepts emerged to realise such a single-base-station positioning system. The first concept makes use of always-present ultrasonic reflections against the walls and ceiling of a room. Considering ultrasonic emissions coming from a source somewhere in the room, reflections can be considered as coming from *virtual sources* at other positions. The virtual sources might replace real sources, thereby reducing the number of BSs. The second idea was to employ acoustic array techniques [8] that are well-known in source localisation research.

The first concept is the topic of this paper. A new method was developed that uses reflections for 3D positioning, requiring a single base station per room. An introduction to the *signature matching* method is given in section 2. An acoustic model that plays a central role in the method, will be developed in section 3. The method will be presented in more detail in section 4. An implementation of the method and experimental work are described in section 5 and conclusions are given in section 6.

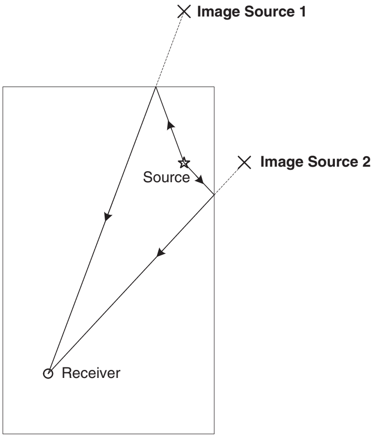


Fig. 1. 2D top view of a room, containing one acoustic source and one receiver. Two acoustic reflections (arrows) and associated image sources (crosses) are shown.

2 Method Overview

The new method is based on the idea of using ultrasonic reflections off surfaces (walls, floor and ceiling) within a room. To clarify this, Fig. 1 shows a top view of a room, with one acoustic source and one receiver. The source emits a direct sound field to the receiver, but sound also reflects off the four walls and arrives at the receiver indirectly. Two such indirect (reflected) acoustic rays are shown in the figure. From ray acoustics theory (to be introduced in section 3.3) it is known that reflections can be modelled as emissions from so-called *image sources* located outside the physical room boundaries. Their positions can be constructed by a mirror-symmetry operation with respect to the room boundary. Two image sources, associated to the example reflected rays, are shown in the figure. Many more image sources exist, like ceiling and floor reflections and higher-order reflections (e.g. wall–ceiling–wall). The combined effect of reflections can be observed in Fig. 2, which shows a processed acoustic measurement. The many peaks in this graph are caused by reflections, arriving at the receiver at different moments in time. Such a pattern of reflections was observed to be dependent on the 3D receiver position and orientation. These patterns were named *signatures* because of the information they contain about receiver position and orientation. The signature shown was obtained using the procedure described in section 5.1.

Let's assume for now that the acoustic source is a base station (BS) fixed at a known position and the receiver is a mobile device, with an unknown position to be measured. It is expected that the image sources can be used as if they are 'virtual base stations' (VBS). The combined information from BS and VBSs might enable calculation of 3D receiver position. The problem with VBSs is that we can neither identify nor distinguish them. For example, for the peaks in Fig.

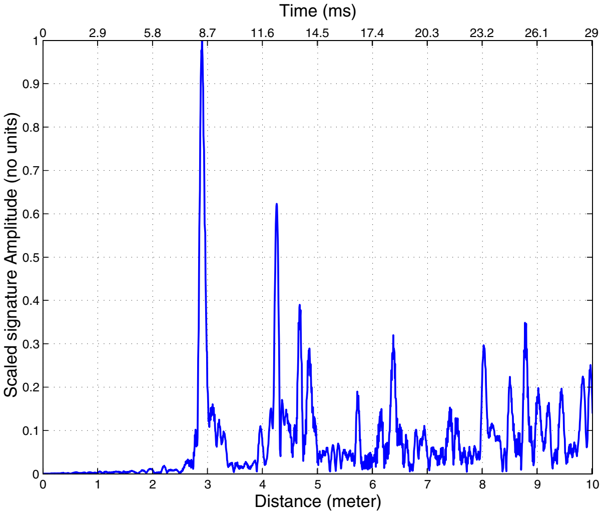


Fig. 2. Measured signature at position $x_R = (2.60, 1.70, 1.27)$. The horizontal axes show time (top) and the corresponding distance interval of $[0, 10]$ m. Signatures are obtained by the procedure in section 5.1.

2 it is not known by which VBS they were 'transmitted'. As a result, standard triangulation algorithms can not be applied. Estimating the identity of peaks would be required first, but this is a very difficult problem.

Concluding, we have not found any method yet that can directly calculate a 3D position from a given signature. However, the reverse is easier: computing a signature, given a 3D position. This fact is exploited by the *signature matching* method. It simply tries many mobile device 3D candidate positions and predicts their signatures. Then, all predicted signatures are compared or *matched* to the measured signature and the best match is selected. The details of the method will be discussed in section 4.

3 Acoustical System Model

Acoustical models of both the positioning system and the room are needed, to be able to predict an acoustic signal at a certain position in a room. Such predicted signals are needed to compare to measured signals, as explained in the previous section. The acoustical model includes a transmitter, a receiver and the room. In section 3.1 the transducers are modelled, in section 3.2 acoustical phenomena indoors are listed, and in section 3.3 we show how a box-shaped room can be modelled. Finally section 3.4 combines these elements into a system model.

3.1 Transducers Model

For the implementation, piezo-electric ultrasound transducers were selected. They are cheap, have a small form factor, and operate on low voltage and low power. Each type of transducer has specific frequency-dependent behaviour. Piezo transducers can be conveniently modelled as linear systems [9], e.g. represented by an impulse response. The transmitter/receiver impulse responses are part of the system model. They can be obtained by measurement, or by modelling [9].

Most ultrasonic transmitters are directional, designed to emit acoustic energy mainly within a narrow beam from the front. Likewise, receivers are most sensitive at the front. This directionality can be captured in the *normalised beampattern function* $D_N(\theta)$ of the transducer, where θ is the angle with respect to the transducer axis. Maximum amplitude occurs on-axis, when $D_N(0) = 1$. The beampattern for disc-shaped piezo transducers can be approximated by that of an ideal circular piston [8]. However, the piston model had to be extended to account for the protective metal casing that surrounds the piezo element. We used an empirical damping factor K_d that was found to correspond well to measured amplitudes:

$$K_d(\theta) = 0.525 + 0.475 \ \cos(\theta) \ . \tag{1}$$

3.2 Acoustical Phenomena Indoors

In this section, the ultrasound phenomena that are part of the acoustic room model are briefly presented.

Ultrasound propagation in air. Distance can be calculated using time-of-flight measurement and the speed of sound c in air. However, c varies with air temperature [10]. Distance errors can be caused due to imperfect knowledge of the current room temperature. We will assume that room temperature is either approximately known or measured.

Acoustic waves are attenuated over distance due to *geometric spreading loss* and *atmospheric absorption loss*. The spreading loss is the well-known acoustic pressure loss of r^{-1} over distance r [9]. The absorption loss is caused by lossy molecular interaction processes [9]. It can be high for ultrasound in air, and limits the usable range of ultrasonic distance measurement. The loss can be calculated using the *absorption loss factor* α in dB/meter. α depends on temperature, relative humidity, and the frequency of sound in a non-linear manner. Equations to calculate α [11] are part of the model. The relative humidity parameter can be measured or simply set to a default value. Automatic calibration of α is possible from base station measurements, but outside the scope of this paper.

Reflection and diffraction. A sound wave in air incident to a boundary surface, such as a wall or ceiling, will cause a reflected wave and a transmitted wave. The latter is usually absorbed within the medium. For ideal homogeneous materials with smooth surfaces, theory predicts *specular reflection* [10]: angle of incidence equals angle of reflection. The reflected wave continues in air, with lower amplitude than the incoming wave due to absorption. The *reflection factor* Γ, reflected wave amplitude divided by incoming wave amplitude, models this. Value for Γ are difficult to calculate precisely, since typical room boundary materials are not smooth and homogeneous. From measurements we observed that for typical building materials there is little reflection loss ($\Gamma \approx 1$) at ultrasonic frequencies around 40 kHz. However, for soft materials such as curtains or carpet (e.g. $\Gamma \approx 0.3$), the loss can be substantial. An estimation $\Gamma = 1$ was used in our model.

Another effect is *diffraction*, the deflecting or 'bending' of acoustic waves around obstacles. In a typical room, diffraction is mostly unpredictable, because we do not know the sizes and positions of the obstacles (like furniture and people). Therefore we can not model diffraction.

3.3 Room Impulse Response Model

Rooms exist in many shapes, but home and office rooms are often approximately box shaped. Another common shape is the L-shaped room, which is acoustically more complicated. We focus on box-shaped rooms, which have six boundary surfaces (four walls, ceiling and floor). The goal of the model presented here is to predict the impulse response of a room $h(t, \mathbf{p})$ as a function of relevant parameters, described by a parameter vector $\mathbf{p}$. It includes the room size and shape (assumed to be known), transmitter and receiver positions and orientations, surface reflection factors, and room temperature and humidity. In practice the room response is a complicated function of other parameters as well, such as people/objects/furniture in the room and room materials. However, a 'minimal' room model of an empty room can be constructed that only includes the six boundary surfaces. This is a standard approach in room acoustics. To model a room the *image method* was applied, because for box-shaped rooms an impulse response for short durations can be calculated efficiently using the model of Allen and Berkley [12]. See section 2 for an explanation of the image sources concept or [12,10] for details.

The image method is based on the *ray acoustics* [10] approximation of acoustic waves. For ray acoustics models of arbitrarily shaped rooms, the room impulse response h in a time interval $[0, t_e]$ can be written as a sum of N independent rays arriving at the receiver:

$$h(t, \mathbf{p}) = \sum_{i=1}^{N} a_i \cdot \delta(t - d_i/c) \tag{2}$$

where ray d_1 is the line-of-sight and (N-1) rays are reflections, d_i is the distance the i-th ray travels, a_i is the amplitude of the i-th ray arriving at the receiver, c is the speed of sound and δ the Dirac delta function. This equation holds for arbitrary transmitted signal shapes. d_i, a_i and c depend on the parameter vector $\mathbf{p}$. For a box-shaped room the details for calculating distances d_i using the image method are given in [12]. The amplitudes a_i can be described in terms of acoustic pressure, taking into account the pressure loss over distance (section 3.2), pressure loss due to reflections (section 3.2) and the attenuation caused by the orientations of both transducers (section 3.1).

3.4 System Impulse Response Model

In sections 3.1 to 3.3 it was shown that impulse responses of transmitter, receiver and the room can be obtained. To obtain a system impulse response we simply concatenate these three by convolution:

$$h(t, \mathbf{p}) = h_T(t) * h_{Rm}(t, \mathbf{p}) * h_R(t) \tag{3}$$

where h_T, h_{Rm} and h_R are the transmitter, room and receiver response respectively. This model is visualised in Fig. 3. For fixed parameter vector $\mathbf{p}$, the model is linear.

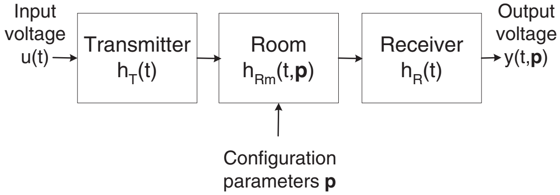


Fig. 3. System model of Eq. 3 shown as a series connection of transmitter, room, and receiver impulse responses.

4 Method

In this section the *signature matching* method will be presented. It can estimate the position of a mobile device in a room, based on a measurement using a single base station. Section 4.1 discusses what information the line-of-sight peak of the measured signal can provide. The algorithm is given in section 4.2. Section 4.3 shows how signatures can be matched, and finally section 4.4 discusses computational complexity and robustness of the method.

4.1 Line-of-Sight Measurement

Measurement of the line-of-sight (LoS) distance between base station and mobile device gives valuable information about the position of the mobile device. To obtain such a measurement, we assume that transmitter and receiver have mutual time synchronisation by an RF link as e.g. in the Cricket [1] system. Figure 4 shows a front view of an empty room. For now, we assume that the fixed transmitter Tx near the ceiling acts as a base station and the mobile device Rx as a receiver. The LoS distance is visualised as a line between them. The partial sphere surface S shows the possible positions of Rx, if nothing but the LoS distance and the coordinates of Tx in the room are known. The LoS distance can be obtained by a straightforward first-peak detection on the measured signature, as demonstrated later in section 5.3.

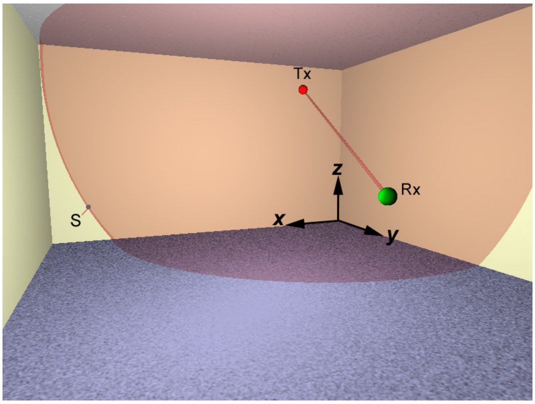


Fig. 4. 3D view of a room with transmitter Tx and receiver Rx.

However, a measurement of the LoS distance may fail due to blocking of the path between transmitter and receiver. Then, a reflection (arriving later) may be mistakenly seen as the LoS. This causes errors in position estimates, just like in current state-of-the-art ultrasonic positioning systems. Within the scope of this paper, a clear LoS path is assured by measuring in an empty room. In non-empty rooms, a higher probability of line-of-sight can be expected when placing the transmitter near the ceiling, because most obstacles are nearer to the floor.

4.2 Signature Matching Algorithm

The signature matching algorithm takes a measured signature vector $\mathbf{s}$ as input and outputs a position estimate $\hat{\mathbf{x}}$ of the mobile device. However, certain parameters must be known before the algorithm can be executed. The first group of parameters are the configuration parameters, describing the physical circumstances within the room. They were represented as parameter vector $\mathbf{p}$ in section 3.3. The room size (part of $\mathbf{p}$) can be obtained by manual input or estimated through echo measurements by the base station. If reflection factors Γ are not exactly known they have to be estimated. Furthermore the 3D position and orientation of the base station should be known. Finally we need the orientation vector $\mathbf{v}_R$ of the transducer mounted on a mobile device to be approximately known. $\mathbf{v}_R$ is a 3D vector that should be seen as the 'pointing direction' of the transducer. There are three options to obtain $\mathbf{v}_R$, as will be discussed in section 4.4. The second group of parameters are the algorithm parameters. These include the grid spacing Δx (see step 2 in the algorithm) and the choice of time interval $[t_0, t_1]$ of the measured signal that we intend to use for matching.

For each measured signature $\mathbf{s}$ the following algorithm is executed:

1. From the measured signature $\mathbf{s}$ detect the line-of-sight peak at a distance d_{LoS} as explained in section 4.1.
2. Construct a partial sphere surface S (as shown in Fig. 4) bounded by the room volume, having radius d_{LoS}. Construct N_p regularly spaced 3D *candidate positions* $\mathbf{x}_i$ with $i = 1 \ldots N_p$ on surface S according to a pre-defined grid. The *grid spacing* parameter Δx represents the euclidean distance between adjacent candidate positions. Δx can be varied, depending on accuracy needs.
3. Start with candidate position $i := 1$. Set vector $\mathbf{m}$ empty.
4. For candidate position $\mathbf{x}_i$ and mobile device orientation vector $\mathbf{v}_R$, calculate the *expected signature* $\mathbf{s}^e$, using the system model (section 3.4) and signature processing (section 5.1).
5. Compare expected signature $\mathbf{s}^e$ to the measured signature $\mathbf{s}$ and compute a *match value* m expressing the amount of matching. Store m into vector element $\mathbf{m}(i)$. Possible matching methods are listed in section 4.3.
6. Proceed to the next candidate position $i := i + 1$, and repeat steps 4-6 while $i \leq N_p$.
7. Find index j with the highest match value in $\mathbf{m}$, $j = \max_i \mathbf{m}(i)$. The estimated position of the mobile device is the candidate position j at position coordinate $\mathbf{x}_j$.

The outcome is a position $\mathbf{x}_j$ whose simulated signature looks most like the measurement. Therefore, $\mathbf{x}_j$ is chosen in step 7 as a likely 3D position of the mobile device.

4.3 Comparison Metrics

We define a *comparison metric* as an expression, involving two signature vectors $\mathbf{x}$ and $\mathbf{y}$ to compare, with a real value outcome m. A good comparison metric has the property that the maximum value of m, for all signatures $\mathbf{x}$ and $\mathbf{y}$ to compare, is associated with a 'best match' of the two signatures. In other words, the higher the outcome, the more $\mathbf{x}$ looks like $\mathbf{y}$. Step 5 of the signature matching algorithm requires such a match value.

Many comparison metrics are possible, for example mean-squared error, probabilistic comparison approaches, or pattern matching. The first metrics we tried were based simply on mean absolute difference between the signatures, because it can be calculated quickly. These metrics can be described by an expression M_q:

$$M_q(x, y) = -\frac{1}{N} \sum_{k=1}^{N} |x(k) - y(k)|^q \tag{4}$$

where q is a parameter to be chosen. The best match occurs when $\mathbf{x} = \mathbf{y}$, yielding maximum match value $M_q = 0$. M_1 is the mean absolute error metric and M_2 the mean squared error metric. Other comparison metrics are currently in development. For example, one such metric is based on the cross-spectrum between two signatures.

4.4 Discussion

The method and algorithm as presented, should be viewed as an initial result. Therefore, a computational complexity analysis was not made yet. Some remarks about the performance of our implementation will follow in section 5.1.

The configuration parameters for the algorithm are contained in parameter vector $\mathbf{p}$ (see section 3.3). Since parameters are measured or estimated, they will likely contain errors. Ideally the sensitivity of the algorithm to such errors should be low. An analysis of sensitivity has not been performed yet. Also for the algorithm parameters (listed in section 4.2) an analysis still needs to be done to determine optimal parameter values.

A drawback of the method is that the mobile device orientation $\mathbf{v}_R$ has to be known. At first sight nothing seems known about this orientation. However, we suggest three methods to obtain it. First, the orientation could be 'fixed by design'. An example is a remote control unit that is mostly lying horizontally on a table surface with the transducer pointing up. A second option would be to estimate an orientation, making use of characteristics of the measured signature. Methods to do so are in development. A third option is to use gravitational and/or inertial orientation sensors within a mobile device.

5 Experimental

The signature matching method has been implemented as a measurement setup, with all processing in software. The current implementation was built for two underlying goals. The first was to validate the room model as described in section 3.3 against a real empty room. The second goal was to realistically emulate an ultrasonic single-base-station positioning system, that makes use of the signature matching method. Then its performance could be tested in several room conditions. Section 5.1 describes the implementation, section 5.2 lists the experimental procedure and section 5.3 gives results.

5.1 Implementation

A choice that had to be made is whether the base station is a transmitter or a receiver. It was chosen to be a transmitter, which allows unlimited mobile receivers without risk of acoustic interference. The base station was implemented as a transmitter attached to a pole. The pole can be placed within a test room at positions close to walls or the ceiling. One mobile device was implemented as a receiver attached to a pole. It can be moved around to measure at various positions and orientations in 3D space.

The remainder of this section will describe the measurement setup, the signal processing operations that produce a *signature*, and finally the implementation of the signature matching algorithm.

Measurement setup. The measurement setup is shown in Fig. 5. One transmitter and one receiver are connected to a measurement PC. Burst-like waveforms for output $u(k)$ are generated by MATLAB® [13] and sent to an output DAC. The analog output drives the 40 kHz Quantelec SQ-40T transmitter within ±3 V at 500 kHz sampling rate. The acoustic signal propagates inside the room, and is recorded by a Quantelec SQ-40R receiver. The weak electrical signal is amplified, and filtered by a 30-100 kHz bandpass filter to remove electrical noise. The ADC samples at 12-bit resolution at 250 kHz and sends the data $y(k)$ to MATLAB. All units are connected by coaxial cables. No special shielding was used for the transducers.

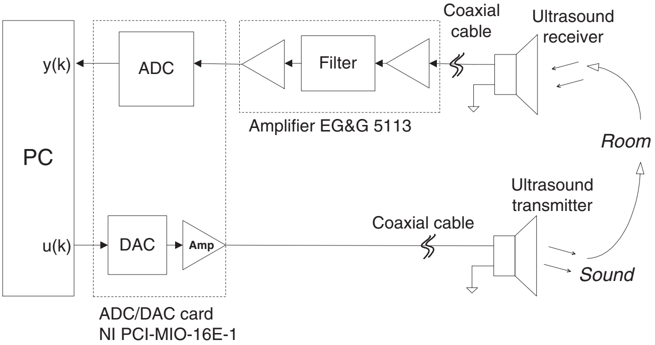


Fig. 5. Measurement setup.

Signal processing. The measured signal $y(k)$ can not be used directly as input to the positioning algorithm. A number of operations are performed to generate *signature data*, which forms the input to a positioning algorithm. The signature contains all relevant information of the measurement in a more convenient and compact form. Figure 6 shows the operations performed on the (discrete) measured data samples $y(k)$. The first step is a cross-correlation filter, that performs a matched filtering operation to remove noise. The template $t(k)$ is the signal as expected to arrive from a single acoustic ray, obtained using the transducer model in section 3.1. The second step demodulates the amplitude envelope from its $f_c = 40$ kHz carrier frequency. Since the bandwidth of the demodulated signal is very low, it can be safely low-pass filtered and downsampled by a factor

5 to a new sampling frequency $f_s = 50$ kHz. The fourth step is attenuation compensation, which compensates the typical drop in signal amplitude of ultrasound over distance. It can be calculated as explained in section 3.2. Without this step, a signature's appearance would be dominated by the first few early-arriving reflections, which are higher in amplitude than late-arriving ones. The compensation step allows for a fair comparison between two signatures when using amplitude-based metrics as in Eq. 4.

The result is a signature $s(k)$ which shows a typical peak and valley pattern, where each peak signifies the arrival of one or more acoustic rays at the receiver at that moment in time. The discrete-time signature $s(k)$ can be written also as a signature vector $\mathbf{s}$. For an example signature see Fig. 2.

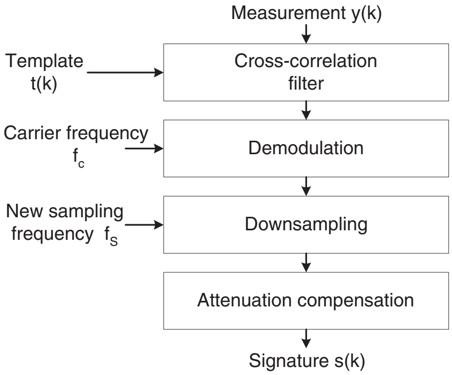


Fig. 6. Signal processing operations to obtain a signature.

Algorithm implementation. The signature matching algorithm was implemented in MATLAB, using the M_1 metric for signature comparison. The most complex part of the algorithm is the simulation of signatures $\mathbf{s}^e$ for each of the candidate positions. Implementation details are not shown, but the major computational bottleneck will be discussed now.

The highest load is imposed by the simulation of the N_p signatures according to Eq. 3. Fourier transforms were used to calculate it in the frequency domain, for increased speed. Two N-point FFT operations and one N-point vector multiplication are then needed per signature, where N=2180 for the current implementation. Typical values for N_p that were tried range from 500 to 20000. To improve performance N_p can be set lower, by choosing a coarser candidate grid size (i.e. $\Delta\mathbf{x}$ higher, see section 4.2) or by excluding certain positions from the candidate set (e.g. positions near the ceiling that are never reached). An interesting improvement would be an iterative approach, that first selects the promising areas in 3D space having high match values, and only executes the algorithm for a limited set of candidate positions located in the promising areas.

5.2 Experimental Procedure

All experiments were performed in an empty office room. An empty room was chosen to verify the acoustic empty room model. Also, an empty room represents the best-case condition for any positioning system. Experiments in a room with obstacles, a more difficult and realistic room condition, are not described in this paper.

The room size is 3.73 m by 7.70 m and 2.97 m high. Some irregularities are present in the form of windows, window-ledges, a door, radiator, a tap and sink, and ceiling-embedded lighting units. A 3D cartesian coordinate system was defined as shown in Fig. 4. The base station position $(0.95, 0.04, 2.95)$ near the ceiling was used, to mimic typical ceiling placement for non-empty rooms, as mentioned in section 4.1. The receiver was placed at several positions in the room which will be shown later in section 5.3. The height of the receiver was set around 1.3 meter, to mimic a typical height for a mobile device that is carried around in a user's hand. The orientation of the receiver was always set parallel to the negative y axis, i.e. $\mathbf{v}_R = (0, -1, 0)$. Measurements were taken at each position. During measurements, no large obstacles or people were present in the room and the door was kept closed.

5.3 Results

The results of initial experiments are presented in this section. First, a single measurement for a receiver position will be examined in detail. A graph will be shown that visualises the output vector $\mathbf{m}$ of signature match values, as generated by the algorithm. Then, for the rest of the measurements only the end results of 3D position estimation will be shown.

One measurement will be examined in detail now. The measured signature $\mathbf{s}$ is shown in Fig. 2. First, the line-of-sight distance is measured (step 1 of the algorithm). A standard peak-detection algorithm estimates the first peak at a distance $d_1 = 2.89$ m, similar to the true distance 2.88 m. A sphere surface S is constructed (step 2) with radius d_1 and centre x_T. Each position on S can now be described in a spherical coordinate system (with origin x_T) by a coordinate (θ, ϕ, r) with $r = d_1$. This way we translate the position estimation problem from 3D cartesian coordinates to a 2D vector (θ, ϕ) to be estimated. A grid spacing $\Delta\theta = 0.017$ is chosen for both θ and ϕ, which corresponds to a variable grid spacing Δx in cartesian coordinates. For this grid spacing it holds that $\Delta x \leq 2 \cdot d_1 \cdot \sin(0.5\ \Delta\theta) = 0.05$ meter. So the candidate positions are at every 5 cm (or closer) within the room. In total $N_p = 11067$ candidate positions exist on the sphere surface within the room.

Starting at candidate position 1, the expected signature is calculated (step 4), a match to the measurement is performed and stored (step 5), and this is repeated for all candidate positions (step 6). The result is a collection of match-values which are a function of the two coordinate parameters θ, ϕ. It is insightful to represent match value as a shade of grey and plot it in a 2D

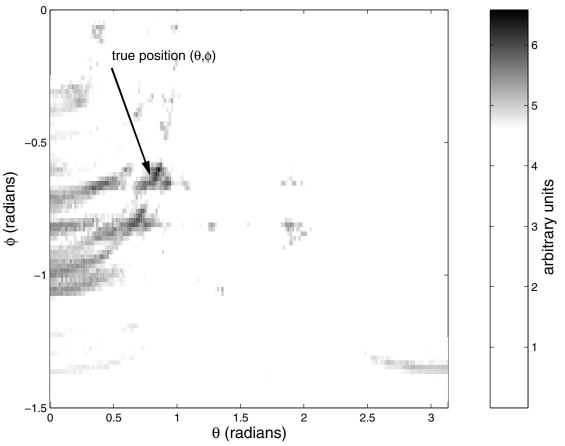


Fig. 7. Visualisation of the result of algorithm steps 1-6. Signature match values are represented by shaded pixels. Darker shades have higher match values. The arrow marks true receiver position.

graph with axes (θ, ϕ). This is shown in Fig. 7, where the higher match values are shown as darker shades of grey. The arrow marks the true mobile receiver position. The area surrounding the true position is darkest grey, which means the signatures there match the measurement best. The maximum match value can now be picked from these results (step 7). Best match is candidate position $(\theta, \phi, r) = (0.849, -0.611, 2.89)$, corresponding to cartesian coordinate $\hat{\mathbf{x}} = (2.51, 1.81, 1.30)$. The 3D distance error $|\hat{\mathbf{x}} - \mathbf{x}_R|$ with respect to the true position $\mathbf{x}_R = (2.60, 1.70, 1.27)$ is just 15 cm.

The algorithm was executed for 20 measurements in total at various positions in the test room. These positions are shown as circles in Fig. 8, which contains an X/Y top view of the room. In the same graph, the estimated positions are shown by lines pointing from the encircled true positions towards the estimated positions. In Fig. 9 the same measurements are shown as a bar graph where the vertical axis represents the 3D estimation error $|\hat{\mathbf{x}} - \mathbf{x}_R|$. It can be seen that accuracy is usually better than 20 centimeters, except for positions 2 and 11 which have a relatively large position error. The reason for these errors is the topic of further research.

6 Conclusions

Based on the experimental work it can be concluded that measured ultrasonic signals contain much more information than just the transmitter-receiver line-of-sight distance. This information is contained in a measured pattern, the *signature*. The signature consists of amplitude peaks, that are caused by acoustic reflections. It was shown that the signature can be predicted by an acoustic

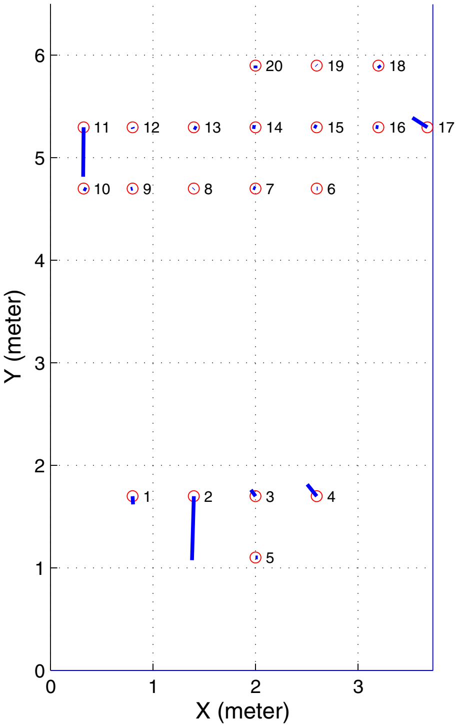


Fig. 8. Top view of the test room, showing 20 measurement locations (encircled). The position estimates per position are shown by the tips of the solid lines.

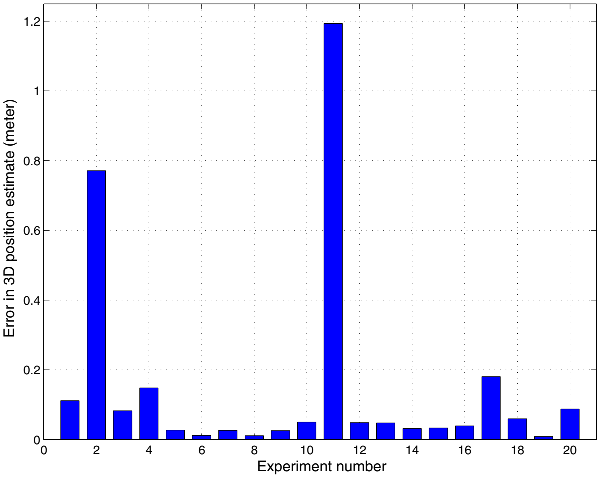


Fig. 9. 3D positioning results over 20 experiments. The vertical axis plots 3D position estimation error.

room model. We propose to use the information contained in the signature to perform 3D device position estimation, using just a single base station per room. A method called *signature matching* was designed and implemented for this purpose. It was shown by initial experiments that the acoustic model is accurate enough to use for 3D position estimation, for the case of an empty room.

The method described in this paper is not yet mature. In future work a number of steps will be taken: Firstly, the method will be tested more thoroughly, specifically in realistic non-empty room conditions. Secondly the computational complexity of the method needs to be improved. Thirdly, a sensitivity analysis has to be performed to find out the effect of errors in the method's input parameters. Fourthly, an extension of the method, based on transducer arrays, is under development. Such an array allows a base station to get more information about the direction of mobile devices, thus enabling more robust position estimates.

Acknowledgments. The authors would like to thank S. Egner for signal processing support, and the other members of the Phenom project team: Y. Burgers, E. van den Hoven, N. de Jong, Y. Qian, D. Teixeira and E. Tuulari.

References

1. Priyantha, N., Miu, A., Balakrishnan, H., Teller, S.: The Cricket Compass for Context-Aware Mobile Applications. In: Proc. ACM Conf. on Mobile Computing and Networking (MOBICOM). (2001) 1–14
2. PHENOM project: www.project-phenom.info (2003)
3. Ni, L., Liu, Y., Lau, Y., Patil, A.: LANDMARC: Indoor Location Sensing Using Active RFID. In: Proc. IEEE Int. Conf. on Pervasive Computing and Communications (PerCom). (2003)
4. Prasithsangaree, P., Krishnamurthy, P., Chrysanthis, P.: On Indoor Position Location with Wireless LANs. In: Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC). (2002) 720–724
5. Addlesee, M., Curwen, R., Hodges, S., Newman, J., Steggles, P., Ward, A., Hopper, A.: Implementing a Sentient Computing System. IEEE Computer **34** (2001) 50–56
6. Hazas, M., Ward, A.: A Novel Broadband Ultrasonic Location System. In: Proc. Int. Conf. on Ubiquitous Computing. (2002) 264–280
7. InterSense: IS series high precision trackers, www.isense.com (2003)
8. Ziomek, L.: Fundamentals of Acoustic Field Theory and Space-Time Signal Processing. CRC press (1995)
9. Crocker, M.: Handbook of Acoustics. J. Wiley & Sons (1998)
10. Kuttru, H.: Room Acoustics. 3rd edn. Elsevier (1991)
11. ISO: Standard 9613–1; Acoustics - Attenuation of sound during propagation outdoors (part 1), www.iso.ch (1993)
12. Allen, J., Berkley, D.: Image Method for Efficiently Simulating Small-Room Acoustics. J. Acoust. Soc. Am. 65 (1979) 943–951
13. Mathworks: MATLAB version 6 (R13), www.mathworks.com (2003)

Wearable Context Aware Terminal for Maintenance Personnel

Heikki Ailisto, Ville Haataja, Vesa Kyllönen, and Mikko Lindholm

VTT Electronics, P.O.Box 1100, FIN-90571, Oulu, Finland
{Heikki.Ailisto, Ville.Haataja, Vesa.Kyllönen,
Mikko.Lindholm}@vtt.fi

Abstract. A wearable context aware terminal with net connection and spoken command input is presented. The context aware terminal can be used, for example, by janitors or other maintenance personnel for retrieving and logging information related to a location, such as a room. The main context cues used are user's identity and location. The user is identified by biometrics which is also used for preventing unauthorized usage of the terminal and information accessible through the terminal. Location information is acquired by using signal strength information of existing Wireless Local Area Network (WLAN) infrastructure. Since the wearable terminal is envisaged to be used by maintenance personnel it was seen important to offer possibility to hands-free operation by using spoken commands. Tentative experiments show that this approach might be useful for maintenance personnel.

1 Introduction

Maintenance personnel, whose task involves taking care of appliances, machines or spaces which are spread physically can not rely on desktop computers for real-time information retrieval and logging. Often, this task is handled by using paper and pen while on the field and separately retrieving or logging information from or into desktop computer. In many cases this is inefficient and error prone way of doing things. Using mobile terminals instead of pen and paper and a desktop has been solicited as a modern solution both by industry and academia [1],[2],[3]. For example, wearable computers with speech input are suggested for large vehicle maintenance [1] and quality inspectors in food processing industry [2]. We take this approach still further by making the terminal more usable and wearable by including context aware (CA) technology for automatic location identification, hands-free spoken command operation and prevention of unauthorized usage by biometric user identification.

Context awareness has been defined by Schilt and Theimer [4] as software adapting to location, collection of people and objects nearby and their change over time. A more general definition of context awareness by Dey [5] states that "A system is context-aware if it uses context to provide relevant information and/or services to the user, relevancy depends on the user's task". Almost any information available at the time of an interaction can be seen as context information. Examples include identity; spatial information - e.g. location, orientation, speed, and acceleration;

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 149–159, 2003.

temporal information - e.g. time of the day, date, and season of the year; environmental information - e.g. temperature, air quality, and light or noise level; social situation - e.g. who you are with, and people that are nearby; resources that are nearby - e.g. accessible devices, and hosts; availability of resources - e.g. battery, display, network, and bandwidth; physiological measurements - e.g. blood pressure, hart rate, respiration rate, muscle activity, and tone of voice; activity - e.g. talking, reading, walking, and running; schedules and agendas [6]. Location is seen as one of the most useful cues [7], [8], [9], [10]. Furthermore, the definition of location is less ambiguous than e.g. that of social situation or activity.

Context aware mobile terminals have been suggested, for example, to be used for tourist guidance [8], [9], and as museum guides [10],[11]. Kindberg and Barton [7] suggest that nomadic users will use Personal Digital Assistants (PDAs) with network access and equipped with special sensors for reading infrared beacons, barcodes or RF-tags for obtaining addresses (URLs) to location specific web resources. It seems that most mobile terminals using CA are suggested for leisure and recreational use [8],[9],[10] or to be used by clerical employees in offices [11],[12].

In most cases the approach suggested in this paper deviates from this mainstream of CA mobile terminal research, since we suggest the use of the terminal by maintenance personnel, who have an actual need for real-time data retrieval and logging as well as a possibility to benefit from clearly defined (and limited) usage of context information, namely the user's identity and location. By including the capability of hands-free operation using spoken commands, the terminal becomes more wearable instead of just mobile or portable. Relying on WLAN infrastructure, which in any case is necessary for wireless terminal and is assumed to have widespread use during coming years, we can avoid the burden and cost of building extra infrastructure, such as IR beacons, barcodes or RF-tags for locating purpose.

The structure and content of the information base for specific maintenance tasks is out of the scope of this study, but we assume it to be intranet or internet based.

In this paper the requirements for a context aware wearable terminal as well as its realization are presented along with tentative experimental results. Some conclusions are drawn and ideas are presented for further research.

2 Context Aware Wearable Terminal

2.1 Requirements

In order to gain insight to what is required, we analyzed the needs set for wearable terminal in maintenance work. We chose a janitor case for detailed study. The work typically involves doing pre-scheduled checks and maintenance tasks on the one hand and sporadic repairs and troubleshooting on the other hand. Both types of work can benefit of using wearable context aware terminal. In pre-scheduled work, the terminal may contain to-do lists for each day or week and location specific information concerning particular rooms, machines or appliances. When a task is completed, it can be checked out in the to-do list and the necessary information can be easily logged in to information base on the spot without the trouble of first writing it down on a paper and then separately feeding it to a computer in the office. An example of this might be the scheduled replacing of fluorescent lamps in a particular hall. Before the work, the correct type of lamps can be checked with the terminal from data base. After the work, replacement date is logged into the data base (e.g. web page) related to the hall. In sporadic tasks, the user complaints or other alarms associated to a certain location, e.g. room, hall, appliance or machine, can be retrieved based on context information.

Identifying the user automatically was seen useful in two ways: first, information security can be ensured without cumbersome and to some extent risky passwords (people tend to write down their passwords) and secondly, user's profile can be evoked automatically.

Possibility to hands-free operation was seen important in some tasks. This implies that interfaces used in typical portable devices such as PDAs and mobile phones are not completely satisfactory, since they rely mostly on keys and touch screens for input. Furthermore, overall wearability was perceived as important.

A similar analysis was done to find out how cleaning personnel could benefit from this type of wearable terminals. The needs were quite similar, except that the wearability issues were even more critical than for janitors. It was also suggested that the terminal could be placed in the cart used by cleaners.

The requirements set for the wearable CA terminal derived from the analysis are:

- wireless data retrieval and logging supported by location awareness (often location = task) and user specific profiles (user = context cue)
- strong User Interface (UI) capabilities: graphical, touch screen, audio signals
- hands-free operation: spoken command input and voice output
- prevention of unauthorized usage: biometric user identification
- communication capability: email and phone calls over Internet

Furthermore, the platform should be open so that it can be configured and programmed. Additional constraints are that terminal should not be too bulky and it should operate reasonable time with onboard batteries.

2.2 Design and Realization

After some considerations, we decided on building the context aware wearable terminal around PDA hardware. This was motivated by the openness, sufficient software support, abundant accessory supply, adequate UI features and reasonable size. We completed the terminal by including a WLAN card and a speaker and spoken command recognition unit. The WLAN card offers wireless Internet/Intranet access, voice over IP calls and indoors locating function. The speaker and spoken command recognition unit is utilized for user identification and authentication as well as for producing the hands-free feature by spoken commands input. A block diagram of the experimental wearable CA terminal is shown in Figure 1.

A Compaq iPaq® was used since we had earlier experience in programming that particular device. The WLAN card is SilverCard by Orinoco. The speaker and spoken command recognition unit is a Sensory Voice Extreme^{TM} with serial connection and a headset interface. The terminal is not as small as we would like it to be, but it is made less awkward by wearing it on the belt and using a headset for commands.

The user interface of the wearable terminal is realized with a browser (Microsoft Internet Explorer). The browser window is divided into four frames. The applet frame is the interface to the speaker and spoken command recognition unit and it gives the starting code. Person, map and info frames are updated according to the results of the speaker recognition, WLAN-positioning and spoken commands, respectively.

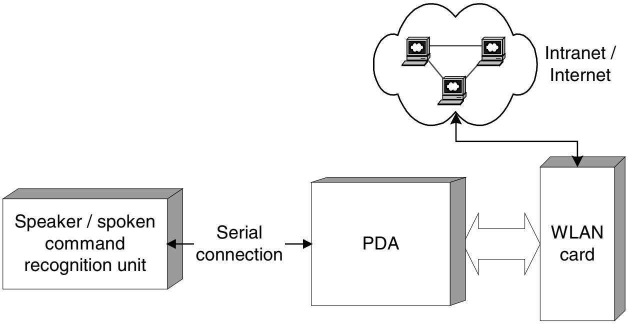


Fig. 1. Block diagram of the wearable CA terminal

The system architecture is presented in Figure 2. The voice recognition module manages connection to the speaker and spoken command recognition unit using serial communication. Messages from the speaker and spoken command recognition unit are sent to the mode reasoner module using TPC/IP socket messages. The voice recognition interface module is implemented in C++.

WLAN-positioning consists of two modules. The measurement module in PDA measures the access point signal strengths and sends them to the server for the calculation. The centralized and versatile positioning module calculates the position of the terminal. The positioning module furthermore sends the position to the mode reasoner module. Both positioning modules are implemented in C++.

The mode reasoner module combines and saves information from the WLAN-positioning, speaker recognition and spoken command recognition. The task of the module is to manage and maintain information about the location, the identity of the user and the last speech command given. It is implemented in Java.

The servlet module asks current positioning and speaker and spoken command information from the mode reasoner module and with this information generates person, map and info frames for the browser. The servlet module is implemented in Java.

For a wearable terminal, wireless connection is necessary and this has been implemented with WLAN connected to the Intranet. WLAN is an existing technology and it has good potential of becoming widely used in public buildings and offices.

WLAN based method for locating was chosen because it can use existing infrastructure, i.e. WLAN access points and a WLAN card needed in the terminal for wireless Internet/intranet access. This approach is less costly than earlier context aware terminals, which use specially built infrastructure, such as IR beacons, RF tags or barcodes for locating [7],[12],[13]. GPS was abandoned, since it requires extra hardware, does not operate well indoors and is not as precise as the WLAN based approach. Solutions based on mobile phone cell identification would not by far be accurate enough. Furthermore these and other operator net-based methods, such as

EOTD supported by the GSM, suffer from the cost imposed by the operators for this service. For these reasons the use of in-house WLAN for acquiring location information is well founded.

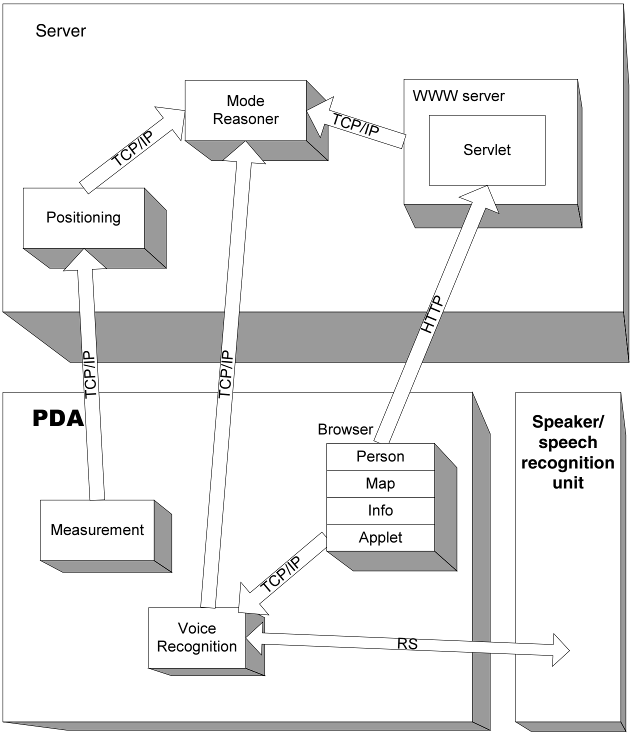


Fig. 2. System architecture

There are two main methods for acquiring location information based on measured WLAN access point signal strengths. Both require measuring the signal strength of at least two, preferably three, WLAN access points. The first method uses theoretical model of attenuation of the signal strength as the basis of the location calculation. In its simplest form, signal strength is assumed to be a function of the distance to the access point. This method can be enhanced by including some assumptions about the effects of antenna, walls and other structures to radio propagation. This makes calculations laborious and some error sources remain, since the effects of e.g. walls are not easily modeled.

The second method, called location fingerprinting, is based on learning. The signal strengths associated with a number of locations are measured in the learning phase and recorded into a table containing the co-ordinates of each location, the signal strengths and possibly some semantic information related to each location. The locating step uses k-nearest neighbor algorithm for determining an unknown location based on measured signal strengths and the table. Kalman filtering is used to estimate the movement of the target and suitable constraints are employed to reject meaningless results. This latter method was chosen since it is more robust and straightforward. Furthermore, this approach had been implemented in VTT Electronics and was readily available for this work.

Voice over internet software transfers voice (Pulse Coded Modulation data) through TCP/IP (Transfer Control Protocol/Internet Protocol) socket. This approach wastes bandwidth, so some compression technique could be used. Recording and playing software is implemented in C module, which is called from Java with Java Native Interface (JNI).

The biometric user identification and user profile selection is based on voice recognition. A pass-phrase is given in the teaching phase and then a template derived of this is later compared with the input phrase. The confidence level of the identification is selectable by software.

Spoken command recognition is also based on learning. The commands are repeated twice in the learning phase. In the recognition phase, a recognition result along with a recognition score - or an error code - is returned after each command.

3 Experimental Results

Experimental results of the operation of each sub-system of the terminal along with experience of tentative trials are concisely described here.

3.1 Performance of the Sub-systems

The sub-systems tested were wireless information retrieval and logging, context awareness based on location, user authentication and profile selection based on bioidentification, hands-free usage by voice commands and voice over internet calls.

When regarding the technical performance of the terminal, it can be concluded that the WLAN based solution was quite satisfactory in this case.

Currently, the context awareness includes user identification and location measurement. Including time-of-the-day information, e.g. for differentiating day-time and night-time check up routines, has also been considered and it would technically be very easy. Here, we concentrate on the location measurement. A test set-up consisting of three WLAN access points situated in a typical office building having long corridors and rooms along them was built. The area was L-shaped with dimensions of ca. 30 x 60 m and an area of 1000 m^2. The learning phase consisted of giving the co-ordinates of 240 points. The performance test included measurements in stabile position and while moving at slow walking pace (0.4 m/s). For the stationary situation, the average error was 0.69 m and worst case error was 1.4 m. The average error increased to 1.5 m when the terminal was moving during measurement. This is due to the fact that the individual measurements are done at 0.3 s intervals and the averaging algorithm uses one to three latest values for calculation. If three values are used, they actually represent three different locations and if only one value is used, the stochastic error of individual measurement is not smoothed by averaging. For this application, since it is most important to locate the room correctly, the accuracy of both stationary and mobile cases is sufficient.

Authentication and profile selection worked as expected. No quantitative experiment of false rejection - false acceptance rates were made in this study.

Hands-free use test included 25 different commands and they were spoken twice in each experiment. The results are summarized in Table 1, where the column "result" contains the number of correctly and falsely recognized and rejected commands in that order. The test was conducted in an office building using the head-set.

Table 1. Recognition of spoken commands

Test condition	Result
Person 1 (male), same day	49 / 0 / 1
Person 1, next day	49 / 0 / 1
Person 2 (male), same day	48 / 1 / 1
Person 2, two days later, sore throat	47 / 0 / 3

3.2 Tentative Experiments

The tentative experiment consisted of a case where a janitor uses wearable context aware terminal for retrieving and logging location specific information relevant to maintenance work. The course of the experiment along with comments and remarks is described here and summarized in Table 2.

Table 2. Course of the experiment

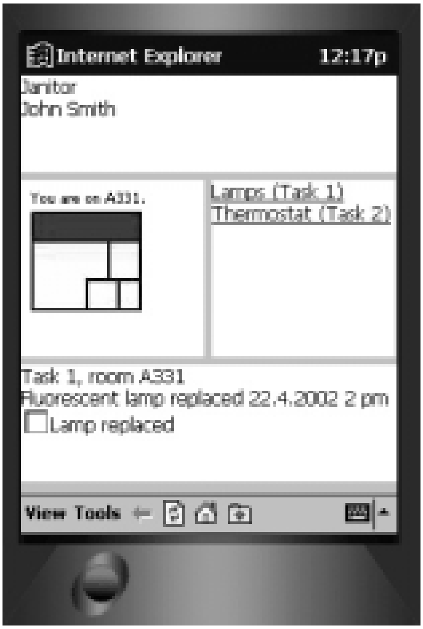
Task or function	Comment
Authentication, voice recognition	Works ok mostly, but may annoy, if there are problems
Spoken commands	Needed for hands-free, may be socially embarrassing still
Context aware information retrieval	This works well, avoids going through menus etc. Good usage design or learning very important
Information logging	Since no keyboard, must be simple, check boxes good
Voice over Internet calls	Not as comfortable as cellular phone

The wearable terminal is activated by giving a spoken password. The activation contains both authentication and user profile selection, when more than one person uses the terminal. The activation can be performed quite fast and since no password or PIN code is required, it is perhaps less stressing than other methods. In cases where other persons were present, talking to a computer was perceived as embarrassing. An alternative method for activating the terminal should be available for such situations or for the cases where persons voice has temporarily altered, e.g. for sore throat.

The hands-free operation using a light-weight head-set was found useful in situations where the user was tied to a manual job or was carrying something. In our experience, a limited set of commands can be handled by voice input, but the information output required a graphical display, since the nature and amount of the information made voice output e.g. by speech synthesis unpractical. Finding a good place for the display was a problem, which was solved by placing the PDA on the belt in a holder which made viewing it possible. The idea of using head-up display was seen as a promising alternative, although that would mean giving up the touch screen. Again, giving spoken commands to the terminal was perceived as embarrassing when other persons were present.

A typical information screen (or web frame) associated to a location may contain a schematic map and a list of tasks related to the room, e.g., TASK 1: Fluorescent lamps (last replacement), TASK 2: Thermostat adjustment. The user can evoke a task by a spoken command, e.g. "Task one", after which a new frame associated with this task comes to display. This example is depicted in Figure 3.

One of the findings of the experiment was that it is important to keep the information offered by the terminal concise, since the size of the display and the nature of the maintenance work (not at your desk!) requires that. Keeping the amount of information to minimum becomes even more important, when we think about information logging. A maintenance person, such as janitor, cleaner or mechanic, does not want to spend any extra time browsing menus and filling in forms. Furthermore, since there is no keyboard in the terminal, most of the inputs should be just check box or radio button type inputs which can easily be used. This puts great weight to good UI and usage design practices.



Janitor
John Smith

You are on A331.

Lamps (Task 1)
Thermostat (Task 2)

Task 1, room A331
Fluorescent lamp replaced 22.4.2002 2 pm
□ Lamp replaced

Fig. 3. Information screen for room A331

Context awareness is often seen as a way for achieving proactive behavior. We noticed that pro-activity should be kept in proportion, for example, it is good to offer applications or information in assumed preference order, but it is annoying, if an application is launched automatically when entering a room (you may be in the room just to pick up a forgotten tool!).

The wearable context aware terminal used by a maintenance person is shown in Figure 4. Note the use of a head-set for spoken commands and audio output, e.g. prompting and acknowledging commands.

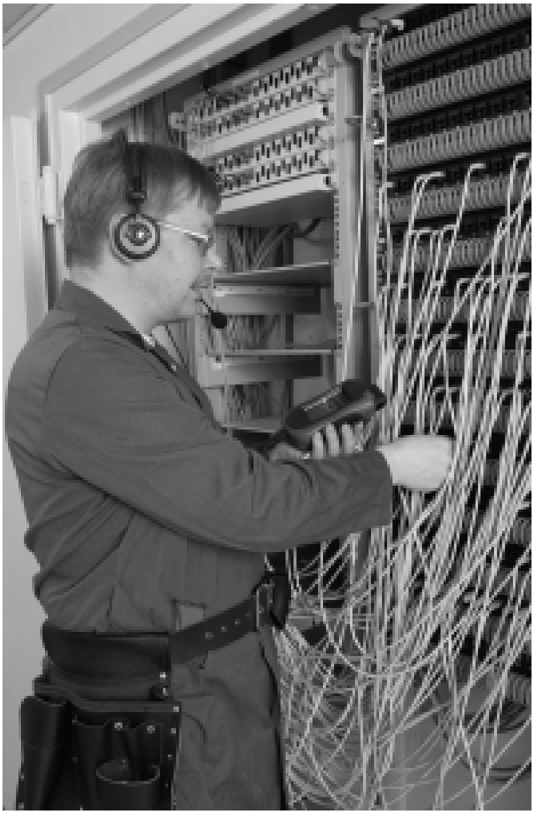


Fig. 4. The terminal in experimental use

4 Conclusion

A wearable context aware terminal with net connection and spoken command input has been presented. The terminal is envisaged to be used by maintenance personnel, who need mobile and wearable terminal for information retrieval and logging, require possibility to hands-free operation and can benefit from context awareness.

The terminal was designed around a PDA equipped with a WLAN card for wireless connection and location measurement. A speaker and spoken command recognition unit was chosen for authentication, user specific profile selection and hands-free operation. The software was implemented in C++ for the primitive modules (serial communication and WLAN-measurements) and in Java for the reasoning. The user interface was implemented in HTML and as a Java applet. The structure and content of the information base for any specific maintenance task is out of the scope of this paper.

After construction, the terminal underwent some performance testing and tentative experimenting. All the sub-systems worked properly. Major concern with the speaker identification and voice command unit is social acceptance of "talking to computer". It may be assumed that this will become more acceptable, the way talking to a mobile phone in public spaces has become. WLAN based locating method was seen as a good choice, since it relies on infrastructure, which is needed anyhow and it offers sufficient accuracy for indoors location. Already the tentative experiments showed that offering - not launching - services, based on location and user cues is useful. Also, it is our view that we must be careful to avoid information overload in retrieval and even more so in data logging. This requires very good design and knowledge of the workflow.

The tentative experiment shows that the requirements set for a wearable context aware terminal can be technically realized with the technology available today. Wireless entrance to Internet/Intranet, location measurement and recognition of spoken commands along with the proliferation of PDAs are recent technological steps facilitating this. Although the size of the display (PDA) and its weight are somewhat problematic, it seems that the largest challenges are with user acceptance and proper design of the applications, i.e. supporting the natural work flow of a maintenance personnel.

Further research is needed on different areas, e.g. using wearable displays, and studies on user needs and work flows in specific tasks such as janitor, mechanic, cleaner and clerical employee.

References

1. Smith, B., Bass, L., Siegel, J.: "On site maintenance using a wearable computer system", in Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, volume 2 of Interactive Posters, 1995, pp. 119–120
2. Najjar, L.J., Thompson, J.C., Ockerman, J.J.: "A wearable computer for quality assurance inspectors in a food processing plant", in Proceedings of the 1st International Symposium on Wearable Computers, 1997, pp. 163–164
3. Ockerman, J.J., Najjar, L.J., Thompson, J.C.: "Wearable computers for performance support: Initial feasibility study", Personal Technology, Vol. 1, no. 4 (1998), 251–259
4. Schilit, B., Theimer, M.: "Disseminating active map information to mobile hosts", IEEE Network 1994, IEEE Press no. 8, pp. 22–32
5. Dey, A.: "Understanding and Using Context", Personal and Ubiquitous Computing (2001), no. 5, pp. 4–7
6. Korkea-aho, M.: "Context-Aware Applications Survey", Technical report, Helsinki University of Technology http://www.hut.fi/~mkorkeaa/doc/context-aware.html, (2001)
7. Kindberg, T., Barton, J.: "A Web-based Nomadic Computing System", Computer Networks, vol. 35, no. 4, March 2001, pp. 443–456

8. Long, S. et al.: "Rapid prototyping of Mobile context-aware applications: The Cyberguide Case Study", in Proceedings of 2nd Annual International Conference on Mobile Computing and Networking (Mobicom '96), ACM Press, New York, 1996, pp. 97–107
9. Davies, N., Cheverst, K., Mitchell, K., Efrat, A.: "Using and determining location in a context-sensitive tour guide", IEEE Computer, (August 2001) pp. 35–42
10. Petrelli, D., Not, E., Zancanaro, M., Strapparava, C., Stock, O.: "Modelling and Adapting to Context", Personal and Ubiquitous Computing, no. 5 (2001), pp. 20–24
11. Not, E., Petrelli, D., Stock, O.,Strapparava, C., Zancanaro, M.: "Augmented space: bringing the physical dimension into play", in Proceedings of the Flexible Hypertext Workshop, held in conjunction with the 8th Int. Hypertext Conference (Hypertext'97), 1997, www.mri.mq.edu.au/ ~mariam/flexht/
12. Want, R. et al.: "An Overview of the ParcTab Ubiquitous Computing Experiment," IEEE Personal Comm., vol. 2, no. 6 (December 1995), pp. 28–43
13. Want, R., Hopper, A., Falcao, V., Gibbons, J.: "The active badge location system", ACM Transactions on Information Systems, vol. 10, no. 1 (January 1992), pp. 91–102

Sensor Fusion for Augmented Reality

Jurjen Caarls, Pieter Jonker, and Stelian Persa

Delft University of Technology
{jurjen, pieter, stelian}@ph.tn.tudelft.nl

Abstract. In this paper we describe in detail our sensor fusion framework for augmented reality applications. We combine inertia sensors with a compass, DGPS and a camera to determine the position of the user's head. We use two separate extended complementary Kalman filters for orientation and position. The orientation filter uses quaternions for stable representation of the orientation.

1 Introduction

Wireless technologies that enable continuous connectivity for mobile devices will lead to new application domains. An important paradigm for continuously connected mobile users based on laptops, PDAs, or mobile phones, is context-awareness. Context is relevant in a mobile environment, as it is dynamic and the user interacts in a different way with an application when the context changes. Context is not limited to the physical world around the user, but also incorporates the user's behavior, his

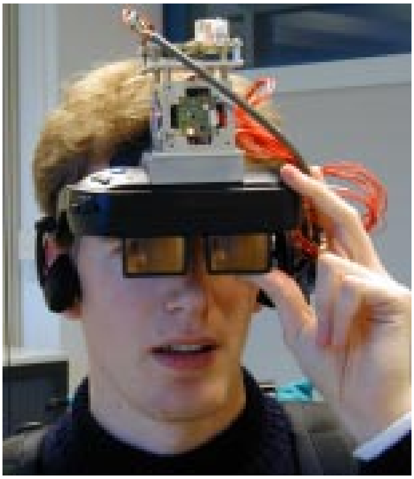


Fig. 1. Augmented Reality Prototype

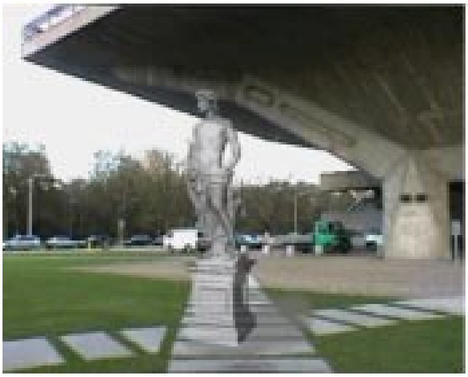


Fig. 2. Example of augmentation of the visual reality (1)

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 160–176, 2003.

terminal and the network characteristics. As an example of a high-end context aware application, we are in the development of an Augmented Reality system that can be connected to a roaming PDA.

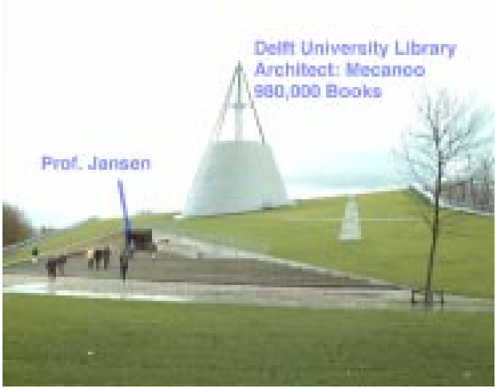


Fig. 3. Example of augmentation of the visual reality (2)

Our user carries a wearable terminal and a see-through display in which the user can see virtual visual information that augments reality (Figure 1). Augmented Reality differs from Virtual Reality in the sense that the virtual objects are rendered on a see-through headset. As with audio headphones, which make it possible to hear sound in private, partly in overlay with the sounds from the environment, see-through headsets can do that for visual information. The virtual objects are in overlay with the real visual world (Figures 2, 3, and 4). It can also be used to place visual information on otherwise empty places, such as white parts of walls of a museum. The 3D vector of position and orientation is referred to as pose. Knowing the pose of those walls and the pose of a person's head, visual data can be perfectly inlayed on specific spots and kept there while the head is moving.

To lock the virtual objects in the scene, the head-movements must be sampled with such a frequency and spatial accuracy that the rendering of virtual images does not cause motion sickness. Our system can be applied in Tour Guiding, Remote Maintenance, Design Visualization and Games.

The wearable system contains a radio link that connects the user to computing resources and the Internet. For outdoor augmented reality, location determination is based on Differential GPS, while WLAN is used for the connection with backbone services. For communication in places without WLAN access-points near-by, GPRS can be used. For indoor applications, a presence and location server can be based on the signal strength of Bluetooth and/or WLAN access-points [4]. Based on the course position and orientation from the servers, a camera on the AR headset can capture the user's environment, which, fused with data from an inertia system: gyroscopes, accelerometers, and compass, can make the PDA fully aware of the absolute position and orientation of the user. Camera images can be sent to the backbone and matched to a 3D description of the environment to determine the user's position and to answer questions of the user and his PDA that relate to the environment. This description can be derived from a GIS database, for outdoor, or a CAD database for indoor applications, see [5]. For simple indoor applications tags can be used that e.g. stick on known positions on the walls. To prevent motion sickness, rendering latencies lower than 10 ms are necessary. In conventional systems with a refresh rate of 50Hz it takes 20 ms to display a single frame. The time to render that frame will add to the total latency. It is clear that it is not possible to reach the required latency for augmented reality (<10ms) by sequentially rendering and displaying a frame. Consequently, our system renders only a part of the frame just ahead of the display's raster beam in four slices, and has a combined rendering and display latency of 8 ms [6].

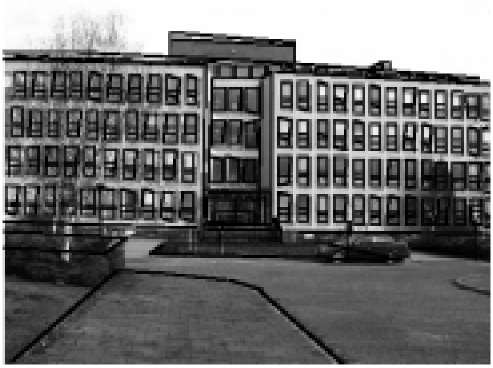


Fig. 4. Example of an overlaid GIS/CAD model of a building.

In this paper we address the problem of the fusion of the sensors that are needed to obtain an accurate, fast and stable rendering system for augmented reality of indoor and outdoor scenes. We fused data of various sensors with different update rates and accuracies, including vision and DGPS, by using extended Kalman Filtering. The novelty in our approach is the use of quaternion descriptions inside the filter.

2 Sensors

To track the pose (position and orientation) of the user's head, we use a combination of sensors, which can be divided into relative sensors (angular velocity and linear acceleration) and absolute sensors (orientation and position). For the relative sensors we used three gyroscopes (Murata) and three accelerometers (ADXL202) combined in one board linked to a LART platform [1] developed at the Delft University of Technology. For the absolute sensors we use a Precision Navigation TCM2 compass, tilt sensor, and a JAI CV-S3300 camera.

The Murata Gyrostar piezoelectric vibrating gyros can measure up to 300 °/s. They are inexpensive but have a large bias that varies with time up to 9 °/s. Consequently, we had to correct for this bias. After our correction, the noise level is around 0.2 °/s when sampled at 100Hz. The accelerometers (ADXL202) have also a varying offset. This offset can be 0.5 m/s^2 and the residual noise level is around 0.06 m/s^2 when sampled at 100Hz. The maximum acceleration that can be measured is 2*g* in both directions.

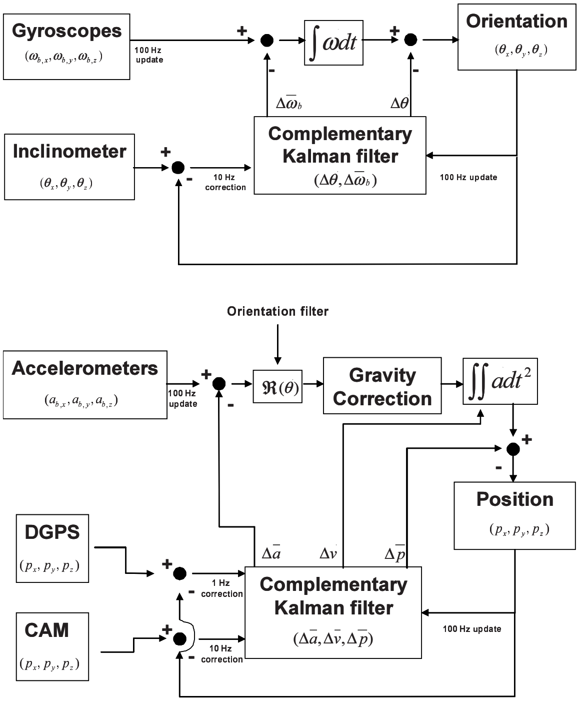


Fig. 5. Fusion of data from the sensors in the pose tracking system. *Top:* orientation. *Bottom*: position

The TCM2-50 liquid inclinometer uses a viscose fluid to measure the inclination with respect to the gravity vector with an accuracy of 0.2°. The heading is calculated using three magnetometers with an accuracy of 0.5-1.5°. Because the liquid will slightly slosh when accelerations are applied, we have to cope with an error of about 20°. The update rate is 16 Hz. The JAI CV-S3300 camera with a resolution of 320 x 240 pixels in grayscale has a wide-angle lens with a 90° opening angle, which introduces spherical distortions. We calibrate the camera using the Zhang algorithm [7]. Images are grabbed at 15 Hz. When the camera sees a specific block pattern in the image it can track the full 6D pose of the camera. The LART platform, that is used for data acquisition and preprocessing, has an 8-channel fast 16-bit AD-converter to acquire synchronous data from the accelerometers, gyros and temperature data. The gyros and the accelerometers are analog devices, which are sampled at 100 Hz by the AD converter. The TCM2 updates at 16 Hz and is read via a serial line. When the TCM2 and the gyros are read out simultaneously, there is an unknown difference in the time of the physical events. We could compensate the relative latencies by attaching a time stamp to the readouts. Figure 5 shows a diagram of the fusion of data from the sensors in the pose tracking system, which is explained in detail in the next section.

2.1 Fusion Framework

For our fusion framework, we use the following coordinate systems:

Ψ_b The body frame. It is attached to the body of the headset for which we need the pose.

Ψ_n The navigation frame. This frame is a rotated body frame. It has the z-axis pointing in the direction of the gravity vector, while the x-axis is pointing to the earth's North Pole.

Ψ_p The pattern frame. This frame is attached to the pattern used to determine the camera's position.

Ψ_w The world frame. This frame is like the navigation frame, but has a fixed origin.

Sub indices like $\Psi_{b,1}$ denote a specific instance of a moving frame.

To be able to combine sensors measurements with different update rates and error characteristics, we have chosen a Kalman filter setup [2]. Due to the rotations, the Kalman equations become non-linear, and hence we need to linearise the filter. We have chosen to use the errors in position and orientation as filter states, as then we can update the real states using nonlinear formulas to obtain a better performance. To overcome singularities when representing orientation in Euler angles, we used the quaternion notation. Quaternions can be used to represent orientations in 3D. The advantage over the use of common Euler angles is that the representation of the orientation is continuous, i.e. without jumps from 2π to 0.

2.2 Quaternions

A quaternion has a scalar part and a vector part:

$$\begin{aligned} q &= \begin{pmatrix} q_0 \\ \vec{q} \end{pmatrix} \text{ and } q^* = \begin{pmatrix} q_0 \\ -\vec{q} \end{pmatrix} \\ q_0 &= \cos(\theta/2) \\ \vec{q} &= \vec{n} \cdot \sin(\theta/2) \end{aligned} \tag{1}$$

in which θ is the angle of rotation around the normalized vector $\vec{n}$, and q^* is the complex conjugate of q. If q represents an orientation, i.e. a unit quaternion, then its inverse becomes:

$$q^{-1} = \frac{q^*}{\|q\|^2} = q^* \tag{2}$$

A quaternion that represents a rotation of a frame Ψ_A expressed in terms of frame Ψ_B is represented by q_A^B, the rotation that rotates frame Ψ_B to frame Ψ_A expressed in Ψ_B. Let the quaternion representation of a vector $\vec{v}_a$ be:

$$q_{\vec{v}_a} = \begin{pmatrix} 0 \\ \vec{v}_a \end{pmatrix} \tag{3}$$

Then the rotation of this vector is obtained by a double quaternion multiplication:

$$q_{\vec{v}_B} = q_A^B \otimes q_{\vec{v}_a} \otimes q_A^{B*} = R_A^B \vec{v}_A \tag{4}$$

in which the operator $\otimes$ is the quaternion multiplication. In matrix form this becomes:

$$\begin{aligned} q_1 \otimes q_2 = \widetilde{q}_1 q_2 &= \begin{pmatrix} q_{1,0} & -q_{1,x} & -q_{1,y} & -q_{1,z} \\ q_{1,x} & q_{1,0} & -q_{1,z} & q_{1,y} \\ q_{1,y} & q_{1,z} & q_{1,0} & -q_{1,x} \\ q_{1,z} & -q_{1,y} & q_{1,x} & q_{1,0} \end{pmatrix} \begin{pmatrix} q_{2,0} \\ q_{2,x} \\ q_{2,y} \\ q_{2,z} \end{pmatrix} \\ = \overset{\frown}{q}_2 q_1 &= \begin{pmatrix} q_{2,0} & -q_{2,x} & -q_{2,y} & -q_{2,z} \\ q_{2,x} & q_{2,0} & q_{2,z} & -q_{2,y} \\ q_{2,y} & -q_{2,z} & q_{2,0} & q_{2,x} \\ q_{2,z} & q_{2,y} & -q_{2,x} & q_{2,0} \end{pmatrix} \begin{pmatrix} q_{1,0} \\ q_{1,x} \\ q_{1,y} \\ q_{1,z} \end{pmatrix} \end{aligned} \tag{5}$$

In which $\tilde{q}$ is the quaternion matrix and $\overline{q}$ the transmuted quaternion matrix.

The representation of angular velocity vectors using quaternions is analogue to the case of rotations of position vectors over angles. The angular velocity of Ψ_i with respect to Ψ_j expressed in Ψ_i is given by:

$$\dot{R}_i^j = R_i^j \cdot \tilde{\omega}_i^{i,j} = R_i^j \cdot \begin{pmatrix} 0 & -\omega_x & \omega_y \\ \omega_x & 0 & -\omega_z \\ -\omega_y & \omega_z & 0 \end{pmatrix} \tag{6}$$

The subscripts i and j in $\tilde{\omega}_i^{i,j}$ are removed for clarity. In quaternion notation this is:

$$\dot{q}_i^j = q_i^j \otimes \tfrac{1}{2} q_{\omega_i^{i,j}} = \tfrac{1}{2} \widehat{q_{\omega_i^{i,j}}} \cdot q_i^j \tag{7}$$

As the scalar part of $q_{\omega_i^{i,j}}$ is zero, the solution to equation (7) is:

$$\begin{aligned} q_i^j(t) &= e^{\frac{1}{2}\widehat{q_{\omega_i^{i,j}}}\cdot t} q_i^j(0) \\ &= \left(I \cdot \cos(\tfrac{1}{2}\left\|\vec{\omega}_i^{\,i,j}\right\| t) + \sin(\tfrac{1}{2}\left\|\vec{\omega}_i^{\,i,j}\right\| t) \cdot \frac{\widehat{q_{\omega_i^{i,j}}}}{\left\|\omega_i^{i,j}\right\|} \right) q_i^j(0) \end{aligned} \tag{8}$$

or:

$$q_i^j(t) = q_i^j(0) \otimes \begin{pmatrix} \cos(\frac{1}{2}\left\|\vec{\omega}_i^{i,j}\right\| t) \\ \sin(\frac{1}{2}\left\|\vec{\omega}_i^{i,j}\right\| t) \dfrac{\vec{\omega}_i^{i,j}}{\left\|\vec{\omega}_i^{i,j}\right\|} \end{pmatrix} \tag{9}$$

A more detailed treatment is given in [3]

2.3 Kalman Filters

The time update for a Kalman filter without control inputs is given by:

$$\begin{aligned} \hat{x}_k^- &= \Theta_{t_k, t_{k-1}} \hat{x}_{k-1}^- \\ P_k^- &= \Theta_{t_k, t_{k-1}} P_{k-1} \Theta_{t_k, t_{k-1}} + Q\left(t_k, t_{k-1}\right) \\ Q\left(t_k, t_{k-1}\right) &= \int_{t_{k-1}}^{t_k} \Theta_{t_k, s} \cdot Q(s) \cdot \Theta_{t_k, s}^T ds \\ &\approx \Theta_{t_k, t_{k-1}} \cdot Q(t_{k-1}) \cdot \Theta_{t_k, t_{k-1}} \cdot \left(t_k - t_{k-1}\right) \\ Q(t) &= \operatorname{cov}\left(w(t), w(\tau)\right) \end{aligned} \tag{10}$$

in which $\hat{x}_k^-$ is the current a-priori estimate of the state x_k, Θ is the state transition matrix that projects a state into the future, and $w(t)$ is the white noise process, which models not modeled changes in the state. When an observation is done, the a-posteriori estimate of the state can be calculated from the estimate of the observation:

$$\begin{aligned} y_k^- &= H\hat{x}_k^- \\ \hat{x}_k &= \hat{x}_k^- + K\left(y_k - y_k^-\right) \\ P_k &= \left(I - KH\right)P_k^- \\ K &= P_k^- H\left(HP_k^- H^T + R\right)^{-1} \\ R &= \operatorname{cov}\left(v(t), v(\tau)\right) \end{aligned} \tag{11}$$

in which y_k^- is the predicted observation, y_k is the real observation, H is the output matrix, K is the Kalman gain, and $v(t)$ white noise that models measurement noise in y_k .

In the sequel we use X for the estimate of the real states (a-posteriori and a-priori) and dX for the error states of the Kalman filter. The state-vector X contains orientation, angular velocity, position, linear velocity, linear acceleration, current gyro and current accelerometer bias:

$$X = \left(q_b^n, \vec{\omega}_b^{b,n}, \vec{p}_b^n, \vec{v}_b^n, \vec{a}_b^b, \vec{b}_{gyro}^b, \vec{b}_{acc}^b\right)^T \tag{12}$$

To simplify the system we use two Kalman filters. One for the orientation, with state:

$$dX_{orient} = \left(dq_b^n, d\vec{b}_{gyro}^b\right)^T \tag{13}$$

being the error in orientation and the drift in the gyroscopes, and one for the position containing the error in position, linear velocity and accelerometer drift:

$$dX_{pos} = \left(d\vec{p}_b^n, d\vec{v}_b^n, d\vec{b}_{acc}^b\right)^T \tag{14}$$

The accelerometer bias and its error state are expressed in body coordinates, and therefore the position filter depends on the orientation. So in this set-up it is essential to have an accurate orientation estimate, as its error will propagate with t^2 into the position.

2.4 Time Update

Each time a measurement is obtained from the inertial sensors, the estimate of the actual state X is updated, using eq.(9) and:

$$\begin{aligned} \vec{v}_b^n(\Delta t) &= \vec{v}_b^n(0) + \left(R_b^n(0)\cdot \vec{a}_b^b(0) - \vec{g}^n\right)\Delta t \\ \vec{p}_b^n(\Delta t) &= \vec{p}_b^n(0) + \tfrac{1}{2}\left(\vec{v}_b^n(0) + \vec{v}_b^n(\Delta t)\right)\Delta t \end{aligned} \tag{15}$$

Note, that we do not take the rotational speed into account, which results in an error in $\vec{a}_b^b(t)$ of about 3 10^{-3} m/s^2 when rotating at 200 deg/s and Δt =1/100s, in a direction perpendicular to the gravity vector.

The state update for the position Kalman filter is based on the formulas (10) and (15), and we neglect rotations errors in R_b^n. The change in rotation should be small due to the high update rate, and the estimate in orientation is rather accurate. The state transition is now given by:

$$
\begin{gathered}
dX_{pos}(t_k) = \Phi dX_{pos}(t_{k-1}) \quad \Rightarrow \\
\begin{pmatrix} d\vec{p}_b^n \\ d\vec{v}_b^n \\ d\vec{b}_{acc}^b \end{pmatrix}_{(t_k)} = \begin{pmatrix} I & I \cdot \Delta t & \frac{1}{2} R_b^n(t_k) \cdot \Delta t^2 \\ 0 & I & R_b^n(t_k) \cdot \Delta t \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} d\vec{p}_b^n \\ d\vec{v}_b^n \\ d\vec{b}_{acc}^b \end{pmatrix}_{(t_{k-1})}
\end{gathered}
\tag{16}
$$

The state update for the orientation Kalman filter is more complicated, because we use the orientation difference in quaternion notation:

$$
q_b^n = q_b^{n-} \otimes dq_b^n \quad \Leftrightarrow dq_b^n = q_b^{n-*} \otimes q_b^n \tag{17}
$$

in which q_b^{n-} is the estimated orientation by integration using eq. (9) and q_b^n is the real state. Using:

$$
\begin{gathered}
\dot{q}_b^{n-*} \otimes q_b^{n-} = 0 \Rightarrow \dot{q}_b^{n-*} \otimes q_b^{n-} + q_b^{n-*} \otimes \dot{q}_b^{n-} = 0 \Rightarrow \\
\dot{q}_b^{n-*} = -q_b^{n-*} \otimes \dot{q}_b^{n-} \otimes q_b^{n-*}
\end{gathered}
\tag{18}
$$

the time derivate of the estimate of dq_b^n becomes:

$$
\begin{aligned}
\dot{dq}_b^n &= \dot{q}_b^{n-*} \otimes q_b^n + q_b^{n-*} \otimes \dot{q}_b^n \\
\dot{dq}_b^n &= -q_b^{n-*} \otimes \dot{q}_b^{n-} \otimes q_b^{n-*} \otimes q_b^n + \tfrac{1}{2} q_b^{n-*} \otimes \left(q_b^n \otimes q_{\omega_b^{b,n}} \right) \\
\dot{dq}_b^n &= -q_b^{n-*} \otimes \left(\tfrac{1}{2} q_b^{n-} \otimes q_{\omega_b^{b,n-}} \right) \otimes dq_b^n + \tfrac{1}{2} dq_b^n \otimes q_{\omega_b^{b,n}} \\
\\
\dot{dq}_b^n &= -\tfrac{1}{2} q_{\omega_b^{b,n-}} \otimes dq_b^n + \tfrac{1}{2} dq_b^n \otimes q_{\omega_b^{b,n}} \\
\dot{dq}_b^n &= -\tfrac{1}{2} \widetilde{q_{\omega_b^{b,n-}}} \cdot dq_b^n + \tfrac{1}{2} \widehat{q_{\omega_b^{b,n}}} \cdot dq_b^n \\
\dot{dq}_b^n &= \tfrac{1}{2} \left(\widehat{q_{\omega_b^{b,n}}} - \widetilde{q_{\omega_b^{b,n-}}} \right) \cdot dq_b^n
\end{aligned}
\tag{19}
$$

We can calculate the true $\omega_b^{b,n}$ from its estimate $\omega_b^{b,n-}$ with:

$$\omega_b^{b,n} = \omega_b^{b,n-} - d\vec{b}_{gyro}^{b} \tag{20}$$

Then equation (19) becomes after some elaboration:

$$\dot{dq}_b^n = \tfrac{1}{2}\left(\widehat{q_\omega} - \widetilde{q_{\omega^-}}\right)\cdot dq = \tfrac{1}{2}\begin{pmatrix} -d\vec{b}\cdot d\vec{q} \\ q_0 \cdot d\vec{b} - d\vec{b}\times d\vec{q} + 2\vec{w}\times d\vec{q} \end{pmatrix} \tag{21}$$

Now this can be linearized around the state dX_{orient} at time t = 0 (the previous estimate). In the indirect filter setup, the error states are reset after every observation update. This means that $d\vec{b}$ will be assumed constant and 0, $d\vec{q}$ will be small, and dq_0 will be approximately 1. Our linearization of the time derivative becomes:

$$\dot{\begin{pmatrix} dq_0 \\ d\vec{q} \\ d\vec{b} \end{pmatrix}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \widetilde{w} & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} dq_0 \\ d\vec{q} \\ d\vec{b} \end{pmatrix} \tag{22}$$

In the case $\tilde{w}$ is assumed constant (zero order hold), we can find the state transition to be:

$$\begin{gathered} dX_{orient}\left(t_k\right) = \Phi dX_{orient}\left(t_{k-1}\right) \Leftrightarrow \\ \begin{pmatrix} dq_0 \\ d\vec{q} \\ d\vec{b} \end{pmatrix}_{\left(t_k\right)} = e^{\begin{pmatrix} 0 & 0 & 0 \\ 0 & \widetilde{w(t_k)}\cdot & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix}\Delta t}\begin{pmatrix} dq_0 \\ d\vec{q} \\ d\vec{b} \end{pmatrix}_{\left(t_{k-1}\right)} \end{gathered} \tag{23}$$

Of course, after this update, dq should be normalized to unity again. The algebraic solution to this exponent of a matrix was found using a mathematical software package.

2.5 Observation Update

When an inclinometer measurement becomes available, it is converted to quaternion notation. Now it becomes easy to determine the observation estimate of the error in orientation:

$$dq_{b,obs\_est}^{n} = q_b^{n-*} \otimes q_{b,inclino}^{n} \tag{24}$$

This observation estimate replaces y_k in eq. (11)

Because both estimates are in quaternions, the output matrix H becomes:

$$y_k^- = H\hat{x}_k^- = \begin{pmatrix} I & 0 \end{pmatrix}\begin{pmatrix} dq \\ d\vec{b} \end{pmatrix} \tag{25}$$

A problem now is the Kalman measurement noise $\vec{v}$, which is dependent on both the measurement and the estimated real state. We only take into account the measurement error, as it will be larger than the estimate error, which is the error in the integration of the gyro measurements.

To find the covariance matrix *R*, we determined the linearised matrix that relates the deviation in dq^n_{b,obs_est} to the inclinometer measurement deviation $\delta\vec{\theta}$:

$$\begin{aligned}
&dq^n_{b,obs}\left(\vec{\theta}_{inclino}+\delta\vec{\theta}\right) \approx dq^n_{b,obs}\left(\vec{\theta}_{inclino}\right)+\frac{\partial dq^n_{b,obs}\left(\vec{\theta}_{inclino}\right)}{\partial\vec{\theta}}\delta\vec{\theta} \\
&\delta dq^n_{b,obs} = \frac{\partial dq^n_{b,obs}\left(\vec{\theta}_{inclino}\right)}{\partial\vec{\theta}}\delta\vec{\theta} = G\cdot\delta\vec{\theta} \\
&\vec{v} = G\cdot\vec{w} \\
&R = \mathrm{cov}\left(\vec{v}(t),\vec{v}(\tau)\right) = G\cdot\mathrm{cov}\left(\vec{u}(t),\vec{u}(\tau)\right)\cdot G^T
\end{aligned} \tag{26}$$

in which $\vec{u}(t)$ is the presumed white noise in the inclinometer measurement.
In the position update, the estimated position error is:

$$d\vec{p}^{\,n}_{b,obs\_est} = \vec{p}^{\,n-}_b - \vec{p}^{\,n}_{b,observation} \tag{27}$$

and using y_k for $d\vec{p}^{\,n}_{b,obs\_est}$ we get for formula (11):

$$y_k^- = H\hat{x}_k^- = \begin{pmatrix} I & 0 & 0 \end{pmatrix}\begin{pmatrix} d\vec{p} \\ d\vec{v} \\ d\vec{b} \end{pmatrix} \tag{28}$$

Matrix *G* that relates the error in $d\vec{p}^{\,n}_{b,obs\_est}$ to the error in $\vec{p}^{\,n}_{b,observation}$ is the negation of the identity matrix.

2.6 Coping with Lag

The inclinometer output is delayed with about 0.375s (or 6 samples @ 16Hz). When we ignore this delay, the orientation Kalman filter will assume an error in orientation and will adjust the current error and bias estimate. After the rotation, the filter needs some time to recover. A bigger problem is that the position filter uses the orientation, and this delay will hence introduce a non-existent acceleration.

In our method we store all the observations of the sensors, as well as the Kalman states and matrixes, at every step and keep a history of 30 steps. When an inclinometer measurement finally arrives, we step back to the position in time of that measurement, and do the filtering in the Kalman state that belongs to that point in time. From here on all the other measurements (such as gyro, camera, GPS) are processed again up to the current time. In this way the best estimate at the current time is achieved.

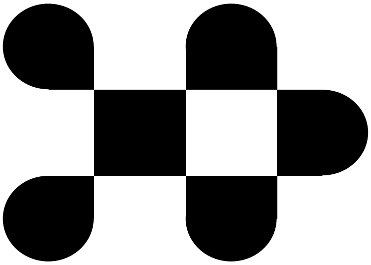


Fig. 6. Pattern with 6 saddle points ordered in two 3-point rows.

3 Camera Positioning

3.1 Finding Markers

For camera positioning we use a pattern like in Figur 6. This pattern consists of 6 saddle points, which are easy to detect using the determinant of the Hessian. The filter output $g(\vec{p})$ of a 2D image $f(\vec{p})$ is given by:

$$g(\vec{p}) = -\det[H[f(\vec{p})]] = -\begin{Vmatrix} \partial_{xx} & \partial_{xy} \\ \partial_{yx} & \partial_{yy} \end{Vmatrix} f(\vec{p}) \tag{29}$$

The derivatives ∂ are implemented using the derivative of a Gaussian with $\sigma = 2.0$. To find the saddle points, we threshold the output with value:

$$th_{\text{detector}} = \frac{1}{2} \max_{\vec{p}} \left[g(\vec{p}) \right] \tag{30}$$

and apply a peak detection filter in a 3 x 3 neighborhood *S*. This leaves us with a set of saddle points:

$$sp = \left\{ \vec{p} \mid g(\vec{p}) = \max_{\vec{q} \in S} [g(\vec{p} + \vec{q})] \wedge g(\vec{p}) > th_{\text{detector}} \right. \tag{31}$$

To obtain sub pixel accuracy at each point a paraboloid is fit. This is possible as the filter has a parabolic shape at a saddle point. The true saddle point is located at an offset with respect to a point in sp. The model is:

$$g(x,y) = d + a\left((x-x_m)^2 + (y-y_m)^2\right) \text{ or}$$
$$\mathbf{y} = A\mathbf{x} = \begin{pmatrix} d + a\left(x_m^2 + y_m^2\right) \\ -2ax_m \\ -2ay_m \\ a \end{pmatrix}^T \begin{pmatrix} 1 \\ x \\ y \\ x^2 + y^2 \end{pmatrix} \tag{32}$$

With the standard least squares method applied to the 3x3 neighborhood of each saddle point we find for *A*:

$$A = \left(\mathbf{x}^T \mathbf{x}\right)^{-1} \mathbf{x}^T \mathbf{y} \tag{33}$$

Note, that if we translate the coordinate system such that the origin is the position of the estimated saddle point, we can pre-calculate $\left(\mathbf{x}^T \mathbf{x}\right)^{-1} \mathbf{x}^T$ to speed up processing.

From *A* the sub-pixel position of the saddle point can be calculated. For convenience we define:

$\vec{p} \in sp$

$\vec{p}$ with subpixel accuracy

$g(\vec{p})$ is still the original filter output

To find all markers (as in Figure 6) we match groups of 6 points. When only marker saddle points are detected and the markers are not too close, we can find the pattern by looking at the 6 nearest neighbors at every point. We use the 10 nearest neighbors to be robust against non-marker saddle points. Observing Figure 6 one can see that the pattern consists of two 3-point lines. These lines will stay perfectly linear under all circumstances when viewed with a calibrated camera.

Consequently, for every point we find all possible 3-points long lines, with that point in the middle. The distance between the point and the line between the two other points should not be greater than some threshold ε:

$$\begin{aligned} L_{p_2} = \{\vec{p}_1, \vec{p}_2, \vec{p}_3 \mid \min_{s=[0,1]}\left[\left\|\vec{p}_2 - \left(\vec{p}_1 + s(\vec{p}_3 - \vec{p}_1)\right)\right\|\right] < \varepsilon, \\ p_i \in sp\} \end{aligned} \tag{34}$$

To find the pattern we consider each pair of two lines that do not have points in common. We find the set of candidate markers that consist of two lines with 6 unique points:

$$C = \{L_1 \in L_{\vec{p}_1}, L_2 \in L_{\vec{p}_2} \mid L_1 \cap L_2 = \varnothing \tag{35}$$

3.2 Determining the Pose from a Markers Feature Points

For each candidate marker a fitness value is calculated. First the position of the camera is estimated using the 6 points and part of the Zhang calibration algorithm [7] as described below. Then the fitness value is calculated as the mean square error of the back-projected marker points. We use a calibrated camera, so we can correct for lens distortion, skewing, scaling, and offset of the center point in the image. For each point we calculate the position on a fictive plane at distance 1 in camera coordinates:

$$p^C = \begin{pmatrix} s_x & 0 & x_{offset} \\ s_{sk} & s_y & y_{offset} \end{pmatrix} \begin{pmatrix} p^i \\ 1 \end{pmatrix} = AP^i \tag{36}$$

An estimate of the position of the camera can be found using the relation between the 6 points in camera coordinates, and the 6 points of the model (a 3D homography):

$$sP^C = s\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \mathbf{T} \end{pmatrix} \begin{pmatrix} p_x^P \\ p_y^P \\ p_z^P \\ 1 \end{pmatrix} \tag{37}$$

in which P^C is the homogeneous image-plane position, p_i^P are the components of the homogeneous position in the pattern-frame (the model), and $\mathbf{R}$ and $\mathbf{T}$ are the rotation and translation from the pattern frame, to the camera frame. This formula can be simplified by the fact that we defined $p_z^P = 0$. When we remove the z-coordinate from formula (37) we obtain:

$$sP^C = s\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} r_1 & r_2 & \mathbf{T} \end{pmatrix} \begin{pmatrix} p_x^P \\ p_y^P \\ 1 \end{pmatrix} = \mathbf{H}P^p \tag{38}$$

This can be rewritten as:

$$s\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} H_1 \\ H_2 \\ H_3 \end{pmatrix} P^p \tag{39}$$

This set of equations can be reordered to:

$$\begin{aligned} & s = H_3 P^P \\ & H_1 P^P - uH_3 P^P = 0 \\ & H_2 P^P - vH_3 P^P = 0, \text{ or} \end{aligned} \tag{40}$$

$$\begin{pmatrix} P^{PT} & 0 & -uP^{PT} \\ & P^{PT} & -vP^{PT} \end{pmatrix} \begin{pmatrix} H_1^T \\ H_2^T \\ H_3^T \end{pmatrix} = L\mathbf{x} = 0$$

in which P^{PT} is the transpose of P^P. The matrix L can be extended downwards for all 6 points, and the solution for **x** is the right singular vector of L, associated with the smallest singular value. The get L numerically well conditioned, data normalization can be used. From **x** we can reconstruct r_1, r_2 and **T**. To complete the rotation matrix **R** we can use:

$$r_3 = r_1 \times r_2 \tag{41}$$

Because R is estimated, the matrix is not orthonormal. We can find the best orthonormal matrix using singular value decomposition:

$$R_{\text{estimate}} = UDV^T \Rightarrow R_{\text{orthonormal}} = UIV^T \tag{42}$$

We found that the resulting camera pose is not very stable, so we applied a Levenberg-Marqardt algorithm that optimizes the pose by minimizing our fitness criterion. The final step is to threshold the fitness value to get rid of candidate markers that do not match well enough. Note that the pose(s) found are in pattern coordinates, for each pattern. This means that we still have to perform a coordinate transformation from pattern coordinates to world coordinates.

3.3 Converting Pattern to World Coordinates

There are two difficulties in converting the position of the camera expressed in the pattern position frame to one expressed in the world position frame. One is, that we might not know the pose of the pattern in the world, and the other is, that the coordinate system of the (pinhole) camera is rotated with respect to the coordinate system of the inclinometer (body frame).

We can find the pattern's orientation expressed in the world frame by measuring three vectors. Each vector can be measured in pattern coordinates and in world coordinates. If the directions of the vectors span all three dimensions, we can find the coordinate transformation matrix:

$$\begin{gathered} \begin{pmatrix} v_1^W & v_1^W & v_1^W \end{pmatrix} = R_P^W \begin{pmatrix} v_1^P & v_1^P & v_1^P \end{pmatrix} \\ R_P^W = \begin{pmatrix} v_1^P & v_1^P & v_1^P \end{pmatrix}^{-1} \begin{pmatrix} v_1^W & v_1^W & v_1^W \end{pmatrix} \end{gathered} \tag{43}$$

These vectors can be found by moving the camera in these three directions, but we do not have an absolute positioning system for measuring the world coordinates. Another method is to turn the device around each of the vectors. The rotation axes can be found by looking at the change in orientation in both coordinate systems. The world origin can be chosen freely, and for now we use the startup position as the world origin. Using that assumption, we can determine the position of the pattern in the world.

This translation is combined with the rotation matrix in a homogeneous matrix $H = \begin{pmatrix} R & T \end{pmatrix}$. We define:

H_A^B The transformation that brings frame Ψ_B to frame Ψ_A expressed in Ψ_B using homogenous coordinates

The pose of the camera – or body frame - is given by:

$$H_b^W = H_P^W H_b^P \tag{44}$$

Unfortunately, we only have the pose H_{CP}^P, of which Ψ_{CP} is the frame at the focal point of the camera, and orientated with the z-axis in the direction of the optical axis. This means that we have to find the unknown relation between Ψ_{CP} and Ψ_b, before we can determine the camera pose by:

$$H_b^W = H_P^W H_{CP}^P H_b^{CP} \tag{45}$$

The transformation H_b^{CP} is constant, and therefore we can just measure the pose of the camera expressed in world coordinates as well as the pose expressed in pattern coordinates. The required transformation is then given by:

$$H_b^{CP} = H_P^{CP} H_W^P H_b^W \tag{46}$$

measured at any one instant.

4 Results and Conclusion

Due to our set-up with a sensor cube with inertia sensors and the TCM2 mounted on top of the camera, the axes of the sensors will not be aligned either. After calibration however, we found that the misalignment in the sensor cube is minimal. Results showed that the orientation Kalman filter converged quickly. The stability of the gyros provided very accurate orientation values during motion, but due to the resolution of the inclinometer (0.2 degrees) the overall accuracy cannot be higher then 0.2 degrees, even when the 0.5 degrees accuracy in the heading has no systematic error.

To verify the camera positioning we tracked an A4 pattern (Figure 6) perpendicular to the optical axis, with a distance from 30-190 cm at 5 cm intervals. The pattern was kept roughly in the middle of the picture. The calibration was done, using the Zhang method. We measured the mean error and RMS value, σ, of the distance and the roll angle in pattern coordinates, by processing 100 images, real-time at 10 fps, at each distance. This was done on an AMD Athlon XP1700+, and the algorithm used about 40 ms per image. Some results are shown in Figure 7.

Looking at the left figure, one can observe that we have 1 cm accuracy up to a distance of about 1.2m. The feature distance is then 11 pixels. This means that if the markers are further away, the marker should increase in size, or multiple markers should be used. In the right figure one can see that the error in the angle is large at distances greater than 70 cm. This has also a negative impact on the accuracy in x and y, so it will be wise to use the orientation output of the Kalman filter to only optimize the position of the camera during the Levenberg-Marquardt step of the algorithm.

Experiments still have to be done to find accuracy measures for general camera positions, and for the position/orientation Kalman filter output during motion.

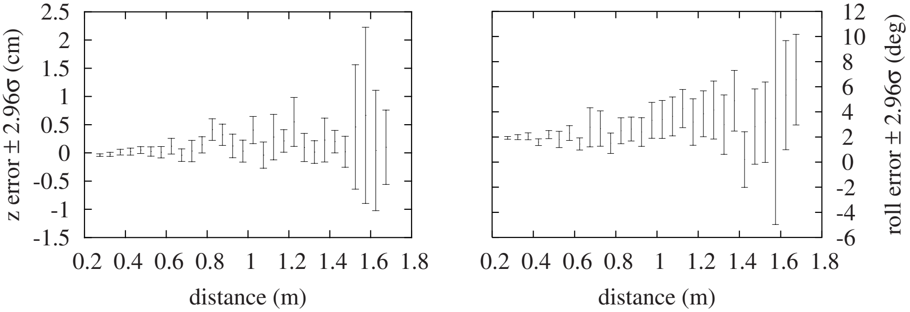


Fig. 7. Experiment in which the camera was moved perpendicular to the pattern. *Left*: error in pattern's z axis with 2.96σ error bars. *Right*: error in roll angle with 2.96 σ error bars. Note that this σ is the RMS of the 100 values (see text).

References

1. Bakker, J.D., Mouw, E., Pouwelse, J., The LART Pages. Delft University of Technology, Faculty of Information Technology and Systems (2000). Available at http://www.lart.tudelft.nl
2. Brookner, E., Tracking and Kalman Filtering Made Easy, John Wiley & Sons Inc. (1998)
3. Caarls, J., Geometric Algebra with Quaternions, Technical Report (2003) http://www.ph.tn.tudelft.nl/Publications/phreports
4. Jonker, P.P., Caarls, J., Eijk, R. van, Peddemors, A., Heer, J. de, Salden, A., Määttä, P., Haataja, V. Augmented Reality Implemented on a Mobile Context Aware Application Framework, submitted to IEEE Computer (2003)
5. Jonker, P.P., Persa, S., Caarls, J., Jong, F. de, Lagendijk, I. Philosophies and Technologies for Ambient Aware Devices in Wearable Computing Grids, Computer Communications Journal, Volume 26, Issue 11, 1 July 2003, Pages 1145–1158. http://www.sciencedirect.com/science/journal/01403664
6. Pasman, W., Schaaf, A. van der, Lagendijk, R. L., & Jansen, F. W. (1999). Accurate overlaying for mobile augmented reality. Computers & Graphics, 23 (6), 875–881. http://www.cg.its/tudelft.nl/~wouter
7. Zhang, Zhengyou. A Flexible New Technique for Camera Calibration. http://www.research.microsoft.com/~zhang/calib

Context Awareness of Everyday Objects in a Household

Elena Vildjiounaite, Esko-Juhani Malm, Jouni Kaartinen, and Petteri Alahuhta

Technical Research Center of Finland, Kaitovayla 1,
P.O.Box 1100, 90571, Oulu, Finland
{Elena.Vildjiounaite, Esko-Juhani.Malm,
Jouni.Kaartinen, Petteri.Alahuhta}@vtt.fi

Abstract. This work studies the context awareness of everyday objects augmented with sensing, communicational and computational capabilities, and presents a prototype context-aware system built for household applications. The system's interaction capabilities help to deal with the challenges of performing context detection with very limited computing resources. These capabilities and the system as such were in general approved by the users. The proposed domain-specific context model supports individual and collective work by objects, so that each collective can fulfil its task independently and needs to communicate with the mobile device only to receive a task and present the results.

1 Introduction

Smart homes (homes which provide services supporting people's everyday activities) have become an active area of research in recent years [1],[2]. Most of the work in the area is concerned with everyday appliances for which computing power is not a critical problem, like televisions, washing machines, lamps or fridges. Accordingly, the methods of context recognition include visual and audio data processing. Our everyday activities, however, involve a lot of interactions with such things as clothes, spectacles, schoolbooks, food packages and so on. Thus there is a need to develop ways of supporting interactions with all kinds of everyday objects. Inexpensive, small-sized hardware enabling computation to be embedded into literally any artefact should be achievable in five to ten years [3],[4], but its computational capabilities will be still very limited.

We suggest here that the above-mentioned smart environments should be complemented by embedding simple sensing, computational and communication capabilities into all artefacts. For example, instead of relying on a vision-based system to check that our children have put all the necessary books into their schoolbags, it would be possible to make the schoolbooks themselves do the job, thus reducing the amount of work for the vision-based system, or to use this as redundant information in order to increase the certainty of context detection. This approach also helps to provide services for mobile users, for example, in the forest (during a sports trip or picnic).

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 177–191, 2003.

The idea of context recognition by collecting information from several simple sensing devices was used in the Mediacup [5] project, for example, where coffee cups were augmented with sensing modules for context detection (so that if somebody was drinking from the cup, playing with it or carrying it around, the context would be derived from temperature and accelerometer data). This would allow the system to detect if a meeting was taking place in a room where many cups had been gathered. It was recently suggested in the Smart-Its project [6] that proactive instructions for furniture assembly [7] should be given by processing data from sensing modules attached to the different parts of the furniture.

The general challenges involved in deploying ubiquitous computing systems are addressed in [3],[8],[9]. The main challenges in the task of making any artefact smart are hardware size and price, a long-lasting energy supply and the need to perform context detection with limited computing resources. Thus it does not seem feasible at first glance to extract anything more than local context data from smart objects (such as movement type and temperature in the case of Mediacups). In both the Mediacup system and the furniture assembly application the sensing modules detect only their own context and send the data to a central computer, which makes conclusions about higher level contexts.

On the other hand, the task of making all the objects around us smart also entails the problem of how to deal with a large quantity of objects, especially in cases where data from many objects are needed at the same moment. This need conflicts with the approach of demanding only local context data from the objects, especially for mobile users. Not all mobile devices are powerful enough to be able to receive and process the local contexts of all smart objects of interest to the user.

We suggest in this work that smart objects can be organised into temporal collectives according to the current task, so that each member of a collective is responsible not only for its own context recognition but also for conclusions about the joint context of the collective. In such a system a central computer is needed mainly as a means of communication between the user and the smart objects and does not need to be powerful, which is important for mobile users. On the other hand, our implementation of collective work of smart objects does not require significant increase in program size on the side of smart objects themselves. We propose a context model for household applications which supports both individual and collective work by objects.

2 System Prototype

2.1 How to Add Smartness to Everyday Objects

In order to make everyday objects smart and able to support the user's everyday activities, it is necessary first to provide the user with the means of interaction with objects. Since objects don't have any screen or keyboard, the main option is wireless communication with a device that has a suitable user interface and the ability to decode wireless messages (we will call this device a central node in this paper, because the user is at the centre of activities). Another option is a tangible user interface. Since the objects' computational abilities are very limited, we have chosen shaking as the most straightforward option for physical interaction with them.

Second, it is necessary to provide the objects with the context model and algorithms of context recognition.

Third, it is necessary to provide the central node with the ability to work as a communication medium between the user and the objects. This means that the central node has to share the same context representation.

2.2 Hardware

Everyday objects become smart after attaching Smart-Its boards to them (see Fig. 1). These boards, which are intended as generic hardware, were developed for research purposes in the Smart-Its project at TecO, University of Karlsruhe (Germany) [6].

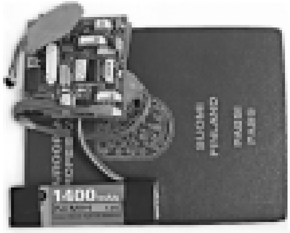


Fig. 1. A smart object

Each Smart-It consists of a sensor board and a core board, each containing a PIC microcontroller. The core board is responsible for RF (radio) communication and the sensor board for sensing and for running the application. The sensor board contains a two-directional accelerometer, light, temperature and pressure sensors, a microphone, three LEDs (Light Emitting Diodes) and a piezo loudspeaker. The two boards are about 4 x 5 cm in size and are mounted on top of each other. The same boards with a smaller communication range serve as beacons for providing location information.

One very important feature of the broadcast communication protocol, which was also developed at TecO, University of Karlsruhe, is that synchronization among the devices divides the operation of the core boards into RF communication and computation phases (the latter also including exchange of data with the sensor board). Synchronization is established in a peer-to-peer manner, i.e. no central node is needed at any time. The broadcast and synchronization capabilities of the protocol concur to support the fast ad-hoc addition of upcoming devices.

The system prototype also includes a desktop computer as a medium of communication between the user and the objects in a home environment and a Compaq iPAQ Pocket PC as a communication medium for mobile users.

3 Household Applications

When we start to analyse what kind of support we would like to have from computing systems in order to make our everyday interactions with personal belongings easier, we find the following problems, for example:

- We would like help in searching for lost things.
- We would like reminders about food products for which the “use-by” date expires soon.
- We would like to check if all the ingredients for a chosen recipe are available at home, and if they are low-fat or gluten-free products, for example.
- We would like to check if any parts of a business suit are waiting in the bathroom to be washed.
- We would like to be warned if a young child starts to play with a wallet or documents.
- We would like to be warned if a red T-shirt with a "wash separately" label is put into the washing machine together with a white blouse.
- We would like reminders to collect all the necessary things for a journey and to bring them back home.

Some of these problems involve interactions with the objects as individuals (e.g. when we are looking for a particular book), while many of the other problems involve interactions with a certain temporal set of objects. When we put on a suit, we create a temporal set of clothes which are all clean and match each other in colour and style. When we put clothes into a washing machine, we create a temporal set of clothes of similar colours and materials, so that the addition of an item of an incompatible colour can cause disaster. The ingredients needed for a certain recipe also create a set, and so on. When leaving home for work, on a journey or for a sports event, we create a set of things which we need for the activity concerned, and which should all leave home together and probably also come back home together. Some of these sets are short-term collectives, as the user just needs to be informed if all the members are present at the moment or not, while others should be able to function during a whole journey, for instance.

In all these cases we need to check as soon as possible if all the members of a temporal set satisfy the task requirements (e.g. all moving together in the case of journey), and help from a computer system would undoubtedly be appreciated in many cases in which information on numerous objects is needed simultaneously.

4 Context Model

The context model for smart objects is very similar to that proposed in [10], but it takes into account the specifics of application domain and the fact that each augmented object is a combination of two parts interacting with each other: a physical object itself and a tiny hardware attached to it. Each context is described using the following properties: context type, context value (symbolic or numerical, depending on the context type), source, confidence and timestamp. The last three properties are optional. For example, the source of the object's movement context is usually the object's accelerometer (which can be understood implicitly), while the source of the information about the object's colour is a memory address (which it is necessary to indicate). Context types form a tree. The context types for smart objects are presented in Figure 2.

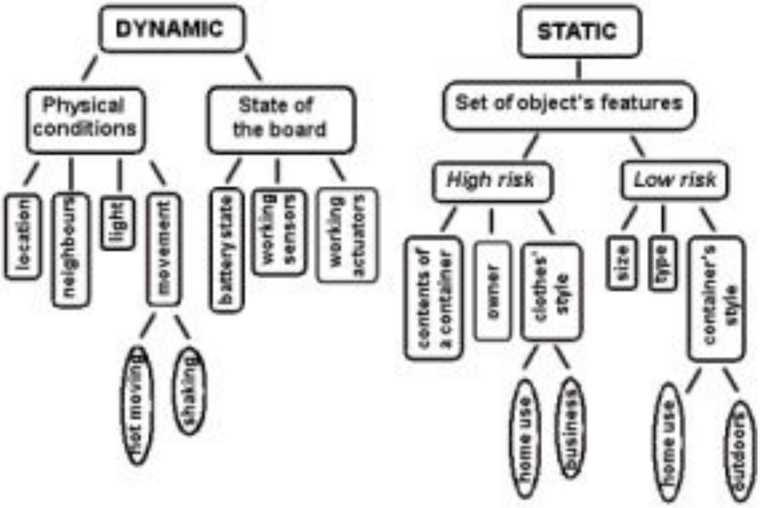


Fig. 2. Context Types

The object's context types are, first, factors which we call physical conditions: movement type, number of other objects which it can hear, light, pressure, temperature and so on. All these factors, which are highly dynamic, normally represent the local context of the object and have an internal source. However, there is one block in the Figure 2 ("neighbours") which has an external source, as it represents the context data of the neighbouring objects. In our applications objects were using this information in their collective work for comparison with their own context data, but it would also be possible for them to use the context data of neighbours to complement or verify their own context data. For example, if some of the objects don't have a temperature sensor, there is a probability that their temperature will be close to that of their neighbours. On the other hand, if the temperature data of one of the objects is quite different from that of its neighbours, there is a probability that the object's temperature sensor is malfunctioning.

Second, an object's context types include the internal state of its board, including battery state and the sensors and actuators working on the board. The battery state is obviously a dynamic factor, while the sensors and actuators included seem at first glance to be static data. However, since these can be broken or switched off for different reasons (e.g. for reasons of privacy in the case of microphones or video cameras, or to save the batteries), we decided to consider this information dynamic as well. The results of our user study [11] support our opinion in that most users would like to have control over switching video cameras and microphones on and off.

Third, factors which we call static and which refer to the internal characteristics of the physical objects. We argue that it is convenient to include features of the object among the context types, because a feature is very often as valid task parameter as physical condition. If we need to find all red clothes waiting in the bathroom to be washed, location and colour are equally important task parameters. We may be also interested in finding certain products in a fridge, and in this case temperature and use-by date become important task parameters.

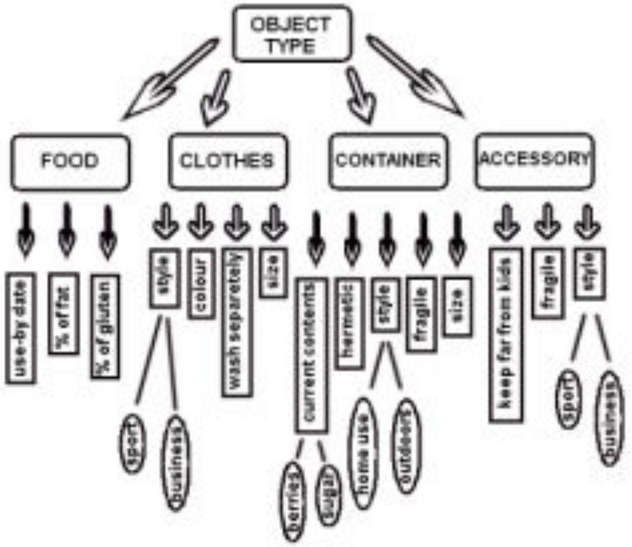


Fig. 3. Different context types for certain object types

The characteristics of an object are stored in its memory and must be different for different types of object. The use-by date is a very important characteristic for a food product, for example, but has a vague meaning for clothes. Static factors may be divided into two groups: high risk (those which are likely to change during the object's life time), and low risk (those which one does not expect to change during the object's life time).

The most important static characteristic of an object is its type: food, clothes, container (examples include bowl, flask, coffee tin), accessory (e.g. keys, glasses, passport) and so on. This is of course a low risk feature. Each type has its own important features, and the same feature can be high risk or low risk depending on the type. For example, clothes change their style from "*business*" to "*home use*" when they are out of fashion, while vases are almost never used in outdoor trips, so that their style remains as "*home use*" throughout their lives.

Some object types and their characteristics (context types) are presented in Figure 3, where the highest risk feature is always in the same memory location for all objects (leftmost feature in Figure 3). For a container, its contents constitute the highest risk feature, while for food the use-by date can be prolonged by storing the product in a deep-freeze. Similarly, a passport that has expired should not be kept away from children any more.

The reason for the division of static object characteristics into high and low risk is to help resolve misunderstandings. Each object is able to "introduce itself" (in response to either a request from the central node or shaking), and then it always sends the highest risk feature together with its type, location and other data, so that the user can see where the problem is.

5 Task Model

The context model presented above is intended to represent the context of each object. The set of tasks which objects have to perform make up the task model. Each task is characterised by four parameters:

1. task ID (identification number), which tells the objects which context types they need to determine
2. list of IDs of group members (or the only ID for individual work)
3. reference context value (with which the objects need to compare their local context data)
4. relater (which tells the objects how to compare their own context with the reference context value)

The last two parameters are important only for certain tasks. For the task of finding objects by location, for example, there could be two meaningful values for the relater: "*in location*" or "*not in location*". For the task of finding gluten free products there is only one meaningful value for the relater, as also for the reference context value ("*is equal to zero*"), so that both can be omitted. For the use-by date task, however, the reference context value can range from one day to many months, and the relater can be "*is equal*", "*more than*", or "*less than*". For a normal journey task there are usually no useful reference context values (although in some situations objects should consider themselves as travelling together only if all of them have moved away from a particular location, e.g. a train compartment). For this reason, on the side of smart objects a relater can be included in the task ID.

This representation has features in common with the context model in [12], which includes the following parameters: <ContextType>, <Subject>, <Relater>, <Object>. In our representation <Subject> is the smart object that finds its ID in the list, and there are can be many <Objects> with which the smart object compares its own context in different ways. In [12] either a reference context value or one of the smart objects in a group can play the role of <Object>, while in our work each group member needs to compare its own context with the contexts of all the other group members.

6 Application Scenario

We have selected the following application scenarios for the system prototype:

1. To check if all the necessary things are packed for a journey and that none of the items are left elsewhere during the journey.
2. To check if all the ingredients for a cake are available at home and (optionally) all of them are low-fat products, for example.
3. To check if all the parts of a business suit are clean (assuming that all dirty clothes will be in the bathroom).

We also included in the system the option "indicate item", which was implemented in the following way: when the user selects an item and presses the corresponding button on the screen of a central node (desktop computer or Pocket PC), it sends a request and the item replies by sending its location information (acquired from radio beacons) and beeps.

The goal of the computer system is to determine if all the members of the current temporal set satisfy the task requirements, and to present conclusions (a list of "bad" items) to the user.

Both central nodes (desktop computer and Pocket PC) keep lists of tasks and lists of items corresponding to each task in their memory, all the tasks and items having names and identification numbers (IDs). After the user has edited or confirmed a task and the items involved, the central node composes a radio message which contains the task ID, a list of IDs of the items involved and other information related to the task and broadcasts this. In the case of group work, the objects with IDs on the list become members of a temporal group.

We assume that the user chooses the items, e.g. those which he or she is planning to take on a journey, from the overall list of things when the task is first given to the computer. Indeed, more than half of the users in our study confirmed that they write lists of things to be collected each time they are going to travel, and some of them added that they kept these lists in their home computer for years. Some others keep their favourite recipes in their computer. In order to help the user create a list of items, smart objects are able to "introduce themselves" by sending their identification number (ID) and certain important context attributes (such as "I am a container", "empty") upon shaking. They also send other context attributes upon receiving a request from the central node. In this way a central node can add the objects to the database. Another way is to make queries in terms of context attributes ("if you are a container, send me your ID").

7 Mediation of Ambiguities

Despite the relative simplicity of the application, we have found that reasoning about the context and the system’s choice of appropriate actions leads to many ambiguous situations. These arise partly from the fact that we are using generic hardware instead of developing a specialised system. We believe, however, that increasing the system’s complexity is not likely to guarantee perfect autonomous behaviour in all cases, and thus mediation methods for resolving ambiguity are inevitably needed in context-aware systems. Dey et al. [13] propose to build "more realistic context-aware applications that can handle ambiguous data through mediation". We suggest that ambiguous situations which can appear in our system should be dealt with by including capabilities for interaction with the user into the system in the following ways:

1. Items can be added and removed from the list at any moment. This saves the users from the need to plan everything carefully in advance and helps to avoid problems with things left somewhere intentionally. For example, if some items are left in the cloakroom of a railway station for two hours, the system should not remind the user about them until this time has expired.
2. The items know of special situations which increase the certainty of context detection. If several items are simultaneously shaken hard, they immediately inform the user about items which show a different movement pattern at the same moment. As stated in [14], this technique is very easy to use, because it does not matter how the user shakes the objects. This option serves two purposes: first, it helps to give an alarm at the right moment. Imagine the following situation, for example: several people are collecting their belongings simultaneously in the same place and some things get mixed up. They then wait for a bus, and one of them jumps onto the bus as it is leaving, carrying a misplaced item. If the user wants to prevent such problems, this can be done easily by shaking the assembled bag. In this case the system can be sure that certain items are not in the bag. Second, this option helps to take into account personal preferences concerning cases in which nothing is missing. In such situations there is always the question of whether the system should remain silent and let the user wonder if it is still functioning, or whether the system should interrupt the user with an "OK" message. (If the user intentionally shakes the bag it means that a message from the system would be desirable.)
3. The system includes explanation and error recovery capabilities intended to deal with malfunctioning items, misplaced beacons and unusual contents in containers. If the user chooses the name of an item from the list and asks the system for explanations, it will present the information that is available concerning that item. For example, since the communication range of the beacons is affected by reflection, it is sometimes necessary to move a beacon half a metre to tune the system, but the user needs to know which beacon's data caused the error in the system. Also, it may sometimes be useful to inform a container about its new contents.
4. The system indicates what sensors are included in the sensor boards and allows the user to switch off cameras and microphones at any moment (this was simulated in the prototype.)

8 Collective and Individual Work by Smart Objects

Each task in our application scenario can be performed by the objects individually, which means that an object determines its local context and sends it to the central node, which reaches conclusions after collecting and processing the local context data from all the objects of interest. This centralized approach reduces the workload for the smart objects and increases that of the central node, which can be critical for mobile devices. Thus it makes sense to increase the workloads of the smart objects slightly (both our centralized and decentralized implementations run on the same PIC microcontroller) by organising them into temporal collectives according to the current task, so that each member of the collective makes conclusions about the joint context of the collective. In this case only one member needs to send the conclusions to the central node, and the central node has to do nothing but to convert the message into a user-friendly form.

Due to the small memory of the PIC microcontroller, the number of members in one group was restricted to 15. We believe that 15 members is a sufficient number for many household applications. For cases where this is not sufficient we included the option of dividing objects performing the same task into several groups, with the central node summarizing the results reported by each group. (In other cases the results from each group are presented separately.)

The following tasks were selected for collective work:

1. To check if all the necessary things are packed for a journey and that none of the items are left elsewhere during the journey. For this task collective work among the items is particularly important, since after the members of the temporal group have received the task, they will continue to work even in the absence of the central node, e.g. in the forest during a sports trip or picnic. This means that the central node does not need to be the same all the time; it can be a desktop computer with a comfortable user interface at home and a wrist computer in the forest. Additionally, there is no need to inform the wrist computer about the running tasks, as the "***result_data***" (see Figure 4) message contains not only the objects' conclusions concerning which items have been forgotten, but also task ID and other important information. It is necessary only to provide the wrist computer with the table for converting lists of IDs into words, which can be done only once.
2. To check if all the ingredients needed for a recipe are available at home and (optionally) that all of them are low-fat products, for instance.
3. To check if all the members of a set are not in a particular location (e.g. in the bathroom or in a train compartment while leaving the train).

The objects perform the following tasks as individuals:

1. Self-introduction. In response to a request from the central node or to shaking, the object will beep and send its ID, type (food, clothes etc), location and movement information and certain higher risk static context data (e.g. a container will indicate its contents).

2. Presentation of context data upon request. In response to a request from the central node, the object will send its ID and the requested static context data together with its location and movement data. This option is used first for explanation purposes in an ambiguous situation, and second, if the user needs to find all the red T-shirts available at home, for example, they can be found by calling objects of type "clothes", style "sport" and colour "red". Another way could be to ask all the T-shirts to distinguish the red ones, but this would not be the optimal way if there were many clothes at home. Third, this option can help to add new objects to a database.
3. Error recovery. Each object can change its own static context data on request, and each object can add and remove items from the group without checking if the other members of the group have succeeded in doing the same.
4. Privacy protection: switching cameras and microphones on and off at the user's request (simulated).

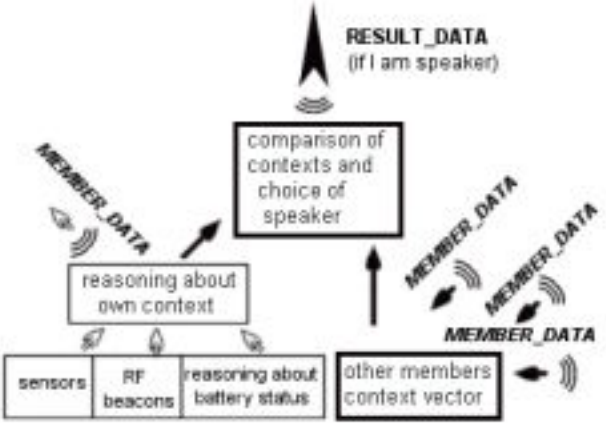

RESULT_DATA
(if I am speaker)
comparison of contexts and choice of speaker
MEMBER_DATA
reasoning about own context
MEMBER_DATA
MEMBER_DATA
MEMBER_DATA
sensors
RF beacons
reasoning about battery status
other members context vector

Fig. 4. Tasks of smart objects in a decentralized system by comparison with a centralized system. Additional information flows and tasks are presented in black

If there are many objects involved in the same task, the first step for each will be to determine its own context and report this by radio. The objects send their contexts in the form of ***member_data*** messages, which contain their IDs, energy and symbolic context values. The next step is to collect the context information from all the other members, compare the details and make the necessary conclusion. The third step is to present this conclusion to the user in an acceptable form, e.g. in words, or to give a beep as an alarm signal. In a centralized implementation, where all the objects work as individuals, a central node performs both the second and third steps, but in the decentralized version (collective work by objects) all the members of a temporal group perform the second step and the central node needs only to decode one resulting message from each group (***result_data*** message in Figure 4).

An object's own context consists of its movement type, location and context attributes such as class (food, clothes, container etc) and features which are important in this respect (e.g. use-by date and percentage of fat for food, contents and size for a container). After an object has detected its own context, it will compare this with the task requirements. In tasks where the object can decide for itself whether it is a "good" or "bad" item, the symbolic context value (included in the ***member_data*** message) is simply this decision. (E.g. for the task of finding which parts of a suit are waiting in the bathroom to be washed, an object is "bad" if it is in the bathroom. Similarly a food product can be "bad" if it contains too much fat.) In the "journey" task the objects cannot decide whether they are "good" or "bad" at this stage but send their movement type as the context value. First we tried to compare the object's accelerometer data, light data and beacon information, but it did not help to exclude ambiguous situations in our application. Thus the final version was implemented with only a comparison of the movement types of the objects.

The next step for each object is to compare the contexts of all group members. This results in the creation of a list of "bad" IDs (objects which are absent or fail to satisfy the task requirements) and the choice of a speaker (decision on whether the item should send this information by radio itself or let another item send it). For the "journey" task, objects can decide which are "bad" (forgotten) with greater or less certainty depending on the user's preferences. Objects are considered "bad" with a high degree of certainty in two cases: 1) after disappearing from the communication range of the other group members; 2) if the movement type of several other group members is "shaking", while they have a different movement pattern. Objects are considered "bad" with less certainty if they stay in the same place while the other group members are leaving. In this case false alarms are more probable, but both this and detection by the "shaking" movement type can help to identify missing objects before they pass out of the communication range.

The energy awareness of Smart-Its is very primitive (the boards do not measure battery status), being based on the fact that all boards have an identical program and the battery status is affected mostly by the number of temporal sets in which the object has taken part and the number of messages sent. Energy awareness was included here in order to help each object to decide whether it should send the conclusion about the joint context of the collective (the ***result_data*** message) by radio itself or let another object send it ("choice of speaker" in Figure 4). The ***result_data*** message is sent either according to the timing requirements or upon shaking of the objects, and contains the Task ID, a list of "bad" items and other task-related information. The ***result_data*** message is sent by the object which has a better energy value than any other group member which it can hear.

9 User Study

Twenty persons (ten men and ten women) of different ages, nationalities and professions took part in the user study. The system was demonstrated in the following scenarios:

- Scenario 1: The user asks the computer system to find whether all the ingredients needed to bake a peach pie are available at home.

- Scenario 2: The user asks the computer system to find whether all the parts of a business suit are clean (assuming that dirty clothes are put in the bathroom).
- Scenario 3: The user selects the items for a journey on the computer screen, packs them into a suitcase and leaves something out, e.g. a wallet. After the system is sure that the wallet is missing, it sends an alarm to the user. This usually happens only after the user has left home.
- Scenario 4: The user packs the selected items into a suitcase and shakes it. The system presents its conclusion immediately.
- Scenario 5: One of the boards (among the items to be packed or the beacons) gets broken. The system presents a wrong conclusion (e.g. that the item is missing), and the user can find the reason by asking it for an explanation.

In all the scenarios the users were also able to use the option "indicate item", which was implemented in the following way: when the user selects an item and presses the corresponding button on the Pocket PC screen, the Pocket PC sends a request and the item replies by sending its location information and beeping.

After that the users answered the following questions:

1. Does the system seem useful to you?
2. Do you usually write lists of things to be taken on a journey?
3. Would you use the "special action" of shaking an assembled bag?
4. Would you use the system’s explanation capabilities?
5. Would you use the option of switching off sensors such as cameras and microphones?

A summary of the user’s opinions is presented in Table 1.

Table 1. Summary of results

Question	1	2	3	4	5
Number of positive answers among 10 men/10 women	8/ 10	6/ 6	5/ 10	7/ 10	7/ 10

One of the reasons for answering "no" to the first question was the lack of trust in computers, and another was unwillingness to use the "memory prosthesis". The option of checking remotely if some parts of a suit are waiting in the bathroom to be washed was very much approved of by two wives of businessmen, and also accepted by users who have many meetings during their working time. Others commented that they don’t have many clothes, or don’t take care of their clothes themselves.

Most of the men who cook at home from time to time very much liked the idea of checking whether the ingredients were available, particularly if these were able to beep. The women were more interested in the general availability of products at home, “use-by” dates and the creation of shopping lists. Several users paid special attention to the options of adding and removing items from the list at any moment and commented that it was very important that they were "not forced to be too systematic". Some users commented that the opportunity to create a list of items to be taken on a journey using a desktop computer (especially if there were pictures of all the clothes), without the need to remember to send this list to all the mobile devices, was a very pleasant prospect.

We understand that users will not necessarily behave in the same fashion in a real situation as during an experiment conducted with a system prototype, but the results are nevertheless encouraging from the point of view that the people (particularly the women) had nothing against explicit interactions with smart objects.

10 Conclusions

The prototype context-aware system was built in order to support people's everyday activities involving interactions with personal belongings, and household applications and the context model for smart everyday objects were introduced. The proposed context model takes into account the domain-specific features of personal belongings and supports collective work by objects. The example applications included collecting things for a journey, checking if all the ingredients for a chosen recipe were available at home and checking if clothes were waiting to be washed. In order to overcome the challenges of context recognition, which are inevitable in context-aware systems, and especially in systems with limited computing resources, certain interaction capabilities were added. These helped us to take personal preferences and privacy issues into account and facilitated the providing of services with a relatively simple system configuration.

The prototype system was implemented in two ways: centralized and decentralized. In the centralized implementation most of the reasoning was performed by the central node (desktop computer or Pocket PC), while the smart objects had to detect their own context and send information on this by radio. In the decentralized implementation the reasoning regarding the joint context of the collective of smart objects was performed by the members of the collective, and the central node had only to process one message from each collective and to present the results to the user. In this case the number of members in one collective was limited, due to the small memory of the PIC microcontroller. For many applications it is simply not necessary to have many members in one temporal set. For the applications where it is necessary, the system was built in such way that it was possible to divide the smart objects into several groups, leaving the central node to summarize the results from these groups.

The main advantage of the decentralized implementation is that after the members of the temporal group have received their task, they work as an independent collective and will continue to do so even in the absence of the central node, e.g. in the forest during a sports trip or picnic. Thus the central node does not need to be same all the time: it can be a desktop computer with a comfortable user interface at home and a small mobile device during a journey.

In general, the users approved of the prototype context-aware system built up for the purpose of helping to interact with things in a smart home and accepted the interaction capabilities included in the system. They also approved of the additional freedom of use provided by the decentralized implementation. Since this implementation did not require a significant increase in the program memory of the smart objects compared with the centralized version, we suggest that this is an appropriate way to deal with a large number of personal belongings.

Acknowledgements. The Smart-Its project is funded in part by the Commission of the European Union under contract IST-2000-25428, and by the Technical Research Centre of Finland. We would like to thank all the people who took part in the user study for their goodwill.

References

1. Brumitt, B., Meyers, B., Krumm, J., Kern, A, Shafer, S.: EasyLiving: Technologies for Intelligent Environments. Proceedings of HUC 2000, Lecture Notes in Computer Science, Vol. 1927, Springer Verlag, Berlin, Germany (2000) 12–29
2. Kidd, C.D., Orr, R., Abowd, G.D., Atkeson, C.G., Essa, I., Mynatt, E.D., Starner, T., Newstetter, W.: The Aware Home: A Living Laboratory for Ubiquitous Computing Research. In Proceedings of CoBuild 99, Lecture Notes in Computer Science, Vol. 1670, Springer Verlag, Berlin, Germany (1999) 191–198
3. Estrin, D., Guller, D., Pister, Ch., Sukhatme, G.: Connecting the Physical World with Pervasive Networks. IEEE Pervasive Computing, Vol. 1 (2002) 59–69
4. http://akseli.tekes.fi/Resource.phx/tivi/elmo/en/rolling.htx
5. Beigl, M., Gellersen, H.W., Schmidt, A.: Mediacups: Experience with Design and Use of Computer-Augmented Everyday Objects. Computer Networks 35(4), March 2001, Elsevier (2001) 401–409
6. Beigl, M., Gellersen, H.-W.: Smart-Its: An embedded platform for Smart Objects. Smart Objects Conference 2003, Grenoble, France
7. Antifakos, S., Michahelles, F., Schiele, B.: Proactive Instructions for Furniture Assembly. In Proceedings of Ubicomp 2002, Goteborg, Sweden (2002)
8. Abowd, B.D., Mynatt, E.D.: The Human Experience. IEEE Pervasive Computing, Vol. 1 (2002) 48–57
9. Davies, N., Gellersen, H.W.: Beyond Prototype: Challenges in Deploying Ubiquitous Systems. IEEE Pervasive Computing, Vol. 1 (2002) 26–35
10. Korpipaa, P., Mantyjarvi, J.: An Ontology for Mobile Device Sensor-Based Context Awarenes. In Proceedings of Context 2003
11. Vildjiounaite E., Malm E.-J., Kaartinen J., Alahuhta P: A Collective of Smart Artefacts Hopes for Collaboration with the Owner. HCII 2003
12. Ranganathan, A., Campbell, R. H., Ravi, A., Mahajan, A.: ConChat: A Context-Aware Chat Program. IEEE Pervasive Computing, July-September 2002, (2002), 51–57
13. Dey, A., Mankoff, J., Abowd, G. & Carter, S.: Distributed mediation of ambiguous context in aware environments. Proceedings of UIST 2002, (2002) 121–130
14. Holmquist, L.E., Mattern, F., Schiele, B., Alahuhta, P., Beigl, M., Gellersen, H.-W.: Smart-Its Friends: A Technique for Users to Easily Establish Connections between Smart Artefacts. Proceedings of Ubicomp 2001, Atlanta, GA, USA (2001) 116–122

Position-Based Interaction for Indoor Ambient Intelligence Environments

Fabrice Blache, Naoufel Chraiet, Olivier Daroux, Frédéric Evennou,
Thibaud Flury, Gilles Privat, and Jean-Paul Viboud

France Télécom R&D, Grenoble-Meylan, France
{Fabrice.Blache,Naoufel.Chraiet,Olivier.Daroux,
Frédéric.Evennou,Thibaud.Flury, Gilles.Privat,
Jean-Paul.Viboud}@rd.francetelecom.com

Abstract. We present a platform and a set of position-adaptive basic services for indoor ambient intelligence environments. This suite of Jini™-based services implement various instances of implicit interaction based on the user's position, mediated by a handheld personal device. An indoor navigation system has been implemented as a prototype application for a structured location-management infrastructure that maintains information about all fixed, movable and mobile features of the ambient environment, according to complementary abstract models of this space. Location and navigation information is presented to the user's handheld device as a fully scalable, dynamic SVG scene. 802.11 fingerprinting is used as a basis for location-estimation at an inter-room scale, while RFID is used for close-range (intra-room) position detection.

1 Introduction

The following scenario illustrates our user requirements for this work.

As a user enters the target indoor space, his/her presence is detected and a directory of the locale gets automatically uploaded to his/her handheld personal device This directory comprises both regular points of interest and those corresponding to ambient-intelligence services/devices that the user may interact with. This menu of points of interest is presented to the user visually on a vector-based map, into which the user may zoom and navigate at will, through which he/she may select these points in usual interactive way. He/she may ask a route to a particular point of interest in direct or inverse fashion, either by pointing it on the map or selecting it in a menu. Yet rather than direct selection by pointing on the PDA screen, the default mode of selection is by proximity detection. As the user gets close enough to a point of interest : depending on the nature of this point, either a more detailed map, additional information, or an software component corresponding to this service/device is uploaded to the user's handheld device. This component makes it possible for the user to interact with this device through his PDA, either to take control of it or to use it as an alternative interface for various services. Examples of such services are presented in section 7.

This scenario may be applied to various environments such as e.g. dense urban precincts, shopping malls, airports, railway stations, exhibition halls, industrial

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 192–207, 2003.

compounds, conference/convention centers, large office buildings, cultural heritage sites, large museums, theme parks, etc, where points of interest may correspond to either specific rooms, exhibits, vending machines, appliances, etc. Potential users are all regular visitors of such places, but also children, elderly persons and people with disabilities (e.g. visually impaired), to the specific needs of which the system should be able to adapt.

Commercial systems combining location-based information with visual presentation and route-finding are widespread for GPS-based vehicular navigation. These systems have evolved from specialized hardware to add-on modules and software packages for PDAs. Similar PDA-based prototype navigation systems have been studied for pedestrians visiting a city, as part of electronic tourist-guide services [1], [2]. Integrating these systems within a more limited and thoroughly modeled ambient intelligence indoor environment makes it possible to integrate a richer repertoire of information relevant to this space, beyond the usual points of interest.

In this view, such a system is not merely a directory or navigation utility, but a personal ambient environment interface. At this stage, ambient intelligence is mediated by a handheld terminal used as a pivotal device, that we call “personal communicator”[1] in the following. This converged PDA-smartphone device is connected to both wide-area networks and personal-area device networks, it may act both as a display for visual presentation of the navigation tool and as a sophisticated universal remote control for all devices which the user may take control of. The long-term evolution of ambient intelligence could make it possible to dispense with this intermediary device when possible (e.g. in the user’s own home), but its use remains relevant for full mobility, wherever alternative ambient interface devices cannot be taken for granted.

We present in the following the work that has been done within the Ambience project towards the implementation of this scenario. Section 2 outlines the system requirements, from which we go on to describe in section 3 the abstract model used for modeling the ambient environment as a space in the more mathematical sense of the term. The location infrastructure based on this model is described briefly in section 4. The two physical location-determination technologies that have been used complementarily are described in section 5 and 6. The different service building blocks making up the ambient environment are described in section 7. We conclude in section 8 with lessons to be learned and perspectives about the integration of these services in a generic service-management infrastructure.

2 System Requirements

The more general user requirements of the above scenario translate into the following system requirements:

- The selection of services should by default occur as an implicit command triggered by proximity detection (it may also of course occur by explicit selection from the user)

[1] The use of this phrase here is unrelated to existing commercial products

- The provisioning and operation of services should not require any manual operation or a pre-configuration of the personal communicator
- The personal communicator should have dual connectivity, with wide-area wireless network for distant service provisioning on the one hand, with other ambient environment devices via WLAN, WPAN or other more means such as inductive coupling using RFID on the other hand)
- The personal communicator should act as a gateway to a chosen service infrastructure, and ideally should make it possible to interoperate between several such infrastructures (e.g. Jini, UPnP, OSGI, SLP, etc.).
- The location infrastructure should be independent of both the location-sensing technologies and the applications, and should make it possible to operate with all of these
- This location infrastructure should be scalable from small-scale location management (ideally a few centimetres for close-range proximity detection) upto large scale location management by interfacing to cellular-network-based or third party location-determination systems using standard protocols (e.g. MLP)
- The location infrastructure should make it possible to maintain information about a plurality of fully mobile devices (personal communicators, wearables, etc.) and also about other ambient devices that may move from time to time (e.g. printers, coffee machines, etc…). We call this latter category of devices “movable” in the following.
- The location infrastructure should make it possible to incorporate a rich set of model information about the environment (fixed features of the environment, points of interest, movable objects, other mobile devices tracked by the system,
- This space modelling information should be obtainable with the assistance of interactive tools from architecture CAD systems

3 Modeling Indoor Ambient Environments

The modeling of location information in this project is based upon a generic template for location information management described in more detail in a previous publication [3]. This project can be seen as a validation of these more abstract ideas about location as applied in an ambient intelligence environment, experimenting how a generic location management service can fulfil the various requirement of the navigation system and other applications, and how it can be implemented to suit the particularities of an indoor environment. According to these considerations, we establish a selection of the key feature of the location requirements to be adapted from the template. Then from this experience we draw a formal and abstract representation of location information using RDF Schema.

3.1 Independence from Location-Sensing Technologies and Applications

Our requirements comprise the independence of the location management service from the location sensing technologies and the applications (such as the navigation service). This requirement can be met by abstracting away both the information provided by sensors and the information to be provided to the application, according to generic location models as put forward in [3]. Based upon these models, location-sensing technologies can be grouped into a few generic families corresponding to the different layers of the template, i.e. set-based, affine-euclidean, graph-based, semantic. Within each of these models, a particular technology having its own capabilities and limitations (such as scale, accuracy, range etc.) could be formally described and integrated in the larger picture. For example, WLAN-based fingerprinting and GPS both provide continuous affine location estimations (i.e. positions relative to a coordinate reference system) with different scales and precisions.

On the other side, these generic location models can match the needs of applications, using location queries for synchronous pull of location information and events for an asynchronous push of information. Queries follow a generic formalism corresponding to each location model. Events are generated by the location service when specified conditions are met. The corresponding location information (from the sensor/to the applications) can be characterized in formal and generic fashion.

3.2 Specificities of Indoor Environments

Set-based model of space is paramount

Indoor environment is mostly structured by the architectural division of space by walls and floors into sets. The resulting sets corresponding to rooms are mutually exclusive and represent the lowest physical level. They may be divided into logical subsets such as cubicles in an open space office. They are likely to be aggregated into supersets, either by their destination at a semantic level, or corresponding to purely architectural divisions (such as suite of rooms, floors, wings of a building, whole building, compound, etc.). At this level, the environment is constituted by sets in which only the presence or the absence of entities (abstracted by points) can be asserted, without any geometric information.

Affine (cartesian-coordinates-based) representation of location information does make sense within a particular undivided set such as a large room, with a coordinate reference system that may be local to this set. Even if not used as the primary reference model for an entire indoor space, a pivot CRS is nevertheless necessary, to combine information from differents sensing technologies, to transform one local CRS to another, and to map the discrete model (using geometrical representation of sets) to the metric/geometrical location model.

Structural relationships derive from the set-based model

The subsets of indoor space have physical and logical relationships together. Physical relationships follow architectural rules. A building is usually divided into floors, a floor is divided into rooms. These rooms have an adjacency relationship between them. These architectural realities lead to two kinds of structural relations (in the mathematical sense of the term) that may be represented, respectively, as containment and adjacency graphs.

Taking into account the division into rooms with the additional information about doorways and other openings makes it possible to derive complementary structural information about possible routes for people moving in the building. A route-finding graph will be built upon this connectivity relation, different from the adjacency relation (as two adjacent rooms may not have a direct route between one another). This route-finding graph will used primarily by the navigation service

Indoor-specific location ontologies for semantic mapping of the set-based model
Space is naturally interpreted at a semantic level. The usual architectural taxonomies used to characterize the subsets of indoor space (hallways, rooms, floors etc.) may in their own right be considered as defining "semantics" of this space related to an architectural ontology, if defined more formally. The same sets will also have a more specific semantic mapping that may be defined through their use (cafeteria, restroom, meeting room). This may be based on a domain-specific ontology, making a possible distinction between home, office, industrial, or public indoor environments.

Other specific ontologies or, more specifically directories, of mobile or moveable entities inhabiting the environment may also serve to characterize subsets of space by way of the temporary or permanent association between the two (childrens room, the boss's office, John's office, etc.)

4 Location Infrastructure

The location management service maintains information about the environment and the positions of mobile or "movable" entities within it. It is formed by a federation of logical modules to retrieve, manage, store and relay location information. These modules have well-defined interfaces to be as much as possible independent from one another.

4.1 Ensuring Scalability

The management of location information has a direct impact on the scalability of the system, a completely dynamic characterization may be more accurate and reactive but cannot be adapted when numerous entities act in a larger place. On the contrary, a static representation of things may not be suited for the dynamic context awareness needs of an ambient intelligence environment. The scalability goal can be reached considering what kind of location information needs to be characterized. Three categories can be distinguished:

- Static description of the environment (geographical information, fixed structure of a building, etc.)
- Dynamic information about highly mobile entities (persons, handheld or wearable devices)
- Partially dynamic information about moveable entities (like printers or office furniture which may be considered part of the environment but may also move and have to be tracked)

A hybrid model can use jointly a static representation of environment information, stored in a persistent distributed database, and dynamic information about location mobile entities as transient information in process memory. Non-location-related attributes of mobile entities will also be mostly static and stored as such in the database. The current location information about a mobile or moveable entity can also be stored into a semi-static fashion whenever it is immobile. But when these entities come to move, a special dedicated process can handle dynamically the information provided by the location-sensors to maintain accurate real-time location information. When the entity stops moving for a sufficient length of time (according to criteria relative to the mobility of the entity and the means to locate it) it is stored again in a semi-static way.

This model can be designed to support the scale factor when a potentially vast amount of location information has to be characterized. If a large number of mobile entities act at the same time the solution can be to distribute the location management service.

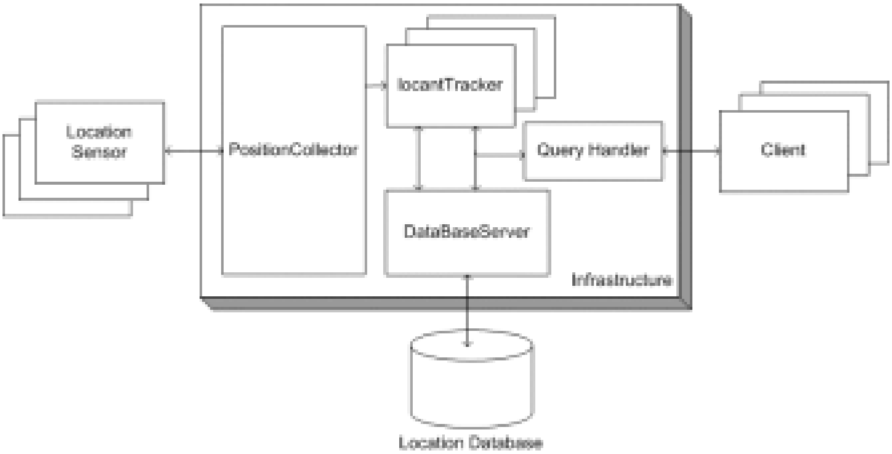


Fig. 1. Location infrastructure

4.2 Position Collector

The position collector has to collect the location information provided by individual location-sensors or complete-location-determination systems. It receives XML-encoded messages containing the identification of the mobile object (depending of the technology, it may be an RFID ID, an IP address etc.) a timestamp, a reference to the technology in use (to retrieve the type and properties) and the measured value(s) (it may be a set of values, with or not an indicator of accuracy…). Theses values are forwarded to Locant trackers and are stored in a log file.

4.3 Locant Trackers

Locant trackers are processes created when a single mobile entity is being detected by a particular sensor or location sensing system. This dynamic process maintains real-time information about the current location of the entity according to the underlying technology of the sensor. Locant trackers belong to general families depending of the location models addressed by the sensors. “Logical” locant trackers may automatically be created to monitor the activity of concurrent "physical" locant trackers, taking into account their characteristics (accuracy, reliability, etc.). If no activity is detected by the locant tracker during a specified lapse of time (depending of the underlying technology), the entity is considered motionless, its current location information is eventually stored into the database and the process is stopped.

4.4 Database Server

The module Database server is an interface between the location information management service and a relational database. This database is used to store the static data about the environment and the static properties of entities and acts as a repository for location information about the currently still entities. This database is only internally used by the location service. Queries from client are addressed to the Query Handler.

4.5 Query Handler

The query handler uses the Database Server and monitors the Locants Tracker to answer queries or generates events (when particular location conditions are met) and forwards them to the subscribers.

5 RFID-Based Position Detection

Not yet fully integrated in the general location infrastructure described above, RFID-based position detection has been used with a specific set-based location-management infrastructure. RFID tag readers, positioned near the points of interest are associated with a Jini™ service.

When a mobile entity is located by its RFID tag, the associated service finds and writes in a central repository (a Javaspace™ service) the reader's id and the tag's id. A specific location service interprets the entry written in the Javaspace™ to determine the neighborhood service around the sensor and calls for an identification service that establishes the mapping between the low-level RFID's id and the high-level client in the ambient intelligence environment.

The neighborhood service then advertises the presence of the client to all services/devices in the area. These services/devices may then react accordingly to the position-based interaction scenarios.

6 WLAN-Based Location Estimation

The use of WLAN networks for obtaining position estimations for mobile devices has become the subject of numerous studies With 802.11 networks, two main methods are used.

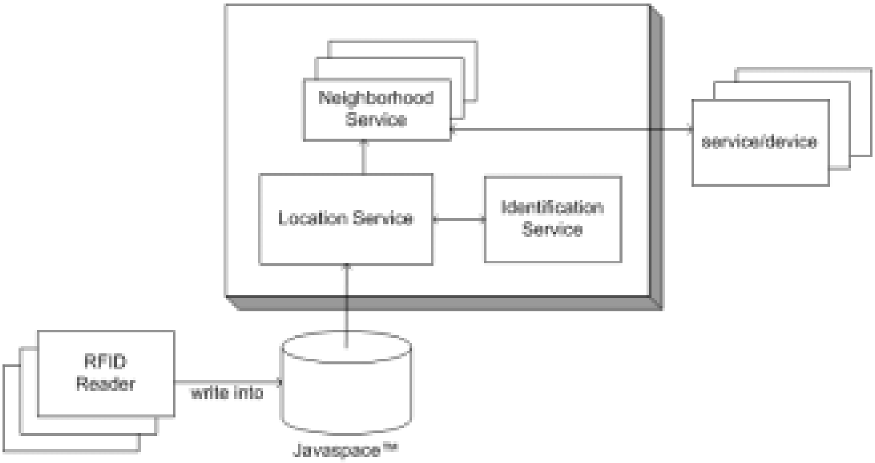


Fig. 2. RFID-based positioning

The first one uses of a propagation model to estimate the distance between the mobile and each network access point from measurements of the received signal strength (RSS [4]). Then a multilateration technique using a Least Mean Square method may be applied to estimate the current position of the mobile. Unfortunately, the received signal strength tends to fluctuates erratically (±10dBm), leading to large errors when the set of equations is solved. We have had a try at this method, but the obtained accuracy is very poor (about 7m for the mean error along a course).

The second and currently the most commonly used method (Microsoft RADAR [5], Bluesoft, Appear Networks,…), is based on a database of the received signal strength ("fingerprinting") over the building. When the mobile requests its location, it makes a measurement of the received signal strength from several access points , and the measurement is compared to the data base. The position can be estimated as the one that minimizes the difference between this measurement and the data base.

This method is more efficient than the direct use of the propagation model, yet signal fluctuations are still present and cause large variations in the brute position estimate. Because of these variations of the signal strength, a leap frog phenomenon can be observed. This can be minimized by applying Kalman filtering or particle filtering [6], [7]. We have used the latter method and present it in the following.
The particle filter is based on two main steps. The first step is a predictive one and the second one adds a correction to this predictive step thanks to a measurement. Thus this filter is based on the use of probability density functions (pdf) that may not be Gaussian and is for this reason a non-deterministic method to get the position of the mobile in comparison to the Kalman filter that is deterministic.

This filter tries to find the most realistic pdf fitting $\Pr\left(x_k \middle| z_{i=0:k}\right)$. where x_k is the current position of the particle, and z_k the measurement of the signal strength, to estimate the most probable places where the mobile can be, taking into account the measurements. This pdf is approximated by a set of particles that randomly explore the environment, to which a weight is assigned taking into account the history of the particles and the current measurement. The weight of each particle is then iteratively estimated by : $w_i = w_{i-1} * \Pr(x_k|x_{k-1}) * \Pr(z_k|x_k)$

Moreover, in this kind of filter, different sources of information can be used to improve the accuracy of the position determination. Here a map of the building has been used. This enables the particles to take into account the environment in which they are :we have chosen that when a particle crosses a wall, then $\Pr(x_k|x_{k-1}) = 0$ otherwise its value is 1.

Here a particle filter using 10.000 particles has been implemented to locate an mobile in a single floor of our premises (a 35*35 m square , see Figure 3).

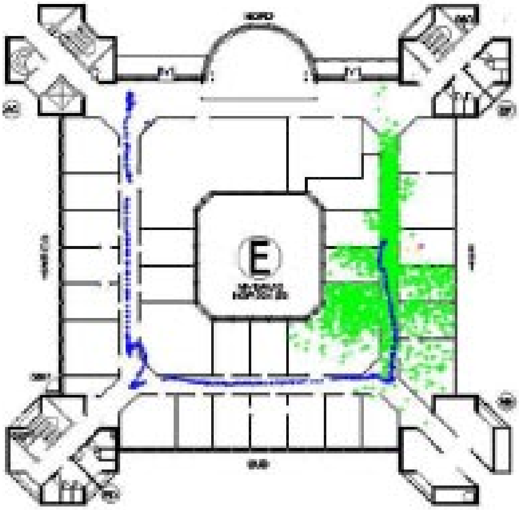


Fig. 3. Estimated trajectory of the mobile (black dots) from 10.000 particles (grey dots)

Even if individual particles may not traverse walls, the position estimate of the mobile, as the barycentre of all particles, is allowed to do so. This makes it possible to take into account the case when the estimate is close to a door : with a sufficient count of particles going through the door, the estimated position of the mobile itself will eventually do so, even if its trajectory had to traverse a wall . If this wall crossing is not a correct estimate (as in the bottom left of Figure 3) it will eventually be corrected. Particle filtering makes it possible to take into account data from other sources of position measurement (such as RFID, see Section 7) to resynchronize the filter. We are currently investigating whether it is more advantageous to merge these two kinds of information at this numeric level, or to integrate them in a “logical tracker” (see section 4.3) in the location management infrastructure, taking into account higher-level information such as the route-finding graph (section 3.2) to which a probabilistic transition model could also be attached.

7 Position-Adaptive Interaction: Basic User Services

All the following services (except the navigation/SVG presentation) have been implemented as Jini services. This means they are spontaneously uploaded to the user's personal communicator as he/she enters the corresponding position-detection perimeter, as a result of an event transmitted to the Jini directory through a specific neighborhood management infrastructure implemented upon JavaSpaces™ as described in section 5 above. The user need not configure anything manually (except, at the outset, the Jini infrastructure itself) and need request explicitly a download to his PDA for these services to be available.

Ultimately, the full-fledged location management infrastructure described above should be seamlessly interfaced with the service discovery infrastructure, so that services may be queried & spawned from location-based queries or events transmitted through this infrastructure. All these services should be considered as building blocks from which an integrated ambient intelligence environment could be composed.

7.1 Navigation

The navigation service may be considered both as a top-level entry point for other ambient services presented afterwards, or as a stand-alone service in its own right.

It comprises a navigation service proper and an SVG-based client presentation module. The navigation module itself comprises the following two modules. The Route Finding module uses a graph-based shortest path algorithm to find a route between two points approximated as nodes of the graph retrieved from the location infrastructure. As the size of the graph is limited, optimization of this algorithm is not critical and a simple Moore-Dijkstra algorithm has been chosen.

The SVG Producer Module retrieves information from the location infrastructure corresponding to the relevant levels of representation : set-based, cartesian and structural (graph-based).These internal representations are converted to SVG used as a presentation format. Ideally a standard XML-based interface to the database should make this process independent from the implementation schema of the database, with the possible use of standard XML transformation tools (DOM or XSLT). Also, the RCC (Rendering Custom Content) SVG feature to be supported with version 1.2 of the language, should make it much easier to have a client-side implementation of this kind of transformations : the original tagset of the structured and semantically rich internal representation language could be directly included with its own namespace in the SVG document describing the rendering. The transformation, corresponding to a model →view mapping in an MVC design pattern, would be specified in an externally referenced SVG "extensionDefs" description. As a possible example of this, we have defined a "Constructive Area Geometry" SVG extension that makes it possible to describe a purely set-theoretic, rather than cartesian, model of space. Using this model, a subset of a plane is described as the union/intersection of other more elementary subsets rather than by the equation of the curve that delimits it. This makes it possible to retain in the client structural information that may be useful for direct interactivity and other features. Obviously, there is here a performance trade-off, as a lower level representation may be faster rendered on a client device with limited capabilities.

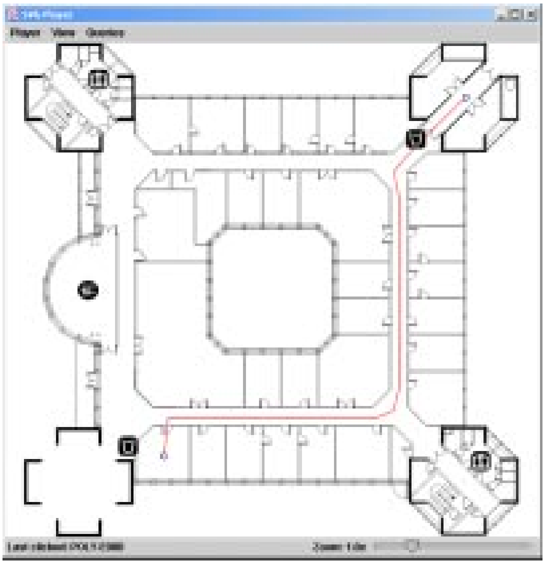


Fig. 4. SVG Player used for the navigation service

This complexity trade-off has to be addressed foremost in the implementation of the SVG player that is the heart of the client presentation module. For the time being we use a Batik java-based player as the most open and portable solution. As this player is based on the Java Swing GUI library, it places a rather high bar on the capabilities of the supporting device, incurring also a severe performance penalty in the process. Clearly we pay a price for the interactivity that such a high-level format enables on the client-side. A fallback version with intermediate content transformation on a client-proxy module will later be implemented, that should make it possible to target low-end devices as final clients.

7.2 Virtual Kiosk

This service is a subset of the previous one, corresponding to a more precise position information as obtained by the detection of the presence of the user in a given neighborhood to which a specific set of services may be attached. It corresponds to a type of query defined above as inverse location query ("what is there around me", "what is there at such location"?)

When the user gets within the perimeter of a designated location, which may be either a particular office, a laboratory, a shop, an airport counter, a booth in a show, an exhibit in a museum, a monument, or whatever point of interest, a menu relative to this location is uploaded to him, presenting the available information and services relevant to this particular location.

This service draws its data from a general location-indexed service directory which should be consolidated between the service discovery infrastructure and the location infrastructure. It is, however, distinct from this general location infrastructure as a standalone client service rather than an infrastructure service.

Being a purely software service, this kiosk need not be implemented as a physical kiosk terminal of the kind illustrated above. The user will just have the interface of the service downloaded to his PDA device. A physical impersonation of the service could nonetheless exist as a physical icon or an alternative interface.

7.3 Book Shelf

This simple scenario exemplifies a very large set of potential "m-commerce" scenarios that empower users to get the best of both worlds by combining on-line and store-based commerce for mobile users. Books are used here as "intelligent" items that sense the user's interest when they are taken out them from the shelf, and provide him with any additional information he/she may wish to get through his/her personal communicator.

Two levels of proximity-based interaction are provided. First, the bookcase locates the user in its vicinity. As for the preceding service, it then provides the user with its own "directory", actually the list of books it contains, as its own service interface, spontaneously uploaded and activated on the user's personal communicator.

The user could select a book from this menu, but the preferred way of doing it is, obviously, to interact with it in a physical way. So, he selects a book by taking it out of the bookcase and of course he will browse it as any regular book, which is the main reason for retaining the book as a richly afforded, culturally endowed physical object, An RFID reader antenna is mounted into the structure of the case, and each book is equipped with a tag in such a way as to be able to detect unambiguously the presence or absence of a book. Triggered by the detection of the book being taken out of the shelf, a new interface, to this single book itself, is then uploaded to the user's personal communicator. This interface may provide the user with additional information such as he may find with an online commerce or aggregator's site. Other media (spoken audio, music, still pictures, video) associated with the book may also be proposed. These media could be played on the user's PDA by uploading the corresponding player if necessary, but a combination with the interface export service presented afterwards makes it possible to display them on the best adapted interface device in the neighbourhood, such a wall-mounted plasma display for video or a five channels surround sound system for audio, that will obviously offer a richer experience.

An m-commerce scenario using these technology solutions is illustrated below, with books used as phicons both in the store and back home.

7.4 Phicon-Activated Services

The books featured in the previous service description are an example of a very generic kind of physical interface device that has come to be described as physical icons, or phicons. Phicons are but a special case of graspable or tangible user interfaces [8] that have iconic (analogue representational) characteristics with regard to the action or service they activate.

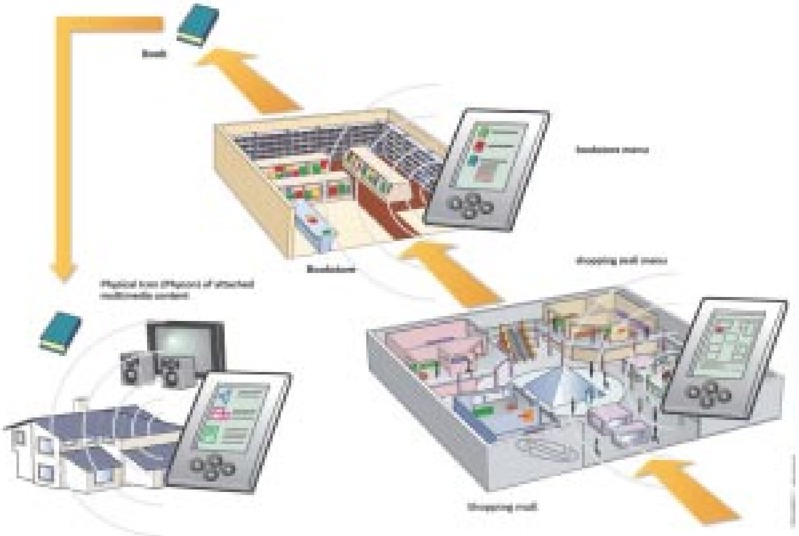


Fig. 5. An"ambient intelligence" shopping mall and bookstore

From this idea, we have diversified the previous example with a set of phicons, passive objects equipped with RFID tags, held in a phicon container into which is mounted the RFID reader used to locate them. This piece of furniture will be used in a similar way as the book case : the phicon-case knows what it contains and sends an event to the infrastructure each time a phicon is taken out of it.

These phicons are used as physical impersonations of purely informational/ communicational services. These may be one of the services proposed above, e.g. distributed bulletin-board or virtual kiosk, or any other service classical such as a videophone, video on demand, etc.

An example we have implemented is a cube with a picture of a person that triggers a SIP phone call to that person (apart from being a nice interface, this could make it possible for very young children or handicapped persons who cannot dial a phone to place a call). This may happen in the simplest way by just taking out of its container; if a voice call is the only possible action associated with this phicon, or in a slightly more sophisticated way, by placing the phicon in the vicinity of an other device such as a regular phone or display screen, indicating that this device is to be used for communication: we have here a nice example of spontaneous service composition made possible by combining the tangible interface idea with a service infrastructure that associates a service to each device.

7.5 Location-Mediated Bulletin-Board Service

In location-mediated communication, one of the end-points of the communication chain is directly a physical location rather than a regular telecom terminal attached to the network, or a person using this particular terminal. This physical location is addressed directly as such, instead of addressing a terminal device relatively to the network and a user relatively to this terminal, as done in classical (fixed or mobile) telecom.

This scenario may be considered as either the virtualization of a physical notice board, the dissemination of virtual post-it notes, or the physical grounding of a purely informational bulletin-board or web-like service, being somewhere in-between.The user may thus attach information to a particular location. This is roughly equivalent to a traditional physical notice board where user may paste notes scribbled on post-it papers : the target of the information is "whoever will be there and will read it"

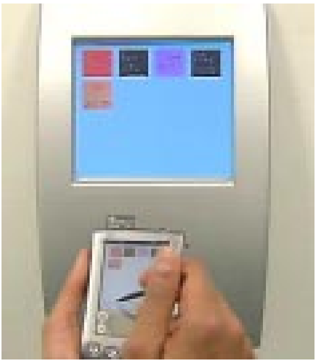


Fig. 6. Bulletin-board with PDA used as an interface

For posting information, interaction with this notice-board may be either local, the interface to the service being uploaded to the PDA, or remote, through a web-service-like interface. For reading information, there are two important cases.

When the service is purely a software service, it is possible to leave a message, or more generally attach information, to whatever addressable "location" : it may be either a piece of furniture, a room, or building. In this case, the person who is potentially interested in this information (as determined through his/her profile) will have it automatically uploaded to his/her PDA upon passing through or close to the destination location.

When the service has its own physical interface, that may be a regular flat-panel display or whatever special-purpose display is deemed suitable to be better integrated in the décor, the user may read directly the information presented on this display, and possibly interact with it.

7.6 Using Distributed Interface Devices

The idea is that the user may, through the mediation of his personal communicator, take advantage of whatever interface device is available in his immediate environment, overcoming the bottleneck of the PC, telecom terminal or PDA as an exclusive interface to the information world.

7.6.1 Interface Import/Export

With this service, the user has the possibility to transfer interfaces from one device to another by what will usually be an implicit command, e.g. pointing to a particular device with another, or waving a device close to another one.

The personal communicator acts again as a pivotal device for this, making it possible to export the interface of purely informational services on the best-adapted devices in the environment, or, conversely, to import the interface of services available in the environment. These may be virtual but physically located services as the "virtual kiosk" and "situated information space" examples, or physical appliances. Of course this latter case may appear at variance with the idea of distributed interfaces, as it amounts to a re-centralization of interfaces on one particular device. This may yet make sense in many cases, such as the adaptation of interfaces to the needs of the user, e.g. from a pushbutton to a speech-recognition-based interface, which the personal communicator may provide. It may also make sense in view of using the personal communicator as an intermediary, the interface imported on it being re-exported on another device.

7.6.2 Follow-Me Interfaces

This idea is the direct application to telecom services of context-awareness, location being the primary element of context information to be exploited. In this case, any information/communication service is aware of the position of the user : he is located at any given time and whatever service he is using at any given time is able to follow him without the user having to carry or wear a specific interface device : the service will just be "handed over" between different interface devices. The example that has been implemented is for a video-on-demand service, that is exported to an external display in the first stage, and then automatically follows the user as he moves through the room, using RFID position detection. The same idea could be applied for a bilateral audio-visual communication service, a follow-me videophone, that would use an ensemble of screens, cameras, speakers and microphones available wherever the user moves.

8 Conclusion and Perspective

We have presented the design rationale and described the implementation of a set of services that make up a good basis for evolving a full scenario of indoor ambient intelligence.

The Jini™ infrastructure on which the first implementation of these services was based is conceptually appealing and, being freely available, has made it possible to jumpstart a quick prototyping effort. Yet Jini has, in its present state, limitations that make it as yet ill-adapted for real-life deployment. First, it sets a too high bar on the capabilities of devices on which it should run. It is still impossible to run a full-fledged Jini peer on even a state of the art PDA. Second, its lookup service is far from being a complete service database that would make it possible to query services from their attributes based on a standard query language. A final, more fundamental limitation of Jini is shared with all other similar infrastructures (e.g. UPnP, Salutation) that require prior standardization of services to make it possible for a client to use them. As long as clients have to be acquainted beforehand with the access primitives of those services they want to use, serendipitous interoperation as promised in ambient intelligence scenarios remains a pipe dream.

For all these reasons, we are currently implementing an alternative service infrastructure that avoids those limitations. Due to this transition from a Jini infrastructure to an OSGI-based one, software integration of these services is not complete yet, and there also remain a number of open issues to fulfill all the requirements outlined at the beginning.

References

1. Banerjee, S. et al. "Rover: Scalable location-aware computing", IEEE Computer, Oct.2002
2. Davies, N. et al. "Using and determining Location in a Context-sensitive Tour Guide", IEEE Computer, Aug 2001
3. Flury T, Privat G; "An infrastructure template for scalable location-based services", Smart Objects Conference, SOS'2003, Grenoble, May 2003 ; http://www.grenoble-soc.org/proceedings03/Pdf/59-Privat.pdf
4. Chen Y., Kobayashi H. , "Signal Strength based indoor geolocation", Proc. IEEE International Conference on Communications, April-May 2002, New-York
5. Bahl, P, Padmanabhan, V.N., "RADAR: An In-building RF-based User location and tracking system", Proceedings IEEE Infocom 2000, Tel Aviv, Israel, vol. 2, pp. 775–784, Mar. 2000
6. Arumpalam, S. et al, "A tutorial on particle filters for non-linear/non-gaussian Bayesian tracking", IEEE Transactions on Signal Processing, vol. 50, no. 2, Feb. 2002
7. Gustafsson, F, et al, "Particle filters for positioning, navigation and tracking", IEEE Transactions on Signal Processing, vol. 50, no. 2, February 2002
8. Ullmer, B., Ishii, H., Emerging frameworks for tangible user interfaces IBM Systems Journal, vol. 39 no 3&4 , 2000

Some Issues on Presentations in Intelligent Environments

Christian Kray, Antonio Krüger, and Christoph Endres

Saarland University
P.O. Box 15 11 50, 66041 Saarbrücken, Germany
{kray, Krueger, endres}@cs.uni-sb.de

Abstract. Intelligent environments frequently embed a varying number of output means. In this paper, we present an analysis of the task of generating a coherent presentation across multiple displays. We review problems and requirements, and propose a multi-layer approach aimed at addressing some of the previously identified issues. We introduce a low-level architecture for the handling of devices as well as mechanism for distributing coherent presentations to multiple displays. We then discuss what further steps are needed on the way from an abstract presentation goal to a concrete realization.

1 Introduction

From a technical point of view the trend of Ubiquitous Computing is tightly bound to the notion of the disappearing computer, where most of the processing power and communication abilities will be embedded behind the scenes. However, from the user's perspective Ubiquitous Computing will lead to ubiquitous input and output facilities in a given environment. New display technologies (e.g. organic displays) will increase the amount of displays in an instrumented environment significantly, allowing intelligent multimedia presentation systems to plan and render presentations for multiple users on multiple displays. Modern airports are already good examples of environments equipped with several different types of displays: Information on arrivals and departures is presented on huge public boards, at gates plasma screens display information on the actual flights, throughout the building small touch screens are used to provide single users with information on the airport facilities. Finally wireless LAN hotspots allow the usage of private PDA screens to access information on the web at several locations. Until now those displays cannot be used together to present information, but several technical suggestions were recently made to move towards a solution for this problem. In [1] a framework is presented that incorporates different devices to render audio and graphics files. An approach that allows users to access publicly available displays (e.g. from an ATM) for personal use is presented in [2]. Some research has been also carried out on the combined usage of a PDA and large screens [2],[4],[7].

However, only little work was done on issues regarding the planning and rendering of multimedia presentations on multiple displays. In this paper we reflect both on the different technical prerequisites and the concepts that will allow for distributed multimedia presentations in instrumented environments. This involves the handling of

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 15–26, 2003.

devices, e.g. their registration and access, the rendering of the presentations, e.g. synchronizing presentations in time and space and the planning of presentations, e.g. how to guide the users attention from device to device. While the processes underlying a presentation using multiple devices are rather complex and hence call for a sophisticated approach that takes into account a number of factors ranging from device control to guiding the attention of the user, the benefits are worth the effort: On the one hand, the users will enjoy consistent and non-interfering presentations that are adapted to the current situation. On the other hand, these presentations will use all the available output means instead of just small limited set. Hence, the system will deal with the complexity while the users benefit from a simplified interface.

We will first review some key issues in the context of intelligent environments in general, and more specifically point out problems related to handling presentations in such a setting. In order to address some of the issues raised, we will then present a low-level infrastructure capable of dynamically handling various devices. Based on this foundation, we will introduce a mechanism for rendering presentations that are distributed across multiple devices in a synchronous way. These presentations are the result of a planning process that can only partially rely on traditional presentation planning. Consequently, we will point out some areas where further reasoning is necessary. After sketching out a possible solution for this problem, we will shortly summarize the main points of this paper, and provide an outlook on future research.

2 Issues

During the process of planning, generating, and rendering a presentation in intelligent environments several issues arise that go beyond those encountered in traditional multimedia presentation [9]. The issues are mainly caused by the dynamicity of the ubiquitous setup and the diversity of the devices used for a presentation. A further factor that complicates the process is the potential presence of multiple users.

A fundamental issue that does not only impact presentations but all services and tasks running ‘on’ an intelligent environment is the discovery, handling, and control of devices. In the case of presentations mainly output devices such as displays or loudspeakers are of interest in this area. The infrastructure has to provide means to determine what devices are present and available, to control these devices, and to dynamically add and remove devices.

The latter point is also closely related to another key issue: the constant change an intelligent environment may be faced with. On the one hand, new devices and users entering or leaving a room may require the system to re-plan all presentations that are currently active. On the other hand, presentations running in parallel may interfere with each other, e.g. two unrelated but simultaneous audio messages. Therefore, a suitable presentation engine has to constantly reassess the current device assignment, and potentially re-plan frequently. In doing so, it must also guarantee a smooth transition from one assignment to another in order to not confuse the users. The requirement for constant re-planning also calls for a representation format that supports this process.

The potential presence of multiple users causes further problems in addition to interference. On the one hand, planning several presentations instead of a single one results in increased complexity. On the other hand, the sharing of devices (such as a large display) adds a new dimension to the planning as well as to the rendering process. An additional issue concerns the actual rendering or output of a presentation. In order to realize a coherent presentation, it is vitally important to properly synchronize the various output devices. In the synchronization process, device-specific properties such as bandwidth, memory, and speed have to be taken into account. A suitable mechanism for synchronized rendering of presentations would therefore have to provide means for prefetching and pre-timed events.

In the following, we present a basic infrastructure for device handling, a rendering mechanism for distributed presentations, and some ideas on the planning process that are aimed at addressing several of the issues raised above.

3 Device Handling

In a dynamic, instrumented environment, there are several requirements for managing the devices. One key factor to take into consideration is the fluctuation of people and their personal devices (e.g. PDAs or mobile phones), the concurrency of several applications/presentations and the differing features of similar devices (e.g. PDAs with color or monochrome displays, or different display sizes).

The task of classifying devices turns out to be rather complex. One possible approach, which was realized in the FLUIDUM project [8], is to assume a different point of view: Instead of classifying devices, we define device feature objects (e.g. video in, audio out, tag reader, etc) and then describe a device using a list of feature objects it possesses.

As underlying infrastructure, we use a device manager with two remotely accessible interfaces. On the one hand, devices can register and announce their availability; on the other hand, services can register and check for devices with certain features. The structure of the device and the architecture of the device manager are shown in Figure 1.

A device consists of a table containing parameter value pairs (for instance "name=camera01") and a collection of property objects (or "features"). The advantage of realizing these objects as remotely accessible APIs is, that we do not only know which features a device possesses, but we can also access it directly. (We decided to use Java remote method invocation (RMI) since it is an advanced technology for object marshalling and sending, but does not have the communication overhead of more sophisticated approaches such as CORBA or JINI.)

The device registers with the device manager over a "device plug adapter" (an RMI interface). The device manager keeps an up-to-date list of registered devices, and provides information to connected services that are interested in knowing about available devices. The device manager thus serves as a matchmaker between services and devices, and also passes events concerning the deregistering of devices to the services. It hence provides a lookup or yellow page service for a given environment such as a room.

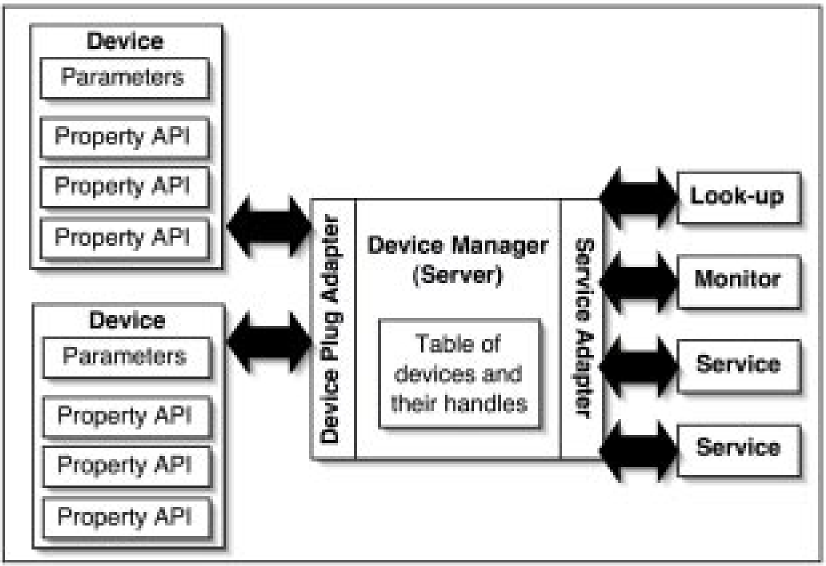


Fig. 1. Architecture of the Device Manager

4 Presentation Rendering

In the previous section, we described an infrastructure aimed at handling the (de-) registration and low-level control of various devices including output means such as displays. While this certainly provides a basis to build upon, further services are needed in order to render presentations in a coherent way. As we have seen previously, the synchronization of a presentation that is spread across multiple displays or output devices is a key concern that still needs to be addressed. Furthermore, the spatial constellation of the devices used for the presentation plays an important role.

For example, we want to avoid to generate a sophisticated presentation and to display it on a screen that is behind the user. In addition, we have to cope with the removal of some devices while a presentation is running, and thus with the incremental and continuous re-planning of the presentation. Therefore, there is a need for a representation format that not only incorporates space and time but that is also modular and can be rendered by a variety of different output devices.

We would like to propose the Synchronized Multi-Media Integration Language (SMIL) [16] to address these issues for several reasons. Firstly, the language incorporates a sophisticated concept of time allowing for the sequential and parallel rendering of various types of media. In addition, it is possible to specify the duration, the beginning and the end of the actual rendering of a part of a presentation both in absolute and relative time. Secondly, there is a similar concept of space, which supports both absolute and relative coordinates. Furthermore, it incorporates a region concept (providing local containers) as well as a simple layer concept that allows for the stacking of regions. Unfortunately, the underlying space is only two-dimensional, which is of limited use in an inherently three-dimensional scenario such as intelligent environments. However, we will point out ways to overcome this limitation later in this section.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
 <smil>
   <head>
     <meta name="title" content="example"/>
     <layout>
       <root-layout title="demo" id="root" width="240" height="300"/>
       <region title="main" height="320" id="main" z-index="1"
         width="240" top="0" left="0"/>
     </layout>
   </head>
   <body>
     <par id="par1">
       <audio begin="0s" region="main" id="audio1" src="song.wav"
         dur="10s"/>
       <seq id="seq1">
         <img begin="0s" region="main" id="img1" src="one.png"
           dur="5s"/>
         <img begin="0s" region="main" id="img2" src="two.png"
           dur="5s"/>
       </seq>
     </par>
   </body>
  </smil>
```

Fig. 2. Example SMIL presentation

A third reason supporting the adoption of SMIL for ambient presentations lies in its inclusion of various media such as text, images, sound and videos at the language level. The short example shown in Figure 2 illustrates the simplicity of including various media. The example describes a presentation that displays a slide show of two pictures (‘‘one.png” and ‘‘two.png”) that are each shown for five seconds while a sound file is playing (‘‘song.wav”). This brings us to a fifth reason for using SMIL, which is the conciseness of the format. Since it is text-based, it can even be compressed and adds very little overhead to the media-data itself. This is especially important in an intelligent environment where many devices will not have a high-speed wired connection but rather an unreliable wireless link with a lower bandwidth.

A final reason favoring the use of SMIL in ambient presentations consists of the availability of the corresponding player software on a wide range of devices. SMIL is a subset of the format used by the REAL player [15], which runs on desktop computers, PDAs and even some mobile phones. Similarly, a slightly outdated version of SMIL is part of the MPEG-4 standard [13] that is at the heart of a variety of services and media players such as QuickTime [11]. The Multimedia Message Service (MMS) [11] -- a recent addition to many mobile phone networks -- is also based on SMIL, and consequently, the current generation of mobile phones supporting MMS provide some basic SMIL rendering capabilities. Furthermore, there are several free players such as S2M2 [14] and X-Smiles [17]. For our purpose, especially the latter one is interesting as it is continuously updated and its source code is available. In addition, it has been specifically designed to run on various devices ranging from desktop computers to mobile phones.

However, before we can actually design a service for controlling and rendering SMIL on multiple output devices, we have to address the problem of SMIL only supporting two dimensions. Fortunately, this is a straightforward task. Assuming that we are dealing with bounded spaces – e.g. we do not consider interstellar space -- we can develop a function that maps the surface of a bounded space, e.g. a room, to a two-dimensional plane. Figure 3 shows an example for such a function. If we just want to deal with rectangular rooms and wall-mounted displays the mapping is very straightforward and consists basically of an `unfolding' of the corresponding three-dimensional box.

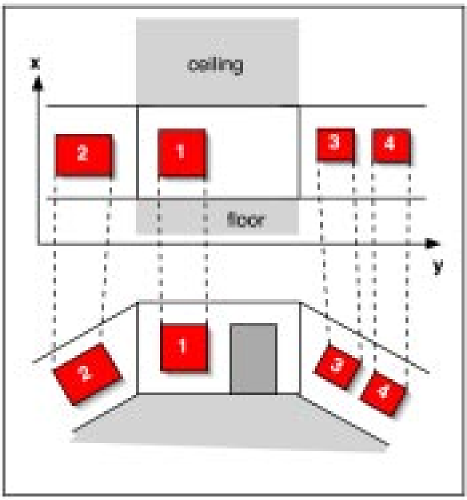


Fig. 3. Mapping of 3D surfaces to a 2D plane

Obviously, the more complex a room is, the harder it is to find an `intuitive' mapping that preserves some of the spatial properties. This is even more so, if we take into account mobile devices such as PDAs or tablet computers. However, as long as the mapping function is bijective, we can always determine which part of a SMIL presentation corresponds to a specific location in 3D space, and vice versa. Even if the mapping does not preserve spatial properties such as neighborhood relations, we can perform spatial reasoning in 3D space by re-projecting 2D parts using the mapping function. Therefore, it is safe to assume the existence of a table-based bijective mapping function based on the basic infrastructure described in the previous section. Since a service maintaining this function, i.e. the presentation-planning component, receives constant updates on the availability and location of various output devices, it can ascertain the integrity of the mapping function. Using such a mapping function, a presentation planner can generate a complete SMIL presentation covering multiple devices in a room. At this stage, we need a piece of software that takes this presentation and sends the various parts of it to the corresponding devices, and that assures that the overall presentation is delivered in a coherent way. We have implemented such a service - the SMIL Manager shown in Figure 4 - that takes care of this part of the presentation pipeline. By introducing a SMIL property object for all devices capable of rendering (parts of) a SMIL presentation, we can rely on the standard service interface for device access and control. The API of the SMIL property object implements the actions of the protocol listed in Table 1.

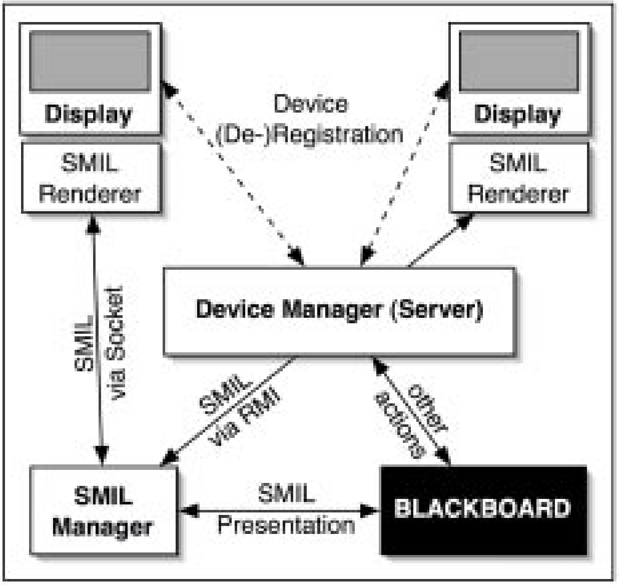


Fig. 4. Architecture of the SMIL Manager

Using the Device Manager (presented in the previous section), the SMIL Manager can hence keep track of all devices capable of rendering SMIL presentations. It relies on the mapping function to identify, which part of the overall presentation is to be rendered on which device. It then sends the various parts of the presentation to the corresponding devices for prefetching. Once all devices have reported back after completing the prefetching of media included in the presentation, the SMIL Manager triggers the simultaneous start of the presentation on all devices.[1]

However, there are a few details in the process that deserve proper attention. In order to assure a synchronized presentation, the SMIL Manager sends out synchronization messages to all attached output devices on a regular basis. Otherwise, the different clocks and time-servers that the various devices rely on would result in unwanted behaviors such as a part of the presentation starting too early or too late. Furthermore, the prefetching of media is vitally important. Especially on a shared wireless network, bandwidth is limited and can result in media not being available when they are needed in the presentation if we do not perform prefetching. But even in case prefetching is included in the process, it is important to actually check whether it has completed on *all* devices since there are great differences in terms of the time that is required to download all media. For example, a screen attached to a wired workstation will probably fetch the required media for a presentation much faster than a PDA connected through WLAN.

Table 1. Protocol for interaction between the SMIL Manager and output devices

Action	Explanation
<LOAD>	Load a presentation specified by an included URL.
<START>	Start a presentation specified by an included URL, optionally a specified time.
<STOP>	Immediately stop a presentation specified by an included URL.
<SYNCHRONIZE>	Synchronize internal clock with time stamp of SMIL Manager.
<REPORT>	Send a message whenever the user activates a link.
<LOADED>	Presentation at included URL has been fully prefetched.
<FINISHED>	Presentation at included URL has finished playing.
<LINK>	User has activated a link pointing to included URL.

Table 1 lists the various actions that the SMIL Manager can perform on the connected devices. These are either transmitted using remote method invocation (RMI) or through a socket connection using plain text messages such as the ones listed in the table. Obviously, there has to be a means to instruct a device to load a presentation, and to start it (at a given time). Additionally, it should be possible to stop a running presentation, and to synchronize the devices with the clock of the SMIL Manager. Finally, there is rudimentary support for interaction (through the <REPORT> action). When enabled, the device reports back to the Manager in case the user clicks on a link embedded in the SMIL presentation. On touch-sensitive screens, for example, this allows for simple interactions. We will discuss some implications of this feature in the

[1] Note that the simultaneous start of the presentation on all devices does not necessarily imply that all devices will actually produce output right away. More often, some will render something directly while others will wait for a predefined time before actually outputting something.

last section of this paper. Furthermore, the devices report back once the current presentation has finished playing.

The SMIL Manager as well as the client application are written in Java and have successfully been run on various desktop computers and PDAs. Since they only require version 1.1 of the Java virtual machine, we are confident that they will run on further devices in the future (such as mobile phones). SMIL rendering relies on the X-Smiles engine [26], and is realized as a dedicated thread that is separate from the communication (either through RMI or a socket connection). Therefore, both the server and the client can communicate while performing other tasks (such as playing a presentation or registering further clients). This is an important feature especially in the context of stopping a currently running application (e.g. when the device has to be used by another service, or when the presentation has to be re-planned due to the removal of some device).

5 Presentation Planning

Now that we have discussed how to control the different devices and how to handle distributed SMIL presentations, there is a final stage still missing in the presentation pipeline: the planning and generation of the presentation. As we have noted in the introduction, there is a lot of research going on in the context of single-user single-device presentations, i.e. for a person sitting in front of a desktop computer. However, when faced with a multi-device (multi-user) presentation, there are some key properties that distinguish the corresponding planning process from traditional setups. Firstly, time and space play a much more prominent role since not only are the various output devices distributed across space but also do they have to be properly synchronized. The latter point refers to the issue discussed in the previous section as well as to the planning stage since the presentation planner has to take into account when to present what on which display. Secondly, it may frequently occur that multiple users are present in a room, which entails several problems not present in single-user scenarios. For example, a number of users may share certain devices, which may interfere with one another, e.g. when one person is making a phone call while a second person uses a voice-based approach to access a database. In addition, security and privacy come into play if there is more than one person involved.

A third important differentiating factor consists of the dynamicity of the environment. While traditionally, it is assumed that the devices used for a presentation do not abruptly disappear, this assumption is not realistic in an intelligent environment. Even more so, it is possible that further output devices actually appear, e.g. a person carrying a PDA enters an intelligent room. All these factors call for an approach that facilitates re-planning on various levels, which in turn is a further difference to more static and traditional presentation planning. In order to take these special requirements into account, we distinguish three different planning problems that are interleaved with each other: (1) the planning of the *content, (2)* the planning of the *media distribution in space and time* and (3) the planning of the *attentional shift* (*focus control)* that has to be supported when users have to focus on different displays in their environment over time.

The first problem is similar to content selection in classical intelligent multimedia presentation systems designed for single display scenarios (e.g. [9]). Hence, we will not discuss it here, as it can be addressed using a classical plan operator based planning technique such as STRIPS [10]. For the planning of the spatio-temporal distribution we plan to adopt a constraint-based mechanism described in [6], where the rendering abilities of each display (e.g. resolution, audio quality, interaction requirements) is modeled and scored with a certain value. The rendering of a video on a PDA display would for example achieve a lower score than on a big plasma screen. Although most of the constraints are soft (e.g. it is possible to render a video on a PDA), some are not: if the presentation contains interactive elements (e.g. buttons) it is not possible to render them on a screen without input capabilities. In this case a plasma screen without a touch screen overlay for example could not be used. One possible solution would consist of rendering the interactive parts of the presentation on a device with a touch-screen (e.g. a PDA) and the rest of the presentation on the plasma screen.

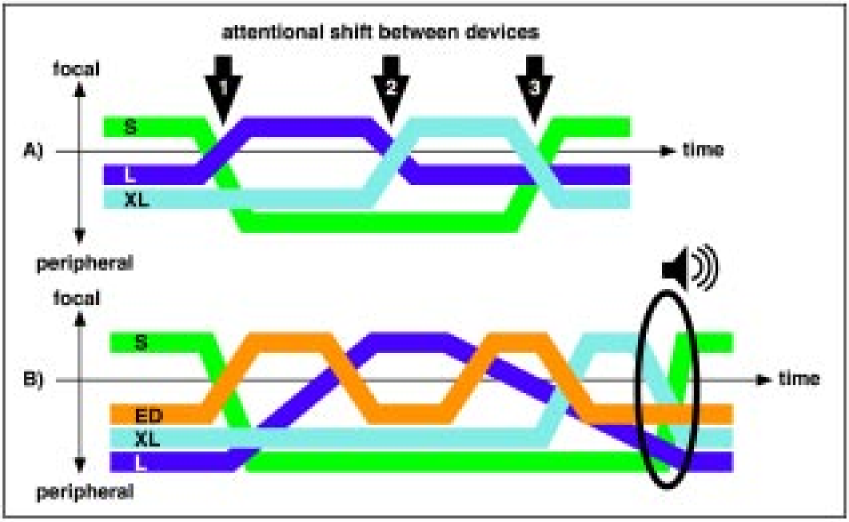


Fig. 5. A distributed presentation: A) without focus control, and B) with focus control using an everywhere display and audio meta-comments

From the users perspective such a distributed presentation can cause problems, if they do not know where to focus their attention at a certain moment of time. One way to overcome this problem is illustrated in Figure 5. The diagram in Figure 5 A) shows an example of a presentation distributed over time on a small ('S' - e.g. a PDA), a large ('L' - e.g. a 17-inch touch-screen) and an extra large ('XL' - e.g. a plasma screen) device. From the user's perspective devices in the environment can be either in focus or peripheral. The three arrows in Figure 5 A) indicate that the user has to perform an attentional shift, whenever a device switches from the focus role to a peripheral role and vice versa. The user's ability to perform this shift may be hindered by several factors, e.g. the spatial distance between the devices and distracting cues in the environment (e.g. noise). If a user is not able to shift the attention to the right device at the right moment in time, the probability increases of the presentation becomes confusing or, even worse, that it is misinterpreted.

Therefore it makes sense to support this shift of attention in the instrumented environment. In Figure 5 B) two additional means are used to facilitate the attentional shift. The shifts 1 and 2 are supported by an everywhere display [18] (ED). An ED is a device that is able to project images on arbitrary surfaces in the environment, i.e. a ceiling-mounted projector attached to a movable stand that can be controlled remotely. Therefore it can be used to guide the attention of the user from one place in the instrumented environment to another place in the environment. If the environment knows the user's position, an ED can even be used to guide the attention from a PDA that is carried by the user towards a fixed display in the environment. Other means to shift the user's attention can be meta-comments that explicitly tell the user where to look next (e.g.: "To stop the video please press the corresponding button on your PDA"). In Figure 5 B) such an audio meta-comment is used to close the third "attentional gap".

We can extend the constraint-based planning approach by introducing another score that represents the difficulty of the users to direct their attention from one display to another in a given situation. Of course these additional elements of the presentation increase the overall presentation time leading to other problems that may result into a backtracking process and a modified distribution of the presentation.

6 Conclusion and Outlook

In this paper, we discussed several issues and possible solutions in the context of generating coherent presentations in an intelligent environment. We presented an analysis of the task of generating a coherent presentation across multiple displays, namely the planning, the rendering and the control of the output devices. We reviewed several problems and requirements such as synchronization and three-dimensionality, and introduced a multi-layer approach aimed at addressing some of the previously identified issues. We described a low-level architecture for the handling of devices as well as mechanism for distributing coherent presentations to multiple displays using the Synchronized Multimedia Interface Language (SMIL). We then discussed the issues related to presentation planning, and provided some ideas for implementing the corresponding process.

Consequently, a major part of future work consists in actually designing a working system for the planning process, and to test it with various services such as navigation, localized information or smart door displays. A second field of major interest is the capture of user interaction. We have already started on this point by incorporating a simple link following action in the protocol used to connect the SMIL Manager and its clients. However, the underlying Device Manager allows for a much more fine-grained interaction, which we intend to include in future versions of the system.

Acknowledgements. The authors would like to acknowledge their funding though the Deutsche Forschungsgemeinschaft (DFG) within the Special Research Center 378: Resource-Adaptive Cognitive Processes and the Fluidum research group.

References

1. T. L. Pham, G. Schneider, S. Goose: A Situated Computing Framework for Mobile and Ubiquitous Multimedia Access Using Small Screen and Composite Devices, in Proc. of the ACM International Conference on Multimedia, ACM Press, 2000
2. Roy Want, Trevor Pering, Gunner Danneels, Muthu Kumar, Murali Sundar, and John Light: The Personal Server: Changing the Way We Think about Ubiquitous Computing, Proc. of Ubicomp 2002, LNCS 2498, Springer, 2002, pp. 192
3. Scott Robertson, Cathleen Wharton, Catherine Ashworth, Marita Franzke: Dual device user interface design: PDAs and interactive television, Proc. of CHI'96, ACM Press, 1996,pp. 79
4. Yasuyuki Sumi , Kenji Mase: AgentSalon: facilitating face-to-face knowledge exchange through conversations among personal agents, Proc. of Autonomous agents 2001, ACM Press, 2001.
5. Brad A. Myers: The pebbles project: using PCs and hand-held computers together, Proc. of CHI 2000, ACM Press, 2000, pp. 14
6. Antonio Krüger, Michael Kruppa, Christian Müller, Rainer Wasinger: Readapting Multimodal Presentations to Heterogeneous User Groups, Notes of the AAAI-Workshop on Intelligent and Situation-Aware Media and Presentations, Technical Report WS-02-08, AAAI Press, 2002
7. Michael Kruppa and Antonio Krüger: Concepts for a combined use of Personal Digital Assistants and large remote displays, Proceedings of SimVis 2003, SCS Verlag, 2003
8. The Fluidum project. Flexible User Interfaces for Distributed Ubiquitous Machinery. Webpage at http://www.fluidum.org/
9. E. Andre, W. Finkler, W. Graf, T. Rist, A. Schauder and W. Wahlster: "WIP: The Automatic Synthesis of Multimodal Presentations" In: M. Maybury (ed.), Intelligent Multimedia Interfaces, pp. 75–93, AAAI Press, 1993
10. R. E. Fikes and N. J. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. Artificial Intelligence, 2: 198–208, 1971
11. The 3rd Generation Partnership Project (3GPP). The Multimedia Message Service (MMS). Available at http://www.3gpp.org/ftp/Specs/html-info/22140.htm
12. Apple Computer Inc. The QuickTime player. Available at http://www.apple.com/quicktime
13. MPEG-4 Industrial Forum. The MPEG-4 standard. Available at http://www.m4if.org
14. The National Institute of Standards and Technology. The S2M2 SMIL Player. Available at http://smil.nist.gov/player/
15. RealNetworks. The REAL player. Available at http://www.real.com
16. The World Wide Web Consortium. The synchronized multimedia integration language (SMIL). Available at http://www.w3.org/AudioVideo/
17. X-Smiles. An open XML-browser for exotic devices. Available at http://www.xsmiles.org
18. Claudio Pinhanez: The Everywhere Displays Projector: A Device to Create Ubiquitous Graphical Interfaces, Proc. of Ubicomp2001, Volume 2201, Lecture Notes in Computer Science, 2001, pp 315

Vision-Based Localization for Mobile Platforms

Josep M. Porta and Ben J.A. Kröse

IAS Group, Informatics Institute, University of Amsterdam,
Kruislaan 403, 1098SJ, Amsterdam, The Netherlands
{porta,krose}@science.uva.nl

Abstract. In this paper, we describe methods to localize a mobile robot in an indoor environment from visual information. An *appearance*-based approach is adopted in which the environment is represented by a large set of images from which features are extracted. We extended the appearance based approach with an *active vision* component, which fits well in our probabilistic framework. We also describe another extension, in which depth information is used next to intensity information. The results of our experiments show that a localization accuracy of less then 50 cm can be achieved even when there are un-modeled changes in the environment or in the lighting conditions.

1 Introduction

Localization and tracking of moving objects is one of the central issues in research in intelligent environments. In many cases, persons need to be localized for security or services. Also the localization of movable intelligent embedded systems is an important issue in wireless local networks or for localizing Personal Digital Assistants (PDA's) which can serve, for instance, as museum guides. In this paper, we focus on the localization of a robot platform (the domestic service robot 'Lino' [1], [4] (see Figure 1), developed within the 'Ambience' project), but the described techniques are applicable in many other domains.

Different solutions have been presented for localizing mobile objects. One class of solutions is to use sensors placed in the environment. Such sensors may be cameras, infra-red detectors or ultrasonic sensors. The main problem here is the *identity uncertainty*: the cameras (or other systems) have to infer the identity of the object/person to be localized from the sensor readings and this is, in general, difficult. For instance, the camera-based localization system described in [6] fails when tracking more than three persons. A second class of solutions is to use radio-frequency signals. These signals are naturally present in mobile computing networks, and their identification is trivial. However, accurate localization is difficult due to reflections, absorption and scattering of the radio waves. A localization accuracy of approximately 1.5 meters is reported using this approach combined with a Hidden Markov Model [7].

Localization without any active beacons or without any environment sensor traditionally takes place in the research area of autonomous robots [2], [5].

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 208–219, 2003.

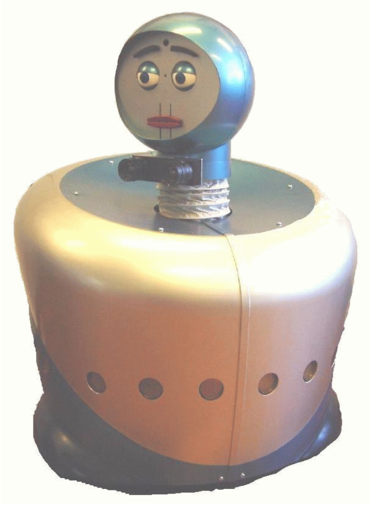


Fig. 1. The user-interface robot Lino.

Generally, robot localization needs some sort of internal model of the environment based on sensor readings from which the pose (position and orientation) of the robot is estimated. Traditionally range sensors (ultrasonic or laser scanners) were used, but recently much progress has been achieved in localization from vision systems. Although the methods show good results, they are not robust to un-modeled changes in the environment and they are quite sensitive to changes in illumination. To cope with these problems, in this paper we describe two modifications to our localization system: an *active* vision method and the use of depth information for localization These two improvements are embedded in a probabilistic framework. In the following section, we will briefly describe our way of modeling the world and, next, our probabilistic localization framework. Then, we introduce our active vision method and the use of depth information for localization. After this, we described the experiments we performed to validate our contributions and the results we obtained from them.

2 Appearance-Based Environment Representation

In the literature on mobile robots, methods for environment representation come in two flavors: *explicit* or *implicit*. The explicit representations are based on geometric maps of free space sometimes augmented with texture information, i.e., CAD models, or maps with locations of distinct observable objects called landmarks. This approach relies on the assumption that geometric information such as the position of obstacles/landmarks can be extracted from the raw sensors readings. However, the transformation from sensor readings to geometric information is, in general, complex and prone to errors, increasing the difficulty of the localization problem.

As a counterpart, the *implicit* (or *appearance-based*) representation of the environment [8] has attracted lot of attention recently. In this paradigm, the environment is not modeled geometrically but as an 'appearance map' that consists of a collection of sensor readings obtained at known poses. The advantage of this representation is that the pose of the robot can be determined directly comparing the sensor readings obtained at a given moment with those in the appearance-based map.

We use a vision-based appearance representation built with many images from the environment. A problem with images is their high dimensionality, resulting in large storage requirements and high computational demands. To alleviate this problem, Murase and Nayar [8] proposed to compress images, z, to low-dimensional feature vectors, y, using a linear projection

$$y = W\,z. \tag{1}$$

The projection matrix W is obtained by Principal Component Analysis (PCA) of a supervised training set ($T = \{(x_i, z_i)|\; i \in [1, N])\}$) consisting of images z_i obtained at known poses x_i. We keep the subset of eigenvectors that represent most of the variance of the images and we use them as rows of the projection matrix W. After the dimensionality reduction, the map used by the robot for localization $M = \{(x_i, y_i)|\; i \in [1, N])\}$ consists of a set of low-dimensional (typically around 20-D) feature vectors y_i corresponding to the images taken at poses x_i. The use of features instead of raw images saves a large amounts of memory space.

For localization, the robot first has to project the image which is observed at the unknown location to a feature representation. Then, the probabilistic model described next is used to localize the system.

3 A Probabilistic Model for Localization

The camera system is mounted on a pan-tilt device, rigidly attached to the mobile robot. We assume that the orientation of the camera with respect to the robot is given with sufficient accuracy by the pan-tilt device. The absolute pose of the camera is considered as a stochastic (hidden) variable x. The localization method aims at improving the estimation of the pose x_t of the camera at time t taking into account the movements of the robot/head $\{u_1, \ldots, u_t\}$ and the observations of the environment taken by the robot $\{y_1, \ldots, y_t\}$ up to that time[1]. Formally, we want to estimate the posterior $p(x_t|\{u_1, y_1, \ldots, u_t, y_t\})$. The Markov assumption states that this probability can be updated from the previous state probability $p(x_{t-1})$ taking into account only the last executed action, u_t, and the last observation, y_t. Thus we only have to estimate $p(x_t|u_t, y_t)$. Applying Bayes we have that

$$p(x_t|u_t, y_t) \propto p(y_t|x_t)\, p(x_t|u_t), \tag{2}$$

[1] In our notation, the Markov process goes through the following sequence: $x_0 \xrightarrow{u_1} (x_1, y_1) \xrightarrow{u_2} \ldots \xrightarrow{u_t} (x_t, y_t)$.

where the probability $p(x_t|u_t)$ can be computed propagating from $p(x_{t-1}|u_{t-1}, y_{t-1})$

$$p(x_t|u_t) = \int p(x_t|u_t, x_{t-1})\, p(x_{t-1}|u_{t-1}, y_{t-1})\, dx_{t-1}. \tag{3}$$

We discretize equation 3 using an *auxiliary particle filter* [9]. In this approach, the continuous posterior $p(x_{t-1}|u_{t-1}, y_{t-1})$ is approximated by a set of I random samples, called particles, that are positioned at points x^i_{t-1} and have weights π^i_{t-1}. Thus, the posterior is

$$p(x_{t-1}|u_{t-1}, y_{t-1}) = \sum_{i=1}^{I} \pi^i_{t-1}\, \delta(x_{t-1}|x^i_{t-1}), \tag{4}$$

where $\delta(x_{t-1}|x^i_{t-1})$ represents the delta function centered at x^i_{t-1}. Using this, the integration of equation 3 becomes discrete

$$p(x_t|u_t) = \sum_{i=1}^{I} \pi^i_{t-1}\, p(x_t|u_t, x^i_{t-1}), \tag{5}$$

and equation 2 reads to

$$p(x_t|u_t, y_t) \propto p(y_t|x_t) \sum_{i=1}^{I} \pi^i_{t-1}\, p(x_t|u_t, x^i_{t-1}). \tag{6}$$

The central issue in the particle filter approach is how to obtain a set of particles (that is, a new set of states x^i_t and weights π^i_t) to approximate $p(x_t|u_t, y_t)$ from the set of particles x^i_{t-1}, $i \in [1, I]$ approximating $p(x_{t-1}|u_{t-1}, y_{t-1})$. In [9] and [12], you can find details on how this is achieved in the approach we use.

The probability $p(x_t|u_t, x_{t-1})$ for any couple of states and for any action is called the *action model* and it is inferred from odometry. On the other hand, $p(y_t|x_t)$ for any observation and state is the *sensor model*. This probability can be approximated using a nearest-neighbor model that takes into account the J points in the appearance-based map that are more similar to the current observation (see [12]).

4 Active Localization

Traditionally, appearance-based localization has problems in dynamic environments: modifications in the environment are not included in the model and can make recognition of the robot's pose from the obtained images very difficult. To alleviate this problem, we introduce the use of an active vision strategy. Modifications in the environment would only be relevant if the camera is pointing towards them. If this is the case, we can rotate the cameras to get features in other orientations hopefully not affected by the environment changes. Therefore,

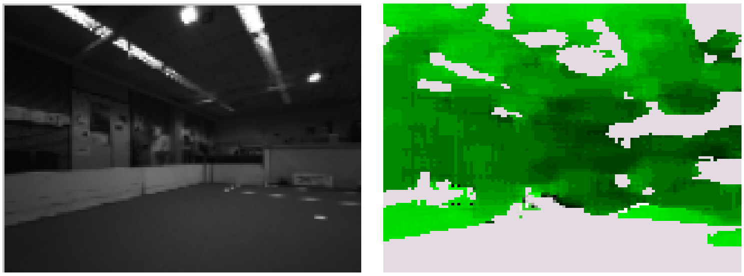


Fig. 2. Plain image (left) and the corresponding disparity map (right). In the disparity map, light gray areas are missing values.

in case of observations that do not match with those in the appearance-based map, the location of the robot can be found out efficiently by issuing the adequate sequence of camera rotations. The question here is how to determine this *adequate* sequence of camera movements.

In [10], we present a method based on the minimization of the estimated entropy $H(u)$ of the stochastic variable x if action u was executed. We describe how to approximate the entropy $H(u)$ by taking advantage of the two main components of our localization system: the particle filter (to get a discrete set of the possible placements of the robot after executing action u) and of the appearance-based map (to get a discrete set of possible observations after executing action u). These two approximations allow to discretize the computation of the entropy $H(u)$, making it feasible.

5 Adding Depth Information

Vision-based appearance localization is sensitive to illumination conditions. An adequate preprocessing of the images such histogram equalization makes the system more robust, but does not solve all problems. Therefore, we decided to complement the visual information with depth information.

For that, we used a commercially available stereo system [3], that provides information of the distance to the nearest object for each pixel in the form of a disparity value obtained matching pixels from the two stereo images. The algorithm we use applies many filters in this matching process both to speed it up and to ensure the quality of the results. For instance, if the area around a given pixel is not textured enough it would be very difficult to find a single corresponding point in the other image: we are likely to end up with many pixels with almost the same probability of being the corresponding point to the pixel we are trying to match. For this reason, pixels on low textured areas are not even considered in the matching process. The result of this and other filtering processes is to produce a sparse disparity map: a disparity map where many pixels don't have a disparity value (see Figure 2). This makes the use of standard PCA to determine the projection matrix unfeasible and we have to use more elaborated techniques such as the EM algorithm we introduced in [11].

Once we have a way to define features from disparity maps, it remains the question of how to combine the information coming from disparity with that obtained from intensity to defined a unified sensor model. Two possible solutions come to mind: to combine them in a conjunctive way or in a disjunctive one.

A conjunctive-like combination can be achieved factorizing the sensor model

$$p(y_d, y_i|x) = p(y_d|x)\, p(y_i|x), \tag{7}$$

with y_d the features obtained for disparity and y_i those for intensity. In this way, only those training points consistent with both the current intensity image and the current disparity map are taken into account to update the robot's position. The problem of this formulation is that wrong matches for intensity or for disparity would result in an almost null sensor model and, thus, the position of the robot would be updated almost without sensory information.

To avoid this, we propose to use a disjunctive-like model that can be implemented defining the global sensor model as linear combination of the intensity and disparity sensors models

$$p(y_d, y_i|x) = w_d \sum_{j=1}^{J} \lambda_j\, \phi(x|x_j) + w_i \sum_{j=1}^{J'} \lambda'_j\, \phi(x|x'_j), \tag{8}$$

where x_j and x'_j are the training points with features more similar to those of the current disparity and intensity image respectively. The weights w_d and w_i can be used to balance the importance of the information obtained with each type of sensor. If both information sources are assumed to be equally reliable, we can set $w_d = w_i = 1/2$.

With this expression for the sensor model, all the hypotheses on the robot position suggested by the current observation are taken into account. The particle filter [9] [12] we use to update the probability on the robot's position takes care of filtering the noise and, thus, of preserving the hypothesis that is more consistent over time.

6 Experiments and Results

In this section, we describe the experiments we performed to validate the our contributions both on active localization and on localization using disparity maps.

6.1 Experiments on Active Localization

We tested our action evaluation system in an office environment. We mapped an area of 800 × 250 cm taken images every 75 cm and every 15 degrees. This makes a total amount of about 400 training images. The short distance between training points make images taken at close positions/orientations to look very

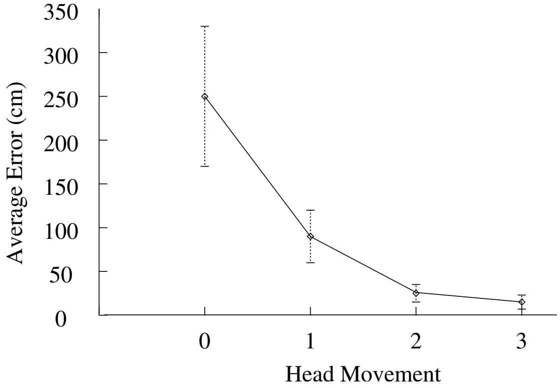


Fig. 3. Evolution of the average error (and the standard deviation) w.r.t. the correct position as we get new images.

similar. In the experiments, we compress the images using PCA keeping 5 feature detectors, we use 10 nearest neighbors to approximate the sensor model $p(y|x)$, and we define the initial distribution $p(x_0)$ uniformly over the configuration space of the robot. We considered 22 different orientations for the camera and we used up to 150 particles to approximate $p(x_t|u_t, y_t)$.

We tested the system placing the robot at positions not included in the training set, rotating the camera as measuring the error and $||c - a||$ with c the correct position and a the position estimated by our system.

Figure 3 shows the decrease on the average positioning error as new actions are issued. The results shown correspond to the average and the standard deviation over ten runs placing the camera in two different testing positions. We can see that the entropy-based action selection allows a fast reduction of the localization error as the head is moved and new images are processed. If we consider the estimation a to be correct if the closest training point to a is the same as the closest training point to the correct position c, then the success ratio in localization after 3 camera movements is over 95%.

Figure 4 shows a typical evolution of particles from a distribution around different hypothesis (left) to the convergence around the correct position (right) achieved as new images are processed. In the figure, each '>' symbol represents a particle (the brighter the larger its weight) and the triangle represents the pose of the camera. The circle represents the standard deviation of particles in the XY plane.

6.2 Experiments on Localization Using Disparity Maps

To test the invariance of the feature detectors obtained from disparity maps to changes in illumination we acquired an appearance-based map in a area of 900 × 500 cm. Images where collected every 50 cm (both along X and Y) and every 10 degrees. This makes a total amount of about 4000 images.

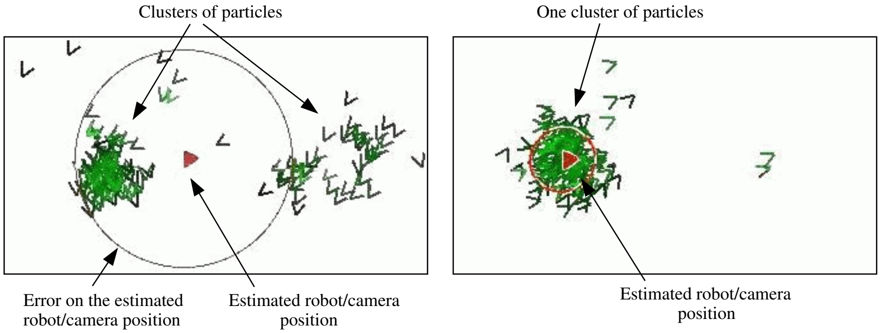


Fig. 4. Convergence toward the correct position estimation: particles around different hypotheses (left) and particles around a single correct hypothesis (right).

Table 1. Ratio of average feature detector variation due to illumination vs. the changes due to a small translations.

Image	**Illumination Setups**		
Process	**Bulb Lights**	**Natural Light**	**Average**
Plain Images	3.85	4.64	4.24
Hist. Equalization	1.50	1.84	1.67
Gradient Filter	1.11	1.37	1.24
Disparity Map	0.68	0.79	0.73

We analyzed the sensitivity of the feature detectors to two of the different factors that can modify them: translations and changes in illumination. For this, we compute the ratio

$$r(a,b,c) = \frac{||y_c - y_a||}{||y_b - y_a||}, \tag{9}$$

with y_a the feature detectors of image at pose a with the illumination setup used to collect the appearance map (tube lights), y_b the feature detectors of the image obtained with the same orientation and the same lighting conditions but 50 cm away from a, and y_c the image obtained at pose a but in different illumination conditions. The greater this ratio, the larger the effect of illumination w.r.t. the effect of translations and, thus, the larger the possible error in localization due to illumination changes.

We used two illumination setups for the test: bulb lights and natural light (opening the curtains of the windows placed all along one wall of the lab). These two tests sets provide changes both in the global intensity of the images and in the distribution of light sources in the scene, that is the situation encountered in real applications.

We computed the above ratio for the feature detectors obtained from plain images, from disparity maps and also for images processes with two usual tech-

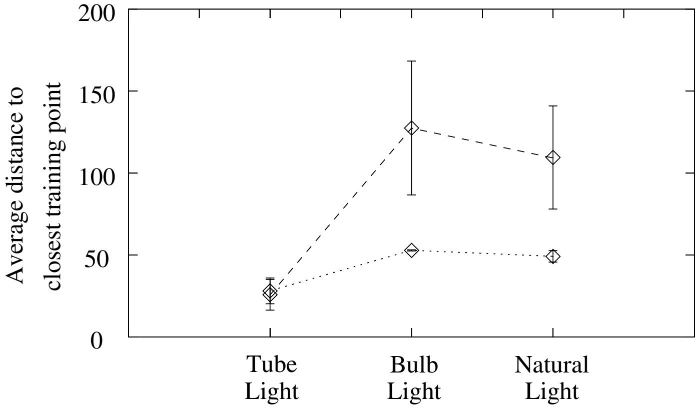


Fig. 5. Error in positioning in three different illumination conditions, using only intensity images (dashed line) and intensity and disparity images (dotted line).

niques for dealing with illumination related problems: histogram equalization and a gradient-based filter.

Table 1 shows the results we obtained for the experiment just described. In this table, we can see that, as expected, using processed images the ratio decreases considerably compared with the ratio using images without any illumination correction. In the case of disparity maps, this ratio is the smallest one meaning that disparity is the best of the three techniques we tested, as far as independence of illumination is concerned.

To assess the contribution of using disparity maps in appearance-based localization, we moved the robot along a pre-defined path in the three different illumination conditions mentioned above: tube lights, bulb lights and natural light. At regular distances along the test path, we took an image and we computed the corresponding sensor model using the J training points with features more similar to those corresponding to the just obtained image. The closer the training points used to define the sensor model to the actual position of the robot, the better the sensor model and, thus, better the update of the robot position estimation.

Figure 5 shows the average and the variance for all the test points all along the path of the error defined as

$$e = \min_{\forall n} \|r - n\|, \tag{10}$$

with $r = (r_x, r_y, r_\phi)$ the poses of the robot at the test position and $n = (n_x, n_y, n_\phi)$ the pose of the points used to define the sensor model, that are different for each test position. An error in the range $[25, 50]$ is quite reasonable since the distance between training points in X and Y dimensions is 50 cm.

We repeat the test in two cases: (a) using only intensity images (dashed line on Figure 5) and (b) using, additionally, disparity maps (dotted line on the figure). In the first case we use $J = 10$ and in the second case we use $J = 5$ but for both intensity and disparity so, we also get 10 nearest-neighbors.

We can see that the use of disparity maps results in a reduction of the error in the sensor model when illumination is different from that in which the training set was obtained (tube lights). Consequently, the use of feature detectors computed from disparity maps increase the quality of the sensor model and, thus, it helps to obtain a more robust localization system.

7 Conclusions

In this paper, we have introduced two extensions to the traditional appearance-based robot localization framework: active selection of robot actions to improve the localization and used of disparity maps for localization. These two extensions are possible thanks to the use of a stereo camera mounted on the mobile head of the service robot Lino.

The experiments we report with our active vision system show that this mechanism effectively helps to find out the location of the robot. This is of great help in dynamic environments, where existing appearance-based localization system exhibited some problems.

Our second contribution is the use of sparse disparity maps to increase the robustness of appearance-based localization to changes in illumination. The results we have presented show that disparity maps provide feature detectors that are less sensible to changes in the lighting conditions than feature detectors obtained from images processed with other techniques: histogram equalization and gradient-based filters. These techniques work well when we have changes in the global illumination but they do not deal properly with different distributions of light sources. Disparity maps are more consistent over changes in the number and in the position of the light sources because only reliable correspondences are taken into account when defining the disparity map. These reliable matches are likely to be detected in different lighting conditions.

We have shown that, using features from disparity maps in addition to those obtained from intensity images, we can improve the quality of the sensor model when illumination conditions are different from those in which the training set is obtained. Thus, disparity maps are a good option to increase the robustness of appearance-based robot localization.

The good results achieved using disparity maps cames at the cost of using a more complex hardware (we need not only one camera but two calibrated ones) and software (the disparity computation process is more complex than the histogram equalization and the gradient filter processes).

The main assumption behind our approach is the existence of a training set obtained off-line and densely sampled over the space where the robot is expected to move. To obtain this training set is not a problem, but it would be desirable the robot to build it on-line. To achieve this improvement, we have to explore the use of incremental techniques to compress the images obtained as the robot moves in the environment.

The type of environment representation underlying our localization system is computationally very cheap: after the dimensionality reduction the appearance-based map is small and the linear projection of images is an efficient process. For this reason, the localization system could be easily implemented in fields other than autonomous robots as, for instance, PDA's or mobile phones provided with cameras. Since the localization would be performed by the same device to be localized, we avoid the *identity uncertainty* problem, and, additionally, the accuracy in localization that could be achieved is better than that reported using radio-frequency techniques.

Acknowledgments. This work has been partially supported by the European (ITEA) project *"Ambience: Context Aware Environments for Ambient Services"*.

We would like to thank Bas Terwijn for helping us to perform the experiments reported on this paper.

References

1. A.J.N. van Breemen, K. Crucq, B.J.A. Kröse, M. Nuttin, J.M. Porta, and E. Demeester A user-interface robot for ambient intelligent environments, In Proceedings of the 1st International Workshop on Advances in Service Robotics (ASER), Bardolino, March 13-15, pages 132–139, 2003.
2. D. Fox, W. Burgard, and S. Thrun Markov localization for Mobile Robots in Dynamic Environments, Journal of Artificial Intelligence Research, 11:391–427, 1999.
3. K. Konolige Small Vision System: Hardware and Implementation, In Proceedings of the 8th International Symposium on Robotics Research, Japan, 1997.
4. B.J.A. Kröse, J.M. Porta, K. Crucq, A.J.N van Breemen, M. Nuttin, and E. Demeester Lino, the User-Interface Robot, In First European Symposium on Ambience Intelligence (EUSAI), 2003.
5. B.J.A. Kröse, N. Vlassis, R. Bunschoten, and Y. Motomura A Probabilistic Model for Appearance-based Robot Localization, Image and Vision Computing, 19(6):381–391, April 2001.
6. J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer Multi-Camera Multi-Person Tracking for EasyLiving, In IEEE Workshop on Visual Surveillance, pages 3–10, July 2000.
7. A. Ladd, K. Bekris, G. Marceau, A. Rudys, D. Wallach, and L. Kavraki Using Wireless Internet for Localization, In Proceedings of the International Conference on Robotics and Intelligent Systems (IROS), Las Vegas, USA, pages 402–408, October 2002.
8. H. Murase and S.K. Nayar Visual Learning and Recognition of 3-D Objects from Appearance, International Journal of Computer Vision, 14:5–24, 1995.
9. M.K. Pitt and N. Shephard Filtering Via Simulation: Auxiliary Particle Filters J. Amer. Statist. Assoc., 94(446):590–599, June 1999.
10. J.M. Porta, B. Terwijn, and B.J.A. Kröse Efficient Entropy-Based Action Selection for Appearance-Based Robot Localization In Proceedings of the International Conference on Robotics and Automation (ICRA), Taiwan, 2003.

11. J.M. Porta, J.J. Verbeek, and B.J.A. Kröse Enhancing Appearance-Based Robot Localization Using Non-Dense Disparity Maps In Proceedings of the International Conference on Robotics and Intelligent Systems (IROS), Las Vegas, USA, 2003.
12. N. Vlassis, B. Terwijn, and B.J.A. Kröse Auxiliary Particle Filter Robot Localization from High-Dimensional Sensor Observations In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Washington D.C., USA, pages 7–12, May 2002.

Multi-sensor Activity Context Detection for Wearable Computing

Nicky Kern[1], Bernt Schiele[1], and Albrecht Schmidt[2]

[1] Perceptual Computing and Computer Vision
ETH Zurich, Switzerland
{kern, schiele}@inf.ethz.ch
[2] Media Informatics Group
University of Munich, Germany
albrecht.schmidt@acm.org

Abstract. For wearable computing applications, human activity is a central part of the user's context. In order to avoid user annoyance it should be acquired automatically using body-worn sensors. We propose to use multiple acceleration sensors that are distributed over the body, because they are lightweight, small and cheap. Furthermore activity can best be measured where it occurs. We present a hardware platform that we developed for the investigation of this issue and results as to where to place the sensors and how to extract the context information.

1 Introduction

For wearable, context-aware applications there are many ways to characterize the user's context. Depending on the application the user's physical activity, his state (e.g. stressed/nervous, etc.) his interaction with others or his location might be of interest. Among all these the user's physical activity is of prime importance. E.g. knowing that the user is writing on a white board tells the application, that he is most likely involved in a discussion with other people and may not be disturbed. Similarly, when the user is sitting and typing on a computer keyboard he is probably working, but may be more open for interruptions.

Since the goal of context-aware applications is to reduce the load of the user and adapt to them seamlessly, context information cannot be supplied by the user. Instead it should be sensed automatically using sensors. While it would be possible to incorporate sensors in the environment, this would make it impossible to use the context information for mobile devices and applications outside the 'augmented' environments. For the recognition in a mobile setting, the sensors should be attached to the body of the user. This also allows for very cheap sensing, since activity is measured directly where it occurs.

For truly wearable applications, the sensors have to satisfy two basic requirements. Firstly they should be unobtrusive to wear, ideally integrated into clothing, so that the user does not need to worry taking them with him. Secondly, they should be small and cheap, so that they can be integrated in many pieces of clothing or wearable devices without adding too much to the cost and size. We

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 220–232, 2003.

propose to use miniaturized accelerometers. They can be produced in MEMS technology making them both very small and cheap. Already today, there are devices available that integrate them [1].

For the application of accelerometers to activity recognition there are two principal questions to answer: firstly, how many sensors are required to recognize a given activity with a desired precision, and secondly where to place these sensors on the user's body.

For general activities, a single sensor will not be sufficient. Accelerometers measure motion, which can only be sensed where it occurs. E.g. the activity 'writing on a white board' includes standing in front of the board, which can well be measured at the legs, and writing on it, which is an activity of the right (or left) hand. Hence, for activities of a certain degree of complexity multiple sensors will be required.

There are two principal contributions in this paper. Firstly we have developed a hardware platform, that allows to record acceleration data from many places on the human body simultaneously (section 3). Using this platform and a naïve Bayes classifier (section 4) we conducted experiments to investigate the required number of sensors and their placements (sections 5 and 6). The paper concludes with a summary and discussion of the findings and future work (section 7).

2 Related Work

Recognizing general human activity or special motions using body-worn acceleration sensors is not a new problem. Apart from the extraction of the actual activity or motion, there are also interesting applications, that use this technology.

Recognizing general user activity has been tried by various authors. Randell and Muller [2] and Farringdon et al. [3] have done early investigations of the problem using only single 2-axis accelerometers. Van Laerhoven et al. [4] try to distinguish user-selected activities using a larger number (32) of 3D acceleration sensors. Unlike in our approach, they try to find recurring patterns in the data and do not model the activity explicitely. Furthermore they do not assume that the sensors have any fixed location, instead their sensors are attached to the clothing and can thus move relative to the user's body. Kern et al. [5] model activities explicitly, but use relatively few sensors and do not address placement issues in their applications. [6] compares the both approaches. Loosli et al. [7] use two acceleration sensors attached to the user's knees to investigate classification issues.

There are also more specialized applications for motion recognition using body-worn acceleration sensors. Chambers et al. [8] have attached a single acceleration sensor to the user's wrist to automatically annotate Kung-Fu video recordings. They focus on the recognition of complex gestures using Hidden Markov Models. Benbasat and Paradiso [9] have used a 3D accelerometer and a 3D gyroscope for recognizing human gestures.

Body-worn acceleration sensors have been used in a variety of applications. Sabelman et al. [10] use them for offline analysis of the user's sense of balance. Morris and Paradiso [11] use a variety of sensor that are built into a shoe for online gait analysis. Golding and Lesh [12] and Lee and Mase [13] both use a variety of different sensors for location tracking. Kern et al. [5] employ multiple acceleration sensors to recognize the user's activity and use that information to annotate meeting recordings. Kern et al. [14] use similar information from a single acceleration sensor to mediate notifications to the user of a wearable computer depending on his context.

3 Acquisition System

In order to acquire useful information in real settings it is inevitable to design, construct and build a wearable sensing platform.

3.1 Use Cases and Requirements

Before constructing the platform we considered potential application domains in which we would like to be able to use the platform. We have seen in previous work that it is feasible to construct systems for use in lab environments [9]. However we wanted to build a platform that allows recording and recognition beyond the lab in real world environments. The anticipated usage scenarios are in the domain of sports (e.g. climbing a wall, playing a badminton match, inline skating, long distance running, and playing a basket ball match) and manual work (delivering goods, rescue workers, production workers). The following set of requirements – many of them more practical as technical – was extracted.

Robustness & Durability. When assessing the scenarios we realized that in all cases a robust and durable platform is paramount.

Mounting Sensors. Attaching the sensors to the body at a desired position and fix them to keep them in this position during an activity became a further vital issue on which the practical usability relied.

Freedom of Movement. In most scenarios it is important not to restrict the user's degree and range of freedom with respect to movements.

Time and Storage. To create useful data sets we recognize that the time interval over which data can be logged has to be fairly long. In our case we decided that we require logging capabilities for more than one hour and potentially for several hours.

Sampling Rate. Based on our own previous experience and work published [4] we aimed for a sampling rate of about 100 Hz per sensor.

Number of Sensors. For estimating a useful number of sensors we assumed having three dimensions of acceleration at each larger body segment. As initial target we set 48 accelerometers and the potential for 192 accelerometers.

Energy Consumption. As the user should be able to work or to do sports with the device over a longer period of time the energy consumption must be low enough to not jeopardize other requirements by the size and weight of batteries.

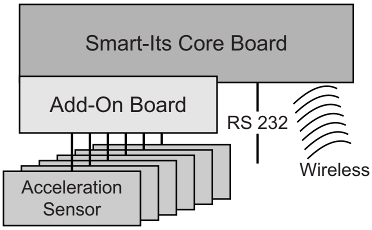


Fig. 1. Architecture of an acquisition module based on one Smart-It

The platform was deliberately designed as an experimental setup. Its main function is to provide us with real data recorded during a particular activity. Thus, properties such as robustness and durability and ease of mounting sensors had a higher influence on the design than unobtrusiveness.

3.2 Components and Architecture

The acquisition system is a modular design based on the smart-its platform [15, 16]. The smart-its core board provides communication (wired or wireless) to a host system which stores the data and has a specifically designed add-on board attached. The actual acceleration sensors are mounted on small boards which are wired up to the add-on board. See figure 1 for an overview of the architecture.

The acquisition system can be attached to a notebook computer or a PDA via serial line. In cases where it is not possible or practical to wear additional components the system can be connected wirelessly to a nearby smart-it connected to a computer (e.g. for a badminton scenario: the player only wears the acquisition system and the data is transmitted wirelessly to a computer next to the field.)

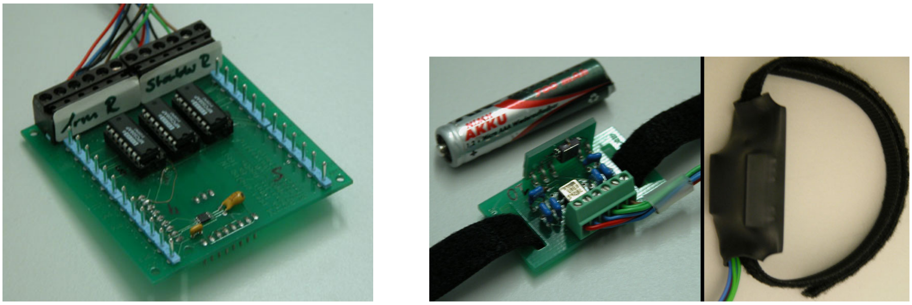
3.3 Smart–Its Core Board

The smart-it core board is a small embedded microcontroller system build around a PIC microcontroller (PIC16F877 or PIC18F252) that offers serial RS232 communication, a wireless link with 19200 bits/s, 8K of non-volatile RAM, a power supply circuit (Batteries or external voltage), and an extension connector for the add-on Board.

For the onboard components there are also libraries available (e.g. RAM, communication). There are also software templates on which further software can be developed. The smart-its core boards are designed as building blocks to ease prototyping of Ubiquitous Computing systems [15,16].

3.4 Multiplexer Add–On Board

To make the system cheaper, ease programming, and to allow a high sampling speed we decided to use acceleration sensors with analog outputs. As the core



(a) add–on board to read 24 analog channels with 2 3D acceleration sensor nodes attached

(b) Left: A single sensor board with 4 channels of acceleration. Right: the sensor with Velcro strap covered by shrink wrap

Fig. 2. Acquisition Platform

board only has 5 analog inputs a multiplexer was required. The microcontroller only offers 10 bit resolution in the analog to digital conversion. To experiment and asses the value of a higher resolution conversion we included a 16 Bit analog digital converter (ADS8320).

The add-on board has 24 analog inputs, set up as 6 groups of 4 inputs. Each group has one connector that also offers power and ground. Each analog input is connected to one of the inputs of one of the 3 analog multiplexers (each with 8 inputs and one output). The output of each multiplexer is connected to an analog input of the microcontroller. For one of the multiplexers the output is also connected to the external analog digital converter. The controls (for the converter and the multiplexer) are connected to the smart-its core board. See figure 2(a).

A library is realized that allows to read each of the external channels. Given the reading time and the time to switch between channels a sampling rate of about 100Hz per channel can be achieved.

3.5 3D Acceleration Sensor Node

For the sensor nodes we used 2 ADXL311 mounted on 2 small PCBs which are attached to each other in a 90 degree angle to effectively obtain 3D acceleration data. See figure 2(b).

The base PCB is about 40mm by 20mm and contains all the signal condition components and one of the accelerometers. The board mounted upright is about 20mm by 10 mm and contains only the ADXL311. The assembled size of a node is 40mm by 20mm by 10mm. We did deliberately not reduce the size of the nodes in order to be able to have very robust screw-on connectors on the board. We also included rectangular holes directly into the PCB to ease fixing of straps.

To increase the robustness the node can be covered (after fixing the straps and cable) by shrink wrap. See figure 2(b) for a picture of a complete, wrapped sensor board.

4 Recognition Algorithm

To classify the acceleration data into distinct classes we employ a Bayesian classifier. In this section we give a brief overview over the classification algorithm and introduce the features we use.

4.1 Bayes Classification

Bayesian classification is based on Bayes' rule from basic probability theory. Other, more complex, classifiers are of course possible for this task and will be investigated as part of future work.

Bayes' rule states, that the probability of a given activity a given an n-dimensional feature vector $x =< x_1, ...x_n >$ can be calculated as follows:

$$p(a|\mathbf{x}) = \frac{p(\mathbf{x}|a)p(a)}{p(\mathbf{x})}$$

$p(a)$ denotes the a-priori probability of the given activity. We assume them to be uniform for the purpose of this paper. The a-priori probability $p(\mathbf{x})$ of the data is just used for normalization. Since we are not interested in the absolute probabilities but rather in the relative likelihoods, we can neglect it.

Assuming, that the different components x_i of the feature vector $\mathbf{x}$ are independent, we obtain a naïve Bayes classifier which can be written as:

$$p(a|\mathbf{x}) = \frac{p(a)}{p(\mathbf{x})} \prod_{i=1}^{n} p(x_i|a)$$

We can compute the likelihoods $p(x_i|a)$ from labelled training data. We represent these probability density functions as 100 bin histograms.

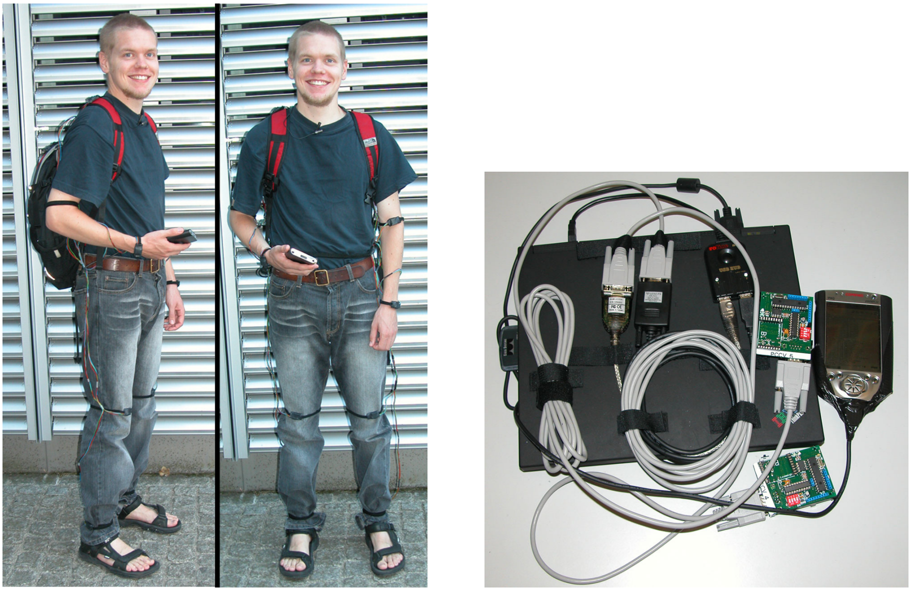
4.2 Features

The above algorithm does not work well just using the raw data samples [12]. Its performance can be considerably increased by the use of appropriate features.

As features we use the running mean and variance, computed over a window of 50 samples. Given that our data is sampled at a rate of 92Hz, this corresponds to roughly 0.5 sec. For every new data vector the window is advanced by one. Thus we can make a new classification every time we receive a new data vector.

5 Experiments

This section describes the experiments we performed. It introduces the experimental setup, including the number and placement of the sensors, and the gathered data in detail. The obtained results are discussed in the next section.



(a) Recording Setup Mounted on a User

(b) Recording Setup: Laptop with IPAQ for Online Annotation and 2 Smart–Its

Fig. 3. Recording Setup

5.1 Experimental Setup

All data is recorded on a laptop that the user carries in a backpack. The sole user interface is a Compaq IPAQ that is attached to the laptop via serial line. It allows to start/stop the recording application and to annotate the data online with the current activity. Figure 3(a) shows the user with the mounted sensors wearing the backpack, holding the IPAQ in his hand.

For the desired number of sensors we need two complete sets of sensors with six sensors each. Each set, consisting of a smart-it, an add-on board, and six 3D acceleration sensor nodes is attached via a serial port to the laptop (see also figure 3(b)). Every sensor is sampled with approx. 92Hz.

Activities. Our goal in this paper is to recognize everyday postures and activities. First of all, this includes basic user postures and movements that allow to roughly classify the user's activity. These are *sitting*, *standing*, and *walking*.

Apart from these basic postures and movements, it would also be interesting to know, what the user is currently occupied with. We hence included *writing on a whiteboard* and *typing on a keyboard*. The former indicates that the user is engaged in a discussion with others, while the latter indicates that the user is working on his computer.

Finally, social interactions are very important and interesting information. We hence include *shaking hands* to determine, if the user is currently interacting with somebody else. Kern et al. [5] use this as one of the cues for annotating meeting recordings.

Number and Placement of Sensors. In order to capture all of the above postures and activities, we decided to add sensors to all major joints on the human body. More precisely on the following six locations: just above the ankle, just above the knee, on the hip, on the wrist, just above the elbow, and on the shoulder.

In order to capture also 'asymmetric' activities, such as writing which use only one hand, we duplicate these six sensors on both sides of the body, resulting in a total of 12 3D acceleration sensors.

The sensors are fixed using Velcro straps, such as depicted in figure 2(b). Figure 3(a) shows the complete setup of all sensors attached to a user.

Experiments. Using the above setup, we have recorded a stretch of 18.7 minutes data. It covers the activities mentioned above namely *sitting*, *standing*, *walking*, *stairs up*, *stairs down*, *shaking hands*, *writing on the whiteboard* and *keyboard typing*. The data can be downloaded under
http://www.vision.ethz.ch/kern/eusai.zip.

6 Results and Discussion

In this section we present and discuss the results we obtained from the experiments that are described in the preceding section.

6.1 Overall Recognition

Figure 4 shows the recognition rates using different sub-sets of the available sensors. 'All Sensor' recognition rates were obtained using all available 12 sensors for recognition. 'Left' and 'Right' use only right and left sensors respectively (six sensors each). While the 'upper body' refers to the sensors on both shoulders, elbows, and wrists, the 'Lower Body' refers to the sensors on both sides of the hip, both knees, and ankles.

The average recognition rate over all eight activities (the last set of bars) shows that the results get better the more sensors are used.

Comparing the upper and the lower parts of the body, we note that the recognition rate for the lower body is significantly lower, because the 'other' activities (*writing on the whiteboard*, *shaking hands*, and *typing on a keyboard*) cannot be recognised well. This is natural, since the main part of these activities does not involve the legs. As expected the 'leg-only' activities (*sitting*, *standing*,

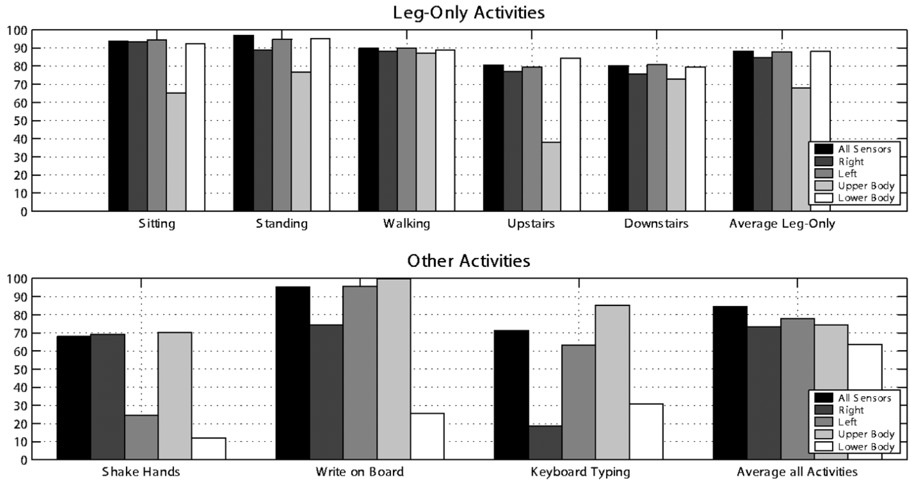


Fig. 4. Recognition Rates for All Sensors and Different Body Parts

walking, *upstairs*, *downstairs*) are better recognised using the lower part of the body. However, the upper part still performs reasonably. Apparently the overall body motion for these activities can be captured using sensors on the upper part of the body.

When comparing the right and left side of the body, we note that for the leg-only activities both sides are nearly equal in recognition rate. However the recognition rates for the other activities are quite different. Since *shaking hands* is a right-handed activity in which the left side of the body plays only a minor role the right set of sensors obtains the best results. Quite interestingly *writing on the white-board* cannot be recognized well with the right set of sensors but rather with the left side, which is due to the position of the left arm which seems to be more discriminative. The low performance of the right side on the *keyboard typing* activity seems also quite interesting: since the right hand was used to annotate the data using the IPAQ the right side is not very discriminant.

6.2 Sensors on the Leg

Figure 5 shows the recognition results for different sensors on the right leg. For the relatively simple motions of *sitting*, *standing*, and *walking* it seems to be sufficient to use one sensor only. Also, the difference between the individual sensors is quite small. Thus the placement can be chosen quite freely.

However, for more complex activities such as walking up- and downstairs, the placement considerably influences the recognition performance. The sensor attached to the ankle is the most discriminative, followed by the hip and (with a little distance) the knee. Combining different sensors, e.g. the hip and the ankle ones, improves the recognition rate. Thus, for more complex activities

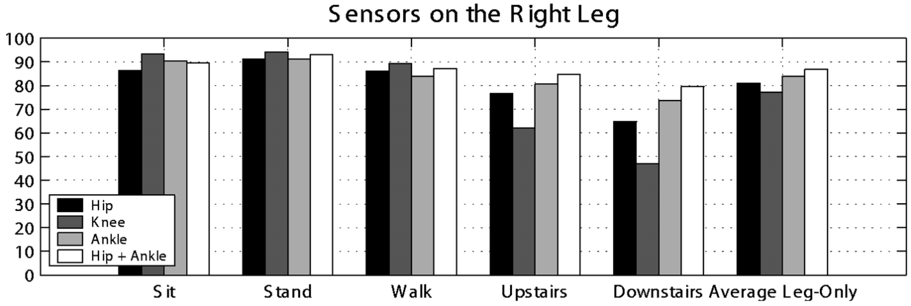


Fig. 5. Recognition Rates for Different Sensors on the Right Leg

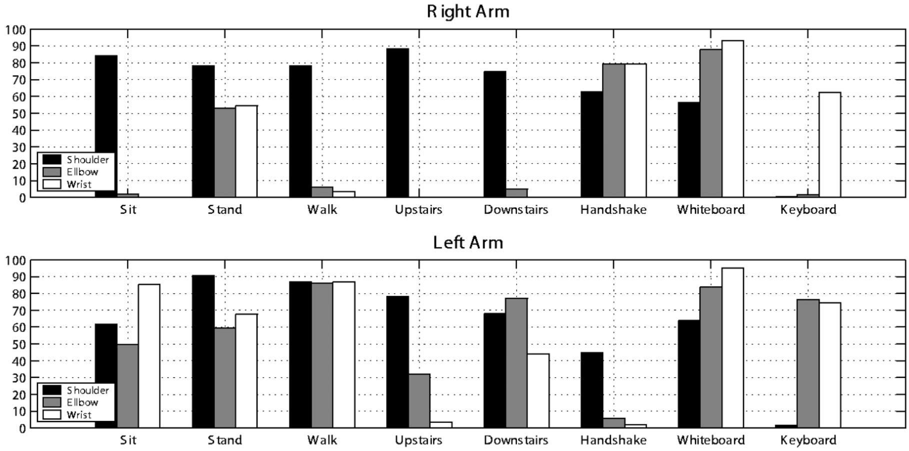


Fig. 6. Recognition Results for Single Sensors on the Arm

than the ones used here, the combination of different sensors might be crucial for successful recognition.

6.3 Sensors on the Arms

Figure 6 shows the recognition results for single sensors on both arms. One of the most interesting results here is that the sensors placed on the shoulders are well suited to recognize the legs-only activities. Furthermore, we note, that typing on a keyboard is best recognized using sensors on the wrists. This seems natural, since it is an activity of the hands only.

When comparing the right and the left arm, the sensors on the elbow and wrist of the right arm perform worse for the leg-only activities. This is again due to the fact that the right hand was used to annotate the data using the IPAQ, which makes the activity of the right arm similar for all leg-only activities.

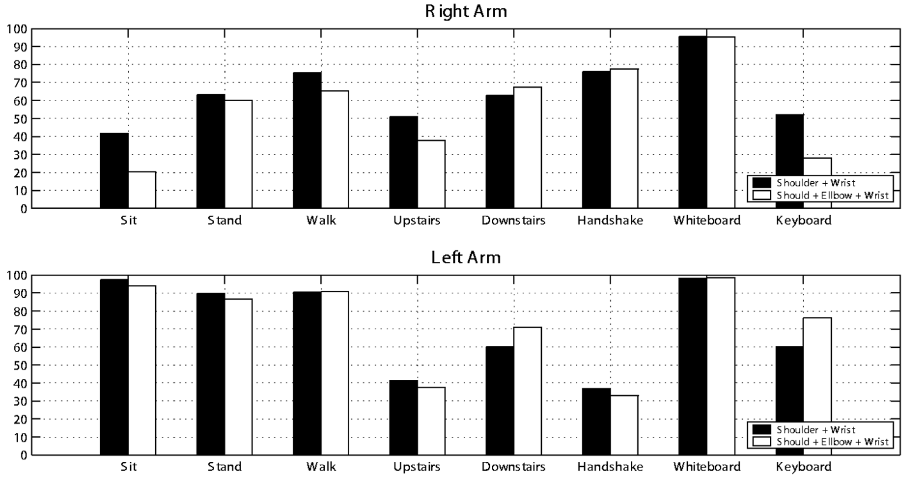


Fig. 7. Recognition Results for Combined Sensors on the Arm

Shaking hands is a right-handed activity and thus cannot be detected well on the left arm.

Considering the sensor placement, the position just above the elbow does not add significant information. Although the recognition rate for the elbow sensor is partly better than either the shoulder or the wrist, it does not contribute to or outperform the combination of the two. Figure 7 shows that the recognition rate using shoulder and wrist sensors cannot be further increased by adding the sensor at the elbow.

7 Conclusion

The user's physical activity is central for context-aware user-centred applications. In this paper we concentrated on context extraction using body-worn sensors. More specifically we propose to use multiple acceleration sensors, since they are lightweight, small, and inexpensive.

We have presented a hardware platform, that allows capturing 3-dimensional acceleration data from up to 48 positions on the human body. It is especially designed for robustness, allowing for recording even very dynamic activities, such as playing badminton or climbing.

We have conducted experiments to investigate the number and placement of sensors. We therefore recorded data of the activities *sitting*, *standing*, *walking*, *stairs up/down*, *writing on a whiteboard*, *shaking hands*, *typing on a keyboard.*

As expected, the combination of multiple sensors generally increases recognition performance. For more complex activities, such as *stairs* or *writing on a whiteboard*, multiple sensors are not only helpful but rather mandatory for good recognition performance.

The placement depends of course very much on the activity. For 'leg-only' activities, such as *walking* or *stairs*, sensors on the legs, e.g. hip and/or ankle, are sufficient. For those activities a single sensor mounted on the shoulder also obtained good recognition performance. For more complex activities such as *writing on a whiteboard*, sensors both on the upper and lower part of the body are required. In our experiments a sensor placed just above the elbow did not seem to add significant information.

Right and left arm work relatively independently. The recognition rate can thus get confused, if either is temporarily engaged in another activity. Our experiments showed that the right arm was not very discriminative, because it was used to hold the IPAQ for annotation. In order not to confuse the recognition by such effects both arms should be equipped with sensors.

Several issues still need to be addressed. E.g. the influence of the features used for recognition and the recognition methodology itself should be addressed. Also, more complex activities should be investigated. At the moment it is not known how much actually can be inferred about the user using acceleration sensors only. The results and the platform presented in this paper are but a first step to investigate this topic.

References

1. Kamijoh, N., Inoue, T., Olsen, C.M., Raghunath, M., Narayanaswami, C.: Energy trade-offs in the ibm wristwatch computer. In: Proc. Sixth International Symposium on Wearable Computers (ISWC). (2001) 133–140
2. Randell, C., Muller, H.: Context awareness by analyzing accelerometer data. In: Proc. Fourth International Symposium on Wearable Computers (ISWC). (2000) 175–176
3. Farringdon, J., Moore, A., Tilbury, N., Church, J., Biemond, P.: Wearable sensor badge & sensor jacket for context awareness. In: Proc. Third International Symposium on Wearable Computers (ISWC), San Francisco (1999) 107–113
4. van Laerhoven, K., Schmidt, A., Gellersen, H.W.: Multi–sensor context–aware clothing. In: Proc. Sixth International Symposium on Wearable Computers (ISWC), Seattle (2002) 49–57
5. Kern, N., Schiele, B., Junker, H., Lukowicz, P., Troester, G.: Wearable sensing to annotate meetings recordings. In: Proc. Sixth International Symposium on Wearable Computers (ISWC). (2002) 186–193
6. van Laerhoven, K., Kern, N., Schiele, B., Gellersen, H.W.: Towards an inertial sensor network. In: Proc. EuroWearable, Birmingham, UK (2003)
7. Loosli, G., Canu, S., Rakotomamonjy, A.: Détection des activités quotidiennes á l'aide des séparateurs á vaste marge. In: Proc. Rencontre Jeunes Chercheurs en IA, Laval, France (2003)
8. Chambers, G., Venkatesh, S., West, G., Bui, H.: Hierarchical recognition of intentional human gestures for sports video annotation. In: Proc. 16th IEEE Conference on Pattern Recognition. Volume 2. (2002) 1082–1085
9. Benbasat, A., Paradiso, J.: Compact, configurable inertial gesture recognition. In: Proc. ACM CHI Conference – Extended Abstracts. (2001) 183–184

10. Sabelman, E., Troy, B., Kenney, D., Yap, R., Lee, B.: Quantitative balance analysis: Accelerometric lateral sway compared to age and mobility status in 60–90 year-olds. In: Proc. RESNA, Reno, NV, USA (2001)
11. Morris, S., Paradiso, J.: Shoe-integrated sensor system for wireless gait analysis and real-time feedback. In: Proceedings of the 2nd Joint IEEE EMBS (Engineering in Medicine and Biology Society) and BMES (the Biomedical Engineering Society) Conference. (2002) 2468–2469
12. Golding, A., Leash, N.: Indoor navigation using a diverse set of cheap, wearable sensors. In: Proc. Third International Symposium on Wearable Computers (ISWC), San Francisco (1999) 29–36
13. Lee, S.W., Mase, K.: Activity and location recognition using wearable sensors. IEEE Pervasive **1** (2002) 24–32
14. Kern, N., Schiele, B.: Context–aware notfication for wearable computing. In: Proc. 7th International Symposium on Wearable Computers (ISWC). (2003)
15. Schmidt, A.: Ubiquitous Computing – Computing in Context. PhD thesis, University of Lancaster (2002)
16. Beigl, M., Zimmer, T., Krohn, A., Decker, C., Robinson, P.: Smart-its – communication and sensing technology for ubicomp environments. technical report issn 1432–7864. Technical report, TeCo, University Karlsruhe, Germany (2003/2)

Towards Computer Understanding of Human Interactions

Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Darren Moore, and Hervé Bourlard

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P.O. Box 592, CH-1920, Martigny, Switzerland
{mccowan,gatica,bengio,moore,bourlard}@idiap.ch,
http://www.idiap.ch/

Abstract. People meet in order to interact - disseminating information, making decisions, and creating new ideas. Automatic analysis of meetings is therefore important from two points of view: extracting the information they contain, and understanding human interaction processes. Based on this view, this article presents an approach in which relevant information content of a meeting is identified from a variety of audio and visual sensor inputs and statistical models of interacting people. We present a framework for computer observation and understanding of interacting people, and discuss particular tasks within this framework, issues in the meeting context, and particular algorithms that we have adopted. We also comment on current developments and the future challenges in automatic meeting analysis.

1 Introduction

The domain of human-computer interaction aims to help humans interact more naturally with computers. A related emerging domain of research instead views the computer as a tool to assist or understand human interactions : putting computers in the human interaction loop [1]. Humans naturally interact with other humans, communicating and generating valuable information. The most natural interface for entering this information into a computing system would therefore be for the computer to extract it directly from observing the human interactions.

The automatic analysis of human interaction is a rich research area. There is growing interest in the automatic understanding of group behaviour, where the interactions are defined by individuals playing and exchanging both similar and complementary roles (e.g. a handshake, a dancing couple, or a children's game) [2,3,4,5,6]. Most of the previous work has relied on visual information and statistical models, and studied three specific scenarios: surveillance in outdoor scenes [5,6], workplaces [3,4], and indoor group entertainment [2]. Beyond the use of visual information, dialogue modelling [7,8] analyses the structure of interactions in conversations.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 235–251, 2003.

While it has only recently become an application domain for computing research, observation of human interactions is not a new field of study - it has been actively researched for over fifty years by a branch of social psychologists [9,10, 11]. For example, research has analysed turn-taking patterns in group discussions [12,13,14], giving insight into issues such as interpersonal trust, cognitive load in interactions, and patterns of dominance and influence [11]. Research has also shown that interactions are fundamentally multimodal, with participants coordinating speaking turns using a variety of cues, such as gaze, speech back-channels, changes in posture, etc. [12,13,15]. In general, visual information can help disambiguate audio information [16], and when the modalities are discrepant, participants appear to be more influenced by visual than by audio cues [11,17].

Motivated therefore by a desire to move towards more natural human-machine interfaces, and building upon findings of social psychologists regarding the mechanisms and significance of human interactions, this article presents an observational framework for computer understanding of human interactions, focussing on small group meetings as a particular instance.

Meetings contain many complex interactions between people, and so automatic meeting analysis presents a challenging case study. Speech is the predominant modality for communication in meetings, and speech-based processing techniques, including speech recognition, speaker identification, topic detection, and dialogue modelling, are being actively researched in the meeting context [18,8, 19,20]. Visual processing, such as tracking people and their focus of attention, has also been examined in [1,21]. Beyond this work, a place for analysis of text, gestures, and facial expressions, as well as many other audio, visual and multimodal processing tasks can be identified within the meeting scenario. While important advances have been made, to date most approaches to automatic meeting analysis have been limited to the application of known technologies to extract information from individual participants (e.g. speech, gaze, identity, etc). Intuitively, the true information of meetings is created from interactions between participants, and true understanding of meetings can only emerge from considering their group nature.

The remainder of this article is organised as follows. Section 2 describes a multi-sensor meeting room that we have installed to enable our research. A framework for computer understanding of human interactions is outlined in Section 3, along with some specific issues and algorithms related to the meeting context. Finally, some perspective on future directions in automatic meeting analysis is given in Section 4, followed by concluding remarks in Section 5.

2 A Multi-sensor Meeting Room

As mentioned above, interactions between people in meetings are generally multimodal in nature. While the *audio* modality is the most obvious source of information in discussions, studies have shown that significant information is conveyed in the *visual* modality, through expressions, gaze, gestures and posture [12,13,15].

In meetings, the *textual* modality is also important, with presentation slides, whiteboard activity, and shared paper documents providing detailed information.

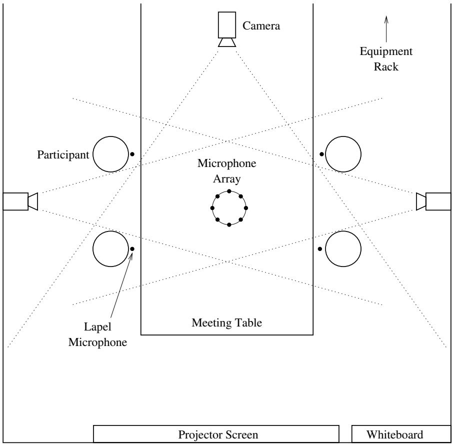


Fig. 1. Meeting recording configuration

To facilitate research into automatic meeting analysis, a meeting room at IDIAP has been equipped with multi-media recording facilities. The audio information is captured using up to 24 microphone channels, including microphone arrays, a binaural manikin, and lapel microphones attached to each meeting participant. Visual information is acquired using three closed circuit television cameras equipped with wide angle lenses. These cameras capture frontal video of meeting participants, as well as wide angle views of the entire meeting room. Textual information is also acquired from presentations and whiteboard usage. Presentation slides are captured from a ceiling-mounted data projector at native VGA resolutions and all whiteboard activity is acquired using transmitting pens and a receiver attached to a standard whiteboard. The acquisition of all modalities is completely synchronised and all data streams are accurately timestamped.

To date, meeting recording efforts at IDIAP have focussed on the compilation of an audio-visual corpus of approximately sixty, five-minute, four-person scripted meetings. The meeting room configuration used for these recordings is illustrated in Figure 1. Two cameras were used to capture frontal views of the meeting participants (including the table region used for note-taking), while the third camera recorded a view of the whiteboard and presentation screen at the front of the room. An eight-element circular equi-spaced microphone array of 20cm diameter was centrally located on the meeting table.

The resulting five hours of multi-channel, audio-visual meeting data is available for public distribution through a MultiModal Media file server at `mmm.idiap.ch`. A new round of meeting collection will be launched in the near future, and will utilise the recently added slide and whiteboard capture capabilities.

3 Multimodal Processing

We propose a framework for computer understanding of human interactions that involves the following basic steps in a processing loop :

```
1. locate and track participants
2. for each located participant
   a) enhance their audio and visual streams
   b) identify them
   c) recognise their individual actions
3. recognise group actions
```

The first step is necessary to determine the number and location of participants. For each person present, we then extract a dedicated enhanced audio and visual stream by focussing on their tracked location. Audio-visual (speech and face) speaker identification techniques can then be applied to determine who the participant is. Individual actions, such as speech activity, gestures or speech words may also be measured or recognised from the audio and visual streams. The ultimate goal of this analysis is then to be able to recognise actions belonging to the group as a whole, by modelling the interactions of the individuals.

Specific issues and algorithms for implementing a number of these steps for the case of meeting analysis are presented in the following sub-sections. A primary focus of our research is the multimodal nature of human interactions in meetings, and this is reflected in the choice of tasks we have included. Naturally, there are many other processing tasks involved in understanding meetings, such as speech recognition and dialogue modelling, that are not covered here.

3.1 Audio-visual Speaker Tracking

The problem in the global view. Locating and tracking speakers represents an important first step towards automatic understanding of human interactions. As mentioned previously, speaker turn patterns convey a rich amount of information about the behaviour of a group and its individual members [10,13]. Furthermore, experimental evidence has highlighted the role that non-verbal behaviour (gaze, facial expressions, and body postures) plays in interactions [13]. Recognising such rich multimodal behaviour first requires reliable localisation and tracking of people.

Challenges in the meeting context. The separate use of audio and video as cues for tracking are classic problems in signal processing and computer vision. However, sound and visual information are jointly generated when people speak, and provide complementary advantages. While initialisation and recovery from failures can be addressed with audio, precise object localisation is better suited to visual processing.

Long-term, reliable tracking of multiple people in meetings is challenging. Meeting rooms pose a number of issues for audio processing, such as reverberation and multiple concurrent speakers, as well as for visual processing, including clutter and variations of illumination. However, the main challenge arises from the behaviour of multiple participants resulting in changes of appearance and pose for each person, and considerable (self)-occlusion. At the same time, meetings in a multi-sensor room present some advantages that ease the location and tracking tasks. Actions usually unfold in specific areas (meeting table, whiteboard, and projector screen), which constrains the group dynamics in the physical space. In addition, the availability of multiple cameras with overlapping fields of view can be exploited to build more reliable person models, and deal with the occlusion problems.

Our approach. We have developed a principled method for speaker tracking, fusing information coming from multiple microphones and uncalibrated cameras [22], based on *Sequential Monte Carlo* (SMC) methods, also known as *particle filters* (PFs) [23]. For a state-space model, a PF recursively approximates the conditional distribution of states given observations using a dynamical model and random sampling by (i) generating candidate configurations from the dynamics (*prediction*), and (ii) measuring their likelihood (*updating*), in a process that amounts to random search in a configuration space. Data fusion can be introduced in both stages of the PF algorithm.

Our work is guided by inherent features of AV data. First, audio is a strong cue to model discontinuities that clearly violate usual assumptions in dynamics (including speaker turns across cameras), and (re)initialisation. Its use for prediction thus brings benefits to modelling real situations. Second, audio can be inaccurate at times, but provides a good initial localisation guess that can be enhanced by visual information. Third, although audio might be imprecise, and visual calibration can be erroneous due to distortion in wide-angle cameras, the joint occurrence of AV information in the constrained physical space in meetings tends to be more consistent, and can be learned from data.

Our methodology exploits the complementary features of the AV modalities. In the first place, we use a 2-D approach in which human heads are visually represented by their silhouette in the image plane, and modelled as elements of a *shape-space*, allowing for the description of a head template and a set of valid geometric transformations (motion). In the second place, we employ a *mixed-state space*, where in addition to the continuous subspace that represents head motion, we include a discrete component that indicates the specific camera plane in which a speaker is present. This formulation helps define a generative model for camera switching. In the third place, we asymmetrically handle audio and video in the PF formulation. Audio localisation information in 3-D space is first estimated by an algorithm that reliably detects speaker changes with low latency, while maintaining good estimation accuracy. Audio and skin-color blob information are then used for prediction, and introduced in the PF via *importance sampling*, a technique which guides the search process of the PF towards regions of the state space likely to contain the true configuration (a speaker). Additionally, audio, color, and shape information are jointly used to compute the likelihood of candidate configurations. Finally, we use an AV calibration procedure to relate audio estimates in 3-D and visual information in 2-D. The procedure uses easily generated training data, and does not require precise geometric calibration of cameras and microphones [22].

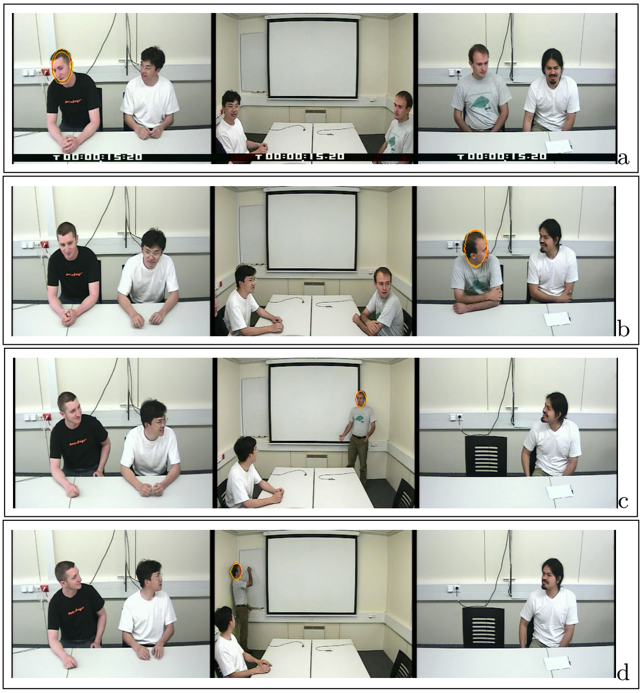


Fig. 2. Tracking speakers in the meeting room. Frames 100, 1100, 1900, and 2700.

The result is a method that can initialise and track a moving speaker, and switch between multiple people across cameras with low delay, while tolerating visual clutter. An example for the setup of Figure 1 is shown in Figure 2. For a two-minute sequence, the system tracked the current speaker in the correct camera in approximately 88% of the frames, while keeping the localisation error in the corresponding image plane within a few pixels. Other AV tracking examples for single- and multi-camera set-ups can be found at `www.idiap.ch/~gatica`.

Open problems. Although the current methodology is useful in its current form, there is much room for improvement. In the following we identify three specific lines of research. We are currently generalising our formulation to a multiple-object AV tracker, which involves the integration of person-dependent appearance models, and the consistent labelling of tracked objects along time and across cameras. Multi-object tracking significantly increases the dimensionality of the state space, which calls for efficient inference mechanisms in the resulting statistical model. Another line of research is the integration of more robust person models. The third line of research is the joint formulation of tracking and recognition. Specifically, we are building head trackers that simultaneously estimate head orientation (a simple form of recognition), which is in turn a strong cue for detection of focus of attention, and useful for higher-level recognisers.

3.2 Speech Segmentation and Enhancement Using Microphone Arrays

The problem in the global view. Having located and tracked each person, it is next necessary to acquire an enhanced dedicated audio channel of their speech. Speech is the predominant communication modality, and thus a rich source of information, in many human interactions.

Most state-of-the-art speech and speaker recognition systems rely on close-talking head-set microphones for speech acquisition, as they naturally provide a higher signal-to-noise ratio (SNR) than single distant microphones. This mode of acquisition may be acceptable for applications such as dictation, however as technology heads towards more pervasive applications, less constraining solutions are required. *Microphone arrays* present a promising alternative to close-talking microphones, as they allow for signal-independent enhancement, localisation and tracking of speakers, and non-intrusive hands-free operation. For these reasons, microphone arrays are being increasingly used for speech acquisition in such applications [24,25].

Challenges in the meeting context. Meetings present a number of interesting challenges for microphone array research. A primary issue is the design of the *array geometry* : how many microphones should be used, and where should they be placed in the room? Naturally a geometry giving high spatial resolution uniformly across a room is desirable for best performance and lowest constraint on the users, however this requires prohibitively large numbers of microphones, and complex installation [26]. For these reasons, more practical solutions with smaller numbers of microphones need to be researched to address computational and economical considerations.

A second challenge in the meeting context is the natural occurrence of overlapping speech. In [27] it was identified that around 10-15% of words, or 50% of speech segments, in a meeting contain a degree of overlapping speech. These overlapped segments are problematic for speaker segmentation, and speech and speaker recognition. For instance, an absolute increase in word error rate of between 15-30% has been observed on overlap speech segments using close-talking microphones [27,8].

Our approach. While it is clear that a large microphone array with many elements would give the best spatial selectivity for localisation and enhancement, for microphone arrays to be employed in practical applications, hardware cost (microphones, processing and memory requirements) must be reduced. For this reason, we focus on the use of small microphone arrays, which can be a viable solution when assumptions can be made about the absolute and relative locations of participants.

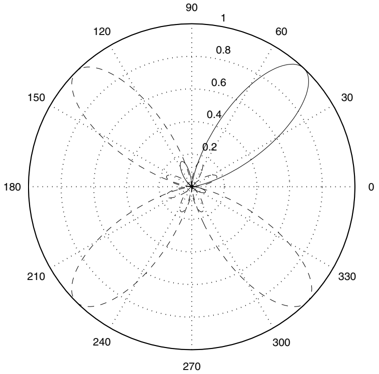


Fig. 3. Microphone array directivity patterns at 1000 Hz (speaker 1 direction in bold)

As shown in Figure 1, the particular array geometry we have chosen is an 8-element circular array (of radius 10cm) placed at the centre of the meeting table. This geometry and placement was selected based on the assumption that a meeting generally consists of small groups of people seated and talking face to face in well-defined regions. Each array is designed to cater for a small group of up to 4 people. In larger meetings, multiple (potentially interacting) small array modules are positioned along the table, where each module is responsible for the people in its local region. The circular geometry was selected as it gives uniform spatial selectivity between people sitting around it, leading to good general performance in separating overlapping speech. This is important for meetings where background noise is generally low, and so overlapping speech is the primary noise source. To illustrate, Figure 3 shows the theoretical *directivity pattern* (array gain as a function of direction) for the array at 1000 Hz for 4 speakers separated by 90 degrees. Having the array on the table also means it is placed in close proximity to participants, leading to naturally high signal levels compared to background noise caused by distant sources.

Given accurate tracking of the speaker locations in the room, the next task is to determine segments of continuous speech from a given speaker location. *Speaker segmentation* in meetings is problematic for traditional techniques based on simple energy or spectral features, as a significant amount of cross-talk from other speakers exists even on close-talking microphones [28,29]. In [30,31] we presented a *location-based segmentation* technique that is capable of providing a smooth speech/silence segmentation for a given room location. As it is based on speech location features from the microphone array, rather than standard spectral features, this location-based segmentation has the important benefit of being able to accurately handle multiple concurrent speakers (identifying which locations are active at any given time). As an example, our current system is capable of segmenting four person meetings, including overlapping speech segments, with over 95% frame accuracy [31].

Once the location of the speakers is known along with their speech activity segmentation, we can then apply microphone array *beamforming* techniques to enhance their speech, attenuating background noise and conflicting speech sources. Beamforming consists of filtering and combining the individual microphone signals in such a way as to enhance signals coming from a particular location. For beamforming filters, we adopt standard *superdirective* filters, which are calculated to maximise the array gain for the desired direction [32]. In addition, we apply a *Wiener post-filter* to the beamformer output to further reduce the broadband noise energy. The post-filter is estimated from the auto- and cross-spectral densities of the microphone array inputs, and is formulated assuming a diffuse background noise field [33]. This post-filter leads to significant improvements in terms of SNR and speech recognition performance in office background noise [33], and has also been shown to out-perform lapel microphones for a small vocabulary recognition task in significant levels of overlapping speech [34].

Open problems. While microphone array speech processing techniques are already relatively mature, a number of open issues remain in this context. As mentioned briefly, larger meetings could be catered for by a series of small microphone array modules working together. A current focus of our research is thus to propose algorithms for these interactions between modules. Further research is also focussing on issues related to the real-time implementation of multiple concurrent beamformers.

3.3 Audio-visual Person Identification

The Problem in the Global View. Identifying participants is important for understanding human interactions. When prior knowledge about the participants is available (such as their preferred way of communicating, topics of interests, levels of language, relative hierarchical levels in a given context, etc), knowing the participants' identities would imply knowing this prior information, which could in turn be used to better tune the algorithms used to analyse the interaction. Fortunately, *biometric authentication* [35], which is the general problem of authenticating or identifying a person using his or her behavioural and physiological characteristics such as the face or the voice, is a growing research domain which has already shown useful results, especially when using more than one of these characteristics, as we propose to do here.

Challenges in the Meeting Context. In order to perform AV identification during a meeting, we need to extract reliably the basic modalities. For the face, we require a face localisation algorithm that is robust to the kind of images available from a video stream (relatively low-quality and low-resolution), robust to the participants' varying head poses, and able to cope with more than one face per image. This could be done using our AV tracking system described in Section 3.1. For the voice, taking into account that several microphones are available in the meeting room, the first challenge is to separate all audio sources and attribute each speech segment to its corresponding participant. Again, this could be done using our speaker segmentation and enhancement techniques, described in Section 3.2. Afterward, classical face and speaker verification algorithms could be applied, followed by a fusion step, which provides robustness to the failure of one or the other modality. Finally, an identification procedure could be applied.

Our Approach. Our identification system is based on an AV biometric verification system. Assuming that we are able to obtain reliable speech segments and localised faces from the meeting raw data, we can then apply our state-of-the-art verification system, which is based on a *speaker verification* system, a *face verification* system, and a *fusion* module.

Our speaker verification system first starts by extracting useful features from the raw speech data: we extract 16 Linear Predictive Cepstral Coefficient (LPCC) features every 10 ms, as well as their first temporal derivative. Then, a silence detector based on an unsupervised 2-Gaussian system is used to remove all silence frames. Finally, the verification system itself is based on the modelling of one Gaussian Mixture Model (GMM) for each individual, adapted using *Maximum A Posteriori* (MAP) techniques from a *World Model* trained by *Expectation-Maximisation* on a large set of prior data. The score for a given access is obtained as the logarithm of the ratio between the likelihood of the data given the individual model and the likelihood given the world model. This system obtains state-of-the-art performance on several benchmark verification databases [36].

Our face verification system is based on a non-holistic view: instead of extracting features from a full face image that are then handled by a classifier, as it is often done [37], we extract Discrete Cosine Transform (DCT) -based features from overlapping blocks of the image (square patches that span the whole face image), using the DCTmod2 technique [38], which has shown state-of-the-art performance in several cases. Finally, we model the obtained feature vectors using GMMs, similarly to the speaker verification system, hence using the same scoring technique.

Our fusion algorithm is based on Multi-layer Perceptrons (experiments with Support Vector Machines give similar performances). The fusion model takes as input the log likelihood scores coming from both the face and the speaker verification systems, and combines them non-linearly in order to obtain a unified and more robust overall score. Optionally, confidence values could also be computed on both the voice and face scores, which then enhance the quality of the fusion model [39].

Finally, in order to identify the correct individual, the whole verification system is run over all previously stored individual models, and the model corresponding to the highest obtained score over a pre-defined threshold (in order to account for unknown individuals) identifies the target individual.

Open Problems. Assuming that speaker segmentation and face tracking have given perfect segmentation, for a given meeting, we will have potentially several minutes of speech and face data per individual. In general, a classical verification system only requires a few face images and less than one minute of speech data to attain acceptable performance. However, the environment is unconstrained, the meeting data may be noisy for different reasons - the individual may not always look at the camera and speak loudly and intelligibly. In this case, rather than using all available data to identify a person, a better solution could be to be more strict on the selection of faces and speaker segments in order to keep only the best *candidates* for identification. Hence, we should try to remove highly noisy or overlapping speech segments, badly tracked face images and faces that are not in a good frontal pose and good lighting condition.

3.4 Group Action Recognition

The problem in the global view. The ultimate goal of automatic analysis of human interactions is to recognise the group actions. As discussed previously, the true information of meetings is created from interactions between participants playing and exchanging roles. In this view, an important goal of automatic meeting analysis is the segmentation of meetings into high-level agenda items which reflect the action of the group as a whole, rather than just the behaviour of individuals (e.g. discussions and presentations, or even higher level notions, like planning, negotiating, and making decisions).

Challenges in the meeting context. Recognition of group actions in meetings entails several important problems for which no satisfactory solutions currently exist. These include (1) devising tractable multi-stream sequence models, where each stream could arise from either a modality (AV) or a participant; (2) modelling asynchronicity between participants' behaviour; (3) extracting features for recognition that are robust to variations in human characteristics and behaviour; (4) designing sequence models that can integrate language features (e.g. keywords or dialog acts) with non-verbal features (e.g. emotion as captured from audio and video); and (5) developing models for recognition of actions that are part of a hierarchy.

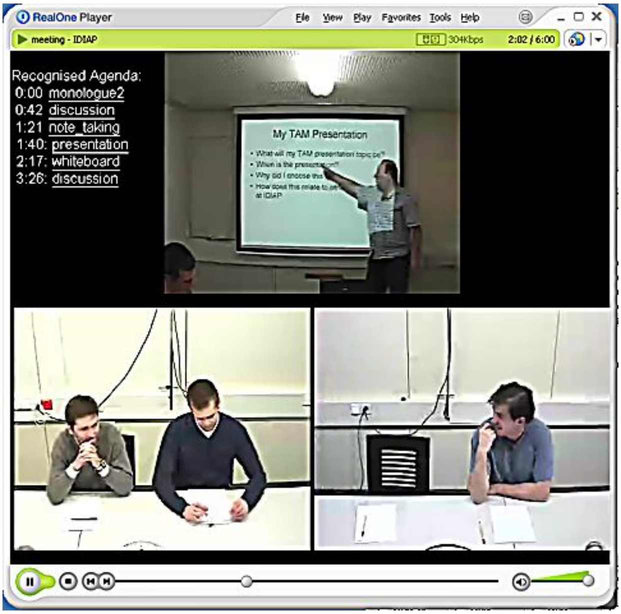
One potentially simplifying advantage to recognise group actions in meetings is that participants usually have some influence on each other's behaviour. For example, a dominant speaker grabbing the floor often makes the other participants go silent, and a presentation will draw most participants' attention in the same direction. The recognition of some group actions can be therefore benefit from the occurrence of these multiple similar individual behaviours.

Our approach. We have addressed meeting group action recognition as the recognition of a continuous, non-overlapping, sequence of lexical entries, analogous to observational approaches in social psychology for analysis of group interaction [10], and to speech or continuous gesture recognition [40,41]. Continuous recognition generates action-based meeting segmentations that can be directly used for browsing. Furthermore, the definition of multiple lexica would provide alternative semantic views of a meeting. Note that in reality, most group actions are characterised by soft (natural) transitions, and specifying their boundaries beyond a certain level of precision has little meaning.

In particular, we have modelled meeting actions based on a set of multimodal turn-taking events. Speaking turns are mainly characterised by audio information, but significant information is also present in non-verbal cues like gaze and posture changes [13], which can also help disambiguate audio information [16]. The specific actions include monologues (one participant speaks continuously without interruption), discussions (all participants engage in a discussion), presentations (one participant at front of room makes a presentation using the projector screen), white-boards (one participant at front of room talks and uses the white-board), and group note-taking (all participants write notes).

To investigate the multimodal and group natures of the actions, we used a variety of Hidden Markov Models (HMMs) [40] to combine the streams of information (with streams representing modalities or people) in different ways. The models include early integration HMMs, multi-stream HMMs [42], and asynchronous HMMs [43]. Furthermore, the individual behaviour of participants was monitored using features from both the audio and visual modalities (including speech activity, pitch, energy, speaking rate, and head and hand location and motion features).

A detailed account of our experiments and results can be found in [44]. For experiments, we used the meeting corpus described in Section 2. Meetings followed a loose script to ensure an adequate amount of examples of all actions, and to facilitate annotation for training and testing, but otherwise the individual and group behaviour is natural. In summary, the best action error rate (equivalent to the word error rate in speech recognition) that we obtained on an independent test set was 5.5% using a two-stream HMM, where one stream modelled audio features, and the other modelled video features coming from all participants. Several other models (audio-only, AV early-integration, and AV asynchronous) produced competitive results. As expected, for this set of actions audio was the main source of information for reliable recognition, while video mostly helped in reducing monologue and discussion errors.



Recognised Agenda:
0:00 monologue2
0:42 discussion
1:21 note_taking
1:40: presentation
2:17: whiteboard
3:26: discussion

Fig. 4. Simple meeting browser interface, showing recognised meeting actions.

An example of the application of the action recognition results for meeting browsing is shown in Figure 4.

Open problems. The experience gained from our results confirms the importance of modelling the interactions between individuals, as well as the advantage of a multimodal approach for recognition. We believe there is much scope for work towards the recognition of different sets of high-level meeting actions, including other multimodal turn-taking events, actions based on participants' mood or level of interest, and multimodal actions motivated by traditional dialogue acts. To achieve this goal, ongoing and future work will investigate richer feature sets, and appropriate models for the interactions of participants. Another task will be to incorporate prior information in the recognition system, based on the participant identities and models of their personal behaviour. We also plan to collect a larger meeting corpus, and work on the development of more flexible assessment methodologies.

4 Future Directions

From the framework outlined in the beginning of Section 3, while much room clearly remains for new techniques and improvements on existing ones, we can see that steps `1-2(c)` are reasonably well understood by the state-of-the-art. In contrast, we are far from making similar claims regarding step 3, recognition of group actions.

The first major goal in computer understanding of group actions, is to clearly identify lexica of such actions that may be recognised. A simple lexicon based on multimodal turn-taking events was discussed in Section 3.4, however there is a need to progress towards recognition of higher level concepts, such as decisions, planning, and disagreements. In this regard, the social psychology literature represents an important source of information for studies on the tasks and processes that arise from human interactions, as was discussed in [44].

Having identified relevant group actions, a further research task is then to select appropriate features for these actions to be recognised. At this moment, features are intuitively selected by hand, which has obvious limitations. Approaches for feature selection could arise from two areas. The first one is human. We require a deeper understanding of human behaviour. Existing work in psychology could provide cues for feature selection towards, for example, multimodal recognition of emotion [45]. The second one is computational. Developments in machine learning applied to problems in vision and signal processing point to various directions [46].

Finally, to recognise the group actions, there is a need to propose models capable of representing the interactions between individuals in a group (see e.g. [47,5,44]). Some particular issues are the need to model multiple data streams, asynchronicity between streams, hierarchies of data and events, as well as features of different nature (e.g. discrete or continuous).

5 Conclusion

This article has discussed a framework for computer understanding of human interactions. A variety of multimodal sensors are used to observe a group and extract useful information from their interactions. By processing the sensor inputs, participants are located, tracked, and identified, and their individual actions recognised. Finally, the actions of the group as a whole may be recognised by modelling the interactions of the individuals.

While initial work in this direction has already shown promising progress and yielded useful results, it is clear that many research challenges remain if we are to advance towards true computer understanding of human interactions.

Acknowledgements. The authors would like to acknowledge our colleagues at IDIAP involved in the research described in this article, in particular Guillaume Lathoud, Johnny Mariéthoz, Sebastien Marcel, Conrad Sanderson, Olivier Masson, Pierre Wellner, Jean Marc Odobez and Florent Monay.

This work was supported by the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The work was also funded by the European project "M4: MultiModal Meeting Manager", through the Swiss Federal Office for Education and Science (OFES).

References

1. A. Waibel, T. Schultz, M. Bett, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, "SMaRT:the Smart Meeting Room Task at ISL," in *Proc. IEEE ICASSP 2003*, 2003.
2. A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, and A. Wilson, "The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment," *PRESENCE: Teleoperators and Virtual Environments*, vol. 8, August 1999.
3. N. Johnson, A. Galata, and D. Hogg, "The acquisition and use of interaction behaviour models," in *Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition*, June 1998.
4. T. Jebara and A. Pentland, "Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour," in *Proc. International Conference on Vision Systems*, January 1999.
5. N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, August 2000.
6. S. Hongeng and R. Nevatia, "Multi-agent event recognition," in *Proc. IEEE Int. Conference on Computer Vision*, (Vancouver), July 2001.
7. J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, "The coding of dialogue structure in a corpus," in *Proceedings of the Twente Workshop on Language Technology: Corpus-based approaches to dialogue modelling* (J. Andernach, S. van de Burgt, and G. van der Hoeven, eds.), Universiteit Twente, 1995.
8. N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. of the Human Language Technology Conference*, (San Diego, CA), March 2001.
9. R. F. Bales, *Interaction Process Analysis: A method for the study of small groups.* Addison-Wesley, 1951.
10. J. E. McGrath, *Groups: Interaction and Performance.* Prentice-Hall, 1984.
11. J. McGrath and D. Kravitz, "Group research," *Annual Review of Psychology*, vol. 33, pp. 195–230, 1982.
12. E. Padilha and J. C. Carletta, "A simulation of small group discussion," in *EDILOG*, 2002.
13. K. C. H. Parker, "Speaking turns in small group interaction: A context-sensitive event sequence model," *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 965–971, 1988.
14. N. Fay, S. Garrod, and J. Carletta, "Group discussion as interactive dialogue or serial monologue: The influence of group size," *Psychological Science*, vol. 11, no. 6, pp. 487–492, 2000.
15. D. Novick, B. Hansen, and K. Ward, "Coordinating turn-taking with gaze," in *Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP-96)*, 1996.
16. R. Krauss, C. Garlock, P. Bricker, and L. McMahon, "The role of audible and visible back-channel responses in interpersonal communication," *Journal of Personality and Social Psychology*, vol. 35, no. 7, pp. 523–529, 1977.
17. B. DePaulo, R. Rosenthal, R. Eisenstat, P. Rogers, and S. Finkelstein, "Decoding discrepant nonverbal cues," *Journal of Personality and Social Psychology*, vol. 36, no. 3, pp. 313–323, 1978.

18. F. Kubala, “Rough’n’ready: a meeting recorder and browser,” *ACM Computing Surveys*, vol. 31, 1999.
19. A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, “Advances in automatic meeting record creation and access,” in *Proc. IEEE ICASSP*, (Salt Lake City, UT), May 2001.
20. S. Renals and D. Ellis, “Audio information access from meeting rooms,” in *Proc. IEEE ICASSP 2003*, 2003.
21. R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, “Distributed meetings: A meeting capture and broadcasting system,” in *Proc. ACM Multimedia Conference*, 2002.
22. D. Gatica-Perez, G. Lathoud, I. McCowan, and J.-M. Odobez, “A mixed-state i-particle filter for multi-camera speaker tracking,” in *Proceedings of WOMTEC*, September 2003.
23. A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, 2001.
24. R. Cutler, “The distributed meetings system,” in *Proceedings of IEEE ICASSP 2003*, 2003.
25. V. Stanford, J. Garofolo, , and M. Michel, “The nist smart space and meeting room projects: Signals, acquisition, annotation, and metrics,” in *Proceedings of IEEE ICASSP 2003*, 2003.
26. H. Silverman, W. Patterson, J. Flanagan, and D. Rabinkin, “A digital processing system for source location and sound capture by large microphone arrays,” in *Proceedings of ICASSP 97*, April 1997.
27. E. Shriberg, A. Stolcke, and D. Baron, “Observations on overlap: findings and implications for automatic processing of multi-party conversation,” in *Proceedings of Eurospeech 2001*, vol. 2, pp. 1359–1362, 2001.
28. T. Pfau, D. Ellis, and A. Stolcke, “Multispeaker speech activity detection for the ICSI meeting recorder,” in *Proceedings of ASRU-01*, 2001.
29. T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, “Strategies for automatic segmentation of audio data,” in *Proceedings of ICASSP-2000*, 2000.
30. G. Lathoud and I. McCowan, “Location based speaker segmentation,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, April 2003.
31. G. Lathoud, I. McCowan, and D. Moore, “Segmenting multiple concurrent speakers using microphone arrays,” in *Proceedings of Eurospeech 2003*, September 2003.
32. J. Bitzer and K. U. Simmer, “Superdirective microphone arrays,” in *Microphone Arrays* (M. Brandstein and D. Ward, eds.), ch. 2, pp. 19–38, Springer, 2001.
33. I. McCowan and H. Bourlard, “Microphone array post-filter based on noise field coherence,” *To appear in IEEE Transactions on Speech and Audio Processing*, November 2003.
34. D. Moore and I. McCowan, “Microphone array speech recognition: Experiments on overlapping speech in meetings,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, April 2003.
35. A. Jain, R. Bolle, and S. Pankanti, *Biometrics: Person Identification in Networked Society.* Kluwer Publications, 1999.
36. J. Mariéthoz and S. Bengio, “A comparative study of adaptation methods for speaker verification,” in *Proceedings of the International Conference on Spoken Language Processing, ICSLP*, 2002.
37. S. Marcel and S. Bengio, “Improving face verification using skin color information,” in *Proceedings of the 16th International Conference on Pattern Recognition, ICPR*, IEEE Computer Society Press, 2002.

38. C. Sanderson and K. Paliwal, “Polynomial Features for Robust Face Authentication,” *Proceedings of International Conference on Image Processing*, vol. 3, pp. 997–1000, 2002.
39. S. Bengio, C. Marcel, S. Marcel, and J. Mariéthoz, “Confidence measures for multimodal identity verification,” *Information Fusion*, vol. 3, no. 4, pp. 267–276, 2002.
40. L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, 1993.
41. T. Starner and A. Pentland, “Visual recognition of american sign language using HMMs,” in *Proc. Int. Work. on Auto. Face and Gesture Recognition*, (Zurich), 1995.
42. S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Transactions on Multimedia*, vol. 2, pp. 141–151, September 2000.
43. S. Bengio, “An asynchronous hidden markov model for audio-visual speech recognition,” in *Advances in Neural Information Processing Systems, NIPS 15* (S. Becker, S. Thrun, and K. Obermayer, eds.), MIT Press, 2003.
44. I. McCowan, D. Gatica-Perez, S. Bengio, and G. Lathoud, “Automatic analysis of multimodal group actions in meetings,” Tech. Rep. RR 03–27, IDIAP, 2003.
45. B. De Gelder and J. Vroomen, “The perception of emotions by ear and by eye,” *Cognition and Emotion*, vol. 14, pp. 289–311, 2002.
46. P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. IEEE Int. Conf. on Computer Vision (CVPR)*, (Kawaii), Dec. 2001.
47. S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, “Learning human interactions with the influence model,” Tech. Rep. 539, MIT Media Laboratory, June 2001.

Experimental Evaluation of Variations in Primary Features Used for Accelerometric Context Recognition

Ernst A. Heinz[1], Kai S. Kunze[1], Stefan Sulistyo,[1]
Holger Junker[2], Paul Lukowicz[2], and Gerhard Tröster[2]

[1] Entertainment and Pervasive Computing Group, International University (IU) in Germany
{ernst.heinz, kaisteven.kunze, stefan.sulistyo}@i-u.de
http://www.i-u.de/schools/heinz/

[2] Wearable Computing Laboratory, Institute for Electronics (IfE), ETH Zürich, Switzerland
{junker, lukowicz, troester}@ife.ee.ethz.ch http://www.wearable.ethz.ch/

Abstract. The paper describes initial results in an ongoing project aimed at providing and analyzing standardized representative data sets for typical context recognition tasks. Such data sets can be used to develop user-independent feature sets and recognition algorithms. In addition, we aim to establish standard benchmark data sets that can be used for quantitative comparisons of different recognition methodologies. Benchmark data sets are commonly used in speech and image recognition, but so far none are available for general context recognition tasks. We outline the experimental considerations and procedures used to record the data in a controlled manner, observing strict experimental standards. We then discuss preliminary results obtained with common features on a well-understood scenario with 8 test subjects. The discussion shows that even for a small sample like this variations between subjects are substantial, thus underscoring the need for large representative data sets.

1 Introduction

Using simple sensors distributed over the user's body has recently emerged as a promising approach to context recognition in mobile and wearable systems. In particular, motion sensors such as accelerometers or gyroscopes have been shown to provide valuable information about user activity. Systems based on a 3 axes have been successfully trained to distinguish between everyday activities like walking, standing, and sitting [6,7,4]. It has also been shown that appropriate combinations of sensors attached to different limbs can provide information necessary to recognize specific complex activities like getting up and greeting an arriving person with a handshake [3] or operating a particular home appliance [5]. An important question that needs to be resolved before such systems can move towards real-life applications is their generalization capability. So far, most successful experiments have been performed in special settings such as a laboratory with a single or just a few randomly selected subjects. At this stage it is not clear how well those results generalize to different locations and other subjects. Thus, a system trained to differentiate between walking up or down a particular staircase with a young female subject might fail for another staircase and / or an older male subject. To a degree,

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 252–263, 2003.

generalization issues can be addressed through additional user- and situation-specific training. However there are limits to how much training effort a user is willing to put up with. Hence, it is desirable to be able to design and train systems to work under a wide variety of circumstances. To this end, at least three issues need to be addressed.

1. For given recognition tasks, it must be determined how sensor signals change for different relevant situations and users.
2. Sets of features as insensitive as possible to the variations above must be found.
3. Standard data sets with the relevant factors varied in a controlled manner over some desired value range must be generated to allow for situation-insensitive training.

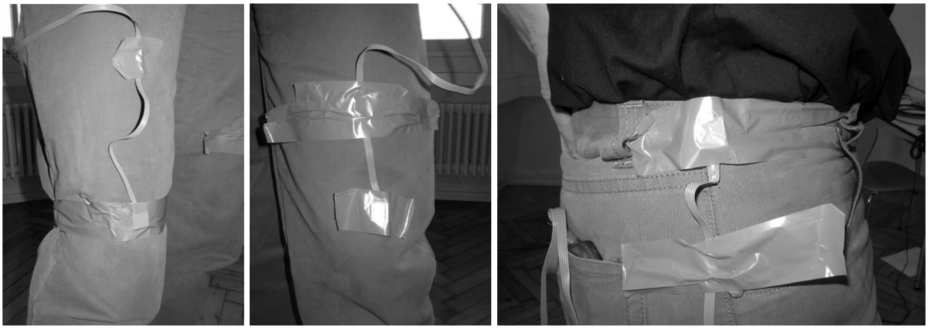
This paper describes some early results in an ongoing project addressing the above issues. The project aims to collect a representative data set for different typical context recognition tasks. The data is to be used to derive person-independent features and recognition algorithms. In addition, we intend to establish a publicly available benchmark database of standard context recognition tasks which can be used to objectively quantify and compare different approaches. Such benchmark databases are widely and successfully used in automatic speech and image recognition. In the remainder of this paper, we first outline the experimental considerations and procedures used to record the data in a controlled manner while observing strict experimental standards. We then discuss preliminary results obtained on simple scenario with 8 test subjects. While no statistically valid conclusions can be drawn from such a small sample, interesting effects can still be observed. In particular we show that variations between subjects do indeed have a strong impact on feature selection and recognition performance. This illustrates the importance of establishing standard data sets as intended by our project.

2 Experimental Procedure

The setup of the experiments was carefully planned and documented based upon trial runs performed in advance. We also explained the whole procedure in considerable detail to the test subjects beforehand, such that they could easily follow the prescribed route and tasks. The overall objective of the experiment was to gather acceleration data of selected body parts in different action contexts for multiple test subjects with varying physical characteristics. Of course, the whole experimental setup should only have a minimal effect on the normal physical movements of the test subjects if at all. The video recordings of the test subjects in action show that we fully achieved this aim.

2.1 Hardware Setup

The sensor system employed in the experiments draws its power from a standard USB connection and interfaces to any RS-232 serial port for control and data signal transmission. We used a Xybernaut Mobile Assistant IV (MA-IV) wearable computer with a 233 MHz Pentium-MMX CPU, 64 MB of RAM, and a 6 GB hard drive running SuSE Linux 6.2 to drive the sensor system. Because of the USB and serial ports needed, the full Xybernaut MA-IV system consisting of both the main unit and the port replicator had to be strapped to the test subjects. Fortunately, there was enough space left here to also



(a) on Right Leg (b) on Left Leg (c) on Right Hip

Fig. 1. Sensor Placement

affix the central components of the sensor system. Due to the professional belt packaging provided by Xybernaut for its wearable computers, this worked quite smoothly and did not hinder the test subjects in any perceivable way during the course of the experiment. Finally, three acceleration sensors were affixed to the test subjects as depicted by the three photographs shown in Figure 1:

(a) on *right leg* about 1 cm above knee cap, tightly fixed, exactly moving like the leg;
(b) on *left leg* about 1 cm above knee cap, lightly fixed, moving like the trousers;
(c) on *right hip*, tightly fixed, exactly moving like the hip.

The sensors on the legs were oriented with their z-axes in the direction of the movement and their y-axes pointing up. The sensors on the hip had their x-axes facing up and their z-axes to the right (i.e., their final orientation differed slightly from the picture above).

Fig. 2. A Single Flight of Stairs

2.2 Action Sequence

After being fully wired and connected, the test subjects were placed at the start of the experimental “trail” that led them through parts of Leibniz Hall, the building housing the School of IT at the International University (IU) in Germany. A small custom program (written in ANSI-C and running directly on the wearable computer) recorded the sensor readings throughout the whole trail by essentially dumping the raw data from the serial input port of the Xybernaut MA-IV. The list below characterizes in detail the stages and actions that the test subjects moved through.

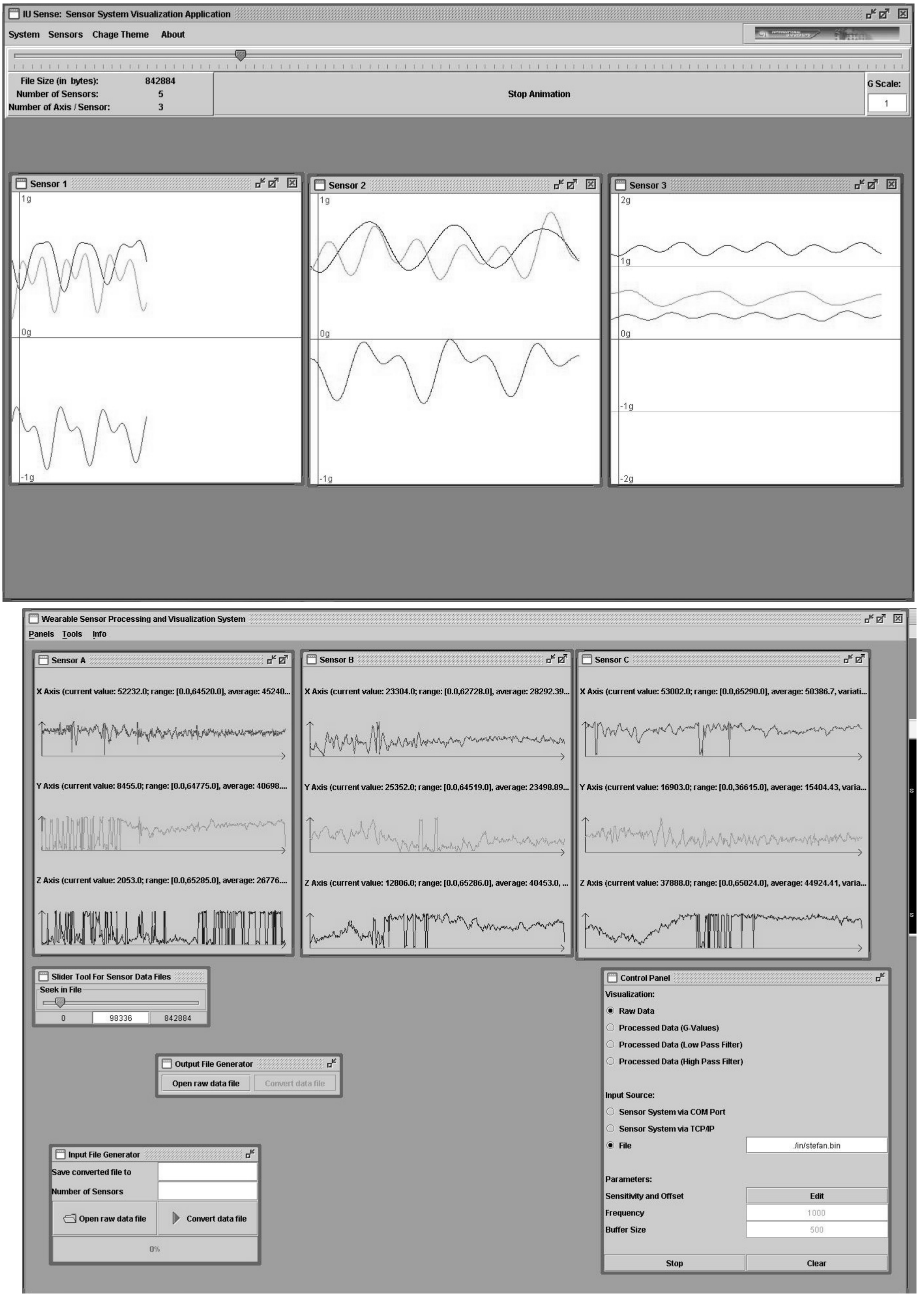
1. Start signal: 3x tapping on right knee.
2. Walk straight for about 9.09 m.
3. Open a door with right hand, walk through, and close it (door swings inwardly with its handle on the left and the hinges on the right side).
4. Turn right and walk straight for about 6.59 m to a staircase as depicted in Figure 2.
5. Climb down 6 flights of 11 stairs each with rightward turns after the first 5 flights (stair width: 30 cm, stair height: 15 cm, stair angle: 27° from the horizontal).
6. Walk straight for about 51.34 m to opposite staircase.
7. Climb up 6 flights of 11 stairs each with left-ward turns after the first 5 flights (same stair specifications as for downward climb).
8. Walk straight at fast pace for about 44.13 m back to the door of the starting room.
9. Stop signal: 3x tapping on right knee again.

2.3 Data Recording and Processing

In order to gain a better understanding of the test subjects themselves and their physical characteristics in particular, we also requested brief individual profiles from them specifying the following details:

- age, gender, height, weight, and right- or left-handedness;
- frequency of sports activities (daily, weekly, monthly, yearly, or never on average);
- additional information such as injuries, handicaps, nationality, and other specials.

The whole procedure was supervised by at least one experiment conductor, who was also responsible for filming the whole sequence with a video camera. Post-experiment activities included verifying that the recorded files actually contained valid data, visually checking the video film for inconsistencies regarding the prescribed sequences of actions, and – last but not least – processing the recorded raw data and transforming it into meaningful visual representations. This was done using IU Sense [1,2] and MatLab (release 12). IU Sense, developed by IU students and written in Java, is an extensible real-time application with graphical data displays in various views and user-configurable layouts (see Figure 3). It also provides numerous possibilities to perform off-line data visualization from files and conversion into several different file formats.



File Size (in bytes): 842884
Number of Sensors: 5
Number of Axis / Sensor: 3

G Scale: 1

Sensor A

X Axis (current value: 52232.0; range: [0.0,64520.0], average: 45240...

Y Axis (current value: 8455.0; range: [0.0,64775.0], average: 40698....

Z Axis (current value: 2053.0; range: [0.0,65285.0], average: 26776....

Sensor B

X Axis (current value: 23304.0; range: [0.0,62728.0], average: 28292.39...

Y Axis (current value: 25352.0; range: [0.0,64519.0], average: 23498.89...

Z Axis (current value: 12806.0; range: [0.0,65286.0], average: 40453.0, ...

Sensor C

X Axis (current value: 53002.0; range: [0.0,65290.0], average: 50386.7, variati...

Y Axis (current value: 16903.0; range: [0.0,36615.0], average: 15404.43, varia...

Z Axis (current value: 37888.0; range: [0.0,65024.0], average: 44924.41, varia...

Seek in File
0 98336 842884

Save converted file to
Number of Sensors
0%

Visualization:
● Raw Data
○ Processed Data (G-Values)
○ Processed Data (Low Pass Filter)
○ Processed Data (High Pass Filter)

Input Source:
○ Sensor System via COM Port
○ Sensor System via TCP/IP
● File ./in/stefan.bin

Parameters:
Sensitivity and Offset: Edit
Frequency: 1000
Buffer Size: 500

Fig. 3. IU SENSE: Graphical Data Display

2.4 Problems and Future Improvements

The trickiest task of the whole experimental setup is the proper fixation of the sensor equipment. This still takes too much time and remains somewhat sensitive to human error. In the long run, the sensors and their interconnect will be woven into some clothes directly. Until then, however, we need to be content with less professional yet nevertheless reliable alternatives. Towards this end, we have discussed and tried many different approaches based on a variety of materials and fixation schemes: strong plaster tape wrapped around the sensors and the test subjects' body parts, strings tied around the leg and knotted together, so-called "velcro straps" as used for portable music players, rubber bands shifted over the leg, etc. Solely the first of these approaches, namely the strong plaster tape, worked reliably enough in practice to be considered adequate by us. Despite its ad-hoc look, the tape solution works surprisingly well. Its main shortcomings are prolonged setup times, minor imprecisions in sensor placement, and some issues of holding strength depending on the test subjects' type of clothes. A better solution than the taping might be achieved by means of an elastic kind of bandage to be strapped over the clothes, with the sensors and their connecting cables permanently affixed to it. This is certainly worth a try.

Future experimental runs may also be improved by wireless access to the wearable computer for remote real-time control and display of the sensor data readings and recordings. Moreover, the Xybernaut MA-IV is still somewhat bulky. A smaller platform, such as a PDA like the HP-Compaq iPAQ for instance, surely impacts the test subjects even less. In fact, the iPAQ platform was our initial first choice. Unfortunately, it did not really work out in practice because some of the available iPAQ devices suffered from hardware problems connecting to the sensor system. A more light-weight setup would also allow for more precise and specific body measurements and sensor fixation within the test subjects' hip areas.

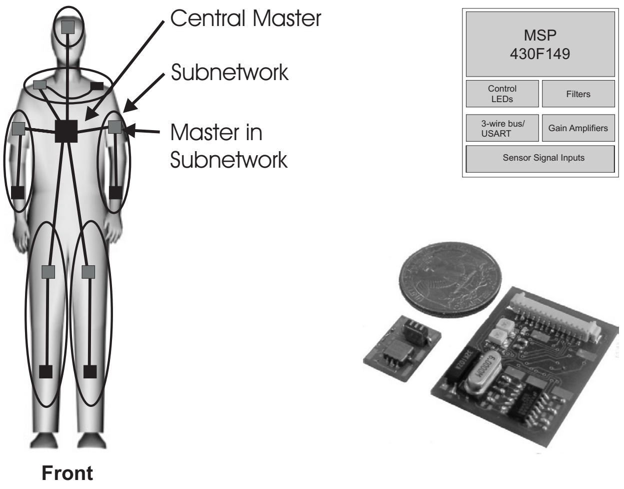


Fig. 4. The PADNET Wearable Sensor System

3 Sensor Platform

The experiments were conducted using PADNET ("Physical Activity Detection Network", see Figure 4), a sensor platform developed for user activity recognition [5]. It has been designed as a wearable system and allows for the easy distribution of multiple sensors over a person's body while being flexible. The platform consists of multiple sensor nodes interconnected in a hierarchical network. The purpose of a sensor node is to provide a physical interface for different sensor types (accelerometers, gyroscopes, magnetic field sensors, etc.), to read out the corresponding sensor signal, to provide certain computational power for signal pre-processing, and to enable communication with all other network components. Figure 4 shows such a sensor node with its logical block diagram. For the experiments, three 3D-accelerometers ADXL202E from Analog Devices were used. The analog signals from the sensors were low-pass filtered (fcutoff = 50 Hz) and A/D-converted with 12-bit resolution at a sampling rate of 100 Hz.

4 Initial Results

In the initial experimental phase of our ongoing project, we recorded data for 8 different test subjects. Of these 8, one data set has proven unusable due to a technical problem. The remaining 7 data sets were then examined with respect to three features typically used in accelerometric activity recognition:

- *root mean square* (RMS) of the signal, giving the average power of the signal;
- *cumulative sum* over the signal (sums);
- *variance* of the signal.

Of course, with just 7 subjects no statistically valid conclusions may be drawn. However, even with such a small data set useful observations can be made. In particular, proving the existence of certain phenomena such as feature variations and their consequences on classification requires isolated examples only.

4.1 Feature Variations

As a first step we investigate the statistical distribution of features for all subjects and four related context classes: walking (class #2), going down stairs (class #3), going up stairs (class #4), and walking fast (class #8). This is done for the downward-pointing axis (which is the most relevant one for all types of walking) of the hip and upper leg sensor. The results can be seen in Figures 5 to 7. The figures show the the mean value (a point) and the variance (an error bar) of each feature for all test subjects and context classes. It is interesting to note that the RMS and sums features seem to show little variations, both for each test subject individually and also between them. The only exception is test subject 6, which is totally out of range. Interestingly, from Table 1 it can be seen that nothing seems to be particularly special about this test subject except for his left-handedness which does not provide a plausible explanation for such a different walking style in our opinion. Taken together, Figures 5 to 7 indicate that variations between subjects can be expected to be a serious problem for context recognition.

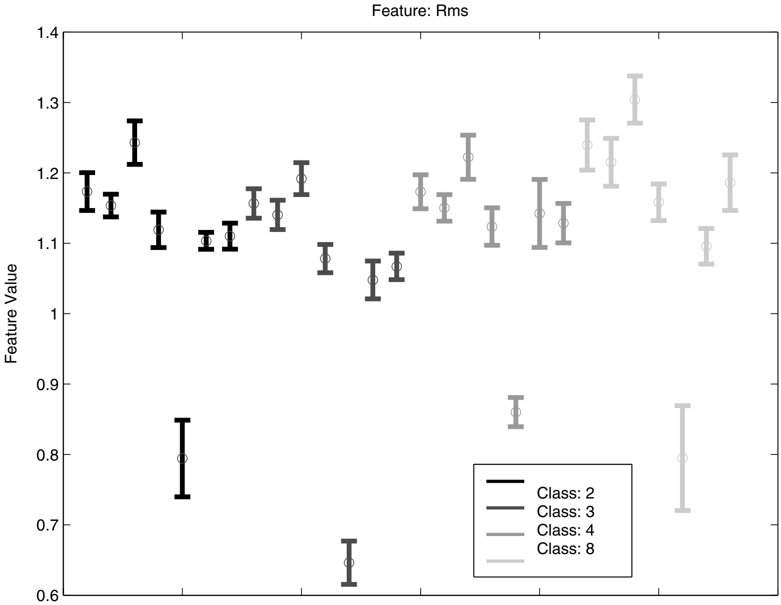


Fig. 5. Variance in the RMS Feature

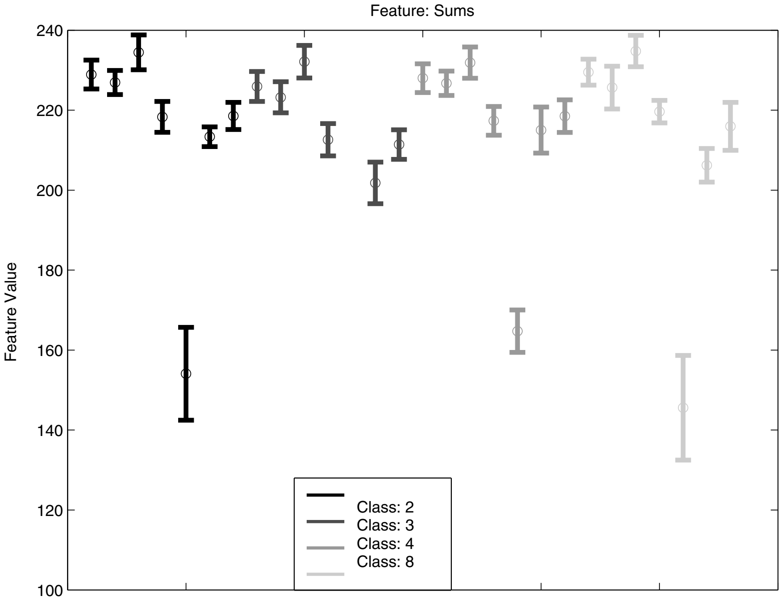


Fig. 6. Variance in the Sums Feature

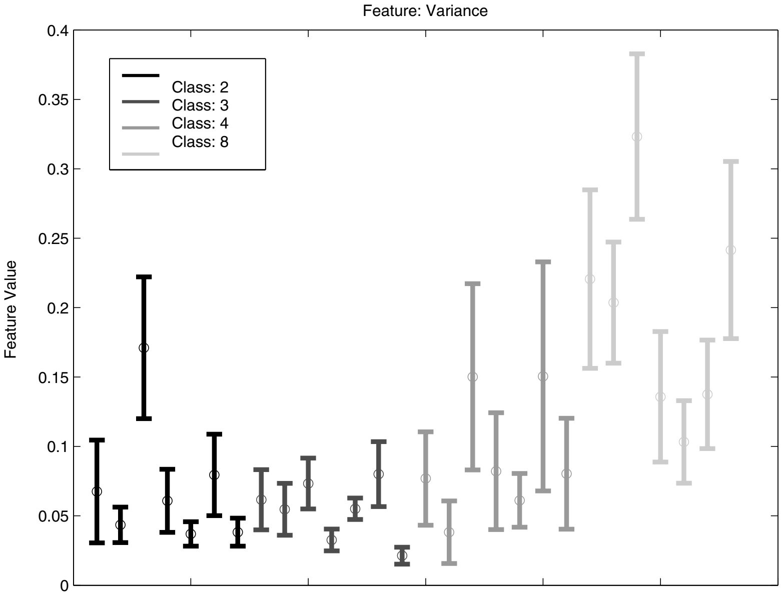


Fig. 7. Variance in the Variance Feature

Table 1. Characteristics of the 8 Test Subjects

No.	**Gender**	**Age** (Years)	**Height** (cm)	**Weight** (kg)	**Sports**	**Handedness**	**Special Notes**
1	male	24	173	75	weekly	right	flat feet
2	male	24	182	78	weekly	right	
3	male	36	180	90–100	monthly	right	slightly out-of-sync right foot (injury)
4	male	65	188	86	weekly	right	
5	female	23	171	62–65	weekly	right	X-like legs, asymmetrical kneecaps, wearing flip-flop sandals
6	male	21	172	70	daily	left	
7	male	23	191	110	weekly	right	wearing flip-flop sandals
8	female	23	170	50	weekly	right	wearing skirt

4.2 Usefulness for Classification

The question of usefulness of a set of features for person-independent classification involves at least two prominent aspects. First, we need to find out whether the set is universal enough to provide separation between the relevant classes for all subjects. As a stronger second requirement, we then need to see if a single separation plane can be found for all subjects enabling a person-independent system design. In our experiment, we have found that for the recognition of any two classes only (in particular, fast walking with any other) reasonable separation is achieved for all subjects by a combination of the variance and sums features. However, combining data from all subjects and then using single separation has not produced satisfactory results. When considering all 4 classes we have found the features to produce excellent separation for test subject 6. For all the others the separation was mixed to poor as illustrated in Figures 8 to 10. These figures show the values of the variance and sums features plotted in different colors using distinct shapes for each class. The main lesson of the above is that features that work very well for some people might not work at all for others. It also shows that widely used features such as RMS or variance are not directly suitable for person-independent systems. Instead, representative data sets are needed to derive more appropriate features sets.

5 Future Work

Our ongoing research project and collaboration intends to collect further data samples by repeating the exact experimental procedure described above with many more test subjects. In the end, we plan to collect enough samples to allow for statistically confident recognition schemes of the features over the whole test population. Moreover, we will make the full data and documentation of all the experiments publicly available on the Internet as soon as possible. Due to the careful design and in-depth documentation of our experimental procedure, we are convinced that our data are of common interest and value. They may even serve the purpose of a general benchmark set.

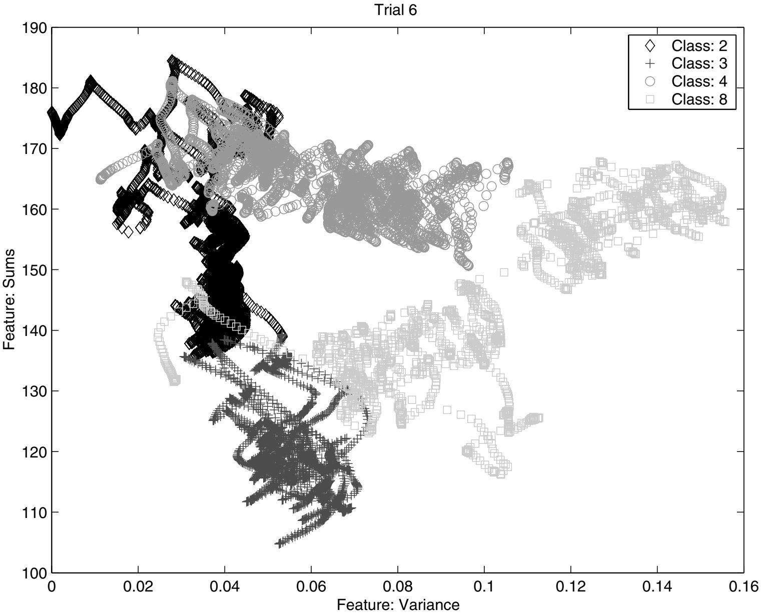


Fig. 8. Good Variance / Sums Classification

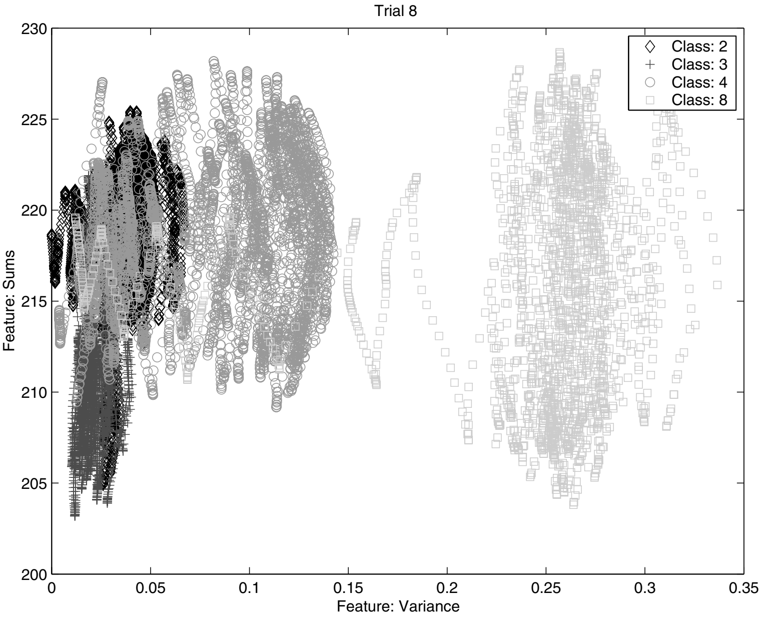


Fig. 9. Mixed Variance / Sums Classification

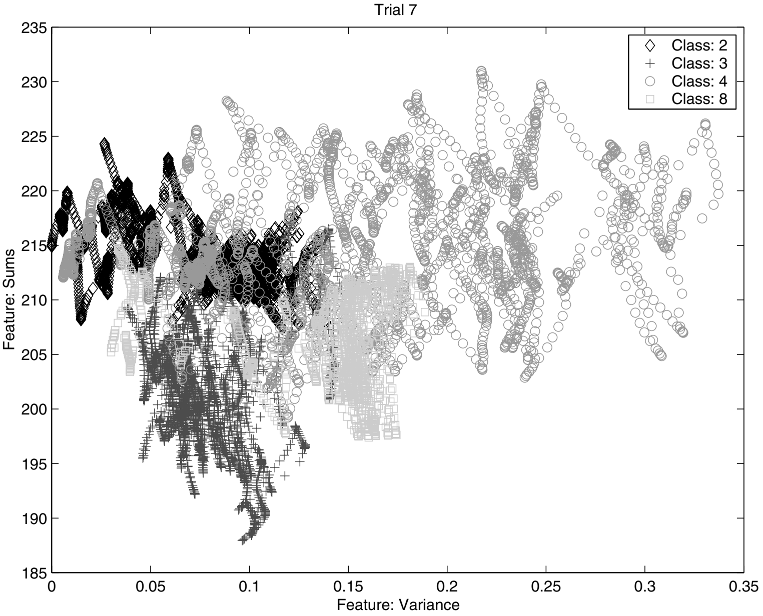


Fig. 10. Poor Variance / Sums Classification

References

1. A. Caracas, E.A. Heinz, P. Robbel, A. Singh, F. Walter, and P. Lukowicz. Real-time sensor processing with graphical data display in Java. Submitted to 3rd International Symposium on Signal Processing and Information Technology (ISSPIT), takes place in December 2003
2. H. Junker, P. Lukowicz, A. Caracas, K.S. Kunze, P. Robbel, A. Singh, S. Sulistyo, F. Walter, and E.A. Heinz. Demonstration of integrated CMOS acceleration sensors and real-time processing software with graphical data display in Java. Description of live demonstration at 1st European Symposium on Ambient Intelligence (EUSAI), November 2003
3. N. Kern, B. Schiele, H. Junker, P. Lukowicz, and G. Tröster. Wearable sensing to annotate meeting recordings. In Proceedings of the 6th International Symposium on Wearable Computers (ISWC), pp. 186–193, IEEE Press, October 2002
4. K. van Laerhoven and O. Cakmakci. What shall we teach our pants? In Proceedings of the 4th International Symposium on Wearable Computers, pp. 77–83, IEEE Press, October 2000
5. P. Lukowicz, H. Junker, M. Stäger, T.v. Büren, and G. Tröster. WearNET: A distributed multi-sensor system for context-aware wearables. In Proceedings of the 4th International Conference on Ubiquitous Computing (UbiComp), G. Borriello and L.E. Holmquist (eds.), pp. 361–370, Springer-Verlag, LNCS 2498, September 2002
6. A. Madabhushi and J. Aggarwal. Using head movement to recognize activity. In Proceedings of the 15th International Conference on Pattern Recognition, Vol. 4, pp. 698–701, IEEE Press, September 2000
7. C. Randell and H. Muller. Context awareness by analysing accelerometer data. In Proceedings of the 4th International Symposium on Wearable Computers, poster paper, pp. 175–176, IEEE Press, October 2000

Lino, the User-Interface Robot

Ben J.A. Kröse[1], Josep M. Porta[1], Albert J.N. van Breemen[2], Ko Crucq[2],
Marnix Nuttin[3], and Eric Demeester[3]

[1] University of Amsterdam,
Kruislaan 403, 1098SJ, Amsterdam, The Netherlands
{krose,porta}@science.uva.nl
[2] Philips Research,
Prof. Holstlaan 4, 5656AA, Eindhoven, The Netherlands
{albert.van.breemen,ko.crucq}@philips.com
[3] Katholieke Universiteit Leuven
Celestijnenlaan 300B, B-3001, Leuven (Heverlee), Belgium
{marnix.nuttin,eric.demeester}@mech.kuleuven.ac.be

Abstract. This paper reports on the development of a domestic user-interface robot that is able to have a natural human interaction by speech and emotional feedback and is able to navigate in a home environment. The natural interaction with the user is achieved by means of a mechanical head able to express emotions. The robot is aware of the position and identities of the users, both from visual and auditory information. The robot estimates its location in the environment with an appearance-based localization method using a stereo camera system. The navigation to the goal is achieved with a hybrid method, combining planning with reactive control. The robot is designed to operate in an intelligent environment, such that external information can be used to localize users and their intentions (context awareness), and that additional information can be retrieved from various databases in the environment. The result is a service robot that can have a simple dialogue with the user, provide information in a natural way (speech and expressions) and can be instructed to navigate to any specific goal in the environment.

1 Introduction

In the last years an increasing effort is spent in research on service and entertainment robots which operate in natural environments and interact with humans. The Sony AIBO is an example of a robot which is meant to play with children: it has a perceptual system (vision, auditory, tactile), plays soccer, and can learn its own behavior [1]. NEC has developed "Papero", a *personal* robot which also is able to entertain the user but has more functionality: it serves as an interfacing with web-services and electronic equipment [18]. Even more functionality is present in various other service robots, such as robot-waiters [11], museum or exhibition robots [21],[3] or care-for-elderly robots [10], all examples of autonomous intelligent systems, operating in a real world.

Parallel to these robotic developments, a new paradigm in information technology is emerging, in which people are served by a digital environment that

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 264–274, 2003.

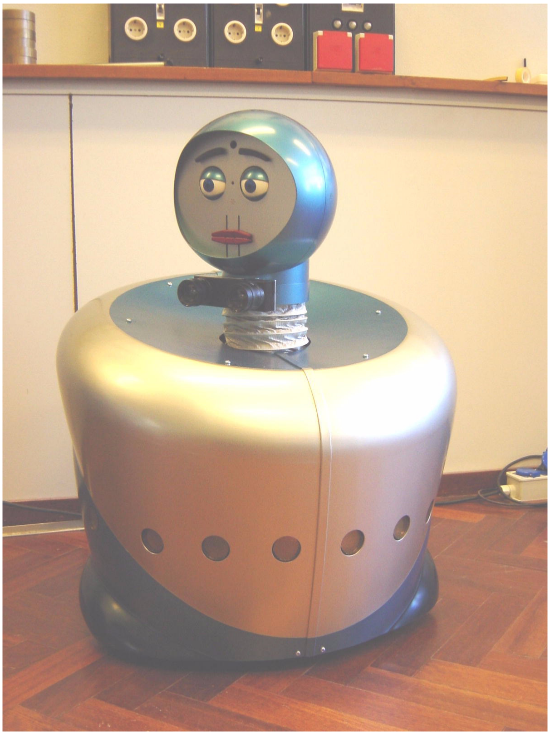


Fig. 1. The robot Lino.

is aware of their presence and context, and is responsive to their needs, habits, gestures and emotions: ambient intelligence. Apart from the many challenges in networking technologies, perception and intelligence, there is an enormous challenge in the field of user interaction: is the user is going to talk to his or her toaster or coffee machine...? We think not.

As a part of the European project "Ambience" [13] we developed a domestic robot (see Figure 1). The robot must be some personification of the intelligent environment, and it must be able to show intelligent behavior, context awareness and natural interaction. The robot exploits the intelligent environment to get information about the user intentions, preferences, etc. In the other way around, the human user must be able to have a natural interaction with the digital world by means of the robot.

Very important for the natural interaction is a nice look of the robot, and the possibility to express some emotional state. Many other robots use a (touch)screen interface, sometimes with an animated face [11],[7]. We decided to use a 'real' face, consisting of dynamic mouth, eyes and eyebrows since this makes the interaction more attractive and also more natural.

The development of software modules for different tasks is carried out by multiple developers. Therefore, we have implemented a dedicated software tool to support the developers. Using this tool, different software modules of the robot application, running on different operating systems and computers, can be connected/disconnected interactively at runtime by means of a graphical user interface. With this approach, integrating the different software components is a matter of "configuration" rather than programming.

The objective of this paper is to introduce the different modules developed for our robot, the software tools used to integrate them, and the preliminary results we have obtained so far.

2 Software Framework

The architecture is depicted in Figure 2. An efficient implementation and integration of all the different functional software components requires a dedicated software framework. We have developed a module-based software framework, called the *Dynamic Module Library* that services this purpose. The basic software construct is that of a *module* that has input and output ports, which can be connected to each other to exchange data. The framework meets the following requirements:

Runtime flexibility. The possibility to change algorithms of modules at runtime, to extend the robot application with new modules at run-time, and to probe ingoing and outgoing data of modules at runtime.

Runtime robustness. Stopping (or crashing) one or more modules of a running robot application should not result into an overall stopping (or crashing) of the robot application.

Runtime configurability. The possibility to define at runtime the configuration of modules (that is, the connections between the modules) that make up the robot application.

Distribution. The possibility to distribute the robot application over several host computers in order to enlarge the computational resources.

Portability. Support for the most popular programming languages (C, C++, Java) and operating systems (Windows, Linux).

Modules are implemented separately from each other and are executed as individual processes in some operating system (both MS Windows and Linux are currently being supported). By using a registry, modules can discover and lookup each other at run-time. Once a module has found an other module, it can make connections to the ports of that other module. It is also possible to externally connect ports of modules. By means of the graphical user interface we can start/stop modules, connect/disconnect ports as well as probing ports. This way, a robot application consisting of modules can be configured at runtime which greatly enhances the interactivity of the development of the robot application.

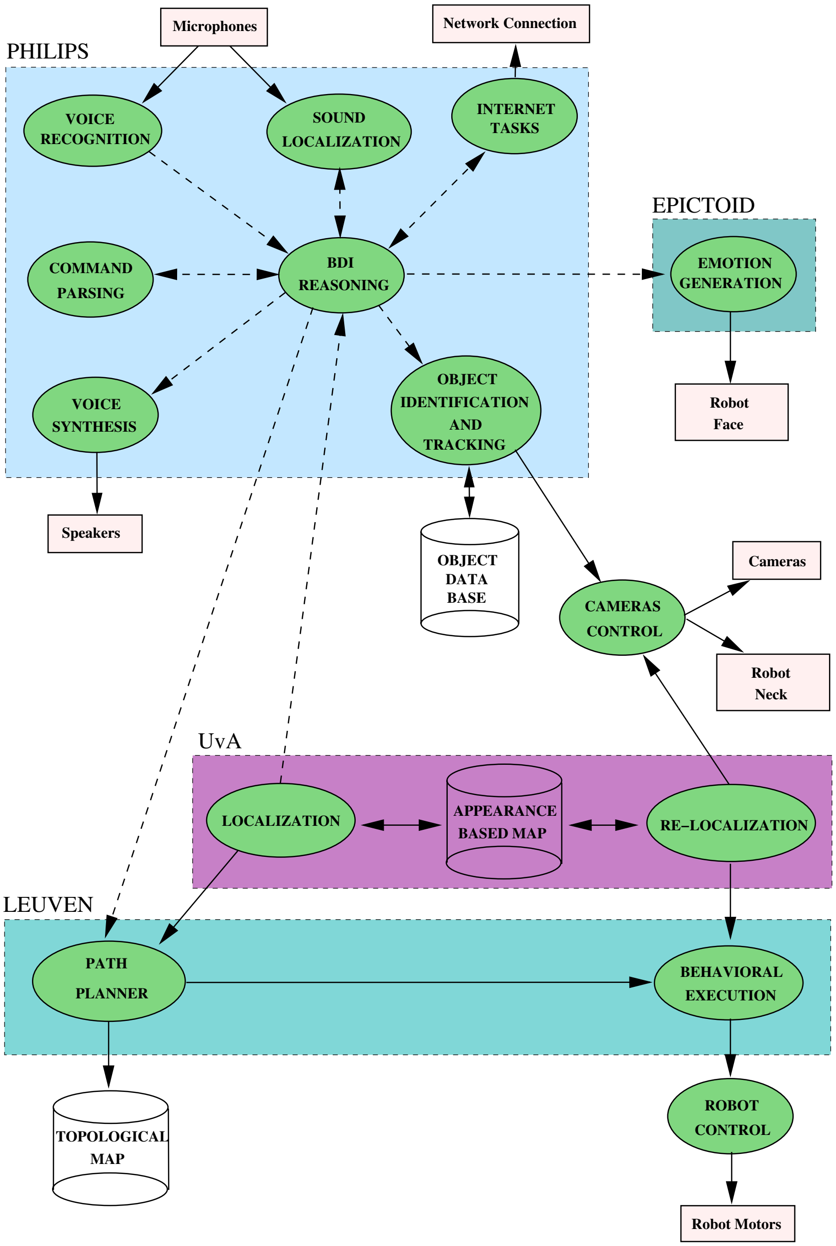


Fig. 2. Set-up of the software modules in Lino

In order to synchronize the tasks of all the modules, each module implements a particular task model. The model consists of five states in which a module can occur, namely `IDLE`, `RUNNING`, `PAUSED`, `SUCCEEDED` and `FAILED`. Each module can read as well as control the state of other modules. For instance, a Reasoning module can send an `execute` message to a Re-localization module to start this one up, and a `pause` message to a pathplanner module to pause its task.

3 User Awareness Module

We have implemented and tested one of the tasks of the robot called "Turn-to-speaker" behavior. This behavior basically determines the direction of a speaker in a 3D space and turns the head and body toward the speaker. The 3D speaker location estimation is determined by means of three mutually perpendicular microphone pairs. These microphones are mounted inside the head and are separated by a distance of 25 cm. Each microphone pair uses a stereo USB audio digitizer for the signal acquisition. We analyze the recorded signals to determine the difference in the time of flight of speech that arrives. The basic problem in this measurement is to get rid of the numerous acoustic reflections in the recorded speech signals. With an adaptive Filtered-Sum Beamformer and an optimization algorithm [4],[5] it is possible to determine the contribution of these reflections and to largely compensate for them in the recorded signals.

The location of the speaker is indicated in the local robot coordinate system by two angles, φ (horizontal plane), θ (tilt). The angle φ is used to turn the robot platform and θ is used to turn the head up to the speaker with the loudest voice. There is a problem when there are many speakers at the same level. In this case the system generates inconsistent values and we pause the turning of the robot. The system does not respond to random acoustic noise. It detects human voices by looking for harmonics (the pitch) in the recorded signal. With this technique we also want to explore "Follow-me" behavior: e.g. advancing small distances in the direction of someone speaking at regular intervals.

4 Emotion Engine

To generate the appropriate facial expressions and body language we have developed an emotion engine. This emotion engine autonomously reasons on the emotional state of the robot and is based on the psychological model of Ortony, Clore and Collins (OCC-model) [17]. The model has been applied successfully in other studies [9],[15]. The OCC-model reasons about the synthesis of emotions based on appraisal of consequences of events, of actions (self or of others) and of aspects of objects. The appraisal is evaluated by comparing the events (occurred or occurring) with the goals, the actions with the standards set and the appealingness of the objects with the attitudes set. Also the history is taken into account. Via a decision tree in which distinction is made between (1) consequences for self or others, (2) action of self or of others, (3) positive or negative aspects, (4) prospect relevant or irrelevant, (5) present or future, (6) desirable or undesirable a total of 22 possible emotional states result. The intensity of the emotion is determined by modelling functions for the desirability, the likelihood of occurrence, the appealingness, the praiseworthiness. Once the emotional state is determined the mapping to facial expressions (22) has to be achieved. This mapping has been carefully researched and devised by Epictoid.

5 Speech

We have implemented an interactive command and control dialogue system based on the SAPI of Microsoft. The speech recognition engine (SRE) is from Lernout & Hauspie (L&H) and the text to speech engine (TTS) is from AT&T. A simple dialogue management system has been devised which basically functions by transitions to different states, e.g. ”sleeping” → ”listing” → ”TV-control” → ”DVD-control” → ”idle”. After start-up the system is still in the ”sleeping” state. Issuing the command ” Lino wake up” will bring the system in the ”listening” state. In this state the system is ready to all kind of services, e.g. ”switch TV on”, ”set channel to ned1” or ”tell me what object do you see”. With the command ”go sleep” the system can be brought back into the sleeping mode. The system automatically switches to the ”idle” state after 20 sec.

Furthermore, the speech signal is also used to identify the current speaker. The algorithm we use is based on a Gaussian Mixture Model (GMM) [16]

The viseme output of the speech synthesizer is used to control and synchronize the lip movement during speech output. This lip synchronization contributes a lot to a more lively appearance.

6 Vision

Object detection, tracking and recognition is a very important capability for a robot. For the implementation of this module we have used the Inca-plus camera [12]. This camera is a stand-alone system and is normally used for machine vision. The complete image processing is locally done in the camera by means of two powerful processors; i.e. the Xetal for the pre-processing and the Trimedia for the actual image processing. In this way we have achieved a throughput rate of 10 Hz and higher. Only the CMOS-sensor chip has been mounted in the forehead of the robot. Currently, the robot is able to detect and recognize simple objects by means of the color and the overall shape (aspect ratio). Once detected, the eyes track the moving object which gives visual feedback that is appreciated very much by user. By means of speech output Lino can report what object it sees. Currently, we are implementing face detection and recognition.

7 Localization and Navigation

In order to navigate to a desired location, the robot must be able to *localize* itself, it has to *plan* a path and it has to *avoid obstacles* while following the path. For

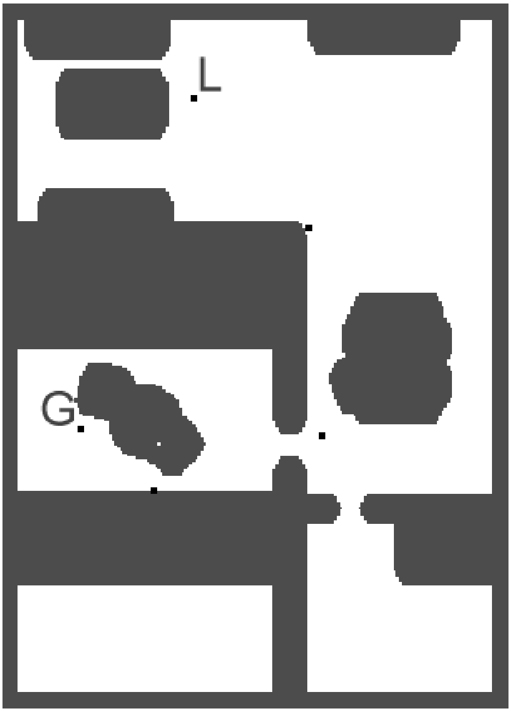


Fig. 3. The planner uses a map to generate subgoals (the small dots in the figure) from the current location *L* to the goal position *G*.

estimating its position, the robot has to compare its sensoric measurements with an internal representation. Of course it can also measure its wheel revolutions (odometry), but this is subjected to large errors in the long term. Consequently, additional sources of information must be used to determine the robot's position. For the Lino robot we use a vision system for this.

Our localization method is an 'appearance-based' method, which departs from a training set of images taken at known positions. The image set is transformed to a set of linear feature vectors and stored on the robot. For a robust localization we use a Markov procedure, where the 'belief' in the location of the robot is updated using new observations. Since the robot can turn its head, we actively acquire the best observations. A description of the probabilistic model, the Monte Carlo implementation and the active vision strategy is given in [14],[23],[19]. An additional advantage of using a stereo vision system is that it can provide depth maps that are less sensitive to change in illumination than usual intensity images. Recently we combined both modalities [20].

The objective of the navigation module is to drive the robot to the desired positions avoiding collision with obstacles. To accomplish this objective, a hybrid architecture is used in the navigation module. There are two internal modules: the planner and the behavioral execution module.

The planner generates subgoals from the robot's current location to the global goal position using a map. In previous work, a computationally efficient planner was developed based on the Wave Front Distance Field (WFDF) algorithm; see [22] for more details. This planner finds the shortest path to the goal and calculates subgoals on this path. The straight path between two consecutive subgoals is obstacle free. In a final step, subgoals that are close to each other are merged. This avoids that subgoals are too close together, which is not desirable in behavioral execution. The efficiency of the algorithm allows to re-plan approximately four times every second, coping with the robot getting off course in front of obstacles. Figure 3 shows the subgoals in the HomeLab (the domestic test environment at Philips Research Eindhoven, The Netherlands), from the robot's current location (L) to a certain goal position (G). The planner outputs the desired change in position in order to reach the first calculated subgoal. By doing so, this ultimately leads the robot to his final goal. The information provided by the localization module is used determine the position of the robot and, thus, to keep track of the change in position already achieved.

The second component of the navigation architecture, the behavioral execution module, receives as input (from the planer) the desired relative displacement for the robot and determines the linear and angular speeds (v and ω, respectively) necessary to perform it. Then, these speeds can be readily transformed to wheel motor commands. The behavioral execution is implemented using a behavior-based control approach. We refer to [2] for an introduction to behavior-based robots. Obstacles which are not in the map, both static and dynamic, possibly show up in front of the robot while moving. To avoid bumping into them, an avoid-obstacle algorithm is implemented. Ultrasonic sensors are used to detect these obstacles.

The cooperation of the fast planner module and the behavioral execution one leads the robot to his goals.

8 High Level Reasoning Module

In order for the robot to realize high level goals it must be capable of reasoning about the information it has about its world. A flexible reasoning mechanism that is dedicate to operate in such a practical problem domain as the domestic user environment is essential for a proper functioning of the robot. We plan to use the Belief, Desires and Intention (BDI) architecture that is well known in the field of agent and multi-agent systems. For the experiments we wrote scenario's in a language "Q" and incorporated this in an expert system (CLIPS).

9 Test Results

As far emotion generation is concerned, Figure 4 show some pictures of the head with different facial expressions. Although the actual scientific evaluation

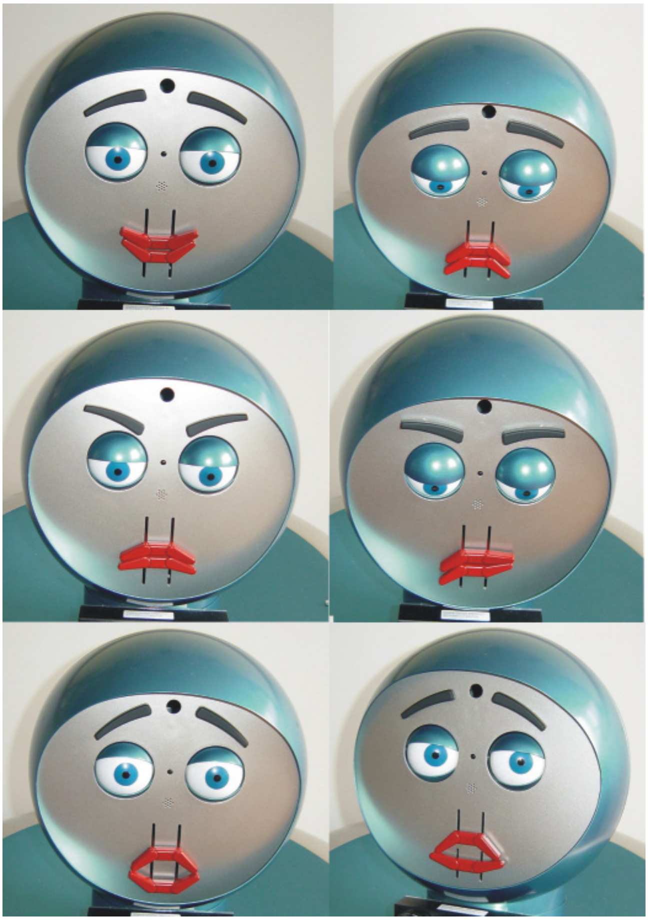


Fig. 4. 3D mechanical head expressing emotions.

still has to be done we have had some first very positive reactions from extensive demonstrations during one week exhibition for totally over 700 people. These demonstrations were conducted by three different relatively inexperienced users (i.e. not the robot developers). By means of some simple dialogues speech recognition, speech synthesis with lip synchronization, emotion generation (facial expressions), object recognition and turn-to-speaker were successfully demonstrated. The general reaction of the observers was appreciation and pleasure.

10 Conclusions

This paper reported the results we have obtained during the on-going development of our domestic user-interface robot. To realize emotional feedback we have built a mechanical 3D head which is controlled by 17 standard RC-servo-motors. The head can express six basic emotional facial expressions. The robot is able to determine the position of the user localizing the origin of any person speaking near him. Additionally, the robot can gather information from the ambient intelligence in which it is assumed to operate and, in the other way around, it can redirect user commands to this environment.

The robot can localize himself in the environment using stereo images and the so-called appearance-based approach. This approach is appealing for its simplicity and due to the stereo vision less sensitive to change in illumination. On the basis of a proper localization, navigation is performed by using two modules: a planner and a behavioral execution module. The planner module calculates subgoal positions for the behavioral execution module in order to prevent getting stuck by obstacles. The Wave Front Distance Field algorithm is used by the planner to calculate the subgoals.

All the modules of our robot are controlled and coordinated in a flexible way using a central controller.

Finally, we presented our software development framework called the Dynamic Module Library. This framework is a state-of-the-art software tool to implement distributed robot applications. An application is runtime configurable by means of a graphical console: the robot application software modules can be probed, started, stopped, removed, added, and connected to each other on-line.

Our project represents a link between two *service to humans* paradigms: service robots and ambient intelligence. Hopefully, other fruitful cooperations would emerge between these two field in the next years.

Acknowledgements. This work has been supported by the European (ITEA) project *“Ambience: Context Aware Environments for Ambient Services”*.

References

1. Aibo, http://www.aibo.com, 2002
2. R. C. Arkin, “Behaviour Based Robotics”, MIT Press (1997)
3. Arras, K.O., Philippsen, R., Tomatis, N., de Battista, M., Schilt, M. and Siegwart, R. “A Navigation Framework for Multiple Mobile Robots and its Application at the Expo.02 Exhibition”, in Proceedings of the IEEE International Conference on Robotics and Automation (2003), Taipei, Taiwan
4. H.J.W. Belt and C.P. Janse, “Audio Processing Arrangement with Multiple Sources”, Patent application PHN 16638 EP-P (1998)
5. H.J.W. Belt and C.P. Janse, “Time delay Estimation from Impulse Responses”, Patent application PHN 017163
6. M.E. Bratman, D.J. Israel & M.E. Pollack, “Plans and Resource-Bounded Practical Reasoning”, Computational Intelligence, 4(4), (1988) 349–355

7. A. Bruce, I. Nourbakhsh and R. Simmons, "The Role of Expressiveness and Attention in Human-Robot Interaction", in Proceedings of the 2002 IEEE International Conference on Robotics and Automation Washington, DC (2002), pp. 4138–4142
8. J. Cassell, "Embodied Conversational Agents: Representation and Intelligence in User Interface", AI magazine,22(3) (2001), 67–83
9. Elliot, C.D., "The Affective Reasoner: A Process model of emotions in a multi-agent system", Ph.D. Thesis, The Institute for the Learning Sciences, Northwestern University, Evanston, Illinois, (1992)
10. A. J. Davison, M. Montemerlo, J. Pineau, N. Roy, S. Thrun and V. Verma, "Experiences with a Mobile Robotic Guide for the Elderly", in Proceedings of the AAAI National Conference on Artificial Intelligence (2002)
11. P. Elinas, J. Hoey, D. Lahey, J. D. Montgomery, D. Murray, S.S. James and J. Little, "Waiting with José, a vision-based mobile robot" in Proceedings of the 2002 IEEE International Conference on Robotics and Automation Washington, DC (2002), pp. 3698–3705
12. Inca plus camera, http://www.cft.philips.com/industrialvision. 2002
13. ITEA Ambience project,
http://www.extra.research.philips.com/ euprojects/ambience/
14. B.J.A. Kröse, N. Vlassis, R. Bunschoten and Y. Motomura, "A probabilistic model for appearance-based robot localization", Image and Vision Computing, 19(6) (2001), 381–391
15. O'Reilly,W.S.N., "Believable Social and Emotional Agents", Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, (1996)
16. D.A. Reynold and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture models", IEE Trans. Speech and Audio processing, vol 3, no1 (1995), pp 72–83
17. A. Ortony, A. Clore and G. Collins, "The Cognitive Structure of Emotions", Cambridge University press, Cambridge, England (1988)
18. Papero, http://www.incx.nec.co.jp/robot/PaPeRo/english/p_index.html. (2002)
19. J.M. Porta, B. Terwijn and B. Kröse, "Efficient Entropy-Based Action Selection for Appearance-Based Robot Localization", In Proc. IEEE Int. Conf. on Robotics and Automation, Taipei, (2003) to appear
20. Josep M. Porta and Ben Kröse, "Enhancing Appearance-based Robot Localization Using Non-Dense Disparity Maps", In Proceedings of the International Conference on Robotics and Intelligent Systems (IROS), Las Vegas, USA, (2003) To appear
21. S. Thrun, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C.R. Rosenberg, N. Roy, J. Schulte and D. Schulz, "MINERVA: A Tour-Guide Robot that Learns", {KI} – Kunstliche Intelligenz, (1999) 14–26
22. J. Vandorpe, "Navigation Techniques for the mobile robot LiAS", PhD Dissertation, PMA, Katholieke Universiteit Leuven, (1997)
23. N. Vlassis, B. Terwijn and B. Kröse, "Auxiliary particle filter robot localization from high-dimensional sensor observations", In Proc. IEEE Int. Conf. on Robotics and Automation, Washington D.C., (2002) pp 7–12

Indexing and Profiling in a Consumer Movies Catalog Browsing Tool

Louis Chevallier, Robert Forthofer, Nour-Eddine Tazine, and
Jean-Ronan Vigouroux

THOMSON
Corporate Research, CSA Lab
1, Avenue de Belle Fontaine
35510 Cesson-Sévigné, France
{louis.chevallier,nour-eddine.tazine,
jean-ronan.vigouroux}@thomson.net

Abstract. This paper discusses the benefits of both indexing and classification techniques combined with natural user interfaces for building consumer browsing tools. It focuses on text indexing and classification techniques used for user profiling.

1 Introduction

This paper presents a consumer application that is concerned with the subject of movies catalogs browsing developed in the framework of the ITEA Ambience project (http://www.extra.research.philips.com/euprojects/ambience). It is based on automatic data indexing and classification techniques and features a so-called natural user interface. We discuss how this contributes to meet user requirements typical to consumer domain.

2 The Movie Catalog Browser

2.1 Concept

The Movie Catalog Browser is a prototype consumer application dedicated to a Video On Demand service (VOD). Here, we put the focus only on the end-user application part of the system that makes it possible for the user to select a particular movie in a catalog. Since the browsing application is dedicated to the consumer domain, the selected functions supported by the user interface are deliberately "easy" in the sense that they should be easy to understand and use. They consists in:

- Standard navigation by genre.
- A step by step navigation in the catalog based on similarity between movies.
- A recommendation section also provides a selection of movies.

A screen shot of the demonstrator is shown on Fig. 1.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 275–287, 2003.

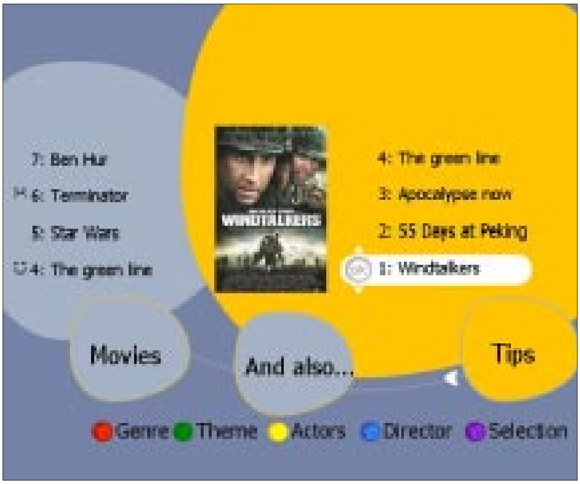


Fig. 1. Snapshot of the demonstrator screen

At any time, a movie is selected (the "current" movie). The user can apply various actions on it, including getting detailed information (summary, actors lists) or, in a real product, playing or buying it. This current movie belongs to the main movies list which is always visible on the left part of the screen. Two other lists of movies can be displayed by the system:

- A list of movies similar to the "current" one.
- A list of recommended movies.

The user can navigate from the main list to the list of similar movies. Selecting one item in this list updates the current list. This let the user move step by step in the catalog (he can go back along the path he has followed).

At any time, the user can rate any movie (like/dislike) present in the main list. The information that is passed along to the profiling engine results in updated user model.

2.2 The IMDB Database

For the constitution of the catalog, we relie on the IMDB movie database. This list comes along with a set of metadata that are used for indexing.

The IMDB (http://www.imdb.com) is a public domain data set describing movies and the people involved in them.

It is based on ASCII files consisting of simple records as shown in Fig. 2. Access to the IMDB is provided via FTP, SMTP and WWW servers mirroring the IMDB data. The IMDB site supports a number of query templates and generates HTML pages on the fly. These are linked by cross-references to films, persons, locations etc.

IMDB contains about 300 000 movies descriptions.

```
  Allen,  Weldon  Dolores  Claiborne  (1994)  [Bartender]  <13>
Allen, William Lawrence Dangerous Touch (1994) [Slim] <3>

  Sioux City (1994) [Dan Larkin] <11>

  Allen, Woody Annie Hall (1977) (AAN) (C:GGN) [Alvy Singer]
<1>

  Bananas  (1971)  [Fielding  Mellish]  <1>  ...  Zelig  (1983)
(C:GGN) [Leonard Zelig] <1>

  MV: Titanic (1943)

  PL: Building the Titanic has been a huge financial effort,
and White Star Line

  PL:  president  Ismay  wants  her  maiden  voyage  to  hit  the
headlines. He urges

  PL: Captain Smith to make the fastest possible crossing to
New York. When

  PL: iceberg warnings come in, the captain must ask himself
if he is willing to

  PL: risk the safety of his ship just to please Ismay.
```

Fig. 2. IMDB data files excerpt (actors, summaries)

However, in order to support the specific indexing and retrieval performed by our application, the IMDB source files must be loaded into a database and query results must be displayed and formatted in our own application.

The available information is the following:

Table 1. Available metadata

Title	
Genre	Romance, Drama, Action, Fantasy, Comedy, Animation, Family, Sci-Fi, Short, Musical, Adult, Documentary, Western, War, Adventure, Horror, Film-Noir, Mystery, Crime, Thriller. A movie can be categorized in several genres.
Year	
Country	
Language	
Director	
Actor:	a list of some actors of the movie
Studio	
Budget	
Summary.	The summaries consist in natural texts that are provided by contributors that volunteered to provide a short report on the movie.

2.3 User Interface

In order to support the very simple operating of the application, the user interface is built on top of multimodal components although the principal functions can still be exercised through regular remote command. Voice recognition (TELISMA) is used to let user asking directly for entities like movies titles or actors names. An action that would otherwise require a keyboard. An automatic user identification is performed by means of face analysis, this saves the user to declare himself to the system. This recognition is done in real-time through analysis of a video stream (VITEC).The user interface is completed by a speaking human agent that conveys feedback to the user (EPICTOID) through spoken messages and gestures.

2.4 Architecture

The architecture is depicted in Fig. 3. The system is split into two main parts : a metadata server and the browser client part. The metadata server handles both metadata searching, retrieval, indexing and profiling. It supports many simultaneous client connections.

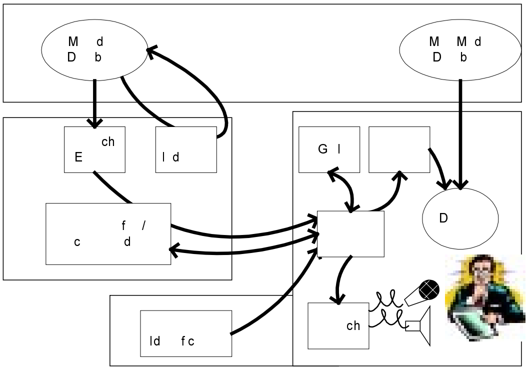


Fig. 3. Architecture

2.5 Indexing

In the current prototype, indexing is applied to the metadata provided by the IMDB database. The following metadata are directly used by the application for user profiling : Genre, Actors, Director, Year, Country. A specific task is devoted to the indexing of summaries. The indexing tries to extract features from these texts for enabling the semantic distance assessment between movies and for contributing to the user profile model.

2.5.1 Extracting Features from Summaries

Summaries consist in relatively short texts written in natural language. The average number of words per summary is 45. Most of them are written in English.

Due to the small length of the texts, pure word based techniques would result in rather poor results regarding recall, since there are many words that can be used for designating a particular event or character and the shorter the texts the fewer of them will appear in the text [1]. It is true, however, that given the application browsing style, the requirement on recall is not as strong as it may be in a tool dedicated to retrieval.

On the other hand, precision is especially important in the consumer domain since users tend to be quickly disappointed by any failure of the system at providing truly relevant document.

The considerations exposed above lead to abandon the idea of using a Vector space model based on raw words. Instead, we have tried to identify more specific features. For this purpose, we have chosen to adopt linguistically based techniques: in a first step, we use a electronic dictionary for filtering terms (words and phrases) extracted from the text: for being retained, each term has to be present in a lists of nouns. The dictionary we use is Wordnet [2]. For improving recall, this term list is then expanded by related terms found in Wordnet. The relation we are using is the synonyms.

Another approach has been taken by MEMODATA another partner involved in the project. They are relying on proprietary modules for extracting entities like place, date and names. In an improved version, terms that are thematically connected could also be added thanks to a dictionary.

This processing results in a set of 20 000 filtered terms and phrases (so called concepts) associated to movies.

2.6 Similarity Distance

The similarity distance has to assess the semantic proximity of two movies. In our case, it is based on the similarity between the summaries texts.Our notion of similarity is quite different from other situations where the task is to detect whether two texts contains common information. This is typically used for topic segmentation, summarization.

There are various definitions of similarity in the literature ranging from information theory to linguistics interpretation. We define informally the similarity of two texts being as high as the number of shared concepts is high and that their mutual closeness is strong. There has been many attempts to develop such a distance in the text processing literature. One can mention linguistically oriented approaches based on linguistic knowledge bases (ontology). These approaches take advantage of a graph data structure that make possible to compute a proximity distance on a term by term basis. These approaches differs by the particular linguistic relations selection and the way the final score is computed [14],[15]. Such techniques are typically used for word sense disambiguation.

Statistics is the other main tool that can be used. In case of non supervised algorithms, it has the advantage of not requiring the (manual) development of linguistics resources or any ground truth. These techniques have proved to be quite effective for document retrieval, clustering or classification. They are mostly based on word sharing or collocations which works as long as indexed documents are long enough for overlapping to be sufficient. A condition that no longer hold with our short summaries. The chose approach is a combination of those two: the linguistic processing is applied at the feature extraction phase while the distance is computed through statistical estimation.

In a first version, the estimation is a computed with the following formula:

$$\mathrm{d}(m_i, m_j) = \sum_k \left(\frac{C_{ik} C_{jk}}{C_k}\right)$$

where C_{ik} is 0 or 1 whether the movie i has the concept k and C_k is the number of movies having the concept k. This amounts to the IDF part of the classical TF*IDF rating applied on concepts.

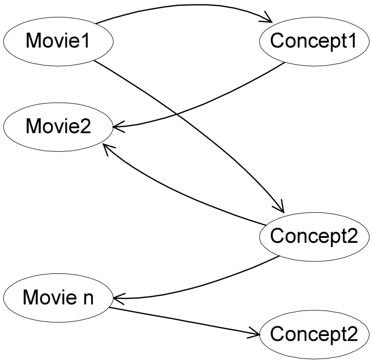


Fig. 4. Movie/Concept graph

In the application, this rate is efficiently computed based on a linked data structure. The ten most similar movies are displayed.

2.7 Profiling

The recommender system needs the definition of a user profile, which will be used for the selection of movies. The user profile contains the data corresponding to the user. In our case, a user profile represents a list of movies, as rated by the user. This enables the classifier to determine the user's tastes in order to generalize and make a selection of movies.

Some research has already been carried out on recommender systems [3],[6]. The principle of analyzing the tastes of somebody in order to predict the movies he wants to watch, or the music CDs he wants to buy, is clearly a tricky issue. Perhaps simply because the user himself cannot always explain why he liked a movie or a piece of music.

Two different ways of tackling the problem exist:

Collaborative method and content-based recommendation. These approaches are basically different. Collaborative methods try to compare the users’ tastes to find similarities, whereas content-based methods are based on similarities between movies or pieces of music.

The collaborative method searches for similarities between different users to recommend a movie. This method is often used. See for example [4] or [5]. A movie is determined by a list of users, who have rated this movie.

This method gives generally good results, but has some disadvantages :

- The system needs to have lot of users’ profiles at its disposal before displaying significant similarities.
- A new user cannot be helped. Nevertheless, these problems occur for content-based recommendations too.
- A new movie can never be assessed and therefore recommended.
- Privacy risks have been a real issue for a couple of years. When using similarities between users, this problem must be taken into account (see [4] and [6] for further information).

This method is most often used, especially, by some sales-websites, which recommend a lot of music CDs or books. For example, “Amazon.com” is known for applying collaborative method.

Content-based recommendation - This method is only based on similarities between movies [5]. Movies are here no more determined by the users who liked it, but by associated metadata. As in the previous analysis, the system cannot help a newcomer user until he gives a list of movies he likes (or does not like). But a new movie can be proposed as early as it is taken into account. Using content-based recommendations, requires the use of a classifier, which will predict if a movie will be appreciated or not.

2.7.1 Applying SVM on Movies

The classification will be computed based on the following features derived from the data coming from IMDB:

Genre
Year
Actors
Origin Country
Concepts

For all the features but the year, these features happen to be boolean values, resulting in a very high dimensional vector space.

When represented in a movie/features matrix, the part corresponding to actors and concepts results in a very sparse matrix.

2.7.2 Movies Selection

Each time the user wants to be helped, 10 movies are pre-selected by the classifier. A few conditions are supposed to be verified by the selected movies:

1. They must, naturally, correspond as well as possible to the user's tastes.
2. They have to be as varied as possible. The user does not want to be proposed 10 similar movies, with the same genres and the same actors.
3. The list must not be limited to only a certain type of movies. The system has to introduce some novelty, in the sense of movies that the user does not usually watch.
4. A movie must not be selected several times.

Nevertheless, these objectives seem to be rather contradictory. By varying the movies, or by introducing new types of films, we are not sure to select movies which correspond to the user's tastes. This will set problems for the evaluation of the program.

These conditions lead to separated processing. Not only each movie can be classified into one group ('likes' or 'does not like'), but the SVM can return a ranked list of the movies (among two movies, we know the one that is the most likely to be appreciated by the user). This will be required for the first condition. Some treatments have been tested to verify the others conditions (see the different following algorithms). Moreover, a chronological list of closer movies is kept in memory, in order to avoid proposing several times the same movie. Several selections, based on SVM classification, have been tested.

2.7.3 Movies Selection 1: Selection of the 10 Best Movies

We have tested a first selection algorithm. The first method consists in selecting the 10 best-placed movies (in the following they will often be called 'the 10 best movies'). In this way the first condition is clearly verified. But the second one clearly not, because two similar movies are geometrically close. Therefore, they are either both selected, or neither. If a user, fond of horror movies, decides to watch one of them, he will then be proposed ten other horror movies with some actors in common. He may feel overcome.

2.7.4 Movies Selection 2: Introduction of Clusters

The need of selecting different types of movies can lead to introduce an unsupervised classification method: the k-means. This algorithm breaks down all the movies into similar groups. In other words, each group contains similar movies, and movies from different groups are as different as possible. We can then select the 'best' movies (in the classifier sense) of each group.

As the k-means groups are intended to contain movies as different as possible, this method results necessarily in a widely 'different' movie selection. But these movies are not, on the other hand, the best selection, as defined above.

The selection is finally made with a 4-cluster k-means, so with four groups of movies. From each group the two best movies are selected. To these eight movies are added two random movies. Thanks to these two last movies, the user is not completely locked up in the genre of movies he likes (or he knows), but can be offered a new film, which could be very interesting as well. This was the third condition.

Likewise, the time computation must be taken into account. The user does not want to wait for several minutes before getting his pre-selection. SVM is quite fast, because the sample size remains very small. On the other hand, the k-means can often be longer. So we decided first to limit the k-means to only one iteration (the goal was not to precisely separate the movies, but only to make different varied groups), and second to apply this classification to a number of pre-selected movies.

Finally, the application chooses 100 movies (which are movies with quite a reasonable probability), and then apply a 4-clusters-k-means on these movies. Two movies are selected from each class. We obtain thus a selection of 10 movies including the 2 random movies.

2.7.5 Evaluation of the Performances

We have at our disposal a list of user profiles (that is to say, for each user, a list of movies with their appreciations) downloaded from MovieLens.com. This base contains 100 000 evaluations, made by 943 users on 1682 movies. Each user has rated at least 20 movies. The method applied by George Karypis [7] to compare different recommender systems is the following : for each user, all but one movie (an appreciated one) served for the learning. The evaluation is based on the presence of the tested movie in the selected list.

Different problems occur, and show the difficulty in evaluating a recommender system:

The absence of the tested movie does not mean that the classifier badly classified this movie. The selected movies could have been preferred to the tested movie by the user. This remains unverifiable.

Only appreciated movies can be tested. Nothing can be deduced by the fact that a not appreciated movie is not in the 10-best movies.

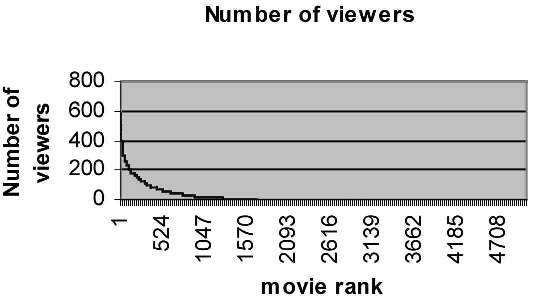
The application is intended to select various movies. This need of variety, in fact, can be prejudicial for such a result.

Moreover, content-based recommendation drastically differs from collaborative methods, and are not easily compared. The method applied by Karypis has given relatively good results on Movielens (about ¼ of the tested movies were selected).

This method, as explained above, consists in making a selection of N movies (N is 10 here), among a list of several hundreds movies. In this list there is one movie (the tested movie), which was appreciated by the user. But how can a system guess which movie the user has seen? The system is only intended to select movies, which are predicted to be liked by the user.

For that problem, collaborative recommendation clearly gives better results. Indeed, movies screening are not uniformly distributed to the movies. Some movies are seen by a huge part of the population, while others are ‘rare’. Therefore, the collaborative recommendation selects often watched movies, what may be the case of the tested movie. That can explain Karypis’ good results. This is not taken into account by the content-based recommendation, which very often selects movies that the user did not watch before.

The following diagram shows the number of viewers for each movie (the movies are ranked by this value):



This shows clearly that movies screening are not uniformly distributed.

This graph shows the curve $\log(number\ of\ viewers) = f(rank)$. We can note that

$$\log(number\ of\ viewer) = a * rank + b$$

On the other hand, a bad result can be explained by a not sufficient number of evaluating movies, and not by a bad classification.

To do so, perhaps an interesting method is to count the number of appreciated movies in the selected movies after each iteration of the application. This can be made by user tests. Tests can be made with user profiles (as behind), but, in this case, only evaluated movies must be taken into consideration. Indeed, what to do with a non-evaluated movie?

We applied the following algorithm to verify if the selected movies converge or not to the movies that the users like : we consider, from the first fifty users of movieLens, the users who rated at least 200 movies. We consider movies used by MovieLens. We found seven such users (users 1, 6, 7, 13, 18, 43 and 49). We consider only these users, because a minimum of ratings is useful to evaluate the convergence of the selected movies. For the evaluation, we consider a training sample and a test sample. The training sample contains movies used in the learning process. The test sample contains the other rated movies. The other movies are not rated. At the beginning, the training sample contains no movie. They are all in the test sample.

Now, we apply the following algorithm to each user i. While the test sample contains movies

1) Select randomly 5 movies rated by user i from the test sample.
2) Add these movies to the training sample. A learning process is realized with the training sample.
3) Make a selection of ten movies from all movies, except those in the training sample.
4) Each of these movies is placed in one of the following categories :

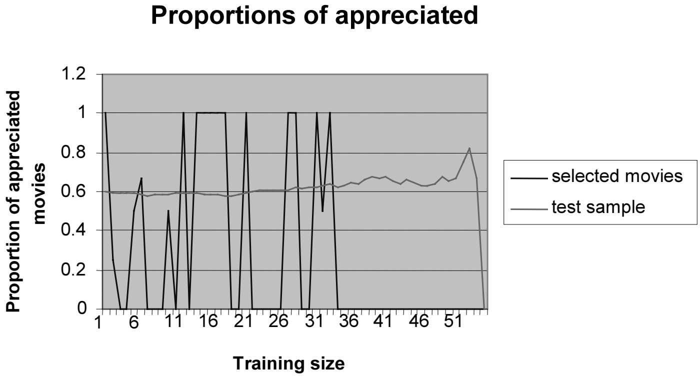
- User i rated and appreciated the movie
- User i rated but did not appreciate the movie
- User i did not rate the movie

5) Compute the following values :

- Number of appreciated movies selected in step 3
- Number of not appreciated movies selected in step 3
- Number of not rated movies selected in step 3
- Number of appreciated movies in the test sample
- Number of non-appreciated movies in the test sample

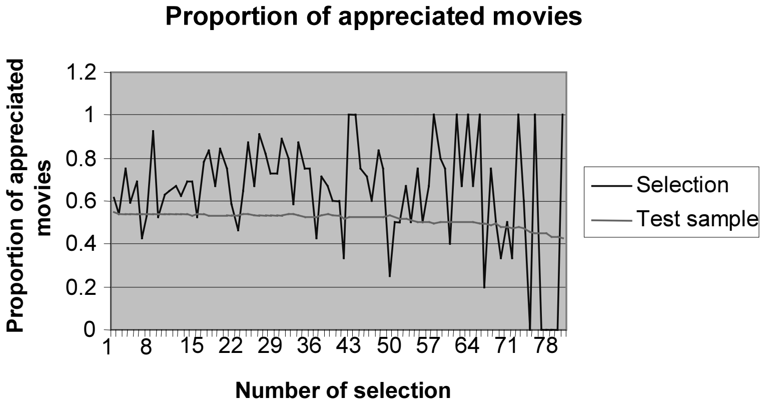
With these values, it is possible to compute the proportion of ‘good’ movies in the list. But this unique value means nothing. It has to be compared to the same percent in the test sample.

We obtain a list of values. We can plot the two values to compare them. But a new problem appears: some movie lists contain no rated movie, or just one. The proportion, in these cases, or does not exist, either is 0% or 100% (this depends of the rating of the unique rated movie). Moreover, when the training size increases, the test size decreases. After some iterations, all selected movies are therefore not rated. These curves are therefore unreadable, as shows the following one (corresponding to user 1):



After one iteration, the classifier has learned, for each user, with 5 rated movies. Each user were proposed ten movies (this represents seventy movies in all). These movies are liked, not liked or not rated by the corresponding user. The first line means that 19 of the seventy movies are liked by the corresponding user, twelve are not liked, and thirty-nine are not rated. 1211 is the total number of appreciated movies, which were on the test sample. 1017 were not appreciated, on the test sample.

This leads to the same curve than behind, but with best values (the curve does not oscillate, going up from 0% to 100%). We obtain the following curve:
The curve begins to oscillate when only one user has still rated movies. We can note that the proportion of appreciated movies is higher in the selected list than in the test sample (the rated movies which were not considered for the learning process). This makes sense to our recommender system.



3 Conclusion

We have presented the indexing and profiling modules that are embedded in a consumer application dedicated to the browsing of movies catalog. Further works includes improving features extraction and collecting user feedback for assessing the relevance of the navigation and recommendation.

Acknowledgments. We are grateful to Ambience partners involved in the video browser demonstrator for their involvement in both the design and realization of the prototype.

References

1. Manning C.D., Schütze H, Foundations of statistical Natural Language Processing The MIT Press 1999
2. Felbaum C., Wordnet, an electronic lexical database, The MIT Press 1998
3. Haym Hirsh, Chumki Basu, and Brian D. Davison : Learning to personalize, Recognizing patterns of behavior helps systems predict your next move. In Communications of the ACM, Volume 43, Issue 8, August 2000, pages 102–106
4. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl : Item-based Collaborative Filtering Recommendation Algorithms. GroupLens Research Group/Army HPC Research Center, Departement of Computer Science and Engineering, University of Minnesota, Appears in WWW10, May 1–5, 2001, Hong-Kong
5. C. Basu, H. Hirsh, and W.W. Cohen: Recommendation as Classification : Using Social and Content-Based Information in Recommendation. In AAAI'98, Proceedings of the Fifteeth National Conference on Artifical Intelligence. AAAI Press, 1998

6. Resnick P., Iacovou N., Suchak M., Bergstrom P., and Riedl J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of CSCW'94, Chapel Hill, NC
7. George Karypis, University of Minnesota, Department of Computer Science/Army HPC research Center, Minneapolis, MN 55455, Technical Report #00-046: Evaluation of Item-Based Top-N Recommendation Algorithms
8. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995
9. V. Vapnik. Statistical Learning Theory, Wiley, 1998
10. M. F. Porter, An algorithm for suffix stripping. In Program, 14(3):130–137
11. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001
12. R. Forthofer, J-R Vigouroux, L. Chevallier, Y. Le Mener, Thomson Multimedia Corporate Research, Presentation of a VOD content-based recommender system, Ambience Workshop, Torino, 09–2002
13. F. Sebastiani, Consiglio Nazionale delle Ricerche, Italy, Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No 1, March 2002, pp 1–47
14. Alexander Budanitsky, Graeme Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures in the North American Chapter of the Association for Computational Linguistics (NAACL-2000, Pittsburgh, PA, June 2001.", (2001)
15. Philip Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy in IJCAI 1995
16. Louis Chevallier, Jean-Ronan Vigouroux, Robert Forthofer, Ted Diamond, Eric Rehm Automatic categorization of short textual metadata of multimedia database in replacement of a rule-based classification system Thomson, Singingfish, ACM SIGIR 2002, Workshop on Operational Text Classification, Tampere, August 2002

An Integrated Framework for Supporting Photo Retrieval Activities in Home Environments

Dario Teixeira[1,2], Wim Verhaegh[2], and Miguel Ferreira[2]

[1] Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands
[2] Philips Research Laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands
teixeira@natlab.research.philips.com
wim.verhaegh@philips.com

Abstract. This paper addresses the content overload problem applied to photo retrieval activities in the home environment. The starting point is an analysis of the main activities that users perform with their photographs. We then discuss two different interaction paradigms, namely browsing assistants and conversational search, which can provide software assistance to help users cope with increasing numbers of photographs. The major contribution is the presentation of an integrated architecture which combines these interaction paradigms to provide the user with a home-friendly, multimodal, and seamless system for photo viewing. We also discuss the user interaction aspects, internal architecture, and algorithms of each of the two paradigms.

1 Introduction

One of the most enduring visions of ambient intelligence portraits the home as a window to the world, enabling users to access all different kinds of information in the comfort of their living rooms [1]. However, the so-called content overload problem presents a real challenge to this vision: given the overwhelming increase in the amount of available information, people must increasingly rely on software assistants to sift through all the choices they are given [2],[3]. Moreover, while significant research has been directed towards solving the problem in domains where computer-like interfaces are the norm, less attention has been paid to content overload in the context of in-home environments. As a consequence, many of the available techniques rely on interfaces which not only contradict the principles of ambient intelligence, but might even alienate some of the less technology-savvy individuals.

Most seriously, the content overload problem is not limited anymore to publicly available content such as films, music, or books. Our statement is that the introduction of new technologies, in particular the generalised use of digital cameras, has the potential to bring about the content overload issues into the domestic front as well. The main reason lies in the convenience and low-cost of digital photography, coupled with ever-growing storage capacities. As a result, we are already witnessing a significant increase in the number of photos that people take, and the consequent difficulty in managing them [4].

In this paper we describe a system designed to assist photo retrieval activities in the context of an in-home environment. The starting point is the characterisation of the various retrieval activities that people engage with their photographs, namely searching,

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 288–303, 2003.

wandering, and recommending. We then describe two different paradigms for accessing information—conversational search and browsing assistants—which can handle those activities. Most importantly, we present the architecture of a system which integrates them in a seamless manner. After a thorough analysis of the particularities of the photo domain, we proceed to describe in detail the user interaction aspects and algorithms used by the browsing assistant and the conversational search engine. Finally, we present some early experimental results, and discuss the following research steps.

2 User Interaction Support

In this section we aim to provide a rationale for the work described in the remainder of the paper. As a first step, we will analyse the kind of activities related to photo retrieval which are performed by users in home environments. We shall then describe the interaction means which can be mapped onto those activities in a natural way. At last, we present an illustrative use-case scenario.

2.1 User Activities with Photographs

For single-user situations we have identified three broad categories where all retrieval activities can be placed: *searching*, *wandering*, and *recommending*. Even though multi-user situations would allow for other possible activities, such as *story-telling*, in the general case these can also be mapped into the three broad categories presented (story-telling, for example, is a form of wandering through the photo collection).

The searching activity refers to the situation where the user is looking for one specific photo or a set of photos. In this case, the user has a concrete goal in mind, and the purpose of the activity is the satisfaction of that goal. The wandering activity, on the contrary, reflects no specific goal on the part of the user. It corresponds to the casual, aimless roaming through the photo collection, without any higher-level purpose other than to enjoy viewing the photos and reliving those memories. At last, we consider asking for recommendations as another possible activity for users to initiate. It differs from the previous two in the sense that rather than being an independent activity, it is more properly classified as complementary to either searching or wandering. In the former case, recommending means that the system is free to suggest some of the search parameters. In the latter, the system would present the user with suggested paths for wandering.

2.2 Interaction Means

The interaction means can be seen as tools or paradigms which support the user to perform a given activity. Our choice of interaction means was guided not only by their efficacy in aiding their designated activity, but most importantly, how well they would fit within the vision of ambient intelligence, especially in what concerns the respect for the specificities of in-home environments. In practical terms, this meant considering means which would not disrupt the normal social interactions present in the home, that could make use of more natural modalities such as speech and objects, and finally, that could provide the users with an ‘experience’ and a feeling of enjoyment when using the system.

The first interaction means we shall consider, the *browser*, is similar in principle to a normal web browser. The latter displays one web page at any time; the page has connections to other pages, which in turn connect to many others, and so on. Instead of web pages, our browser deals only with photographs, and the connections to other photos are calculated by the system and presented to the user as thumbnails. This is an interaction means where the user is in control, using the photo connections as association aids to traverse the photo collection.

As depicted by Fig. 1, the browser could be used for both searching and wandering activities, even though using it to search for a photograph through a potentially huge photo collection is likely to be inefficient. However, it does lend itself naturally to the wandering activity, and this particular combination (coupled with presenting recommendations to the user) fits well into the general description of a *browsing assistant* [5]. This sort of system is designed to run continuously in the background, looking for items related to the user’s current selection, and suggesting them in a non-intrusive manner. Since it may keep track of the user’s preferences and browsing habits, one expects it to be able to produce very relevant suggestions. We will elaborate further on the browsing assistant in Sect. 4.

A *conversational interface*, on the other hand, is an interaction means which can only be mapped naturally onto a searching activity (again, with or without recommendation). Conversational interfaces emerged from attempts at improving classic query-based search and recommendation processes. Traditionally, query-based systems were synonymous with the so called *ranked list* approach, whose prime example can be found in the familiar web search engines. A conversational interface aims to provide better results by mimicking human dialogue interaction as the means to iteratively fill in the variables for the query [6],[7],[8]. This solution has several advantages over the ranked list approach: first of all, the user is not overwhelmed with a list of possible alternatives; instead, she can have the system play the role of an all-knowing mentor, guiding her through a series of well-thought questions, and eliminating all irrelevant items step-by-step. Moreover, it is our expectation that having a dialogue with the system will prove to be a very natural way of interaction, thus satisfying our primary criteria. To this process, where a conversational interface is used to progressively narrow down the search space until just a handful of matching items remain, we have given the name *conversational search*.

At this stage, one could point out that even though both these interaction means serve their purpose competently, we would only be replacing a computer-like interface with a two-headed system that would force the user into different ‘modes’ for different tasks. The answer to this problem is based on the insight that it is possible to present the user with a unified system where the transition between the different activities is as seamless as possible. The key idea is to have the browsing assistant navigate solely through the subset of photographs which match the current criteria as defined by the conversational search process.

Figure 2 presents an architecture which meets the broad requirements we have just outlined. One can see the division between the conversational search engine and the

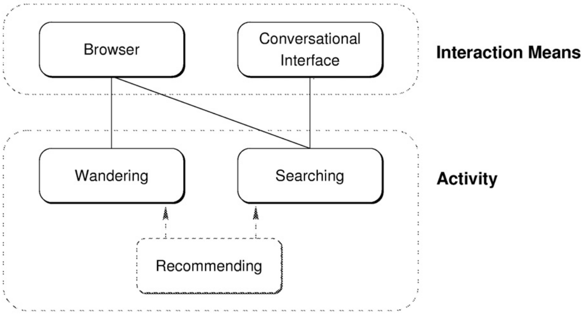


Fig. 1. User activities with photographs, and the proposed interaction means to provide computer assistance.

browsing assistant, and most importantly, the points where the two paradigms interconnect. Not only do they share the same user profile and the same domain database (the photographs and their meta-data), but the conversational search engine also signals the browsing assistant anytime the search space changes.

2.3 Use-Case Scenario

The following use-case scenario elucidates how the described system would be used in a real-world situation.

It's a lazy Sunday afternoon; Claire has picked up her portable photo browser, and is now sitting in the couch wandering through her photo collection. Over the years she has accumulated over ten thousand photographs—way more than what she would be able to handle without the aid of intelligent software assistance.

Claire ponders over a photo of the Sagrada Familia that she took in Barcelona many years ago. The browsing assistant allows her to see all related photographs across multiple dimensions: by choosing the contents *attribute, the system displays the thumbnails of all the other photographs which also contain the Sagrada Familia, closely followed by photographs which contain other cathedrals; choosing the* location *attribute presents her with all photos which were also taken in Barcelona, and likewise for the other attributes.*

At this point Claire remembers that she took a brilliant photograph of her pet dog Bello playing on a beach somewhere in Spain. She asks the system for some help in finding it: 'James, could you help me find a photograph? It was a photo of Bello on a beach'. The system—known as James by Claire—promptly reduces the search space to those photos containing both Bello and a beach. Claire can see this change immediately, because the browsing assistant only displays the photos which meet the specified criteria.

The search is not yet complete, however, as there are still roughly fifty photographs left. Thankfully, the system knows that Claire is usually quite good at remembering also the location of her photos. 'Claire, do you remember where it was taken?'. To which Claire provides a positive reply. 'The photo was taken in Spain'. At this point, the search space is reduced to a couple of dozen photographs, and the system determines that the

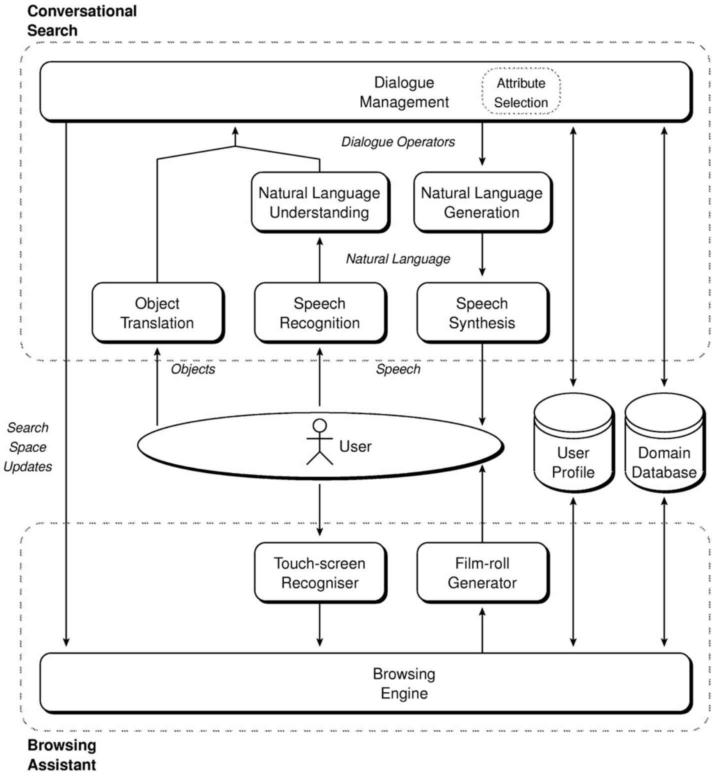


Fig. 2. The architecture of the system. One can see the major components of both the conversational search system and the browsing assistant. Also noteworthy are the contact points between the paradigms, and the central role played by the user.

date *attribute provides the best chances of reducing it further. 'Claire, do you remember when it was taken?', asks the system. Claire does not remember, but she is already satisfied with the current set of photos: 'Thanks James, and please stop!'*

3 Characterisation of the Photo Domain

Before we proceed to the interaction paradigms and the associated algorithms, it is important to understand the structure of the data we will be handling. In this section we will describe the details surrounding the meta-data about the photos.

There are several problems concerning the definition of a structure for the meta-data of domestic photographs. The most pressing issue is user-related: each person has a different perception of what matters about a photograph and how it should be classified. As an example, let us consider what attributes are needed to describe what is visible on a photograph. Most users would want an attribute like *people*, since that is one of the focus points of anyone's life. However, if we allow *people*, why not also consider *pets*, or *monuments*, or whatever is important for different users? Even though this problem seems insurmountable without providing each user with their own personal meta-data categorisation, a middle-ground solution can be attained. We have classified the meta-data about each photo into five generic attributes; the internal structure of each attribute can then be tailored to suit the user's wishes.

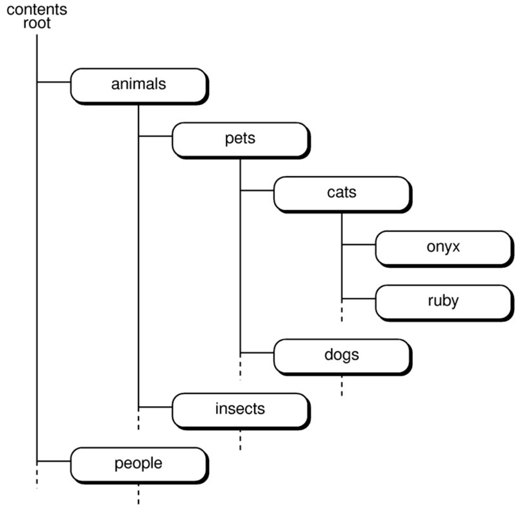


Fig. 3. A subset of the ontology tree for the contents attribute.

Table 1. The attributes constituting each photo's meta-data.

Attribute	Cardinality	Orderability	Structure	Similarity	Example
Contents	multiple	unorderable	hierarchical	partial match	/plants/flowers/tulips
Location	single	unorderable	hierarchical	partial match	/Europe/Spain/Barcelona
Topics	multiple	unorderable	hierarchical	partial match	/holidays/camping trips
Date	single	orderable	flat	temporal distance	21^{st} of June, 1999
Album	single	unorderable	flat	total match	"Summer 1999 in Spain"

The first attribute we considered, *contents*, stores all the information which is visible in the photograph, including categories such as *people* or *pets* which we have discussed previously. In order not to lose the conceptual division between the different categories, we have opted for the construction of an ontology tree capable of grasping the inter-connections between the various elements found in the real world (a subset of that tree is shown in Fig. 3). Likewise, the *location* attribute also requires an ontology tree, in this case to store the physical location where the photograph was shot. The tree we used follows a division of the planet into continents, countries, regions, and cities. More refined branches are possible, of course, since in many cases users will be aware of those subdivisions. The rationale is to provide always enough structure that allows for any possible query from the user, e.g., 'show me photos taken in Africa', 'show me photos taken in Spain', 'I want to see photos from Montparnasse', etc.

The *topics* attribute represents the high-level category where the photo belongs, such as 'holidays' or 'parties'. It also requires an ontology tree because users might consider subcategories such as 'birthday parties' or 'camping holidays'. Simpler is the meta-data stored by the *date* attribute: it just contains the time stamp of when the photo was taken. Finally, the *album* attribute contains the single event where the photo belongs. In the days of 35mm film, it would correspond to one film roll or set of film rolls taken sequentially.

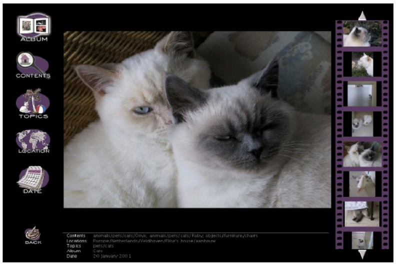
Table 1 provides more detailed information about the properties of the five attributes. The *cardinality* property indicates whether the attribute will have single or multiple instances for a photograph; *orderability* is the property that allows us to define comparison operators between instances; the *structure* will be hierarchical for those attributes which allow the definition of an ontology tree, or flat for those that do not; and at last the *similarity* column merely indicates which distance measure is used to compute the similarity between photos.

As we have explained in Sect. 2.1, photographs are a form of domestic content. This fact has one important consequence as far as the meta-data is concerned: contrary to what happens for publicly available content, there will be no professionally produced meta-data. Either the users have to manually annotate all their photographs, or we must rely on automatic techniques to produce it. The former possibility is not only unrealistic (especially considering the huge increase in the amount of available photos), but would also represent a contradiction as far as any vision of ambient intelligence is concerned. Fortunately, the latter possibility is becoming more and more feasible. For some attributes, such as the date and the location, modern digital cameras with built-in clocks and GPS receivers can already provide the meta-data. Regarding the contents, there is abundant research on automatic feature extraction techniques, showing promising results [9],[10].

4 The Browsing Assistant

The browsing assistant was implemented according to the general principles outlined in Sect. 2.2. In consonance with the requirement that the system should fit in a living room environment, the software was designed to run on a portable webpad, making use of the touch-sensitive screen for input. As one can see from the screenshot in Fig. 4, the main elements of the user interface are as follows: in the centre, a large area where the user can display a photograph for viewing; to the left, a panel with all the meta-data attributes, allowing the user to choose the preferred dimension for wandering (see Fig. 5); to the right, the film-roll with thumbnails of photographs related to the current one according to the selected dimension; at last, in the bottom, a text area displaying the meta-data of the current selection.

Whenever a new photograph is selected by being dragged from the film-roll onto the centre, the film-roll is updated with the thumbnails of the most similar photos according the currently selected dimension. Since only a limited number of thumbnails can be displayed, the user can scroll the film-roll in order to view the thumbnails of the second-best suggestions, the third-best, etc. Changing the current dimension is done by clicking on any of dimension icons. The thumbnail suggestions will immediately be changed in tune with the switch of dimension.



Contents animals/pets/cats/Onyx; animals/pets/cats/Ruby; objects/furniture/chairs
Locations Europe/Netherlands/Veldhoven/Elise's house/aanbouw
Topics pets/cats
Album Cats
Date 20 January 2001

Fig. 4. The user interface of the browsing assistant.

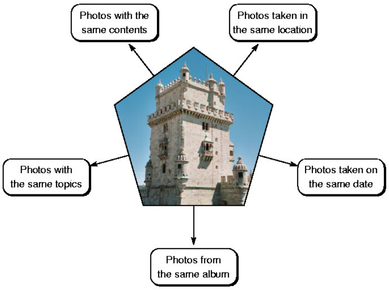

Photos with the same contents
Photos taken in the same location
Photos with the same topics
Photos taken on the same date
Photos from the same album

Fig. 5. The browsing assistant provides multiple dimensions for wandering. These correspond to the available meta-data attributes.

4.1 Computing Photo Similarity

In order to present the user with suggestions of photos somehow related to the one currently being displayed (which shall forth be named the *pivot*), one must determine which photos are good candidates. Moreover, since the number of thumbnails which can be displayed simultaneously must be small—not only because of limited screen-estate but also to avoid overwhelming the user—it is necessary to rank the thumbnails by order of similarity and to display the top matches first.

The algorithms used to compute the similarity measure differ according to the attribute (these are listed in Table 1). In the case of the album attribute, the comparison is straightforward: a photograph either belongs to the same album as the pivot or it does not. Also for the date attribute the calculation is simple: the similarity is simply the temporal distance between the two photographs.

The three hierarchical attributes present a bigger challenge however. The reason is that even non-perfect matches are relevant if they share part of the tree structure. Let us consider for example the contents attribute, and take a look at the photos stylised in Fig. 6. If photo A is the pivot, which of the other photographs should be ranked higher? Photo B contains two strong partial matches, but photo C contains one perfect match. In other words, should the algorithm take a generalist or a specialist bias? We believe that the user should make the ultimate choice. With this in mind we have constructed an algorithm which can handle the partial matches of hierarchical attributes and allows for the fine-tuning of the generalist/specialist bias.

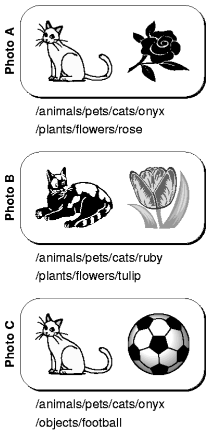


Fig. 6. The photographs used as examples for content similarity.

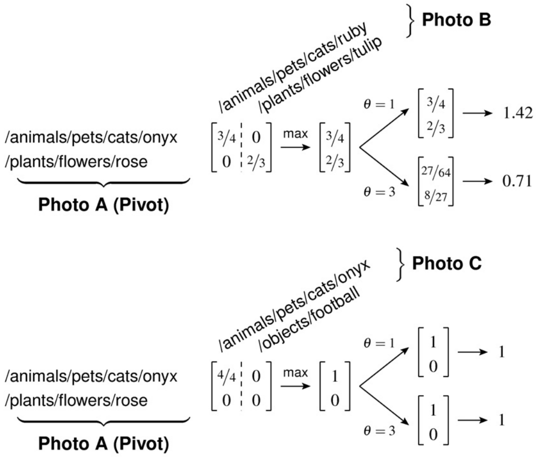


Fig. 7. The partial matches algorithm.

The *partial matches* algorithm works as follows. The first step is to construct a matrix where the row and the column headers are the meta-data of respectively the pivot and the photo to be compared. Each element in the matrix is computed as the fraction of the number of components from the test photo which match the pivot, divided by the length of the pivot tree. The next step is to build a vector where each element is the maximum value (the best match) of each of the matrix rows.

The following step is the key to the fine-tuning between the generalist and specialist preferences. Each element in the vector is raised to the power θ. This operation introduces a non-linear bias to the calculation, being the fractions least affected those close to the unity (which is not affected at all). In practical terms, the higher the value of θ, the higher the penalty imposed on partial matches, and the more specialist the bias. The last step of the algorithm consists of simply adding the elements of the vector: the obtained sum is the similarity measure.

Figure 7 shows the various steps of the algorithm, using photo A as the pivot and photos B and C as the test photos. Notice that in accordance to the above description, for $\theta = 1$, photo B is the most similar, while for $\theta = 3$, the more specialist photo C takes that role. Also noteworthy is the fact that the similarity measure is not necessarily commutative.

5 Conversational Search

We have defined conversational search as the process where a conversational interface is used to assist a search activity. As stated, the principle consists of having a dialogue between the system and the user as the means to construct a query iteratively and thus to progressively reduce the search space. Since people are so well-versed in conversation, one can deduce that it is critical for the system to engage in a dialogue which feels natural and consistent. Also, since the system plays the role of an all-knowing assistant, leading the conversation in the direction of a quick discovery of the user's target, it is likewise important that the system asks the right questions to the user. In this section we will discuss these and other issues essential for the good performance of a conversational search engine.

5.1 Dialogue Operators

Human language provides virtually infinite ways for people to express their intentions. As far as the conversational engine is concerned, all this diversity is unmanageable and largely redundant. For this reason, the engine does not process raw human dialogue (either in speech or text form), but rather the high-level dialogue operators associated with it. These are largely derived from research conducted on *speech acts* [11], and intend to capture the essence of each dialogue utterance. As depicted in Fig. 2, a number of low-level components—those handling speech recognition and synthesis, plus natural language understanding and generation—take care of the back and forth translation between speech and dialogue operators.

Table 2 lists some of the most important dialogue operators used by the system. These are classified into different categories according to the role they play in the dialogue. The core operators are the ones dealing with the *constraining*, *relaxation*, and *suggestion* operations. The meta-operators deal primarily with the control of the dialogue itself, and the clarification operators are mainly used to provide extra information to the user [7].

Table 2. A selection of the dialogue operators used by the system (*attr* represents the attribute, and *conf* the confidence level on the user response).

Operator	Category	Source	Parameters	Example
ASK_CONSTRAINT	constrain	system	attr	'Do you remember the location?'
ACCEPT_CONSTRAINT	constrain	user	attr, value, conf	'The photo was taken in Spain.'
REJECT_CONSTRAINT	constrain	user	attr, conf	'I don't remember.'
ASK_RELAXATION	relax	system	attr, value	'Perhaps it was not taken in Spain?'
ACCEPT_RELAXATION	relax	user	attr, value	'Yes, you are right.'
QUERY_VALUES	clarify	user	attr	'To which places have I been?'
START_DIALOGUE	meta	user	—	'Help me find a photo, please.'

5.2 Use of Objects

Using high-level dialogue operators to encode the user's intentions has another added benefit: the system becomes modality-agnostic, relying on any input/output translators we wish to implement. While speech was the first choice for the reasons we have already stated, it is possible to experiment with other modalities. In particular, previous experiences with object interaction within our project suggested that using physical objects as filters for photographs could also provide a natural and fun way for users to interact with the system [12],[13]. Moreover, research on graspable interfaces presents them as valid but still largely unexplored input modalities [14]. For all these reasons, we have also constructed an input translator for object interaction, as depicted in Fig. 2.

The hardware necessary for the translation is based on RFID (Radio Frequency Identification) technology, and was already available and ready-to-use. The basic idea consists of embedding small RFID tags in the objects we wish to identify, and to hide the detection coil beneath a table. Each tag has a unique identifier, allowing us to know precisely which objects the user places on top of the table.

The translator converts the low-level events of placing and removing objects from the table into the dialogue operators ACCEPT_CONSTRAINT and ACCEPT_RELAXATION, respectively. Each object also has an associated attribute/value pair from the photo database, enabling the user to express the equivalent of 'Show me photos taken in Barcelona' by simply placing her souvenir from Barcelona on the table. Since one can associate anything with an object, people and pets present themselves as obvious candidates. Furthermore, since the system recognises multiple objects, complex queries and filters can be constructed in a very intuitive and tangible manner: as an example, the user could add a puppet dog to the table to express her desire to see the photos of Barcelona that also contain her pet dog.

5.3 Dialogue Management

Dialogue management is the problem of handling the high-level dialogue intentions in such a way as to provide the users with the feeling that they are interacting with an intelligent entity. In practical terms, the dialogue manager must interpret the user dialogue intentions and select which dialogue operation to perform at each step.

In many telephony applications of conversational interfaces, the dialogue tends to follow a fairly rigid structure of question-answer sequences. Breaks to the sequence are allowed, but they are treated as exceptions that must be resolved before the sequence can proceed again. It is our view that this often called *ping-pong* dialogue structure does not fit well in the definition of an ambient intelligent conversational engine. Rather, one crucial aspect of dialogue management for in-home situations is what we call the *asynchronicity* of the dialogue. The system might still direct the communication and ask questions to the user, but these do not follow a rigid programme. The user is also free to ignore the system's questions, to provide answers out-of-sequence, or to provide information that was never asked. The rationale is that home environments tend to be fairly chaotic places, and the primary objective of the dialogue is to constrain the search space, not the user. Furthermore, the concept of using objects as an interaction modality would be defeated if the user were forced to use them only as a reply to a question from the system.

Concerning the implementation, the asynchronicity of the dialogue is achieved by giving the user enough time to provide multiple dialogue intentions in a single iteration, and by always reassessing the state before a new question is posed. Moreover, the dialogue manager does not rely on the user to immediately provide answers to questions asked. The heuristics that decide which operation to perform at any moment are straightforward and can be hard-coded into the system: while the search space is large, the decision is always made to *constrain* it; once we have narrowed it down to just a handful of elements, then we can *suggest* them to the user; should it happen that the search space becomes over-constrained—meaning that there are no photos which match all search criteria—then a decision is made to *relax* one of the constraints. It should be noted that even though plenty of research exists on the optimisation of dialogue strategies, most results suggest that the heuristics we use are optimal in these circumstances [15],[16].

5.4 Attribute Selection

Even after the dialogue management heuristics have made a decision regarding which operation is the most appropriate at any given moment, the system might have another problem to solve before it can ask a new question to the user. In particular, the operations which suggest to the user a constraint or a relaxation of the search space demand one extra parameter: the target attribute for constraining, or the constraint subject to relaxation, respectively. The problem of attribute selection lies precisely in the determination of this extra parameter. The decision must take a number of different factors into account, namely the characteristics of the search space, and the knowledge and preferences of the user.

Regarding the first factor, the analysis of the search space is based on the *maximum entropy method*, which in the case of a constraining operation chooses the attribute which maximises the potential reduction of the search space, or in the case of a relaxation operation, the constraint which increases it the least. This procedure is based on the standard formula from information theory for the calculation of entropy: $H(a) = \sum_{i=1}^{n(a)} -P_i(a) \times \log P_i(a)$, where $H(a)$ denotes the entropy for attribute a, and $P_i(a)$ is the probability of finding a value i belonging to attribute a, if we were to uniformly select a photo at random.

As far as the second factor is concerned, modelling the user is justified by the realisation that we can avoid asking the wrong questions (the ones which the user knows nothing about) if we have some insight about the typical behaviour of the user. Of the several facets that can be modelled, the ones most relevant for a conversational search system include the user's interests and the user's knowledge about the domain. The former assumes that people are most likely to search for items that they prefer[1]; in the case of the latter, the assumption is that people's knowledge is not uniform across all the domain's attributes and items. At the current time we have limited the profile to only keep track of the user's knowledge.

[1] User preferences could be accomodated by considering non-uniform item probabilities in the calculation of the entropy, for example.

5.5 The Objective Function

Particularly in the context of an ambient intelligence scenario, improving the quality of a dialogue system means foremost providing a better experience to the users. This realisation has a direct impact on the definition of the objective function which will assess the performance of the algorithms used for attribute selection.

Even though a thorough evaluation of user satisfaction can only be performed by extensive user testing, we believe that it is possible to determine beforehand a number of objective metrics which are likely to play a significant role in the overall user satisfaction. The most obvious is the total number of steps necessary to attain the final goal, the assumption being that users will prefer to locate their photographs as quickly as possible. Another possible metric considers that long dialogues might not be detrimental to user satisfaction as long as they feel that progress is being made. To be more explicit, this metric would consider *rejections*—defined as questions which the user is forced to decline because she does not have an answer for them—as being quite frustrating, and should thus be avoided, even if at the expense of a slightly longer (but safer) dialogue. Obviously, many other metrics are possible and even likely to be relevant. One could say that, for example, a single rejection is not harmful if intermixed with multiple positive responses. Again, only extensive user tests will provide a categorical answer to this question.

It should be noted that any entropy-based heuristic will attempt to minimise the length of the dialogue, while a profile-based heuristic strives to reduce the number of rejections. While there is a correlation between these two measures, it is not linear and far from obvious. A more detailed discussion on the issue can be found in [17].

5.6 User Profiles and Stereotypes

We have already commented on the potential benefits one can derive from knowing the typical behaviour of the user. However, one must also take into account that any profile-based heuristic will suffer from the well-known *blank slate* problem whenever a new user is brought into the system. Basically, since the heuristic will need time to adapt to the particularities of the user, the initial performance will be far from optimal. Using stereotypes provides an elegant solution to this problem, and was therefore one of the primary considerations when designing a framework for attribute selection.

The basic principle consists of having multiple profile heuristics in the system. One of them will be the user's personal profile, and all the others will represent a variety of different stereotypes. By designing the system in such a way that the importance of a heuristic automatically adapts in accordance to its performance, we can rely mainly on the stereotypes in the beginning, and later switch to the personal profile once it has learned enough about the user. Obviously, the expectation is that at least one of the stereotypes will be close enough to the user to provide good results.

Figure 8 depicts a framework for the process of attribute selection which includes support for stereotypes. Furthermore, it also enables the fine-tuning between profile-based and entropy-based heuristics, thus allowing one to adjust the objective function to suit the user. At last, it also enables the enforcement of policies which might run against the dictates of both the profile or entropy-based heuristics. As an example, consider an application where a certain attribute a must always be asked first. Even though our conversational search engine for photographs does not make use of such policies, one could imagine other applications that require them.

Adaptation takes place by changing the weights of the profile heuristics. After each round, the heuristics which provided good suggestions are rewarded (their weights increase), while the others are penalised (their weights decrease). Eventually, the weight of a heuristic will reflect its probability of being right, i.e., its reliability. It is this adaptation mechanism which allows the smooth transition between the initial estimates based on stereotypes and the later ones relying on the user's personal profile. Again, a thorough description of this framework can be found in [17].

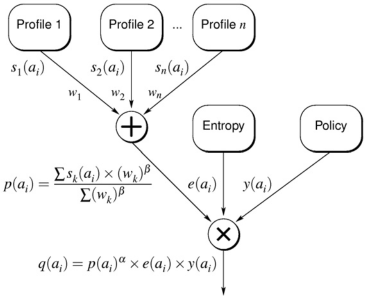

Profile 1
Profile 2
...
Profile n
$s_1(a_i)$
$s_2(a_i)$
$s_n(a_i)$
w_1
w_2
w_n
Entropy
Policy
$p(a_i) = \frac{\sum s_k(a_i) \times (w_k)^\beta}{\sum (w_k)^\beta}$
$e(a_i)$
$y(a_i)$
$q(a_i) = p(a_i)^\alpha \times e(a_i) \times y(a_i)$

Fig. 8. The framework for attribute selection. Typically the system will have one personal profile heuristic and various profile heuristics based on stereotypes. The combination of the estimates of all profile heuristics is given by $p(a_i)$, the estimate of the entropy-based heuristic is defined by $e(a_i)$, and the suggestion from the policy is $y(a_i)$. The final 'quality' value of attribute a_i is given by $q(a_i)$.

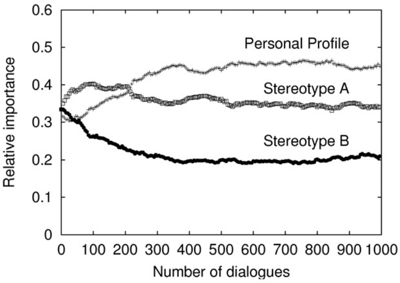


Fig. 9. Adaptation of the profiles. The advantages of using stereotypes can be clearly seen in this graph: they provide good results in the short term, compensating for the relatively slow learning process required by the user's personal profile.

6 Experimental Results

The evaluation of a system like the one described poses a number of challenges. While the browsing assistant can be assessed by conventional methods for user testing, the evaluation of the conversational search system is complicated by the reliance on user models: the importance of learning cannot be forgotten, and requires that the test be conducted over a long period. Furthermore, using stereotypes poses another issue: how can we construct enough of them to have a 'pool of stereotypes' large enough to be effective?

Nevertheless, fine-tuning many of the heuristics and parameters used in the framework can be done by relying on simulated users. A simulated user is a mathematical model which attempts to capture the essence of how a real person would behave. It has the advantage that one is not limited by constraints such as user fatigue, boredom, and the relatively slow speed of real-time interaction. On the other hand, simulated users can be dangerous, and pose a number of potential issues that one must be aware of. Foremost, no matter how intricate the model, it will always fail to capture all the subtleties of a real person. The consequences will range from skewing the results towards the model, to the possibility of overfitting the data.

At the moment, the simulated users we implemented capture a single aspect of the user knowledge, namely the fact that each attribute will have a different probability of being answered positively by the user. However limited, it has enabled us to conduct preliminary tests on the feasibility of the proposed engine for conversational search. One of the most interesting results corresponds to the evaluation of the mechanism used to incorporate stereotypes: the outcome can be seen in Fig. 9. It represents the evolution of the relative importance of three different profile heuristics: a personal profile, a relatively close stereotype A, and a more distant stereotype B. As one can see, the advantage of using stereotypes is obvious: in the short term they are indispensable, quickly providing good results, and compensating for the fact that the personal profile needs some time to learn about the user. The adaptation mechanism also comes out in evidence: as the number of dialogues increases and the personal profile improves its performance, the weights slowly adapt to reflect this change. As a result, the personal profile takes the predominant role in the long term. Also, as one could expect, stereotype A is better than stereotype B.

Obviously, one cannot rely on simulated users forever, and as such, one of the next steps in our research will be to evaluate the presented system in a real-world setting.

7 Conclusion

In this paper we have taken a look at photo viewing in home environments. We have argued that given the present increase in the number of photographs that people take, the issue of content overload will soon be brought into the domestic arena. With the goal of mitigating this problem, we have first identified the sort of activities that people engage in with their photographs. We have then proposed computer assistance in the form of two interaction paradigms: conversational search and the browsing assistant. Furthermore, we have shown how these paradigms could be brought under the umbrella of an unifying architecture, providing the user with a seamless experience.

The core of the paper focused on a detailed description of each of the two interaction paradigms and how they integrate with one another. One of the most important conclusions we derived was that user modelling plays a prevalent role in a system that aims to be ambient intelligent. On that note, part of our future work will address precisely the improvement of the user models used by the conversational search engine. Moreover, given the importance of the user side of the equation, work of this nature will never be validated without thorough tests with real users. This also shall be addressed soon.

Acknowledgements. The authors would like to thank the members of the PHENOM project team: Esko Dijk, Elise van den Hoven, Nick de Jong, Evert van Loenen, and Yuechen Qian.

References

1. Aarts, E., Marzano, S., eds.: The New Everyday. 010 Publishers, Rotterdam, The Netherlands (2003)
2. Maes, P.: Agents that reduce work and information overload. Communications of the ACM 37 (1994) 31–40
3. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM 40 (1997)
4. Rodden, K., Wood, K.R.: How do people manage their digital photographs? In: Proceedings of the ACM Conference on Human Factors in Computing Systems (ACM CHI 2003), Fort Lauderdale, Florida, USA (2003)
5. Lieberman, H.: Autonomous interface agents. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (ACM CHI 1997), Atlanta, Georgia, USA (1997)
6. Zue, V.: Conversational interfaces: Advances and challenges. Proceedings of the IEEE 2000 (2000)
7. Langley, P., Thompson, C., Elio, R., Haddadi, A.: An adaptive conversational interface for destination advice. In: Proceedings of the Third International Workshop on Cooperative Information Agents, Uppsala, Sweden (1999)
8. Göker, M.H., Thompson, C.A.: Personalized conversational case-based recommendation. In: Proceedings of the 5th European Workshop on Case Based Reasoning, Trento, Italy (2000)
9. Rui, Y., Huang, T.S., Chang, S.F.: Image retrieval: Current techniques, promising directions and open issues. Journal of Visual Communication and Image Representation 10 (1999) 39–62
10. Wang, J.Z., Li, J.: Learning-based linguistic indexing of pictures with 2-D MHMMs. In: Proceedings of the 10th ACM International Conference on Multimedia, Juan-les-Pins, France (2002) 436–445
11. Searle, J.R.: Speech Acts: An Essay in the Philosophy of Language. Cambridge Unversity Press (1969)
12. van Loenen, E.: On the role of graspable objects in the ambient intelligence paradigm. In: Proceedings of the Smart Objects Conference, Grenoble, France (2003) 3–7
13. van den Hoven, E., Eggen, B.: Digital photo browsing with souvenirs. In: Proceedings of Interact2003, Zurich, Switzerland (2003)
14. Fitzmaurice, G.W.: Graspable User Interfaces. PhD thesis, Dept. Of Computer Science, University of Toronto (1996)
15. Levin, E., Pieraccini, R., Eckert, W.: Using markov decision process for learning dialogue strategies. In: Proceedings of ICASSP98, Seattle, Washington, USA (1998)
16. Litman, D.J., Kearns, M.S., Singh, S., Walker, M.A.: Automatic optimization of dialogue management. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000), Saarbrucken, Germany (2000)
17. Teixeira, D., Verhaegh, W.: Optimising attribute selection in conversational search. In: Proceedings of the Sixth International Conference on Text, Speech, and Dialogue—TSD 2003, České Budějovice, Czech Republic (2003)

A Robotic Assistant for Ambient Intelligent Meeting Rooms

Marnix Nuttin[1], Dirk Vanhooydonck[1], Eric Demeester[1], Hendrik Van Brussel[1], Karel Buijsse[2], Luc Desimpelaere[2], Peter Ramon[2], and Tony Verschelden[2]

[1] K.U.Leuven, PMA, Celestijnenlaan 300 B, B-3001 Heverlee, Belgium.
marnix.nuttin@mech.kuleuven.ac.be
[2] BARCO Noordlaan 5, Industriezone, B-8520 Kuurne, Belgium
karel.buijsse@barco.com

Abstract. This paper reports on a robotic assistant for ambient intelligent meeting rooms and also for human-centred environments in general. The usefulness of such an “embodied” assistant as a video conferencing tool for sites not possessing an intelligent meeting room and as a mobile extension of an "intelliroom" is discussed, along with possible scenarios. The most important benefit is probably the more natural interaction between the human and the intelligent environment through this "embodied" assistant. This paper also proposes a hybrid approach for moving around in human-centred environments.

1 Introduction

A recent trend in information technology is the increasing research on ambient intelligent environments. In the framework of the ITEA project AMBIENCE, amongst other activities, the concept is explored of a robotic assistant for ambient intelligent meeting rooms. Several challenges and research questions arise in this context, including the questions how this environment should interact with the user, what benefit an “embodied” assistant offers, how users react to a robotic assistant, and how to control the robotic assistant. This paper reports on on-going research on this subject.

Research on AMIS's (autonomous mobile intelligent systems) nowadays is moving towards operation in human-centred environments and on interaction with these humans, especially in the area of service and entertainment robotics. A well-known example is the Sony AIBO robot that is meant to be an intelligent pet; it has a sensory system that allows it to see, hear and feel for itself. It can even learn from experience and adapt its behaviour or its personality accordingly [4]. Another example is the NEC PaPeRo robot: “NEC is conducting research on a personal robot to really break through the barrier between people and computers” [5].

Examples of service robots can, amongst others, be found in the area of elderly and handicapped people: on the one hand there exist personal assistants like MOVAID (MObility and actiVity AssIstance system for the Disabled) [6], on the other hand there are intelligent wheelchairs like Sharioto [7],[8]. Other examples of service ro

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 304–317, 2003.

bots also appear in the area of telepresence. In this case the robot has to be equipped with a (wireless) link to a (local) network, a camera, and microphones. Possible applications are surveillance, video conferencing, etc. The latter is the topic of this paper. The "Personal Roving Presence" [3] is an example of a platform with a camera, an LCD screen and a user interface. Service robots are also appearing as interactive guides in e.g. musea [9].

It is our belief that the human interaction with an ambient intelligent environment can occur more naturally via an AMIS. Natural interaction can occur through speech, gesture and emotional feedback (e.g. by facial expressions and body language), that are all possible via an appropriately embodied assistant.

We can distinguish different human-centred intelligent environments that could benefit from a personal robotic assistant: the home environment, the professional environment and the public environment.

1.1 Home Environment

The domestic robot in this case is a personification of the intelligent environment. In that sense it could incorporate a lot of functionality. In the first place it is an interface for environmental control: it can switch on/off the television for you, it can start the coffee machine, dial to friends… It is connected to the internet and can buy you some tickets for the opera, or allow telepresence so that, from your office, you can open the door for e.g. the repairman of your washing machine while keeping an eye on what is happening over there. It can be an interactive playmate or companion that recognises persons of the family, gives them information (e.g. about tomorrow's weather) and keeps their personal settings, and e.g. suggests a movie based on their preferences. Moreover, the intuitive/emotional interaction can be just fun, which probably is a very important feature. In the Ambience framework, a domestic user interface robot is under development [1].

1.2 Public Environment

The robot can be your interactive guide in a public building, e.g. a museum: it shows you around, starts a documentary on a nearby television, shows pictures on a big screen, etc. Another application is the robot as a receptionist in a company who guides (important) visitors to a meeting room, a demo centre or to the manager's office.

1.3 Professional/Office Environment

Intelligent meeting rooms enable effective sharing of ideas and documents while physically being far apart, see Figure 1. However, not every company possesses such a room. An intelligent robotic assistant with a camera, microphones, a user interface and eventually a screen, might overcome this problem by creating the required telepresence.

Fig. 1. The intelliroom at BARCO

Another use of a robotic assistant can be in the intelligent meeting room itself. The assistant can e.g. present the user interface of the meeting room when it is called, or it can make close-ups of certain documents when these are presented to it.

Note that an embodied telepresence, where a person can move around in a remote space, could be experienced as a stronger form of telepresence, that in the case where the person is merely being projected on a screen.

1.4 Outline

The remainder of this paper presents a robotic assistant for ambient intelligent meeting rooms. The next paragraphs continue with the motivation for a robotic assistant and give a description of our assistant along with selected scenarios. Moreover, a software-architecture is proposed suitable for operation in human-centred environments. We also refer to [1].

2 Rationale of the Approach

In our opinion, the only way to assess the usefulness of a robotic assistant for the professional domain, is to build prototypes and test them and iterate. This section motivates our approach.

As already mentioned above, the use of a robotic assistant can allow video conferencing via the AMIS for partners not possessing an intelligent meeting room. It is easy to install (just contact its docking station and let it drive to your office) and requires no building nor construction works nor cabling to install cameras and required infrastructure in your room. It simply can be shipped by courier service to a desired location.

A possible benefit compared to the standard video conferencing room is the more 'natural' point of view of the camera. The camera on the AMIS is at eye level, so the person at the other side of the video conferencing-line gets a more natural interaction with his partner, instead of a bird-perspective view. It can even be convenient that the camera has no fixed place in the room, so the camera can take nearly any point of view of the speaker. The assistant can also track a person while he is moving (e.g. during a presentation or during a meeting at an intelligent wall). In this scenario we can say that the AMIS acts as a camera-man with a microphone. This camera-man should also respond to questions that are asked to him (speech/voice recognition): e.g. zoom to some document a person has brought with him.

Let us consider another possible benefit. The knowledge that every object in the environment (omnipresent cameras) is watching him or her, could give the user an unpleasant "Big Brother" experience. We believe that an embodied personal robotic assistant - as a mediator between the human and the intelligent environment - might overcome these barriers and make the ambient presence of digital assistance more "localised" and acceptable.

When the presence of cameras is not suitable (e.g. in a confidential meeting, e.g. a meeting with the trade-union) the AMIS can easily be taken away. In principle the AMIS can go autonomously to a cupboard (privacy). When it is not used (during night, breaks, etc.) it can automatically dock at a recharging station.

In another scenario, the AMIS could be a robotic assistant in the intelligent meeting room itself. It could for example solve the problem where to put the user interface to the video conference system: should this user interface be placed somewhere on the table or in the neighbourhood of the screen? A possible alternative is to integrate all user interface functions onto the AMIS that acts as a mobile user interface. The robotic assistant can also extend the possibilities of the conference room by providing an additional mobile camera, e.g. to zoom to some document a person has brought with him, or to provide a better close-up of certain persons.

3 Description of the Meeting Assistant

3.1 The Assistant "Maktub"

The main specifications for the robotic assistant are:

- Audio/video - input;
- Audio/(video) - output;
- Camera at eye-level;
- Ability to move appropriately in a standard meeting room or office, i.e. the robot should not be too wide, but should be higher than a table;
- Sensors and software for obstacle avoidance and navigation;

- Wireless connection to a local network or the internet;
- Software for speech recognition, for visual detection of objects to be zoomed at, for detection of the direction of speech, for tracking persons, etc. (future work).

In our lab, we are developing the meeting room assistant called "Maktub". Maktub is an AMIS that is based on the ActivMedia Peoplebot platform. Fig. 2 and Fig. 3 give an overview of the hardware components of this platform, which show that Maktub is suited for his task.

Maktub is equipped with a laser scanner and ultrasonic sensors for obstacle avoidance and navigation. Maktub also has inertial sensors and a compass for measurement of the heading direction. Infrared sensors take care of table detection. Furthermore, Maktub has a PTZ (pan tilt zoom) camera at eye level and microphones for audiovisual input. Speakers at the top allow users of the robotic assistant to hear their partners at the other side of the line. A wireless ethernet connection interfaces Maktub with a local network or with the intelligent environment. The audiovisual signal can also be transmitted to a remote station by the A/V transmitter to be processed there. Maktub has an onboard computer for autonomous behaviour. All software that is required for the autonomous assistant runs on this computer. This computer is connected to the microcontroller of the platform by a client-server link. The microcontroller (= server) takes care of the low-level actuator commands and the sensor signals.

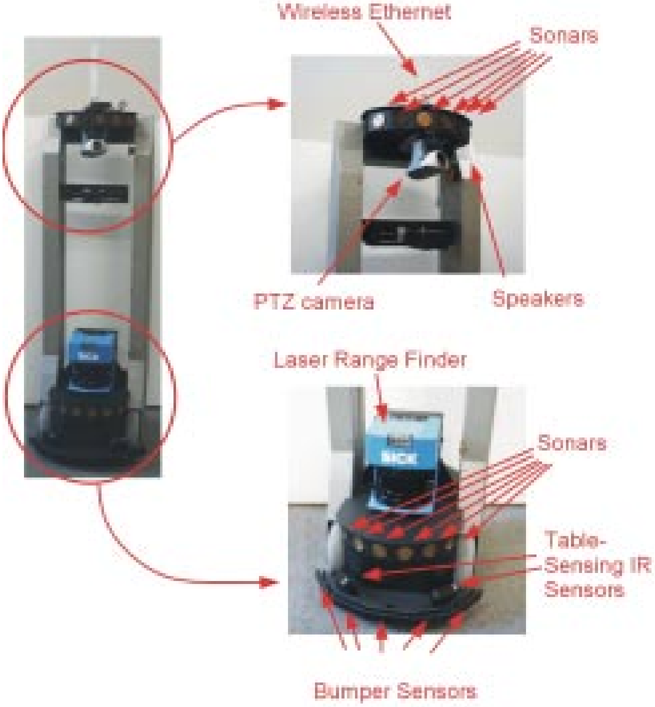


Fig. 2. Description of Maktub

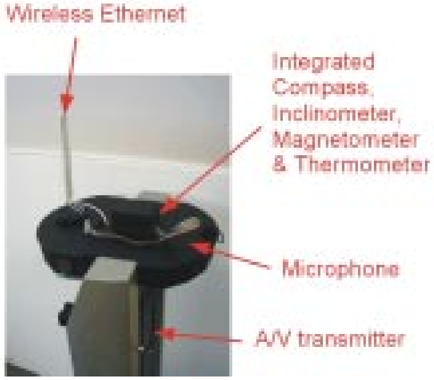

Wireless Ethernet
Integrated Compass, Inclinometer, Magnetometer & Thermometer
Microphone
A/V transmitter

Fig. 3. Description of Maktub (c'd)

From a practical point of view, the Peoplebot is a slender robot, which allows it to manoeuvre easily in narrow environments in between tables and chairs as shown in Fig. 4.

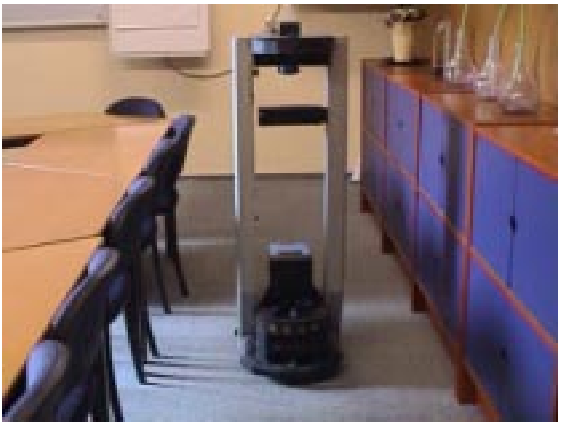


Fig. 4. The robot Maktub in a meeting room

3.2 Services to Be Offered by "Maktub"

Maktub can be used as a substitute for a video-conferencing room. A possible scenario is the one that is demonstrated in Figure 5. Maktub observes the person who is explaining, and captures the audiovisual signal. This signal is sent to a local network through the wireless Ethernet link. From the local network, it travels via the internet to the other side of the video-conferencing line. The partner at this other side can receive the audiovisual signal on his computer, on "Maktub 2" or on the walls of the intelligent meeting room. In this scenario, Maktub could be:

1. either remotely controlled by the partner at the other side (e.g. simply with a joystick); this partner controls the pan-tilt-zoom camera, and the position and point of view of the robot; in this control scheme, a basic form of shared control is required: the partner performs global navigation of the robot, while the robot itself takes care of the local obstacle avoidance; this raises some ethical question e.g. on the privacy of the observed person;
2. either controlled under shared autonomy; both the local and the remote user control the robot; priorities need to be given to both users to decide who controls the robot at which moment, or to decide which user controls which parameters; of course this form of control still requires the basic level of obstacle avoidance;
3. either (fully) autonomous; the robot identifies and automatically tracks the person he should observe, plans the required navigation to get to the best point of view, zooms to objects when they are shown to him, comes closer when he is called, etc.

This functionality is also useful in a scenario where the robotic assistant is not a standalone video conferencing tool but a mobile extension in an intelligent meeting room.

In each scenario (extension of intelligent meeting room vs. standalone / shared autonomy vs. fully autonomous), the robot requires the appropriate software skills for human-centered mobility. The next paragraph deals with the proposed software architecture.

4 The Proposed Hybrid Architecture for Moving in Human-Centred Environments

The next paragraphs present a hybrid architecture used for performing the tasks explained above. This architecture consists of a deliberative and a reactive part. In this way, the advantages of both approaches are combined. The robot is able to reason about how to reach a certain goal position, taking a priori knowledge about the environment into account. At the same time it is able to react very quickly to unmodelled obstacles in the environment, by adopting a more direct coupling between sensors and actuators.

4.1 The Architecture

Figure 6 depicts the proposed architecture. The navigation module as a whole calculates the linear and rotational velocities v and ω of the robot, given the current robot location (x, y, θ) and its uncertainty, the robot's global goal (x_g, y_g), the measured ranges from the exterioceptive sensors, and the odometry values.

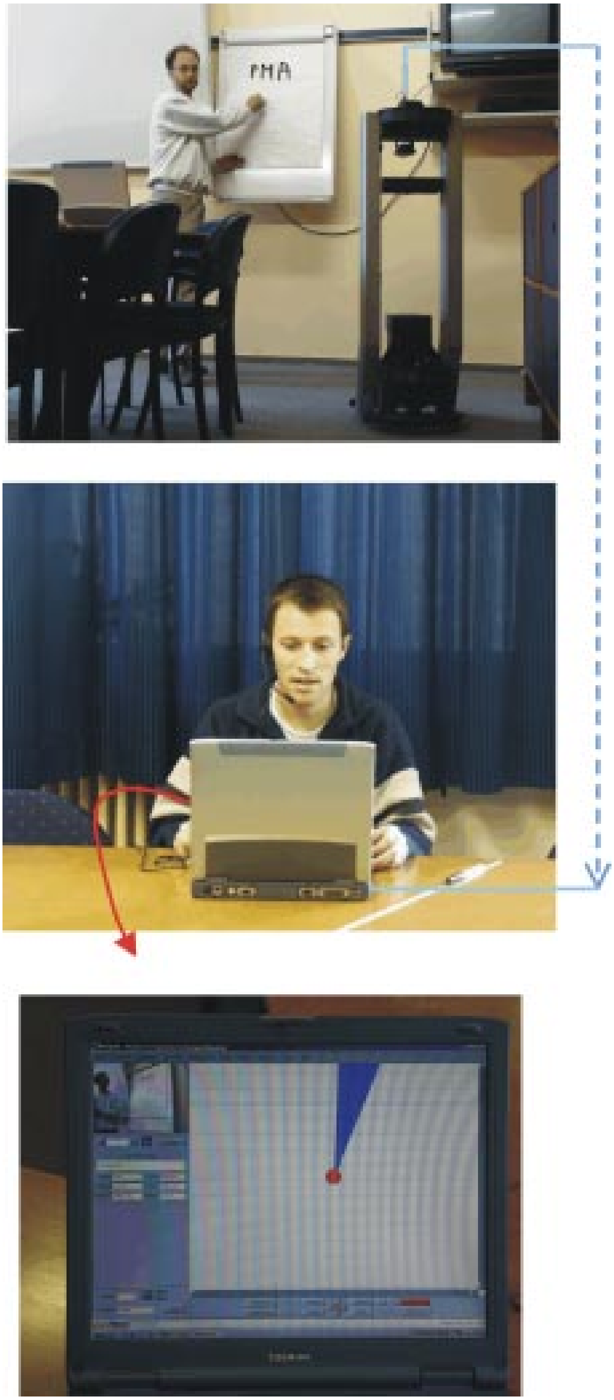


Fig. 5. The scenario

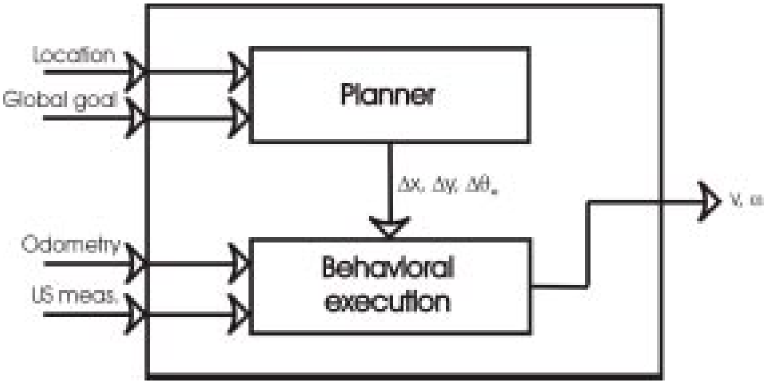


Fig. 6. Architecture of the navigation software

During navigation, a global planner and a behavioural execution unit co-operate. These basic components are explained in the next sections.

4.2 The Global Planner (Deliberative Part)

The global planner adopts a two-dimensional grid based representation of the environment, with a grid cell size of 5 cm square. Currently, only a priori information of walls, doors and furniture is incorporated. The left hand side of Fig. 7 shows a grid map of the office environment in which we performed tests.

A Manhattan path planner uses this grid to generate subgoals or via-points from a starting position to a goal position. The direct path between two subgoals is obstacle free. In order to generate the list of subgoals, the algorithm goes through the following steps:

1. The obstacles (occupied cells) are grown up, such that the robot (that moves as a point in the algorithm) doesn't collide with them. The result is shown in the right hand side of Fig. 7. The grow-up radius around occupied cells is taken somewhat larger than the robot's radius. The figure also shows a possible robot's current location *L* and a global goal *G*.
2. Since the robot motion and the sensor measurements always contain some uncertainties, and in order not to put too many unnecessary constraints on the obstacle avoidance, the clearance to obstacles is taken into account while finding a global path from start to end position. The approach followed here is inspired by the one in [10]. The strength of this approach is that it takes the obstacle clearance into account without suffering from the problem of local minima. In order to do so, a wave front distance field (WFDF) algorithm starts at the grid cells containing obstacles, setting their values at 0. Next, the value of this grid cell's (unoccupied) neighbours is set to one. After that, their (free) neighbours get value two, and so on, spreading like a wave. The values set in the grid cells represent the (Manhattan) distance from every cell to the most nearby obstacle. The WFDF algorithm stops

when all free cells have been appointed an obstacle clearance value. The left hand side of Fig. 8 shows this obstacle clearance grid for the environment shown in Fig. 7. Dark colours indicate cells that are lying close to obstacles. A very efficient implementation of the WFDF algorithm allows for fast calculation of this obstacle clearance grid.

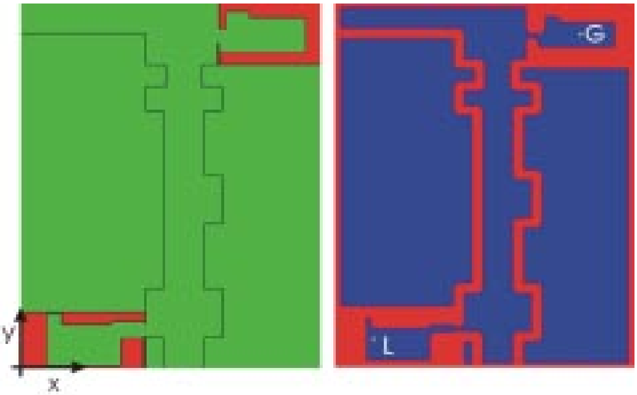


Fig. 7. (left) Grid map of an office environment and grown-up map of the environment

3. Next, the same WFDF algorithm is used to spread a wave front starting from the goal position *G* using the following cost function:

$$\min_{p \in P}\left(length(p) + \sum_{c_i \in P} \alpha \cdot obstacle(c_i) \right),$$

   where *P* is the set of all possible paths *p* from the cell *c* to the goal. The function *obstacle(c)* is a cost function generated using the values of the obstacle transform. It represents the degree of discomfort the nearest obstacle exerts on a cell *c*. A linear cost function that ranges from zero at a fixed distance (set by the user) to a maximum cost at zero distance, yields good results. The weight α is a constant ≥ 0 that determines by how strongly the WFDF will avoid obstacles. The algorithm stops when the current position *L* has been reached.
4. In this wave grid, a Manhattan path is calculated. The path to the goal is found by tracing the path of steepest descent, starting at the current position *L*. Fig. 8 shows an example. The colour changes from dark grey at *G* to light grey at cells that are as far from the goal as *L*.
5. The last step is to find subgoals on this path. Beginning at the starting point, the Manhattan path is walked through, stopping at a cell which cannot be reached (in a direct way from the starting point) without bumping into obstacles and without exceeding a minimal obstacle clearance threshold. The cell before is taken as a subgoal. From this point, a next subgoal is found in the same way. The last subgoal will be the endpoint. Fig. 8 shows these subgoals on the Manhattan path.

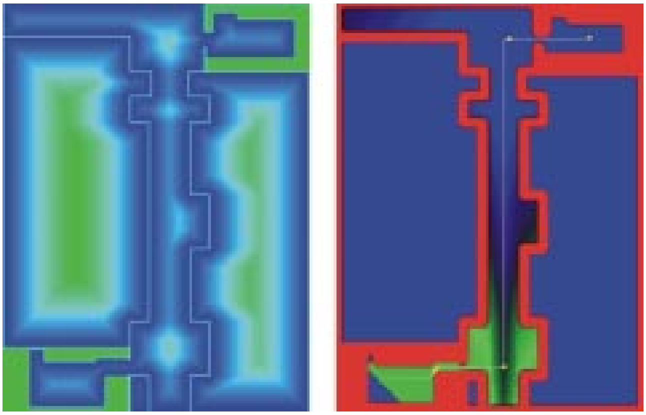


Fig. 8. (left) Obstacle clearance grid and (right) planned path from start to goal location

The list of subgoals is calculated very quickly (refresh rate of 4 Hz), with an update during each cycle of the current location and possibly, but less likely, of the global goal. Of importance for the robot is the first subgoal of this list, because this is the position it has to go to at the moment of calculation. Based on this first subgoal, the planner calculates (*Δx, Δy, Δθ_e*), i.e. the requested changes in position and orientation. These are expressed in the local robot frame. The values must be interpreted as follows: the robot first turns over the necessary angle $\Delta\theta_i$ in order to see the subgoal, then it goes to a position which is (*Δx, Δy*) from its current position, and eventually it turns such that its orientation is $\Delta\theta_e$ farther than the starting orientation (see also Fig. 9). These values are the input for the behavioural execution unit.

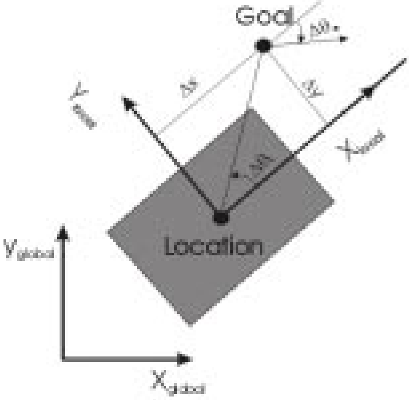


Fig. 9. Explanation of the inputs of the behavioural execution unit

4.3 The Behavioural Execution Unit (Reactive Part)

The general idea behind the behaviour-based approach is to decompose a task in simpler tasks that are easier to implement and test. The challenge of this approach remains in how to combine these different subtasks such that the global task is executed in a robust manner.

We developed a multi-agent framework in which behaviours can be specified conveniently. The currently developed behavioural execution unit consists of a move-to-goal behaviour and an obstacle-avoidance behaviour. Fig. 10 gives a schematic description of the behavioural execution unit. The "input" module receives the deltas (Δx, Δy, $\Delta\theta_e$) from the planner, and updates these deltas while driving using odometry. Furthermore, it processes the deltas, and puts the processed deltas at the agents' disposal together with the sensor values.

The agents react upon these values: the "turn" agents bring the robot on its course to the next (sub)goal by turning the robot ($\omega = \alpha\cdot\omega_{max}$ and $\omega = -\alpha\cdot\omega_{max}$ respectively). The "mtcurrgoal" agent goes straightforward ($v = \alpha\cdot v_{max}$) when the robot is on course, thus going to its next (sub)goal. The "ObstAvoid" agent avoids obstacles while taking the current subgoal into account. Each of these agents decides to which extent it has to be active. For example, when not on course, either the counter clockwise (ccw) or the clockwise (cw) turn agent wants to be very active. Its activation α is near 1. The other turn agent is not active at all (activation is zero), and the mtcurrgoal agent's activation evolves gradually from 0 to 1 as soon as the robot is on course. These activation values are the weights in the "weighted addition" co-ordination object, which calculates v and ω. The "lazy" agent basically does not do anything ($v = \omega = 0$), but always has activation equal to one, in order to avoid a division by zero in the weighted addition. The weighted addition co-ordination object receives α, v, and ω from the turn, mtcurrgoal and lazy agents. The resulting v and ω are then calculated as follows:

$$v = \frac{\alpha_{cw}v_{cw} + \alpha_{ccw}v_{ccw} + \alpha_{mt}v_{mt} + \alpha_{lazy}v_{lazy}}{\alpha_{cw} + \alpha_{ccw} + \alpha_{mt} + \alpha_{lazy}}$$

$$\omega = \frac{\alpha_{cw}\omega_{cw} + \alpha_{ccw}\omega_{ccw} + \alpha_{mt}\omega_{mt} + \alpha_{lazy}\omega_{lazy}}{\alpha_{cw} + \alpha_{ccw} + \alpha_{mt} + \alpha_{lazy}}$$

In order to avoid collisions, the move-to-goal and the obstacle-avoidance behaviour should not be active at the same time. Therefore, a fixed priority co-ordination scheme is used to combine both behaviours: if the obstacle-avoidance activation is above a certain threshold, then obstacle avoidance is active. Otherwise, the move-to-goal behaviour is active. These results are then given to the wheel system output, which is the global output of the behavioural execution unit.

5 Results

Experiments (see e.g. [1]) showed that the performance of the Manhattan planner is very reliable and robust. A major contribution of our approach is the very fast implementation of the planner. This allows for almost real-time replanning of the path on a global scale (i.e. replanning in large environments such as offices of e.g. *15m x 20m*).

Furthermore, the fact that obstacle clearance is taken into account during planning adds to the robustness of the approach. In the future, a bidirectional communication may be considered between the reactive and the deliberative part, so that the behavioural execution unit may e.g. notify the planner of subgoals that cannot be reached.

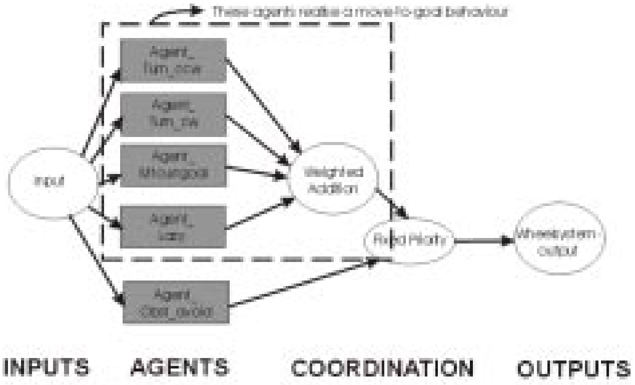


Fig. 10. Co-ordination scheme for the behaviours

During tests with this navigation architecture, we also experienced that it is necessary to tune the different parameters that are introduced in the various behaviours and their activation functions. The parameters may not only depend on the robot's geometry, its kinematics and dynamics, but also on the environment in which it travels. Therefore, part of our future research will focus on how to automatically tune the parameters by a learning process to enable robust navigation.

6 Concluding Remarks

This paper presents on-going research in the framework of the ITEA project AMBIENCE. This research explores how intelligent digital environments can serve humans more naturally through a robotic assistant. The concept of an embodied assistant for an intelligent meeting room is explored and a hybrid approach is proposed for human-centred mobility.

Acknowledgments. This research is carried out with financial support of IWT-Vlaanderen in the ITEA project AMBIENCE. The research group further acknowledges the GOA project 99/04, and the Inter-University Attraction Pole, IUAP/AMS, Office of the Prime Minister. We thank the partners of the AMBIENCE project.

References

1. A.J.N van Breemen, K. Crucq, B.J.A. Krose, M. Nuttin, J.M. Porta, and E. Demeester, "A User-Interface Robot for Ambient Intelligent Environments", ASER 2003, March 13–15, 2003, Bardolino, Italy, pp. 132–139
2. AMBIENCE project: http://www.extra.research.philips.com/euprojects/ambience/
3. Personal Roving Presence, http://www.paulos.net/
4. Sony AIBO, http://www.aibo.com/
5. NEC Personal Robot PaPeRo, http://www.incx.nec.co.jp/robot/PaPeRo/english/p_index.html
6. Movaid (MObility and actiVity AssIstance system for the Disabled), http://www-arts.sssup.it/old_site/Research/movaid.html/
7. Vanhooydonck, D., Demeester, E., Nuttin, M., Van Brussel, H. Shared Control for Intelligent Wheelchairs: an Implicit Estimation of the User Intention. Proceedings of the ASER '03 1st International Workshop on Advances in Service Robotics, Bardolino, Italy, March 13–15, 2003
8. http://www.mech.kuleuven.ac.be/pma/research/mlr
9. R. Siegwart R., K.O. Arras, B. Jensen, R. Philippsen, N. Tomatis Design, Implementation and Exploitation of a New Fully Autonomous Tour Guide Robot. Proceedings of the ASER '03 1st International Workshop on Advances in Service Robotics, Bardolino, Italy, March 13–15, 2003
10. A. Zelinsky, Using Path Transforms to Guide the Search for Findpath in 2D. International Journal of Robotics Research, Vol. 13, n. 4, p. 315–325, August

Correlating Sensors and Activities in an Intelligent Environment: A Logistic Regression Approach

Fahd Al-Bin-Ali[1, 2], Prasad Boddupalli[1, 2], Nigel Davies[1, 2], and Adrian Friday[2]

[1] Computer Science Department, University of Arizona, Tucson,
AZ 85721, USA
{Albinali, Bprasad}@cs.arizona.edu

[2] Computing Department, Lancaster University,
Lancaster, LA1 4YR, UK
{Nigel, Adrian}@comp.lancs.ac.uk

Abstract. An important problem in intelligent environments is how the system can identify and model users' activities. This paper describes a new technique for identifying correlations between sensors and activities in an intelligent environment. Intelligent systems can then use these correlations to recognize the activities in a space. The proposed approach is motivated by the need for distinguishing the critical set of sensors that identifies a specific activity from others that do not. We compare several correlation techniques and show that logistic regression is a suitable solution. Finally, we describe our approach and report preliminary results.

1 Introduction

In his classic paper "The Computer for the 21st Century" [14] Weiser envisions a world of intelligent environments that are highly aware of their inhabitants. In this vision, physical spaces are enhanced with computing capabilities to act more intelligently: they observe, interact with and react to humans in meaningful ways. They understand human reasoning, analyze behaviors and infer intentions. Furthermore, intelligent environments actively collaborate with their inhabitants to assist them in making their surroundings more pleasant. Intelligent environments even take decisions and execute actions on their own. They become integral participants in the daily human activity.

A critical element that Weiser anticipated, yet has not been achieved, is the invisibility of pervasive systems. The ability of such systems to disappear into the background of everyday life is dependant on their ability to correctly interpret the state of the environment and to act accordingly: intelligent systems that incorrectly interpret the state of the world or the intentions of users are likely to take inappropriate actions that are not naturally anticipated by users [6]. Such incorrect actions could become very disruptive and intrusive to users, they distract the inhabitants of intelligent spaces from their ongoing activity and therefore, they make them more aware of the

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 318–333, 2003.

system. This paper begins to address the challenge of designing less intrusive intelligent environments that can engage in richer and more meaningful interactions with users. We believe that such systems must have a deep understanding of user context and, specifically, should have an understanding of activities that a user is engaged in. Our approach is thus inspired by concepts from activity theory [9] and requires support for three basic system functions:

- *Sensing context:* By observing and monitoring users' context, intelligent systems can collect information about the intelligent space and its inhabitants.
- *Analyzing context:* By analyzing users' context, intelligent systems can estimate and interpret users' activities.
- *Gracefully reacting to the inhabitants:* By understanding users' activities, intelligent systems can react unobtrusively to their inhabitants and therefore can potentially become more invisible.

In this paper, we focus on one aspect of our system design, i.e. how to identify sensors that correlate with activities in an intelligent space. First, we motivate our use of an activity-centric approach and justify the need for precisely identifying correlations between sensors and activities. Second, we identify a number of desirable properties for activity-aware intelligent systems. We then analyze different techniques for identifying the correlations between sensors and activities and show that statistical logistic regression has the desired properties. Third, we describe in detail our regression technique. Finally, we report preliminary results and state our conclusions.

2 Why an Activity-Centric Approach?

Intelligent environments are inherently social and collaborative spaces. Understanding the "behavioral-level" interaction in such environments require modeling the context in which the inhabitants of the space interact [6]. Early research [3],[12] in intelligent environments focused on establishing simple relationships between tangible context and appropriate actions, for example, switching on and off devices based on user proximity. Intangible context such as activities, human moods and human intentions and complex relationships between sensor data and actions have not received significant attention to date. However, to be invisible, intelligent systems must understand both tangible and intangible aspects of context and the complex relationships between sensors and actions.

We believe that the best method for capturing these complex relationships is using the notion of 'activities'. Many earlier projects acknowledge a need for such a capability. For example, MIT utilizes an activity based approach in their second generation iRoom [11]. EasyLiving [3] from Microsoft acknowledges the need for tracking activity in an intelligent environment. Responsive Offices [8] from Xerox PARC identifies activity as an essential ingredient for determining appropriate reactive behaviors. Moreover, numerous studies in psychology [9] advocate that individual and group behavior should be interpreted in relation to the activities people participate in. Indeed, recent work on groupware [2] has employed many of these concepts (in par-

ticular concepts from activity theory) for modeling collaborative tasks. Such systems interpret behaviors by considering the activity as the fundamental unit of analysis.
Figure 1 shows a high-level view of our activity analysis system. Initially, sensors in the intelligent space are correlated with activities that interest the inhabitants. The system uses empirical data (collected from the space) to derive causal correlations between activities and sensors. The correlations are then used to create a *correlation matrix* that captures all the correlations between activities and sensors in an intelligent space. Subsequently, the intelligent system can use the matrix to interpret the activities in the intelligent space, for example, a probabilistic reasoner can use the matrix for building a Bayesian network to analyze the activities. This might involve assessing the uncertainties in the reasoner's inferences or establishing a dialogue with the inhabitants of the space to disambiguate activities in situations of high uncertainty as proposed by [4],[5].

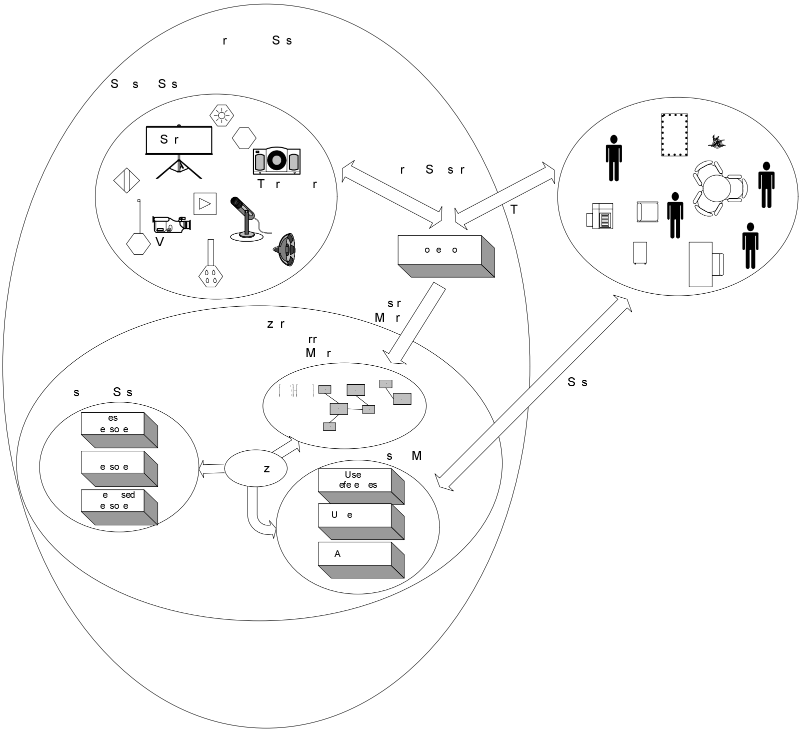


Fig. 1. Activity-aware intelligent space

It is important to emphasize that in a ubiquitous environment that is saturated with sensors, it is extremely important to distinguish the *critical set* of sensors that correlate with a specific activity from others that do not. For example, imagine constructing a Bayesian network for the activities in a ubiquitous space without knowing the dependencies between sensor readings and the activities. Including uncorrelated sensors in the Bayesian network will result in inaccuracies that can potentially misguide the reasoner. Similarly, excluding correlated sensors from the Bayesian network could result in ignoring some important aspects of the activities that can also misguide the reasoner. This paper focuses on how to determine the *critical set* of sensors that correlate to activities and proposes a new technique for accomplishing that. We begin our discussion by examining some of the desirable properties for activity-aware intelligent systems.

3 Desirable Properties for Activity-Aware Intelligent Systems

Few intelligent environments exist, and those that do are confined within research labs. Therefore, to identify the desirable properties for activity-aware intelligent systems, we examined recent work on intelligent environments [3],[12],[13], studies from psychology on individual and group behavior [9], work on natural and multimodal human-computer interaction (HCI) [6,7] and connectionist and statistical modeling techniques [5],[10],[12]. These efforts led us to the following desirable properties:

3.1 Transparency and Comprehensibility

Intelligent systems must support transparent activity modeling. Transparent modeling enables intelligent environments to reason in ways that are comprehensible to their inhabitants. Such a property is critical in order that it is possible to formulate precisely how systems reached particular decisions. Subsequently, this information could be relayed to the inhabitants of an intelligent space to support a dialogue with the system as proposed by [4,5] to fix any incorrect actions.

3.2 Adaptability

Intelligent systems must be adaptable to endure the highly dynamic nature of ubiquitous environments. Such adaptability must apply to both physical reconfiguration of spaces (e.g. changes in the availability of sensors) and to changes in activity patterns within these spaces. Different systems will require different forms of adaptability including offline adaptability in which sensor data is logged for later analysis and online adaptability in which sensor data is examined and adaptation is performed while the system is in use.

3.3 Accuracy

Clearly, achieving high accuracy in terms of identifying the activity in an intelligent environment from a given set of sensor data is crucial. However, it should be noted that the exact requirements in terms of accuracy are actually a property of the entire system and are influenced by the significance of the actions that will be triggered: users will perceive the activity analysis process as accurate and indeed as invisible when the system's reactions are correct. However, this does not necessarily mean that the system has identified the users' activities correctly. For example, imagine a user having a nap while watching TV. An intelligent system might detect a reduction in the overall mobility in the space and therefore infers that no one is in the room; resulting in switching off the TV and the lights. Clearly, the analysis process misdiagnosed the activity, but the outcome is still considered correct by the user.

3.4 Knowledge Portability

It is important that knowledge about users and their activity patterns can be moved between intelligent environments, reflecting user mobility inherent in the real world. This will require a clear separation between the models that represent the system's knowledge about activities and the system-specific assumptions and mechanisms. In practice, achieving portability is likely to be extremely complex, raising many technical challenges (e.g. determining the equivalence between sensors in different environments) and non-technical challenges in areas such as legal and social ethics (e.g. can models about activity patterns be exchanged between private places and public places without violating the privacy of people?).

So far, we have described 4 desirable properties for activity-aware intelligent systems. It should be clear that the above properties are not exhaustive, but we have deliberately chosen them because of their importance in the context of intelligent environments. It should also be noted that many of the properties discussed above are greatly exacerbated when multiple people are participating in an activity.

4 Techniques for Correlating Activities and Sensors

Several techniques can be conceived for correlating activities and sensors including: expert correlation, statistical correlation and connectionist correlation. We briefly describe these approaches and we analyze their merits and demerits.

4.1 Expert Correlation

The easiest way to correlate activities and sensors in an intelligent space is to use the opinion of an expert who is familiar with the space. For example, in a smart classroom, a teacher can identify different activities that students participate in such as pop quiz, discussion, on-board problem solving exercise etc. Subsequently, a rough mapping could be made between these activities and the available sensors in the classroom. Gaia [13] uses this approach for activity analysis where the inhabitants of the intelligent space identify the correlations and construct a belief network that models their activities. This network is then used by a Bayesian reasoner to identify the activities.

Expert correlation suffers from several limitations. Firstly, it does not scale well: as more activities and sensors are introduced, it becomes harder for human experts to assess the correlations. Secondly, people might have different views about the degrees of correlation between sensors and activities. Therefore, relying on the subjective assessment of a particular individual might lead to inaccuracies. Thirdly, adapting the correlations to the dynamic nature of a ubiquitous space requires a human expert: an undesirable proposition especially when intelligent spaces host rapidly changing activities. Hence, we believe that expert correlation is of limited use in ubiquitous environments that are heavily saturated with sensors.

4.2 Connectionist and Statistical Correlation

Alternatively, connectionist or statistical techniques can be used to identify correlations between sensors and activities. Connectionist correlation relies on neural network analysis that identifies patterns between different inputs. This approach has been used in the neural house project [12] where a neural network observes the lifestyle of the inhabitants of a house and programs itself accordingly. Similarly, statistical techniques such as regression can identify potential causal relationships between different variables. In ubiquitous environments, these two techniques can certainly handle large amounts of data that human experts find cumbersome. Several research studies [5],[12] have affirmed that neural techniques are more accurate than regression techniques owing to their ability to capture non-linear correlations automatically. However, they suffer from the incomprehensibility of the decision making process, i.e. it is very hard to reconstruct the rationale of a neural network of why a particular correlation between a sensor and an activity is strong. In contrast, statistical techniques are based upon “well understood models of behavior” and therefore, it is usually easier to reconstruct the rationale behind their decision making process [5]. Moreover, adapting neural networks to the continuous changes in an intelligent space might often require retraining the whole network which can be an expensive process especially in cases that require on-line adaptation.

4.3 Analysis

In light of the above discussion, we can see that expert correlation is not a viable solution due to its vulnerability to inaccuracies and its inability to deal with the abundance of sensor information in ubiquitous environments. In contrast, regression and neural networks can deal with the richness of sensor information in such environments. However, regression provides a more comprehensible framework than neural based techniques thereby making it more suitable for supporting transparent modeling where users can establish a dialogue with the system. Moreover, reapplying regression to adapt to the dynamic nature of a ubiquitous space is likely to incur less overhead than retraining a neural network. Finally, it should be noted that although regression is less accurate than neural techniques, its outcome is still comparable [5]. However, it would be unfair to give the impression that neural analysis is unusable, while regression is completely without problems. The major conceptual limitation in regression is that it can never identify the underlying causal mechanism. For example, one would find a strong positive correlation between the number of users attached to a particular access point in a conference hall and the presentation activities taking place. Do we conclude that a presentation activity causes an increase in the number of users attached to an access point? Even though that might be the case in this simple example, in many other cases, the causal explanations might not be obvious. Moreover, as the number of variables increase, more empirical observations are required to avoid having significant correlations while in fact one or more variables are capitalizing on chance. Finally, even though the rationale behind correlations is potentially easier to reconstruct using regression, it is unclear how easy it is to relay that information to regular inhabitants of an intelligent space that have no prior knowledge of statistics. Undoubtedly, friendly means should be developed to enable such system-user dialogues. We acknowledge these problems and recognize the need for exploring them further.

5 Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) [10] is a statistical technique that investigates and models relationships between a dependent variable and one or more independent variables. It is typically used when a dependent variable has the following properties:

- *Categorical*: The dependent has a limited set of values (e.g. for an activity {presentation=0, break=1, lunch=2}) that could be ordinal (e.g. {strongly agree, agree, disagree}) or non-ordinal.
- *Mutually Exclusive*: Any instance of a dependent cannot be classified as belonging to more than one category. For example, considering an activity as a dependant, an instance of an activity cannot be a presentation and a break at the same time.
- *Polychotomous:* The dependent can have 2 or more categories. A special case of MLR is the binomial logistic regression that deals with the dependent when it is a dichotomy.

MLR can deal with independents of any type (e.g. continuous, discrete, dichotomous, polychotomous etc.). Generally speaking, MLR has less stringent requirements than conventional regression techniques including:

1. It does not assume linearity of relationship between the independent variables and the dependent.
2. It does not require normally distributed variables.
3. It does not assume homoscedasticity (i.e. the variance around the regression fit is the same).

Details of logistic regression techniques can be found in [10], below we explain only those aspects critical to our discussion. In particular, we explain how to assess the adequacy of a logistic regression.

5.1 Logistic R^2

The logistic R^2 measures the strength of the association between the dependent variable and the independents. It should be noted that the logistic R^2 is different from the R^2 in conventional regression. The latter measures the goodness of fit relying on the variance around the regression fit. However, the variance of categorical dependent variables depends on the frequency distribution of that variable and therefore logistic R^2 just reflects the strength of the association.

5.2 Classification Percentage

The classification percentage reflects how good a logistic regression formula is in estimating the correct categories of a dependant. In a perfect model, the estimated values are the exact actual values making the overall classification percentage 100%. It should be noted that the classification process relies on a probability cutoff where higher cutoffs mean more sensitivity in the classification process.

5.3 Model Chi-Square Test

It is very important to determine the effect of each independent in the logistic formula. For example, the formula might show better correlation without some independents or with some additional independents. Model Chi-Square is a technique that measures the improvement in a fit that an independent variable makes compared to the null model (i.e. model without independents). This technique uses the null hypothesis to test for individual significance. The null hypothesis says that an independent variable coefficient has no effect on the dependent variable. Therefore, rejecting the hypothesis means that the independent should not be deleted from the formula because it has a significant contribution. While accepting the null hypothesis means that the independent variable is insignificant and therefore should be deleted. Generally speaking, when the probability of the Model Chi-Square is less than 0.05, the null hypothesis is rejected.

So far, we have described some important concepts for our following discussion. Next, we describe how to use logistic regression for identifying the correlations in an intelligent environment.

6 Correlation in Intelligent Environments

In the context of intelligent environments, we are using MLR for identifying the *critical set* of sensors that highly correlate with activities in an intelligent space. Sensor data is collected for some period of time while users are required to record their activities. The system records this information along with statistical data from all sensors. The collected data is then analyzed by a logistic regression engine to identify the sensors that are showing high correlation with the activities. The output of the regression engine takes the form of a correlation vector.

Definition 1. In an activity-aware environment with the following properties:

- A is a set of n+1 activities defined by $\{a_1, a_2, \cdots, a_n\} \cup \{a_\varphi\}$ where a_φ denotes the unrecognized activity, and
- S is a set of k sensors in the intelligent space defined by $\{s_1, s_2, \cdots, s_k\}$,

a Correlation Vector (CV) identifies the critical set of sensors that highly correlate with 1 or more activities (and is thus influential in identifying the activities). A CV has the following form:

$$CV(A') = < c_1, c_2, ..., c_k > \quad where\ A' \subset A\ and \\ c_i \in \{0,1\}\ and\ 1 \le i \le k \tag{1}$$

where c_i reflects the correlation between the activities belonging to A' and sensor s_i such that: $c_i = 0$ indicates no or insignificant correlation and $c_i = 1$ indicates significant correlation. For example, in a space with three sensors { $s_1 = temperature\ sensor$, $s_2 = projector\ \ sensor$, $s_3 = people\ count\ sensor$ }, a correlation vector for a presentation activity might look as follows:

$$CV(A' = \{presentation\}) = < 0,1,1 > \tag{2}$$

This indicates that a presentation activity is highly correlated with the projector sensor and the people count sensor but not with the temperature sensor.

6.1 Sensor Selection and the Correlation Matrix

Our current regression engine relies on the Chi-Square test and the classification percentage to determine the CV. We can configure the engine to select the highly correlated sensors in one of two ways:

1. *Forward Selection:* In this procedure, the best sensor is found. Next, the sensor that adds the most to the logistic fit is included. This process continues until specific cutoff thresholds (in the Chi-Square and in the classification percentage) are reached or none of the sensors add a significant value to the strength of the association.

2. *Backward Selection:* In this procedure, all the sensors are initially included in the logistic model. Subsequently, sensors are deleted from the model based on their level of significance. Again, this process continues until all the sensors left are at a specific significance level.

It should be noted that higher cutoff values reduce the number of sensors that correlate with particular activities. This potentially simplifies the reasoner's logic, for example, a rule-based reasoner that relies on a small set of sensors is likely to generate simpler rules than reasoners that account for a big set of sensors.

Furthermore, sensor selection is directly influenced by the number of categories of the dependent. When the logistic engine is given some empirical data, it tries to account for all the categories of the dependent using one single logistic formula. For example, imagine a space with the configuration shown in Figure 2 where the links between sensors and activities indicate the presence of a correlation.

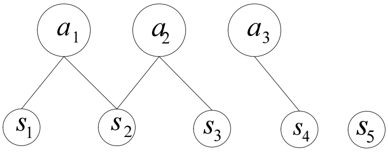


Fig. 2. Example intelligent environment

Analyzing empirical readings from the above space, the logistic regression engine produces the following CV:

$$CV(A' = \{a_1, a_2, a_3\}) =< 1,1,1,1,0 > \tag{3}$$

Obviously, the CV fails to reflect the exact dependencies shown in Figure 2. This can potentially result in inaccuracies when disambiguating activities. For example, suppose an intelligent system wants to disambiguate two particular activities $\{a_1, a_2\}$, relying on the above CV includes s_4 which is uncorrelated to the two activities. Clearly, this can potentially misguide the reasoner. To resolve this issue, we use binomial logistic regression to identify the CV of sensors for each activity with respect to a_ϕ (i.e. the no activity state). We give the resulting CV a special name: the *Reference Correlation Vector* (RCV). In the example shown in Figure 2, the RCVs are:

$$\begin{aligned} RCV(a_1) &= CV(A' = \{a_1, a_\phi\}) =< 1,1,0,0,0 > \\ RCV(a_2) &= CV(A' = \{a_2, a_\varphi\}) =< 0,1,1,0,0 > \\ RCV(a_3) &= CV(A' = \{a_3, a_\phi\}) =< 0,0,0,1,0 > \end{aligned} \tag{4}$$

Notice that the RCVs precisely reflect the dependencies shown in Figure 2. More importantly, the disjunction of RCVs is the CV for the union of their activities. For example, the disjunction of the above 3 equations is the correlation vector for all the activities shown in Figure 2:

$$\begin{aligned} & RCV(a_1) \vee RCV(a_2) \vee RCV(a_3) \\ & =< 1,1,0,0,0 > \vee < 0,1,1,0,0 > \vee < 0,0,0,1,0 > \\ & =< 1,1,1,1,0 > \\ & = CV(A' = \{a_1, a_2, a_3\}) \end{aligned} \tag{5}$$

Furthermore, combining RCVs of all activities is simply a matrix that represents all the *critical correlations* in an intelligent environment. We call this a *correlation matrix.* The following equation shows the general form of this matrix:

$$\begin{bmatrix} RCV(a_1) \\ RCV(a_2) \\ \vdots \\ RCV(a_n) \end{bmatrix} = \begin{bmatrix} c_1^1 & c_1^2 & \cdots & c_1^k \\ c_2^1 & c_2^2 & \cdots & c_2^k \\ \vdots & \vdots & \vdots & \vdots \\ c_n^1 & c_n^2 & \cdots & c_n^k \end{bmatrix} \tag{6}$$

$$\textit{where} \quad n = (|\,A\,| - 1))\ \textit{and}\ k = |\,S\,|\ \textit{and}\ c_j^i \in \{0,1\}\ \textit{and}\ 1 \le i \le k\ \textit{and}\ 1 \le j \le n$$

The above matrix can also be represented as a simple correlation graph. Figure 3 shows an example of a graph that correlates 4 sensors with 3 activities.

Finally, we note that a correlation matrix does not reflect the exact degree of correlation between a particular activity and its sensors. However, the degree of correlation can be roughly estimated using the reduction in the logistic R^2 of the model as a result of omitting the term of a particular sensor from the regression formula. We further elaborate on this particular issue in the results section.

$$\begin{bmatrix} RCV(a_1) \\ RCV(a_2) \\ RCV(a\ ) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

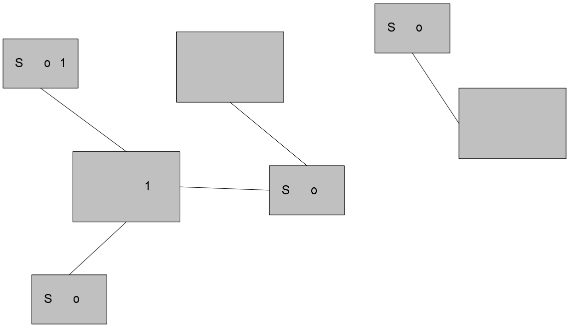


Fig. 3. Correlation matrix and correlation graph

6.2 Using the Correlation Matrix

Referring back to our discussion about an activity centric approach, we highlighted the need to identify the activities in an intelligent space. We explained that the abundance of sensors in ubicomp environments complicates activity analysis: including uncorrelated sensors or excluding correlated sensors from the decision making process can potentially mislead any reasoner. The correlation matrix (described above) serves as a filter that reflects the strong dependencies between activities and sensors in an intelligent space. Reasoners that use the correlation matrix will deal with a reduced set of sensors that are highly correlated with the activities they are trying to recognize. Clearly, this simplifies the task of a reasoner.

In situations of high uncertainty, intelligent systems fail to identify the activities with reasonable confidence. The logistic engine can be used with higher cutoff values to determine the sensors that show the highest correlation and therefore could be considered more reliable. These sensors can then be used to identify the activities. Moreover, when sensors are removed from the space, their values are replaced with zeros in the matrix. Depending on the accuracy of the classification process and the weight of the removed sensors, the system might decide to include one or more correlated sensors to compensate for the removed sensors. Similarly, when sensors are added to the space, the system gathers empirical data from the new sensors. Subsequently, activities that are frequently misclassified can be reexamined with the new sensors included for potentially improving the classification process.

7 Preliminary Results

In this section, we describe preliminary results of our approach. We use publicly available traces recorded over three days at the ACM SIGCOMM'01 conference (held at U.C. San Diego in August 2001) to demonstrate that logistic regression is effective in correlating sensors with activities. A detailed description of these traces can be found in [1]. The traces record data samples from wireless access points serving the conference. Note that due to the lack of availability of information on the no activity state (a_ϕ), we are unable to calculate the RCV and therefore we present measurements based on analysis of the CV.

Two important pieces of information can be identified: the number of mobile nodes attached to a particular access point and the load on each access point. These two quantities will serve as sensors for our experiment. In addition, two different activities can be identified including: sessions and breaks. Intuitively, we would expect a corre lation between these sensors and the activities. For example, during breaks the load over an access point is likely to drop and therefore the load sensor will show a negative correlation with breaks. Our experiments used 100 sample readings from one day to identify the logistic regression formulae. We also performed 3 experiments to classify 100 activities using the regression formula with the throughput sensor only, the number of nodes sensor only and both sensors.

7.1 Throughput and Activities

First, we characterize the strength of the correlation between the throughput at the access point and the activities in the conference hall. Figure 4 shows the proportion of the correctly classified activities using the regression formula for different cutoffs. From the figure, we see that the throughput sensor can indeed classify all the activities correctly when the cutoff is very low (i.e. we accept classifications with a broad error margin). However, its accuracy decreases rapidly as the cutoff is increased (i.e. demanding less deviation from the categories). With a 0.5 cutoff the regression formula classifies 83% of our test cases correctly.

We also found that the logistic R^2 for the regression formula is equal to 0.55. This reflects a moderate association between the throughput and the activities.

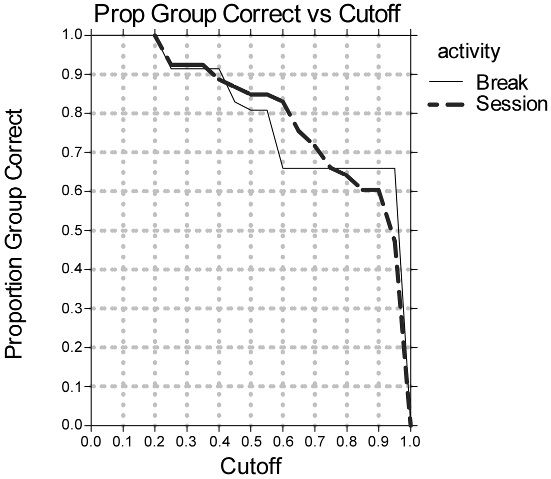


Fig. 4. Proportion of correct classification Vs. cutoff (using throughput)

7.2 Number of Nodes and Activities

Our second experiment characterizes the strength of the correlation between the number of nodes attached to the access point and the activities in the conference hall. Figure 5 shows the proportion of the correctly classified activities using the regression formula for different cutoffs. From the figure, we see that the number of nodes sensor can also classify all the activities correctly for low cutoffs. However, the sensor is more robust to higher cutoffs than the throughput sensor. In other words, its accuracy decreases more slowly than that in the throughput case as the cutoff increases. With a 0.5 cutoff the regression formula classifies 92% of our test cases correctly. We also found that the logistic R^2 for the regression formula is equal to 0.966. This reflects a strong association between the number of nodes and the activities.

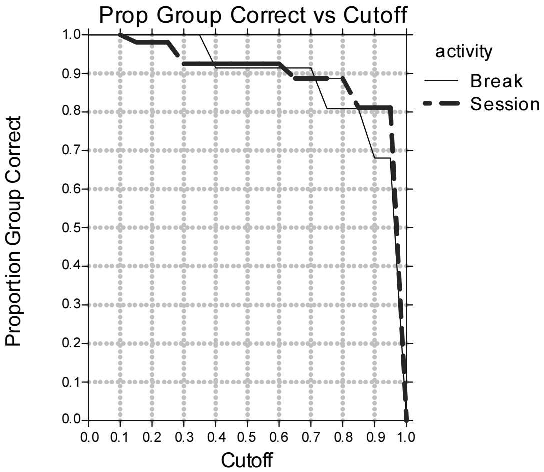


Fig. 5. Proportion of correct classification Vs. cutoff (using number of users)

7.3 Number of Nodes, Throughput, and Activities

Finally, our third experiment characterizes the strength of the correlation between both the number of nodes and their throughput, and the activities in the conference hall. Figure 6 shows the proportion of the correctly classified activities using the regression formula for different cutoffs. From the figure, we see that the regression formula can still classify all the activities correctly with low cutoffs. However, the classification percentage does not seem to improve from the one that uses the number of nodes only. With a 0.5 cutoff the regression formula classifies 92% of our test cases correctly.

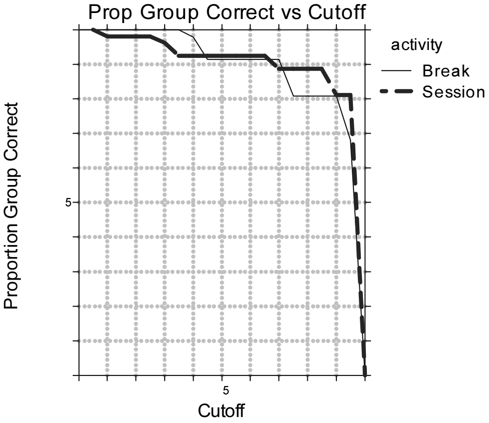


Fig. 6. Proportion of correct classification Vs. cutoff (using throughput and number of users)

We also found that the logistic R^2 for the regression formula is equal to 0.80. This means that the strength of the association between the number of nodes and the throughput, and the activities is significant.

7.4 Sensor Selection and the Correlation Matrix

When the logistic regression engine performed forward and backward selection on the (throughput and number of nodes) regression, it omitted the throughput in both cases. First, the engine measured the reduction in R^2 when omitting the throughput term. This reduced the strength of the association by 0.00172. Second, the engine measured the reduction in R^2 when omitting the number of nodes term. This resulted in a reduction of 0.24. Clearly, including the throughput sensor does not improve the strength of the association between the activities and the independents significantly. Moreover, including the throughput has not improved the classification percentage beyond 92%. Therefore, the regression engine omitted the throughput from the correlation matrix:

$$\begin{bmatrix} CV(break) \\ CV(session) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad where \quad \begin{array}{l} s_1 = numberOfUsers \\ \\ s_2 = throughput \end{array} \tag{7}$$

Finally, we note that the reduction in R^2 can be used as a rough estimate for the weights of sensors.

8 Discussion and Future Work

In this paper, we have highlighted the importance of correlating sensors with activities in an intelligent space. Our approach uses logistic regression. We described some desirable properties for activity-aware environments including: transparency and comprehensibility, adaptability, accuracy and knowledge portability. In light of these properties, we analyzed several techniques for correlating activities with sensors including: expert correlation, regression correlation and connectionist correlation. We concluded that regression provides a more comprehensible framework for correlating activities than the other approaches. We then described in detail our logistic regression approach. Finally, we reported preliminary results.

Our plan for future work is to assess our approach in the intelligent environment in our research lab. We are currently developing software components for hardware and software sensors to use them for collecting empirical data. We are also working on building a probabilistic reasoning system that will use our correlation matrix to identify activities in the intelligent space. In addition, we are developing techniques for exporting and importing contextual knowledge across intelligent environments to allow spaces to identify unfamiliar activities using imported knowledge from other spaces. Finally, we intend to deploy all these components in our research lab and to make our system accessible to a user community that can report on the impact of our system on user perceptions of activity analysis.

References

1. Balachandran, A., Voelker, G., Bahl, P., and Rangan, P.: Characterizing User Behavior and Network Performance in a Public Wireless LAN. In: Proc. ACM SIGMETRICS (2002)
2. Barthelmess, P., and Anderson, K.: View of Software Development Environments Based on Activity Theory. Computer Supported Cooperative Work. Vol. 11 Issue 1–2 (2002)
3. Brumitt, B., Meyers B., Krumm, J., Kern, A., and Shafer, S.: EasyLiving: Technologies for Intelligent Environments. In: HUC (2000)
4. Dey, A., Mankoff, J., Abowd, G. and Carter, S.: Distributed mediation of ambiguous context in aware environments. In: Proc. of UIST (2002) 121–130
5. Dix, A.: Human Issues in the Use of Pattern Recognition Techniques. Workshop on Neural Networks and Pattern Recognition in Human Computer Interaction. King's Manor, York (1991)
6. Dix, A.: Managing the Ecology of Interaction. In: Proc. of Tamodia, First International Workshop on Task Models and User Interface Design. Bucharest, Romania (2002)
7. Dix, A., Finaly, J., Abowd, G., and Beale, R.: Human-Computer Interaction. Prentice Hall (1998)
8. Elrod, S., Hall, G., Costanza, R., Dixon, M., and Des Rivieres, J.: Responsive Office Environments. Communications ACM Vol. 36, 7 (1993) 84–85
9. Engeström, Y.: Learning by Expanding: An Activity-Theoretical Approach to Developmental Research. Helsinki: Orienta-Konsultit (1987)
10. Hosmer, D., and Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons, 2nd Edition (2000)
11. Kulkarni, A.: A Reactive Behavioral System for the Intelligent Room. Master's Thesis in Computer Science and Engineering at the Massachusetts Institute of Technology. Cambridge, MA (2002)
12. Mozer, M. C.: The Neural Network House: An Environment that Adapts to its Inhabitants. In: Proc. of AAAI Spring Symposium. Menlo, Park, CA. (1998) 110-114
13. Roman, M., Ziebart, B., and Campbell, R.: Dynamic Application Composition: Customizing the Behavior of an Active Space. In: PerCom 2003, Dallas-Fort Worth, Texas (2003)
14. Weiser M.: The Computer for the Twenty-First Century. Scientific American (1991)

Methods for Online Management of AmI Capabilities Relative to Users' Goals

Ittai Flascher[1,2], Robert E. Shaw[2], Claire F. Michaels[1,2], and Oded M. Flascher[2,3]

[1] Free University Amsterdam, Faculty of Human Movement Sciences, van der Boechorststraat 9, 1081 BT Amsterdam, The Netherlands
{i.flascher,C_F_Michaels}@fbw.vu.nl
http://www.fbw.vu.nl/index.htm

[2] University of Connecticut, The Einstein Institute, 843 Bolton Road, Unit 1182, Storrs, CT 06269, U.S.A.
roberteshaw@aol.com
http://www.ia.uconn.edu/einstein.html

[3] Visteon Corporation 17000 Rotunda Drive, Dearborn, MI, 48120, U.S.A.
oflasche@visteon.com
http://www.visteon.com

Abstract. Managing the various capabilities of computing environments to best support users' goals has proven a difficult problem in transportation systems and hand-held devices. In the case of mobile users, the goal of safe and efficient navigation is a persistent part of the users' context and therefore in the online decisions on what information and services to provide. We present a feasibility test of general methods for measuring and predicting actors' goal-directed performance, and outline their use in effecting decisions with regard to initiating and halting interactions with users, anticipation of users' needs, and the evaluation of Ambient Intelligence designs.

1 Introduction

A current challenge to the development of the envisioned seamless integration of users' computational and ecological (physical) environments is to manage all the available capabilities of Ambient Intelligence (AmI) to meet the demanding physical goals. In cars, aircrafts, and hand-held computing devices the need to manage the flux of available information and the use of communications and infotainment during operations has been most pressing [4],[5]. Several information management systems have been recently developed, and some related products are expected to reach the market in a few years [9],[17].

Underlying problems such as deciding when to initiate and halt interaction with the user, anticipating users' actions, and evaluating the ability of designs to support goal-directed actions still persist however, and are the current subject of intense research and development efforts [2].

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 334–348, 2003.

Further challenges arise in the need to achieve AmI for mobile users across tasks and domains (e.g., home, workplace, a car). Current methods of information management are therefore required to become increasingly general to handle a spectrum of goals that may arise in users' everyday lives, as well as be adaptable to each individual's changing needs[1].

In the following we briefly outline some current approaches to information management and identify where the proposed methods can be of service in overcoming the above challenges.

1.1 Current Challenges in AmI Information Management

Initiation and Cessation of AmI Activities. Many currently available approaches to online information management effect a decision on questions such as whether to pass an incoming phone call to the user while the user is walking in a busy street in the following manner. First, the automated manager assembles all the available performance data (e.g., speed of vehicles, the average reaction-time of the user, workload etc.). It then identifies the most relevant measures for the achievement of the goal (e.g., reach the office quickly and safely). Finally, it decides on how to combine those into an overall measure of performance by which the current "state" of goal achievement is determined [e.g., 8]. Such a measure forms the basis for AmI decisions on initiating, halting, and prioritizing interactions with the user.

Partial measures, however, such as task duration, number of mistakes, among many others cover different aspects (dimensions) of performance and require a method to reconcile time, cognitive load, force, and number as components of goal achievement. Methods that can robustly quantify the connection between users' dynamics and constraints of different goals are needed.

Evaluation of AmI Designs. Currently available approaches attempt to establish the value of a given design through both subjective and cognitive measures (e.g., "expert" evaluation, users questionnaires), as well as by using partial performance measures. The proliferation of measures, however, opens door for a given design to be shown as better than a competitor by some of the measures and worse by others. Therefore, an overall measure of goal-relevant performance is additionally required by designers of AmI.

Anticipation of Goal-Relevant Performance. Prediction of users' performance on partial performance measures can be of great value. For example, predicting the effect that a phone ring will have on a given walker's deviation from a straight line can

[1] Much research is currently dedicated to developing interfaces that allow users to explicitly specify their goals to the AmI environment more easily [e.g., 15]. Other efforts are dedicated to developing automated AmI systems that can infer users' goals from their gestures and expressions [e.g., 5]. In this presentation we assume a goal has been specified and tackle the problem of managing AmI capabilities to support users in its achievement.

serve in the decision of whether to signal the user while the user is about to cross a street. Goals, such as reaching the office as quickly as possible, however, require a more general prediction. Predictions of whole dynamical paths are necessary, for example, to reschedule an incoming call to a later time or situation that would be optimal from the perspective of achieving the user's goal.

In the following we outline the proposed methods and measures, demonstrate their potential utility in solving the discussed problems in AmI implementation, and finally, we present empirical results from a recent experiment testing the feasibility of the framework.

2 Methods of Intentional Dynamics

2.1 Rationale

In the approach we take, goals are treated as a set of constraints on the outcome and/or process of performing tasks. For example, a person in the bedroom intending to go to the kitchen faces the task of walking (i.e., transportation). Getting to the kitchen as quickly as possible (i.e., minimal duration) sets a constraint on the task of walking and the observed dynamics will reflect that change to some degree. If a method can be found by which to quantify the influence (coupling) goal-constraints have on the dynamics, it will open three possibilities: First, a measure sensitive to every constraint on the dynamics should have properties by which determine when the actor is meeting the goal in all respects. Second, by the same argument the measure might also open the way to rank goal-performance for complete processes. Third, such an encompassing performance measure may reveal predictable regularities in goal-constrained dynamics where partial ones could not.

2.2 Methods[2]

To make the calculations explicit, we present an experiment where four volunteers controlled the motion of a graphically displayed sphere with a force-feedback joystick (i.e., a forcestick) [11].

The goal of the task (conveyed verbally) was to bring the controlled sphere to coincide with the target sphere in the shortest duration they can manage on every trial. Participants pressed a trigger to "release" the sphere and terminate the trial. Each participant performed ten sessions (in separate meetings) each consisting of five hundred trials. Data from a trial is termed a path. It consists of a sequence of samples of the forcestick's handle positions along two linear dimensions in time,

[2] In the following we give but a brief and intuitive outline of the methods. The reader is referred to [10] for further details on the approach. For further details on Simulated Annealing and related Monte Carlo techniques see e.g., [1,16].

$[\mathbf{q}] = \{\mathbf{q}_0, \ldots, \mathbf{q}_c, \ldots, \mathbf{q}_f\}$, where $0 \le c \le f$ stands for the current time-slice and f the final one in a completed path.

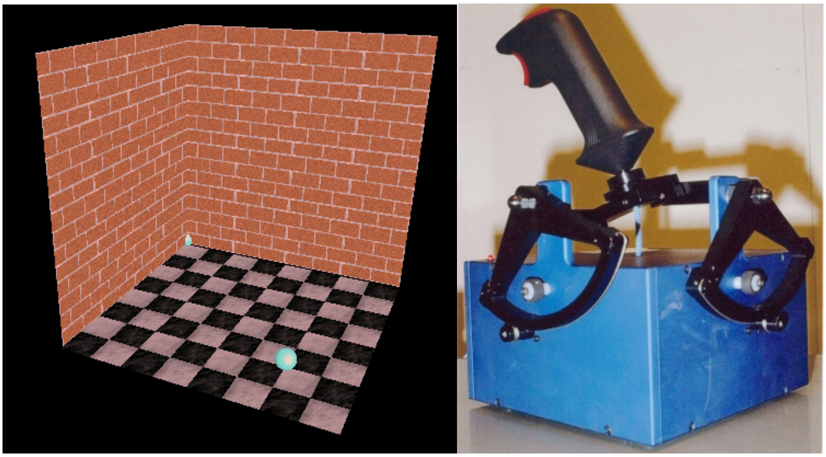


Fig. 1. The experimental setup. The graphics display (*left*), and Immersion's Impulse Engine 2000 force-feedback joystick (*right*)

For the purpose of presentation, the task we chose is simple relative to activities found in practice. Nevertheless, the task does involve all the basic components of the general transportation problem and the methods we present are applicable to the more complex cases.

Initiation and Cessation of AmI Activities. To support users' goal achievement, AmI requires a measure by which to determine online whether users' actions do not fulfill all the goal constraints and therefore assistance should be initiated. The same measure should also indicate when performance is adequate so that assistance could be properly terminated. In the following we outline the two computational stages by which the required indicator is achieved.

1. Path identification

The first step in the determination of whether a participant's performance satisfies the goal of the task is to identify at each time-slice (sample) all the paths that can reach the target and satisfy all the constraints. To achieve that we use a *Simulated Annealing* algorithm, which is a general method of sampling high-dimensional spaces [12]. The method is particularly advantageous in identifying paths when the number of constraints is large, and can do so for constraints originating from essentially any task.

The algorithm starts its search for goal-paths by constructing a random path. It then constructs a new path from the old one by randomly changing (perturbing) the old positions on some of the time-slices. This first out of two stages of the algorithm is termed the *generating* stage. Any known user or environmental constraints on path creation are treated here. For example, if the average maximal speed for the actor is known through any means, the algorithm directs its search to regions of the path space that include realistic paths for the actors' current capabilities.

In the *acceptance* stage, the algorithm selects which one of the paths will be recorded in a frequency table of goal-paths according to the following rule: if the new path is of shorter duration than the old, as the goal required, then the new path's counter is increased by one. If the new path is of longer duration it is not always rejected, rather, it is put to the following test:

$$\texttt{If,}\quad \exp\left(\frac{-t_{old}+t_{new}}{\sigma_t}\right) > rand[0,1)\texttt{, then accept.}$$

The test is known as the *Metropolis criterion* [14], and in the case of our goal says: if the new path is not "much longer" in time than the old one then it will still be accepted. "Much longer" is quantified in the denominator by the standard deviation of observed trial duration for that actor, and the relative probability of acceptance is given by the negative exponent of the ratio. If the new path is still rejected the old path is recorded in the frequency table.

When set up properly, the algorithm approximates to a high degree the globally minimal (goal) path distribution within a few thousand samples. Correspondingly, the frequency table is divided by the number of samples yielding the probability distributions of goal-paths scaled to the known constraints on the user and the environment.

This stage of the analysis provides a set of paths in space and time that satisfy the goal to different degrees and may be executed by the user. In step 2. we develop a measure that indicates which path the user is actually following and ranks the feasible alternatives that AmI may promote.

2. Dynamics quantification

As all the goal and task constraints are implicit in the simulated paths, we need to quantify the dynamics that will be involved in producing such goal paths from the current time-slice. To get an intuition into what "quantifying dynamics" means in this context we first recall that the goal was to reach the target in minimal duration. We would like to consider all the forces and energy that might lead to the violation of the goal-constraints. For example, if the sphere controlled by the actor does not travel along the shortest distance to the target it will not be reached in minimal time. Similarly, if the sphere changes directions or speed unnecessarily, travel time will be prolonged. The same applies to any other goal (e.g., transport with minimal accelerations). We need to quantify all those relevant dimensions into a single scalar.

To achieve that, we sum three *action* terms along the mean simulated path $\left[\mathbf{q}^*\right]=\left\{\mathbf{q}_c,\ldots,\mathbf{q}_{f^*}\right\}$, from the current time-slice to the final time-slice when the sphere would have reached the target[3],

$$S^*[c]=\sum_{t=c+1}^{t=f^*}\left[\frac{m}{2}\cdot\left\|\dot{\mathbf{q}}_t\right\|^2\right]\cdot\Delta t+\sum_{t=c+1}^{t=f^*}\left[F_t^{\perp}\cdot\left\|\mathbf{q}_t\right\|\right]\cdot\Delta t+\sum_{t=c+1}^{t=f^*}\left[F_t^{\|}\cdot\left\|\mathbf{q}_t\right\|\right]\cdot\Delta t. \tag{1}$$

Action is the highest level dynamical variable and can be computed in several ways. The first term on the right side of the equation computes the action associated with the kinetic energy of the sphere's motion (where $\left\|\dot{\mathbf{q}}_t\right\|$ is the speed). The second term computes the action associated with the component of the resultant force signifying changes to the direction of motion (where $F_t^{\perp}$ is the Normal component of the force, and $\left\|\mathbf{q}_t\right\|$ is the distance along the path). The third quantifies the action arising from the changes to speed along the path. We term the sum of these components *prospective* action.

Similarly, at each time-slice we compute the amount of action already exerted up to the current one along the path traveled:

$$S[c]=\sum_{t=0}^{t=c}\left[\frac{m}{2}\cdot\left\|\dot{\mathbf{q}}_t\right\|^2\right]\cdot\Delta t+\sum_{t=0}^{t=c}\left[F_t^{\perp}\cdot\left\|\mathbf{q}_t\right\|\right]\cdot\Delta t+\sum_{t=0}^{t=c}\left[F_t^{\|}\cdot\left\|\mathbf{q}_t\right\|\right]\cdot\Delta t. \tag{2}$$

We finally arrive at the quantity we were after by summing the *prospective* and *retrospective* actions into *generalized action* (GA) at time-slice *c*:

$$\tilde{S}[c]=S^*[c]+S[c]. \tag{3}$$

It is the constructed property of this quantity that serves as an indicator of the state of performance relative to the goal. More explicitly, in constructing the simulated paths we have essentially set up a new tracking task at each time-slice. Therefore, as long as the tracking is precise from one time-slice to the next the retrospective and prospective actions are complements and leave the values of GA invariant. In other words, if the actions of the user do not bring about the necessary dynamics required to satisfy the goal, invariance is not maintained. Such an indicator can therefore be used by AmI to determine when interventions are needed and when they are no longer required.

[3] In general, there is a fourth term associated with a change in the orientation of the sphere, (i.e., a rotation around an axis). In our setup, the forcestick does not allow that degree of freedom.

For presentation purposes we set the target at $x = y = 40$mm, and clamp the maximal speed $\dot{q}_{\max}$ of the forcestick at 4mm/hundredth of a second along each axis (i.e., $\dot{q}_{\max} \approx 5.66$ in the Euclidean sense). By further setting the mass of the simulated object at 2 units, we get GA constant at 416.0.

Table 1. Action values for each time-slice in a goal-path

Time-slice	*x,y*	GA	Retro	Prosp
0	0,0	416.0	0.00	416.0
1	4,4	416.0	128.0	288.0
2	8,8	416.0	160.0	256.0
3	12,12	416.0	192.0	224.0
4	16,16	416.0	224.0	192.0
5	20,20	416.0	256.0	160.0
6	24,24	416.0	288.0	128.0
7	28,28	416.0	320.0	96.0
8	32,32	416.0	352.0	64.0
9	36,36	416.0	384.0	32.0
10	40,40	416.0	416.0	0.0
	TGA	**4,576**		

Table 2. Constraint violation and the breakdown of invariance

Time-slice	*x,y*	GA	Retro	Prosp
0	0,0	416.0	0.0	416.0
1	4,4	416.0	128.0	288.0
2	8,8	416.0	160.0	256.0
3	12,12	416.0	192.0	224.0
4	16,16	416.0	224.0	192.0
5	20,20	416.0	256.0	160.0
6	24,24	416.0	288.0	128.0
7	28,28	416.0	320.0	96.0
8	32,32	416.0	352.0	64.0
9	35,35	400.0	364.0	36.0
10	39,39	400.0	404.0	4.0
11	40,40	400.0	400.0	0.0
	TGA	**4,944**		

Invariance disappears for any violation such as the slowdown occurring on time-slice 9 in Table 2 below. Invariance returns as all the constraints are met and the motion is at maximal speed in the direction of the target from the tenth time-slice.

As we can see, this invariance of GA under goal-directed dynamics gives the sought after indicator to determine whether participants satisfy their goal while in the process.

Evaluation of AmI Designs. The sum of GA, Total-Generalized-Action (TGA) for a complete path *j*,

$$\widehat{S}_j = \sum_{c=0}^{c=f^*} \widetilde{S}[c], \tag{4}$$

is minimal for goal-paths as can be seen from the previous tables and the next one.

Table 3. TGA calculations[4]

Time-slice	*x,y*	GA	Retro	Prosp
0	0,0	416.00	0.00	416.00
1	4,3	398.66	100.00	298.66
2	6,5	386.85	104.13	282.73
3	10,7	421.46	144.29	277.17
4	12,11	438.67	189.59	249.08
5	16,15	438.67	248.80	189.87
6	18,18	443.74	259.52	184.22
7	21,22	445.79	294.27	151.52
8	25,24	482.30	330.29	152.01
9	29,27	470.77	366.91	103.86
10	33,31	470.77	411.70	59.07
11	37,35	470.77	443.70	27.07
$f^* - 1$	40,39	470.77	472.50	1.74
f^*	40,40	470.77	470.77	0.00
	TGA	**6,226**		

Therefore, TGA is an overall measure of success in satisfying the goal for a complete process. As the constraints of the goal are violated TGA increases and can therefore serve to rank the observed paths according to their merit in goal achievement. Comparing the effectiveness of competing AmI designs in facilitating users goal-relevant behaviour can therefore be carried out through the TGA measure.

[4] In case the sphere is triggered away from the target the simulation completes the path; hence f^* instead of f .

Anticipation of Goal-Relevant Performance. Predictions of goal-directed behaviour may arise from knowledge of human cognitive and physical abilities and constraints. In the following we present a test of feasibility for a complementary approach to the human-centered approach. In the goal-centered approach we search for a principle capturing the predictable regularities in actors' task dynamics under goal-constraints. More specifically, we would like to formulate a principle that predicts for any given actor the probability distribution of TGA (i.e., the relative frequency of occurrence in repeated experiments). We formulate a "least-TGA principle" given by the Boltzmann distribution:

$$\Pr\left\{\widehat{S}_j = \widehat{s}\right\} \equiv p_j = \frac{1}{Z(\omega)} \exp\left(-\frac{\widehat{S}_j}{\omega}\right), \tag{5}$$

where,

$$Z(\omega) = \sum_j \exp\left(-\frac{\widehat{S}_j}{\omega}\right), \tag{6}$$

and,

$$\omega \equiv \sigma_S \cdot T \ . \tag{7}$$

In words, the probability of observing a path (in repeated experiments) is an exponential function of the (negative) value of TGA (i.e., $\widehat{s}$); the higher the value of TGA for a path, the less likely the path is.

In addition to TGA, there are two quantities in the denominator affecting the distribution of TGA.

$\sigma_S = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(s_j[f] - \left\langle s_j[f]\right\rangle\right)^2}$, where $\langle \ \rangle$ signifies the mean (average), is the standard deviation of (retrospective) action distribution observed for a participant up to the time of prediction (N trials)[5].

The standard deviation of the observed action distribution serves as the standard unit by which we measure TGA, and $0 < T < \infty$ is a multiplying factor which estimates the magnitude of the effect the goal-constraints had on the actor's performance (i.e., goal-coupling strength). More explicitly, when $T \to \infty$, the effect of the minimization principle disappears and the resulting distribution is the consequence of a random-walk under the constraints of the generating stage of the algorithm alone. As $T \to 0$, the shape of the distribution shifts towards the exponential of the Boltzmann distribution as can be seen by comparing the next two figures. Improving skill quantified by the generating parameters (e.g., maximal observed speed) may allow a user to produce paths with smaller TGA values. The observed change in the distribution due

[5] In the implementation presented we measure the action after each trial (i.e., at the final time-slice f).

to that is a displacement towards the origin. T on the other hand, influences the relative frequencies with which paths with smaller values are performed. It is that type of improvement that is associated with developing goal-relevant expertise and gauged by T .

An observed TGA distribution of an experimental session (500 trials) is plotted in red. Given the large value of T , the predicted distribution in blue gives the random-walk result.

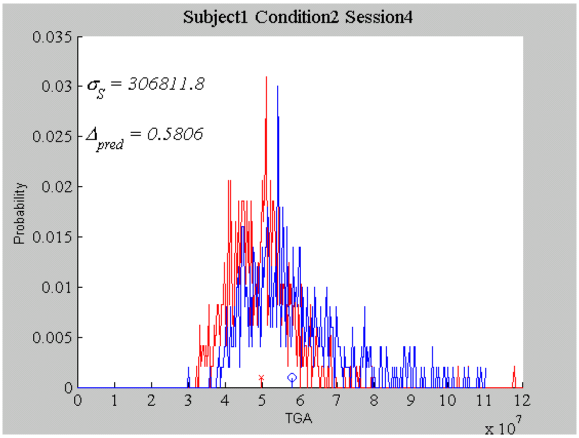


Fig. 2. $T \cong 5{,}000$

At lower values, the predicted distribution changes its shape towards the exponential.

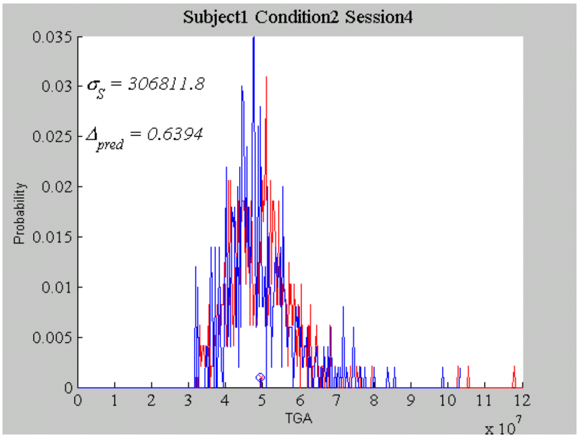


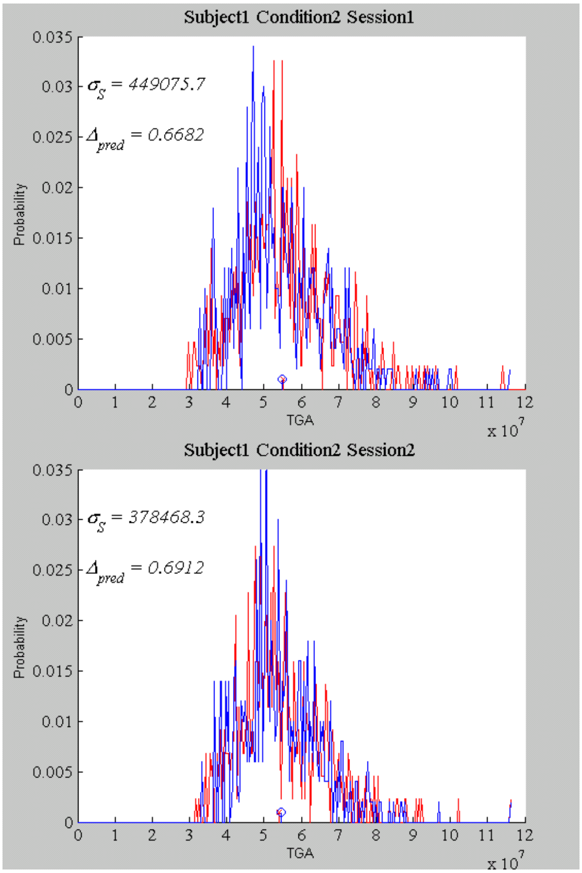
Fig. 3. $T \cong 500$

T is therefore estimated by a separate Simulated Annealing algorithm searching for a value which brings the *means* of the distributions to coincide.

The feasibility test of the above principle should boost confidence in the approach by showing that the least-TGA principle both predicts the data of every participant to a large extent, as well as by demonstrating that participants converge to the Boltzmann equilibrium distribution of that measure as the number of session increases.

1. *First experimental hypothesis*, $H_1^{(1)}: \quad \Delta >> 0$.

Using only three generating parameters of maximal speed observed, mean reaction-time to trial onset, and the standard deviation of final (Euclidean) distance from the target, close to 60% of the total distribution of every session of every participant in the study were predicted. Due to space constraints we show only a sequence of a few sessions of one participant in the following figures.



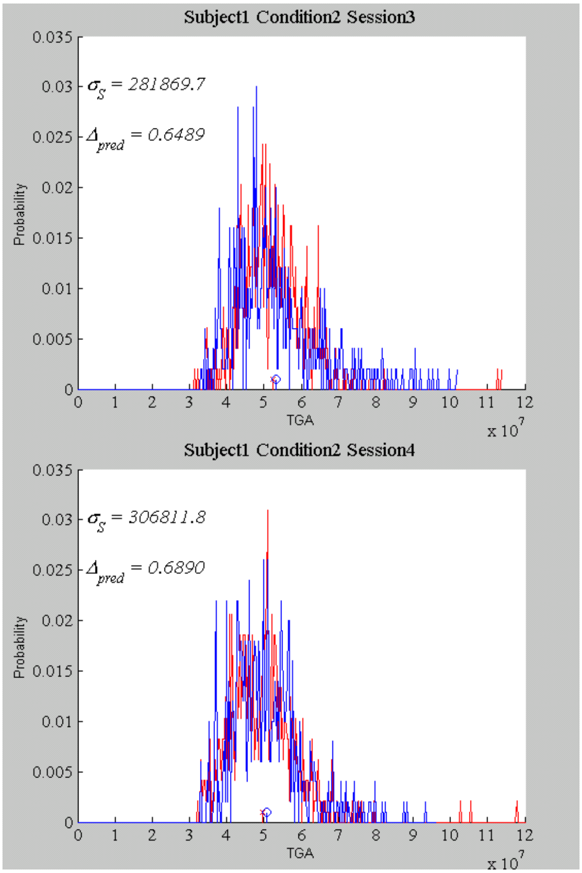


Fig. 4. A sequence of predicted sessions

We use the percentage similarity measure, also known as *Weitzman's Δ* (delta) measure [6], to quantify the level of prediction. As its name suggests, the measure yields the percentage of the observed distribution intersected (overlapped) by the predicted one,

$$\Delta = \sum_i \min\left(p_i^{obs}, p_i^{pred}\right), \tag{8}$$

where *i* are the values of TGA for which both the observed and predicted distributions have probability larger than zero.

As a first approximation, the approach seems promising, predicting most of every session in the study. Much improvement can be expected with the introduction of more elaborate path construction schemes and additional constraining parameters.

2. *Second experimental hypothesis*, $H_1^{(2)}: \underset{N\to\infty}{D} \to 0$.

When researchers use the simulated annealing algorithm as an integration or simulation tool they try to make sure the algorithm converges to the globally optimal (equilibrium) distribution they seek. One of the components in achieving that result is to construct paths at the generating stage which are independent of one another. The (Metropolis) acceptance criterion can then make sure that the path sequence (chain) converges to the optimal distribution. In our use of the algorithm as the model for users' *learning* process, the paths are strongly dependent and convergence is not guaranteed. Participants in our case are in charge of constraining their control to meet the goal requirements as much as they can, given their *skill*-level measured by the generating parameters (e.g., maximal observed speed).

The second experimental hypothesis is therefore that participants will show a decreasing distance from the predicted equilibrium distribution (i.e., the optimal distribution) as the number of trials (sessions) is increased. In the next figure we compare the red observed distribution with the black equilibrium distribution generated (sampled) at the same level of standard deviation $\sigma_{\hat{S}}$ (i.e., bin-size along the horizontal axis). The latter distribution signifies the best goal-relevant performance this participant may achieve given his/her current sensitivity to the goal-constraints.

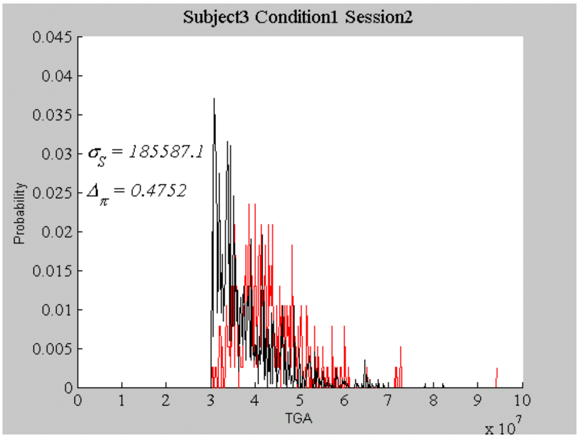


Fig. 5. The observed and the equilibrium distributions

To test whether participants' distributions show convergence to the Boltzmann equilibrium distribution specified by our principle, we measure for each of them the (variation) distance of the observed distribution from the $\pi(\sigma_{\hat{S}})$ distribution at each session [e.g., 7]:

$$D = \frac{1}{2}\sum_i \left|\pi_i(\sigma_{\hat{S}}) - p_i^{obs}\right| \ . \tag{9}$$

Distance values between the distributions were fitted with a least-square line. The slope of that line was tested through a reshuffling technique [13] to statistically determine whether it is significantly different from zero. As can be seen in the figures below, convergence of modest rate was detected for three out of the four participants.

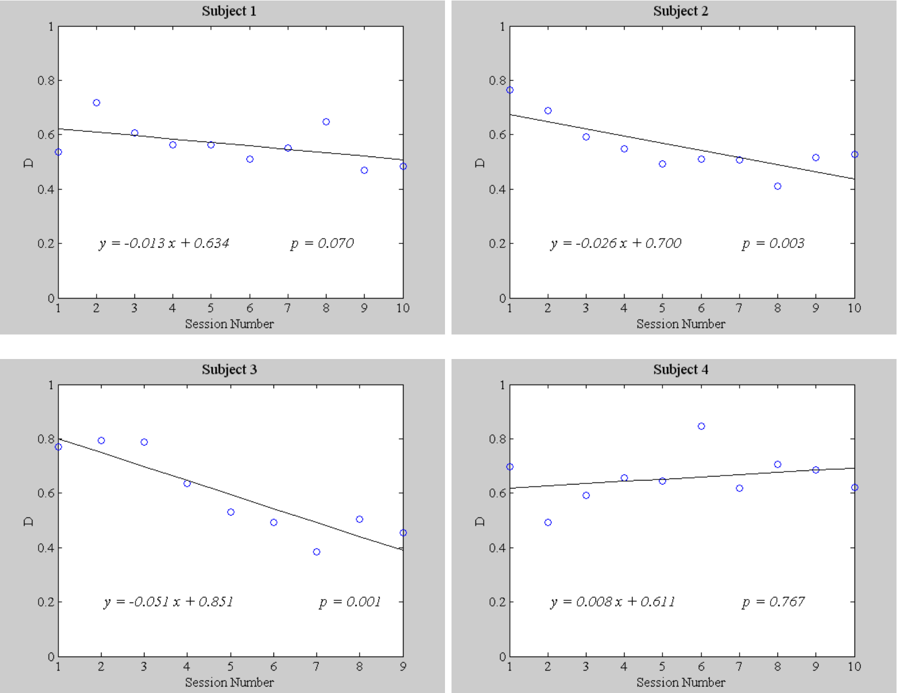


Fig. 6. Convergence assessment

These preliminary results are quite promising in showing good data fits and the possibility of convergence. Of course, further work is required and testing the methods' performance on more realistically complex tasks is necessary. However, the evidence is quite strong that the approach is feasible and that further work is warranted.

3 Summary of Contributions

Methods were outlined for the solution of three general problems currently impeding the development of management systems of AmI. We outlined the methods' execution of decisions on the initiation and cessation of AmI interventions, overall design assessment through the evaluation of the level of goal achievement by the TGA measure, and the prediction of goal-directed performance. The methods and measures demonstrated are generally applicable and are particularly suited for adapting to different users' changing capabilities and goals. If shown valid for the complex tasks found in practice, the offered methods promise to facilitate the solution to some of most hindering problems in AmI's development.

References

1. Aarts, E. H. L., and Korst, J. H. M. Simulated Annealing and Boltzmann machines. John Wiley & Sons. New York. (1989)
2. Brookhuis, K. A., de Waard, D., and Fairclough, S. H. Criteria for driver impairment. *Ergonomics, 46,* 5 (2003), 433–445
3. Camurri, A., Lagerlöf, I., and Volpe, G. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *Int. J. Human-Computer Studies, 59,* (2003), 213–225
4. Chávez, E., Ide, R., and Kirste, T. Interactive applications of personal situation-aware assistants. *Computers & Graphics, 23,* (1999), 903–915
5. Chen, F-S., Fu, C-M., and Huang, C-L. Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing,* (In press)
6. Clemons, T. E., & Bradley Jr., E. L. A. nonparametric measure of the overlapping coefficient. *Computational Statistics & Data Analysis*, *34* (2000). 51–61
7. Denuit, M., and Bellegem, S. van. On the stop-loss and total variation distances between random sums. *Statistics & Probability Letters, 53* (2001). 153–165
8. de Waard, D., Hernández-Gress, N. and Brookhuis, K. A. The feasibility of detecting phone-use related driver distraction. *Int. J. Vehicle Design, 26*, 1 (2001). 85–95
9. EVAID. Trademark of QinetiQ Limited, Farnborough, Hants, GU14 0LX U.K. (2003). http://webdb4.patent.gov.uk/tm/number?detailsrequested=C&trademark=2299745
10. Flascher, I. Goal-centered approach to the measurement of human-systems performance. Ph.D. dissertation. University of Connecticut. (In press)
11. Impulse Engine 2000, Software Development Kit, release 4.2, (January, 2001). Immersion Corp.
12. Laarhoven, P. J. M. van, Aarts, E. H. L. Simulated Annealing: Theory and applications. Reidel, Dordrecht. (1987)
13. Lunneborg, C. E. Data analysis by resampling: Concepts and applications. Duxbury Press, USA. (2000)
14. Metropolis, M., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, *21* (1953), 1087–1092
15. Prekop, P., and Burnett, M. Activities, context and ubiquitous computing. *Computer Communications 26,* (2003), 1168–1176.
16. Robert, C. P., and Casella, G. Monte Carlo statistical methods. Springer-Verlag, New York. (1999)
17. Wood, C., Leivian, R., Massey, N., Bieker, J., and Summers, J. Driver Advocate^{TM} Tool. In *Proceedings of Driving Assessment 2001: International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Aspen, Colorado, August 14–17, 2001*

Interaction Design for the Disappearing Computer

Norbert Streitz

AMBIENTE – Workspaces of the Future
Fraunhofer IPSI
64293 Darmstadt Germany
streitz@ipsi.fraunhofer.de
http://www.ipsi.fraunhofer.de/ambiente

Abstract. This invited talk starts out with a review of the previously developed Roomware® concept and sample prototypes as an approach for designing new forms of interaction and collaboration in future work environments. This is followed by presenting the EU-funded proactive initiative "The Disappearing Computer" (DC), a cluster of 17 related projects designing new people-friendly environments in which the "computer-as-we-know-it" has no role. Finally, a specific example of the DC-initiative is presented, the project "Ambient Agoras: Dynamic Information Clouds in a Hybrid World". It aims at transforming places into social marketplaces ('agoras') of ideas and information, providing situated services and feeling of the place ('genius loci') by creating new social architectural spaces. This is achieved by developing combinations of ambient displays and mobile devices that require and provide new forms of natural and intuitive interaction.

1 Roomware®: Beyond Desktop PCs and WIMP Interaction

The introduction of information technology has caused a shift away from the real objects we were and still are used to in our physical environment as the sources of information towards desktop computers as the interfaces to information that is now (re)presented in a digital format in virtual environments. Associated with this shift is a style of interacting with information that is known as WIMP – windows, icons, mouse (or menus), and pointers; and, of course, not to forget the keyboard. In this paper, I argue for returning to the real world as the starting point for designing future information and collaboration environments. The goal is to design environments that exploit the affordances provided by real objects but at the same make use of the potential of computer-based support that is available via the virtual world. Taking the best of both worlds requires an integration of real and virtual worlds resulting in hybrid worlds.

Since 1997, we have developed a new approach for the design of work environments where the "world around us" is the interface to information and for the cooperation of people [4]. In this approach, the computer as a device disappears and is almost "invisible" (see also "The Disappearing Computer" in the next section) but its functionality is ubiquitously available via new forms of interacting with information. The environments consist of what we call "roomware" components. We define

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 351–355, 2003.

Roomware® [6] as the result of integrating information and communication technology in room elements such as doors, walls, and furniture [www.roomware.de]. Thus, the roomware approach moves beyond the limits of standard desktop environments on several dimensions.

At the beginning of these efforts, we designed and built a testbed called i-LAND [4],[5] with a range of different roomware components as, e.g., the DynaWall, the InteracTable, the ConnecTables, and the CommChairs. There are two generations of Roomware components; the second one was developed in the context of the R&D consortium “Future Office Dynamics”. We also developed the dedicated software infrastructure BEACH [6],[10] and applications on top of it as, for example, MagNets, BeachMap, PalmBeach [3], and the Passage mechanism [1] in order to exploit the full potential of roomware. Furthermore, we used non-speech audio in order to provide sound augmentation for the different types of interaction and collaboration [2].

The *DynaWall* is an interactive electronic wall, representing a touch-sensitive vertical information display and interaction device that is 4.50 m wide and 1.10 m high. The availability of sufficient display space enables teams to display and to interact with large information structures collaboratively in new ways. Two or more persons can either work individually in parallel or they share the entire display space. The size of the DynaWall provides challenges as well as opportunities for new forms of human-computer interaction that are provided by the BEACH software.

The *InteracTable* is an interactive table for informal group discussion and planned cooperation. It is 90 cm high with a display size of 63 cm x 110 cm. The horizontal workspace is realized with a touch-sensitive plasma-display that is integrated into the tabletop. People can use pens and fingers for gesture-based interaction with information objects. Using BEACH, they can create and annotate information objects that can also be shuffled and rotated to accommodate different view orientations around the table.

The *CommChair* combines the mobility and comfort of armchairs with the functionality of a pen-based computer. It has an independent power supply and is connected to all other roomware components via a wireless network. The BEACH software provides a private workspace for personal notes and a public workspace that allows moving them to other roomware components, for example to the DynaWall. Using the CommChair, one can interact remotely with all objects displayed on the DynaWall.

The *ConnecTable* is a modular version of the CommChair and can be used in different positions: either sitting in front of it on a regular chair or using it in a stand-up position as a high desk. Its particular name, ConnecTable, results from the functionality that its workspace area can be easily extended by "connecting" several ConnecTables [9]. The coupling of the individual displays resulting in a common shared workspace is achieved by simply moving the ConnecTables together in physical space which is detected by sensors. No additional login or typing of IP addresses is needed.

The *"Passage"* mechanism [1] provides an intuitive way for the physical transportation of virtual information structures using arbitrary physical objects, so called

"Passengers". The assignment is done via a simple gesture moving the information object to (and for retrieval from) the "virtual" part of the so called "Bridge" that is activated by placing the Passenger object on the physical part of the Bridge. No electronic tagging is needed. Passengers can be viewed as "physical bookmarks" into the virtual world.

The cooperative hypermedia environment *BEACH* [6],[10] provides new, intuitive forms of human-computer interaction based on using only fingers and pens and new ways of cooperative sharing for multiple device interaction. It provides a modeless user-interface allowing to scribble and to gesture (for commands) without having to switch modes. The incremental gesture recognition detects the type of input and provides feedback via different colors or sounds.

The context of these development is our more comprehensive notion of so called Cooperative Buildings that we introduced some time ago [4]. We used the term “building” (and not “spaces”) in order to emphasize that the starting point of the design should be the real, architectural environment. By calling it a “cooperative” building, we wanted to indicate that the building serves the purpose of cooperation and communication. At the same time, it is also “cooperative” towards its users, inhabitants, and visitors by employing active, attentive and adaptive components. This is to say that the building does not only provide facilities but it can also (re)act “on its own” after having identified certain conditions. It is part of our vision that it will adapt to changing situations and provide context-aware information and services. Our roomware components are examples of the major constituents of these cooperative buildings.

2 The Disappearing Computer

"The Disappearing Computer" (DC) [www.disappearing-computer.net] is an EU-funded proactive initiative of the Future and Emerging Technologies (FET) activity of the Information Society Technologies (IST) research program. The goal of the DC-initiative is to explore how everyday life can be supported and enhanced through the use of collections of interacting smart artefacts. Together, these artefacts will form new people-friendly environments in which the “computer-as-we-know-it” has no role. There are three main objectives:

1. Developing new tools and methods for the embedding of computation in everyday objects so as to create artefacts.
2. Research on how new functionality and new use can emerge from collections of interacting artefacts.
3. Ensuring that people's experience of these environments is both coherent and engaging in space and time.

These objectives are addressed with a cluster of 17 related projects under the umbrella theme of the DC-initiative. The cluster is complemented by a variety of support activities provided by the DC-Network and coordinated by the DC Steering Group, an elected representation of all projects.

3 Ambient Agoras

"Ambient Agoras: Dynamic Information Clouds in a Hybrid World" is a project of the “Disappearing Computer" initiative [www.ambient-agoras.org]. It addresses the office environment as an integrated organisation located in a physical environment and having particular information needs both at the collective level of the organisation, and at the personal level of the worker.

This project promotes an approach of designing interactions in physical environments using augmented (smart) physical artefacts and corresponding software to support collaboration, social awareness, and to enhance the quality of life in the working environment. Ambient Agoras combines a set of interaction design objectives (mental disappearance of computing devices, communicating awareness and atmospheres) with sensing technologies, smart artefacts (walls, tables, ambient displays, and mobile devices) and the functionality of artefacts working together. "Ambient Agoras" aims at transforming places into social marketplaces of ideas and information ('agoras') and provides situated services, place-relevant information, and feeling of the place ('genius loci'). It adds a layer of information-based services to the place and provides the environment with 'memory' accessible to users. This is achieved by providing better affordances and information processing to existing places and objects. In this way, it aims at creating a social architectural space [8] that facilitates novel interactions and experiences by augmenting existing architectural spaces.

We designed and realized a set of different smart artefacts that function not only as independent artefacts but also in combination [7],[8]. Examples are the Hello.Wall and the ViewPort. The *Hello.Wall* is a large (1.80 m x 2.00 m) ambient display with sensing technology. We use changes in light patterns for conveying information about different states of people and of the physical as well as the virtual environment in an office building as, e.g., atmospheres. The *ViewPort* is a handheld compact artefact with a pen-based interactive display and provided with sensing technology and a wireless network. It can be used as a personal, a temporarily personal or public device for creating and visualizing information. It provides also the functionality of visualizing information “transmitted” from other artefacts that do not have displays of their own and are "borrowing" this display as, e.g., the Hello.Wall. A more recent application is the combination and coupling of two Hello.Walls at distributed locations.

Acknowledgements. The work reported is supported in part by the European Commission (contract IST–2000-25134) as part of the “Disappearing Computer” initiative. Thanks are due to our cooperation partners in the different projects and to Carsten Magerkurth, Christian Müller-Tomfelde, Thorsten Prante, Carsten Röcker, Richard Stenzel, Peter Tandler, Daniel van Alphen, Daniela Plewe, members and all students of the research division AMBIENTE for their contributions in realizing the different components and environments [www.ipsi.fraunhofer.de/ambiente].

References

1. Konomi, S., Müller-Tomfelde, C., Streitz, N. (1999). Passage: Physical Transportation of Digital Information in Cooperative Buildings. In: N. Streitz, J. Siegel, V. Hartkopf, S. Konomi (Eds.), Cooperative Buildings - Integrating Information, Organizations, and Architecture. Proceedings of Second International Workshop (CoBuild'99). Springer LNCS 1670, pp. 45–54
2. Müller-Tomfelde, C., Streitz, N., Steinmetz, R. (2003). Sounds@Work - Auditory Displays for Interaction in Cooperative and Hybrid Environments. In: C. Stephanidis, J. Jacko (Eds.), Human-Computer Interaction: Theory and Practice (Part II). Lawrence Erlbaum Publishers. Mahwah, June 22–27, 2003. pp. 751–755
3. Prante, T., Magerkurth, C., Streitz, N. (2002). Developing CSCW tools for idea finding – Empirical results and implications for design. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2002) (New Orleans, USA). pp. 106–115
4. Streitz, N. Geißler, J., Holmer, T. (1998). Roomware for Cooperative Buildings: Integrated Design of Architectural Spaces and Information Spaces. In: N. Streitz, S. Konomi, H. Burkhardt, H. (Eds.), Cooperative Buildings - Integrating Information, Organization, and Architecture. Proceedings of the First International Workshop (CoBuild '98). Springer LNCS Vol. 1370, pp. 4–21
5. Streitz, N., Geißler, J., Holmer, T., Konomi, S., Müller-Tomfelde, C., Reischl, W. Rexroth, P., Seitz, P., Steinmetz, R. (1999). i-LAND: an Interactive Landscape for Creativity and Innovation. In: Proceedings of ACM Conference CHI’99 (Pittsburgh, USA). pp. 120–127
6. Streitz, N., Tandler, P., Müller-Tomfelde, C., Konomi, S. (2001). Roomware: Towards the Next Generation of Human-Computer Interaction based on an Integrated Design of Real and Virtual Worlds. In: J. Carroll (Ed.), Human-Computer Interaction in the New Millennium. Addison-Wesley, pp. 553–578
7. Streitz, N., Röcker, C., Prante, Th., Stenzel, R., van Alphen, D. (2003). Situated Interaction with Ambient Information: Facilitating Awareness and Communication in Ubiquitous Work Environments. In: D. Harris, V. Duffy, M. Smith, C. Stephanidis (Eds.), Human-Centred Computing: Cognitive, Social, and Ergonomic Aspects. New Jersey, Lawrence Erlbaum Publishers. Mahwah, June 22–27, 2003. pp. 133–137
8. Streitz, N., Prante, T., Röcker, C., van Alphen, D., Magerkurth, D., Stenzel, R., Plewe, D. (2003). Ambient Displays and Mobile Devices for the Creation of Social Architectural Spaces: Supporting informal communication and social awareness in organizations. In: K. O’Hara, M. Perry, E. Churchill, D. Russell (Eds.), Public and Situated Displays: Social and Interactional Aspects of Shared Display Technologies. Kluwer Publishers (to appear in fall 2003)
9. Tandler, P., Prante, T., Müller-Tomfelde, C., Streitz, N., Steinmetz, R. (2001). ConnecTables: Dynamic Coupling of Displays for the Flexible Creation of Shared Workspaces. In: Proceedings of the 14. Annual ACM Symposium on User Interface Software and Technology (UIST'01), ACM Press (CHI Letters 3 (2)), 2001. pp. 11–20
10. Tandler, P., Streitz, N., Prante, T. (2002). Roomware: Towards Ubiquitous Computers In: IEEE Micro (November/December 2002). pp. 36–47

Natural Language Processing and Multimedia Browsing Concrete and Potential Contributions

Dominique Dutoit[1], Yann Picand[2], Patrick de Torcy[2], and Geoffrey Roger[2]

[1] CNRS CRISCO, Memodata, 17, rue Dumont d'Urville,
14000 Caen, France
dutoit@info.unicaen.fr
[2] Memodata, 17, rue Dumont d'Urville
memodata@wanadoo.fr

Abstract. Considering an "Intelligent Multimedia Browsing at Home (MB)", what should be the main benefits of "Natural Language Processing (NLP)" technologies? In that document, MB is limited to ergonomics in which the user cannot or refuses to type a keyword to define a query with its input device (remote control, voice). Despite that, NLP can be concretely or potentially used in many areas. We describe some of these components and the main results.

1 Introduction

Often, NLP technologies are used to compare a query with a full-text index. In that application, we study a hypothesis where the user can't type any query. In that approach, the user can browse with his/her voice or his/her remote control the different available items.

If the total number of such items is limited to 50 simultaneous movies, the browsing should be fully done. But if we consider thousands and thousands downloadable movies in the TV, the impossibility of such a full browsing appears clearly.

This critical context for NLP, where a speaker cannot finally use natural language to communicate, was provided to us by a project supported by Thomson Multimedia. In this project called "Intelligent Multimedia Browsing at Home (MB)", many partners were involved; their main contributions are quoted below:
As we can see, text analysis is not the core of this Ambient Intelligence project. Though this technology seems to stand at the periphery, we will see that text analysis may be used in many peripheral areas so as to enrich global quality.

To do this, we describe some techniques where NLP should contribute to a satisfactory browsing. In a first section, we describe the concrete input data used for the MB prototype. In the second section, we detail some possible contributions of NLP to the browsing of these data. The last section details some of these techniques and results.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 356–371, 2003.

Table 1. MB prototype's partner

Partner	Country	Task
Thomson Multimedia	France	- UI specification - graphical design - database - user profile - user test
Telisma	France	- speech recognition
Epictoid	The Netherlands	- avatar
Vitec	France	- user identification
Memodata	France	- text Analysis
VTT	Finland	- data classification

2 The Data

In this section, we describe our corpus (set of computed text) and our suggestions for NLP computing.

2.1 The Corpus

As prototype leader, Thomson has selected the "Internet Movie Database (IMDb)" and obtained a large part of the database. IMDb's url is: www.imdb.com. This site presents itself: "*the IMDb is the ultimate online movie database covering over 325,000 titles and over 1,000,000 people with facts, trivia, reviews plus multimedia links from the earliest films to the latest releases.*"

In the project, we have only selected the movies database (pictures and reports). The database version contains 301.908 movies documented by 48.871 summaries. There are other kinds of data, but we chose to use only the summaries: other partners, as VTT, may compute the whole data to a Kohonen map. In that way, we voluntarily evade data such as: name of film-maker or actor (important when they are well-known as specialized in a kind of movie), movies category etc. It was also possible to work with movie title but this data has not been investigated. Finally, the following study concerns only the summaries and how to extract/reformulate significant information from them. The size of the 48.871 summaries maximally reaches 21 Mo. The average size of the summaries in terms of word units is: 71.

2.2 Our Work Schedule

In the first place, we have to say that, in NLP, we are specialized in lexical semantics; that is the study of the links between meaning and lexicon. Because the corpus contains 95% of English summaries, as we will see it, such a specialization is not advantageous: indeed we mainly work with the French language. Thus, at the beginning of this work, we decided to consecrate a large amount of manpower to overlap the English. It was also possible to use WordNet (a large semantic net from Princeton University [10]), but it was not convenient enough in terms of semantic links variety. Then, we preferred to update our conceptual dictionary called “Integral Dictionary” (ID) [8],[9].

After that preliminary decision, we began to study manually and automatically the content of the corpora to define a coherent strategy. The automatic study was mainly based on statistics and the manual work was based on human reading of samples.

Processing. This analysis concerned the distribution of words in the whole corpora: it was interesting to direct the manual lexicological task to the most common words occurring in IMDb that were absent from our dictionary. Secondly, we had to check if the IMDb corpus was roughly "neutral" or, on the other hand, specialized in a particular knowledge domain. To find an answer to this question, we compared the lexical statistics to the very large statistical corpora given by [13].

The result of this task was that:

- It was possible to select 15000 most relevant words, particularly frequent in IMDb and absent in ID. For these words, we added semantic information using the French semantic model (explained below) and morphological data (used to compute inflected forms, conjugation, plural etc.). At present, 11257 word-meanings were added to our dictionary, some results are thus incomplete.
- Concerning the specialization of the corpora, it occurred that some words appear abnormally often (in the point of view of the Binomial distributions model). Of course, some terms like *movie*, *film*, *interview* etc. are numerous, but there is also a rich vocabulary corresponding to the topical movies: e.g. LAPD, i.e. Los Angeles Police District etc. Last, we were surprised by other strange words, e.g. *la, che, una, une, le, de, der, im, nach* etc. After checking, we discovered that some summaries were not written in English.

Considering these results, we decided to accomplish the following tasks:

- Develop and add to our set of NLP APIs called "The Semiograph", a language recognition tool.
- Develop and add to our set of APIs, a part-of-speech analyzer for the English language. In English, though it's true that the inflected forms are not numerous, it's also true that a lot of words are ambiguous. Consider *see*. It can be a noun. As noun, *see* is a *bishopric* or something approaching. Consider *back*. It can also be a noun: a debt.
- Because IMDb contains a very large number of proper nouns, because these proper nouns may be very useful for classifications, we need to develop and add to our set of APIs a tool to extract and define the nature of each proper noun existing in the corpora. This tool will be able to decide if the proper noun is a title inside a summary, a place, a person, a company, an event etc.

- Because IMDb does not only contain English words and English summaries, we concluded that IMDb's users do not have the English as a mother tongue. Thus, we decided to help readers of English summaries.

Human analysis. The human analysis was not really fruitful. As it is volunteers without professional training that write IMDb, it frequently occurs that summaries do not describe the content of the movie but only give a subjective evaluation. It seems that the case occurred when the movie was famous or insignificant. We give below one example.

The 1997 "Wine" is the first attempt to realize a modern symposium , i.e. a drunken philosophical discussion about love. This attempt generated very little of interest , and the producers halted the project , indicating that a new version would be made eventually.

In that example, the kind of movie (*intellectual discussion*) and the subject (*love*) can be identified, but the latter is quite vague: the phases of the discussion, the opinions (etc.) don't appear at all.

The rate of these little-relevant summaries seems to be high. We then decided to identify these summaries.

Thomson Multimedia provided us a detailed thesaurus to organize the movies. Occasionally, it is possible for a human (using only the summary, and provided this human does not know the movie) to index one summary in the thesaurus. Because, it is generally not the case, we did not develop this application but we will show, in this paper, some possible good results computed by our APIs. Concerning this application, the last difficulty is the elapsed time of computing, as we have not optimized this API.

Let us now describe the main results. We will describe the technology when we use an original and uncommon one; when it is not so, we only give references.

3 The APIs and the Results

We describe below:

- The language recognition API
- The POS tagger for English
- The named entities
- The multilingual dictionary
- The English semantic net inside the Integral Dictionary
- The keyword extractor and the clustering of the lexicon
- The mapping to the thesaurus

3.1 The Language Recognition API

This tool is able to identify up to 70 different languages. This tool is not based on the lexicon but on a stochastic language model (n-Gram). For example, to identify an English text, the recognition of frequent sequences of characters proves efficient. For example, *_t, _th, _the, _the_, th, the, the_, he, he_, e_* are very frequent sequences in English, mainly due to some words like *the* and *he*. *The* is the most common word in English: there is 5.776.397 occurrences of *the* in an English corpora of 88.704.926 words. The algorithm knows for each language the statistics of sequences and evaluates the distance between any text and these vectors of sequences. The results are 99,9 % correct when the size of text is greater than 70 characters.

In IMDb we found the following results: *English* (46899), *Italian* (690), *Gaelic* (415), *Hungarian* (224), *Spanish* (188), *German* (97), *Norwegian* (94), *Catalan* (74), *French* (74), *Frisian* (69), *Danish* (17), *Portuguese* (13), *Swedish* (8), *Dutch* (4), *Bosnian* (1), *Romansch* (1), *Serb-ascii* (1).

Let us extract the beginning of an Italian summary:

<Summary Idfilm=”103157” language=”Italian”>Il film è il montaggio delle sequenze girate con un 16 mm. e con una videocamera da tre ragazzi nel bosco di Blair nel 1994...

The uses of this data in the prototype should be:

- Automatically select a summary with a relevant language for the user (when a movie has a summary in Norwegian, it also has a summary in English).
- Automatically select the right source and target language when the user asks for a translation.

3.2 The Part-of-Speech Tagger for the English

We have developed a complete POS tagger based on computing the evidence found in an English corpus. Thus, the technology uses mainly statistics. The algorithm was firstly developed for French to obtain a fast shallow parser that would not output important missing even if some ambiguities were unsolved.

The learning corpus was not disambiguated, so the system learned only the unambiguous sequences of POS and/or words. Because we studied IMDb, we decided to learn the sequences contained in the IMDb itself.

Table 2. Details in named entities tags

Tag	Sub Tag	Meaning
C	-	Company, institute etc.
Pe	Attrib **F** : **Form** Attrib **T** : **type** **…**	="P-N" : first name, last name ="N" : first name ="A" : actor …
Ti	Attrib F…	Date
Pl	Attrib F…	Location, place
T		Title
E		Event
D		Date

To check if the general results were correct or not, we compared them to the statistics given by [13] from the National British Corpus. The manual verification of deviation did not show big errors on our account so we concluded that results were helpful.

The IMDb's English summaries shallow parsing took 1 H 35 min with an AMD 1800 processor.

3.3 The Named Entities

Firstly, we have semi-automatically annotated a significant percentage of the IMDB summaries.

The main tags are given below :

After that, we taught the precedent algorithm these tags and their contexts. Faced with the sentence *Yesterday, G. Straussy said that ...* the system receives the following tags:

G. --> MajLetter, Point **or** *abbreviated First Name*
Straussy --> Unknown First_letter_Cap
said --> verb preterit, **often after** *a proper noun of person...*
and decides to mark out *G. Straussy* as a proper noun of person.

The main statistics of IMDb's proper noun are:

Table 3. Count of named entities in IMDb

Named entity	Count
Person	
Character	93957
Actor	14118
Title	2612
Event	1855
Date	2590
Corps list	3480
Place	24625

These results confirmed our first feeling about the richness of proper nouns in the summaries. But, also, these results are expressing that the work in that entity has to be continued. Indeed, a lot of proper nouns have several occurrences in IMDb. The following table gives the total number of entity counted only one, two, three times etc.

Table 4. Recurrence of a same named entity in IMDb

The entity is counted X times	Total number of entities
1 time	41733
2 times	4039
3 times	1314
4 to 10 times	636 to 102
11 to 30 times	114 to 21
31 to 250 times	14 to 1
More than 251	25

These results confirmed our first feeling about the richness of proper nouns in the summaries. But, also, these results are expressing that the work in that entity has to be continued. Indeed, a lot of proper nouns have several occurrences in IMDb. The following table gives the total number of entity counted only one, two, three times etc.

Named entities are similar to common words:

- Some of them have homonymous: Paris(France)/Paris(Texas)/Paris(person), Georges W. Bush (father)/ Georges W. Bush(son).
- Others have different spelling Georges W. Bush, the president Bush, G. W. Bush, WW2, world war II, World War II etc.
- The distribution is strongly concentrated. *New York, England, America, Califormia, world war II* have more than 250 occurrences
- A specific technology has to be developed to manage that phenomenon, probably in relation with the semantic lexicon to disambiguate the references (Georges. W. Bush Junior may be automatically attached to the World Trade Center, Twice Sisters, war in Iraq, war against terror etc.). We have no result concerning this more complex sub-task.

We give below a complete example of a tagged summary with named entity:

<Summary Idfilm="15979" language="English"><Pe F="P-N">Shurik Timofeev</Pe> builds a working model of a time machine. By accident, <Pe F="P-N">Ivan Bunsha</Pe> , an apartment complex manager , and <Pe F="P-N">George Miloslavsky</Pe> , a petty burglar , are transferred to the <Ti F="Xth-C">16th century</Ti> <Lo>Moscow</Lo> , while <Pe F="Mr-P-The-X">Tsar Ivan The Terrible</Pe> goes into the <Ti>year 1973</Ti>. </Summary>

3.4 The Multilingual Dictionary

Our multilingual resources are not a direct result of this project: we have had them for 5 years and each new project increases the richness and the quality of the resource.

We summarize here the current content. The first table shows the number of words for each language. The second table gives the number of translations in a two dimensions table.

Table 5. Number of entries (May 2003)

Language	Root forms	Inflected form management
German	67.288	Many
English	90.330	All
French	215.346	All
Dutch	54.568	Many
Italian	59.573	Most
Portuguese	46.965	Most
Spanish	55.506	Most
Swedish	28.450	Some

Table 6. Multilingual translation bridges (May 2003)

German	**G**						
English	56761	**E**					
French	66253	79256	**F**				
Dutch	51980	51986	51995	**D**			
Italian	55826	57754	58234	50092	**I**		
Portuguese	43265	44925	44262	44298	40925	**P**	
Spanish	52824	51985	53254	53241	53240	43256	**Sp**
Swedish	22853	22847	25845	22847	24857	21453	23587

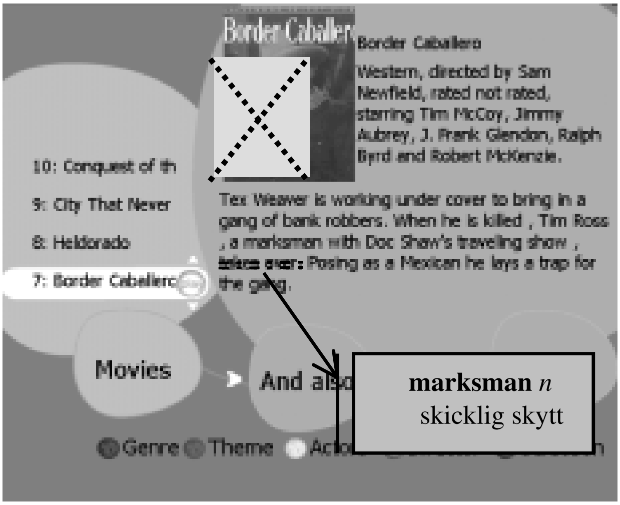
These tables show how useful these data may be for a user of the prototype confronted with an unknown word in a second language.

The following picture shows the interface helping a Swede television viewer confronted with an uncommon word written in English.

3.5 The English Semantic Net inside the Integral Dictionary

To extract some keywords from the summaries, one has to understand the structure of the ID.

Location of the English words. As the table 5 shows it, a lot of English word-meanings have direct links to the French (and vice versa). The other English words are still attached to our semantic net but without any direct translation. The translation of some traditional English dishes cannot be offered by a French word (and vice versa). In these cases, if you have something like *coleslaw*, the conceptual links to *salade* (salad) and *choux* (cabbage) might be useful. The link to *salade* may be annotated with an IsA relation, and the other link with a MadeOf relation.



Border Caballero

Western, directed by Sam Newfield, rated not rated, starring Tim McCoy, Jimmy Aubrey, J. Frank Glendon, Ralph Byrd and Robert McKenzie.

Tex Weaver is working under cover to bring in a gang of bank robbers. When he is killed , Tim Ross , a marksman with Doc Shaw's traveling show , takes over: Posing as a Mexican he lays a trap for the gang.

Fig. 1. Helping a reader

Bearing in mind this important point about the interlingual structure, we now need to have a look to the ID records.

The ID is a semantic network combined with a lexicon. The total number of entries is comparable to that of major lexical networks available in English such as WordNet or MindNet [17].

The ID organises words into varieties of concepts and uses semantic lexical functions. Concepts definitions are based on the componential semantic theory (see [12] and [15]) and lexical functions are inspired by the Meaning-Text theory (see [14]).

Both lexical functions and componential semantics are accessed in the ID using a Java Application Programming Interface (API).

Detailed organization. The basic component of the ID is the **concept**. The concept has a gloss of few words to identify its content. When the concept is entirely lexicalized, it gives the definition of a word (this case is marked with a particular kind of relation between the concept and the word: *Generic*). The concept may sometimes be only partially lexicalized. Basically, concepts, in the ID, are not sets of synonyms (synsets). In WordNet, a synset gathers synonyms. In the ID, a concept gathers words that share part of a meaning. A graph of concepts forms then a structure around which the words are organized.

A starting \ denotes a concept as in \human being or \fur animal. The concepts are classified into different categories. This paper describes only to of the easiest and main ones: classes and themes. Classes form a hierarchy and are annotated with their part of speech such as [\N] or [\V]. Themes are concepts that can be used as a predicate to the hierarchies of classes. They are denoted by a [T].

Words in the dictionary appear as terminal nodes in the hierarchical graph of concepts as shown in Figure2 for the word *fleur* (*flower*). Relations annotate arcs between concepts – themes and classes – and between words and concepts. Major relations are hypernymy (Gen), hyponymy (Spec), various forms of synonymy, ToTheme, and ToClass.

The way in which we organize words and concepts in the ID is the crucial difference with WordNet. In WordNet, concepts are most of the time lexicalized under the form of synonym sets – synsets. They are then tied to words of a specific language.

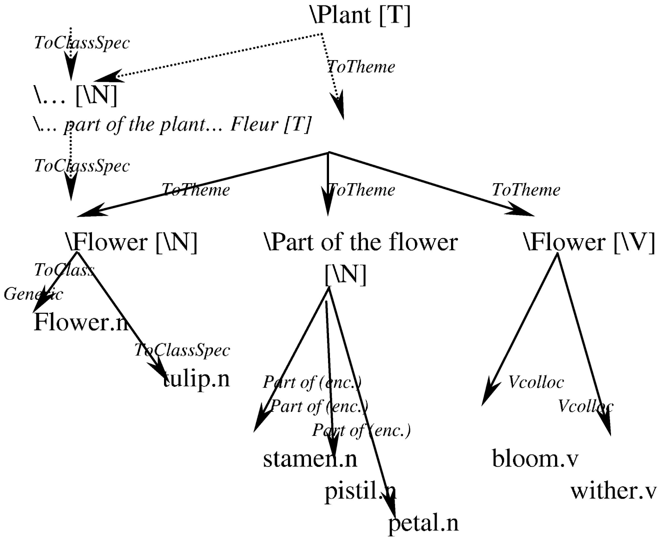


Fig. 2. Graph of concepts for the word flower

In the ID, Theme and Classes do not depend on language's words and it is even possible to create a concept without any words. This is useful for example to build a node in the graph so that a semantic feature that is not entirely lexicalized is shared.

Size. The Integral Dictionary contains approximately 16,000 themes, 25,000 classes, the equivalent of 12,000 WordNet synsets (with more than one term in the content) There is a total of 389,000 arcs in the graph.

Componential Semantics. Componential semantics consists in the decomposition of the words into a set of smaller units of meaning: the semes.

The term ‘seme’ is not very common in English although this concept may prove quite effective and instrumental in the construction of a semantic network. English-speaking linguists prefer the phrases *semantic feature* or *semantic component*, which are not exactly equivalent.

According to the French semantics tradition, the interpretation of a text is made possible by the semes distributed among the words. The repetition of semes in a text ensures its homogeneity and coherence and forms an isotopy.

One problem raised by the semic approach is the choice of primitives. Although, there is no consensus concerning them, a well-shared idea is that the primitives should be a small set of symbolic and atomic terms.

This viewpoint may prove too restrictive and misleading in many cases. Indeed, there are multiple ways to decompose a word according to its possible paraphrases and to its different contexts, as for *florist*:

Semes(florist) = [person] [sell] [flower]
Semes(florist) = [seller] [flower]
Semes(florist) = [person] [work] [shop] [sell] [flower]

The ID adopts a componentional viewpoint but the decomposition is not limited to a handful of primitives. Any concept is a potential primitive and the possible semes of a word correspond to the whole set of concepts connected to this word. Word semes can easily be retrieved from the graph of themes and classes. This approach gives more flexibility to the decomposition while retaining the possibility to restrain the seme set to specific concepts.

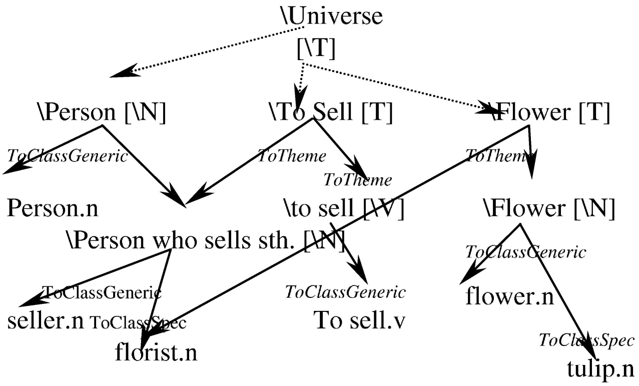


Fig. 3. A part of the semantic decomposition of florist

Lexical Semantic Functions. Lexical semantic functions generate word senses from another word sense given as an input. Functions are divided into subsets. Among the most significant ones, a subset, S0, S1, and S2, carries out the semantic derivations of verbs. These functions could be compared to nominalization in derivational morphology but they operate in the semantic domain and are applied to a specific verb case:

S0(buy) = purchasing/buying (morphological nominalization)
S1(buy) = buyer (subject nominalization),
S2(buy) = {purchase, goods, service} (object nominalization),

The ID has implemented 66 lexical functions for the whole. They correspond to 96,000 links between words. The links between adjectives and nouns are among the most productive ones in the French part.

Algorithm of the distance. The ID superimposes two graphs. A first one forms an acyclic graph whose terminal nodes are the words, the other nodes are concepts, and arcs correspond to relations. A second one connects the words using a lexical function. Figure 4 shows a simplified picture of this structure. The distance between phrases is derived from the graph.

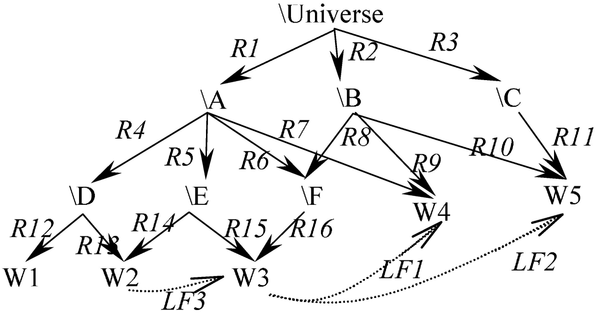


Fig. 4. The graph of concepts, words, relations, and lexical functions

In figure 4, nodes beginning with a backslash '\' are concepts while W1, W2, W3, etc. are words. The root node of the graph is the \Universe label, which is the ancestor of all the concepts. It has a three children respectively \A, \B, and \C, which can either classes or themes.

Arc labels Rn are relations linking the concepts and LFn are lexical functions. In figure 4, *W3* has two **parents** connected by arcs representing two different relations: R15(\E) = W3 and R16(\F) = W3. LF1 is a lexical function linking W3 to W4: LF1(W3) = W4. Inverse relations are implemented so that a parent can be found from its **child**.

The average number of parents of a word or a concept in the ID is 2.1. The average depth of the graph starting from the root is 15. From these numbers, we can evaluate the average number of concepts a word can be member of: $15^{2.1}$ = 294.

The distance between two words or phrases is derived from the graph topology as shaped by the relations. It is the sum of two terms that we call respectively the semantic activation distance and the semantic proximity distance. We describe here a simplified version of this distance based only on the semantic activation.

The semantic activation of two words, M and N, is defined by their set of least common ancestors (LCA) in the graph (see [1]). The semantic activation paths correspond to paths linking both words M and N through each node in the set of least common ancestors.

In Figure 4, we have LCA(*W*2, *W*3) = {\E} and LCA(W3, W4) = {\A, \B}. The activation path between *W*2 and *W*3 consists in the nodes *W*2 \E *W*3 with the functions $R14^{-1}$ and R15. The path between *W*3 and *W*4 consists in *W*3 \E \A *W*4 and *W*3 \F \B *W*4.

We define the semantic activation distance as the number of arcs in theses paths divided by the number of paths. We denote it d^ In Figure 4 :

d^(*W*2, *W*3) = (1 + 1) / 1 = 2
d^(*W*3, *W*4) = ((2 + 1) + (2 + 1)) / 2 = 3

Conceptually, the least common ancestors delimit small concept sets – small worlds (see [11]) – and provide a convenient access mode to them. They enable us to extract a search space of potential semes together with a metric.

There is another kind of distance, which is asymmetric. As we didn't use them a very much in the summaries; we do not describe this distance. Let us now describe the results of the algorithm applied to the corpus.

3.6 The Keyword Extractor and the Clustering to the Lexicon

We use the previous algorithm to extract some keywords from the summaries. More precisely, the idea is the extraction of some important words with their relevant collocation.

Let's take two examples. Suppose a summary presenting a movie about the family life. Words like *father, mother, son, child, cousin, born, baby, childnearing* etc. should be selected. Moreover, the extraction should be roughly organised as clusters:

father:	*1)mother, son, baby, 2) cousin, born*
mother:	*2)father, son, baby, childbearing 2) cousin, born*
son:	*1)father, mother 2) baby, cousin*
baby:	*1) father, mother, born, childbearing 2) cousin, son*
etc.	

The system of keyword extraction parsed the 48.871 summaries in 90 minutes. The total number of extracted clusters is 177.181. The table 7 shows the number of clusters for summaries.

Table 7. Number of clusters (May 2003)

Number of clusters	Number of summaries
1	1.453
2 to 5	10.000
6 to 20	22.000
21 to 31	3.000
>31 <89	3.000

Here is an example for the movie *Border Caballero* (see the summary figure 1).

bank:2172	*working:787;gang:1640;gang;1640*
gang:1342	*bank:1754;robbers:1426;killed:1437; marksman:1634;trap:1774;gang:290;*
robber:1243	*gang:1426;bank:1754;killed:855; marksman:1034;trap:711; gang:1426;*
trap:1330	*robbers:711*
marksman:1582	*robber:1034;gang:1342*

In the following table, the scores express the importance of each term. The lower the score is, the higher the importance is.

The left column shows a global score for the keyword. The right column shows for each keyword the individual score to the other keyword.

The following screen gives an idea of the Semiograph interface.

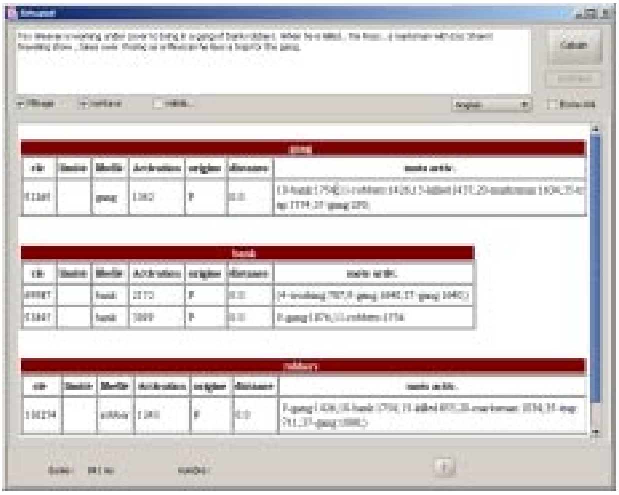


Fig. 5. The Semiograph graphical interface of keywords extraction

For this movie, the final XML files looks like:

```
<Summary Idfilm=”2564” language=”English”> <Pe F="P-N">Tex Weaver</Pe>
is working under cover to bring in a gang of bank robbers. When he is killed, <Pe
F="P-N">Tim Ross</Pe>, a marksman with Doc Shaw's traveling show , takes over.
Posing as a Mexican he lays a trap for the gang.
<PersonsList>
        <Pe>Tim Ross</Pe>
        <Pe>Tex Weaver</Pe>
</PersonsList>
<ClustersList>
        <CL word="bank" value=2172>
                <ClItem>working:787</ClItem>
                <ClItem>gang:1640</ClItem>
                <ClItem> gang;1640</ClItem>
        </CL>
....
</ClustersList>
</Summary>
```

One use of this text indexing might be clusters computing between summaries by using this semantic trace. As, it is not our main task in this project, we did not compute these matches.

3.7 The Matching towards the Thesaurus

The matching towards the Thomson Multimedia thesaurus is not finished yet. The main reason is that we have to introduce this thesaurus to the Semiograph itself. This procedure is automatic but the functionality is limited to the French language.

When the thesaurus items will be integrated to the Semiograph, we assume that computing the activation distance between each extracted clusters and the thesaurus wordings should be a good strategy.

4 Conclusion

This paper tried to show some possible contributions of NLP to an Ambient Intelligent project. NLP can precisely study parts of texts and extract relevant tags and marks. Moreover, NLP components could be in the core of the interface to simplify the access to translation and facilitate the learning of foreign languages.
In that project, we provided:

- language identification tags
- named entities extraction
- shallow parsing
- translation services
- semantic tags to enrich the summaries

In fact, the NLP works may be integrated to most of Ambient Intelligence projects. Though this paper describes particular features and improvements, we must open it to the whole debate. Our company, working mainly in semantics, is involved in many hard to please applications. In very interactive ambient intelligence projects, dialogs should be at the core of the applications. Why not contribute with us to this kind of technology?

Acknowledgments. We wish to thank Louis Chevallier from Thomson Multimedia for having suggested this study to us. We also thank the other members involved on this demonstrator for their contribution. Special thanks go to Evert van Loenen, from Philips, for his kind and efficient organization.

References

1. V. Aho, J. E. Hopcroft, J. D. Ullman, ‘On computing least common ancestors in trees’, Proc. 5th Annual ACM Symposium on Theory of Computing, 1973, pp. 253–265
2. Dutoit D., Poibeau T.: Inferring knowledge from a large semantic network, Août 2002, full paper, acte de Conference on Computational linguistics, COLING TAIWAN

3. Dutoit D., Poibeau T.: Generating extraction patterns from a large semantic network and an untagged corpora, Août 2002, acte de Workshop, COLING, TAIWAN
4. Dutoit D, Nugues P.: A lexical network and an algorithm to find words from definitions, acte de European Conference on Artificial Intelligence, ECAI 2002, LYON
5. Dutoit D, Poibeau T.: Évaluer l'acquisition semi-automatique de classes sémantiques, acte de TALN 2002
6. Dutoit D., Poibeau T. : Evaluating resource acquisition tools for information extraction, May 2002, full paper, acte de Language resource and evaluation, LREC, Las Palmas
7. Dutoit D, Nugues P. : The right word, May 2002, full paper, acte de Language resource and evaluation, LREC, Las Palmas
8. Dominique Dutoit, 'Quelques opérations sens→texte et texte→sens utilisant une sémantique linguistique universaliste a priori', PhD thesis, Université de Caen, 2000
9. Dominique Dutoit, 'A Set-Theoretic Approach to Lexical Semantics', Proceedings of COLING, 1992
10. Christiane Fellbaum (ed), 'WordNet: An electronic lexical database,' MIT Press, 1998
11. Ramon Ferrer I Cancho, Ricard V. Solé, 'The small-world of human language', Proceedings of the Royal Society of London, B 268, 2261–22661, 2001
12. Algirdas Julien Greimas, 'Sémantique structurale', Coll. Champs sémiotiques, PUF, 1986
13. Adam Kilgarrif, 'BNC database and word frequency lists'. URL: www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html
14. Igor Mel'cuk, 'Dictionnaire Explicatif et Combinatoire du français contemporain (DEC), Recherche Lexico-sémantiques III', Presses de l'Université de Montréal, Québec, 1992
15. Bernard Pottier, 'Linguistique générale, Théorie et description', Klincksieck, 1974
16. James Pustejovsky, 'The Generative Lexicon', MIT Press, 1995
17. Stephen D Richardson, William B. Dolan, Lucy Vanderwende, 'MindNet: acquiring and structuring semantic information from text', Proceeding of COLING'98, 1998

A Physical Selection Paradigm for Ubiquitous Computing

Heikki Ailisto, Johan Plomp, Lauri Pohjanheimo, and Esko Strömmer

VTT Electronics, P.O.Box 1100, FIN-90571, Oulu, Finland
Heikki.Ailisto@vtt.fi

Abstract. More natural communication with ubiquitous digital devices requires new ways of interaction between humans and computers. Although the desktop metaphor and the windows, icons, menu, and pointing device (WIMP) paradigm work well in the office computer, different means of interaction might be more suitable for mobile terminals and their communication with ambient devices, objects and services. We present three cases where physical selection[1] may prove advantageous over more traditional ways of interaction. Also, we suggest different ways of realising physical selection and compare their characteristics. Finally, we give an example of physical selection in the case of activating and reading a temperature sensor wirelessly. In the future, we shall investigate the possibility of implementing the physical selection paradigm with mobile phones and Personal Digital Assistants.

1 Introduction

The vision of ubiquitous computing inherently includes natural interaction between humans and digital devices embedded in their environment. The desktop metaphor [12] works well in the office, but it is not so well suited to ubiquitous and mobile computing [15].

Myers et al. describe how the advances in computing will set new requirements for user interface (UI) development [8]. Current interaction with (desktop) computers is implemented by means of an interface paradigm generally referred to as WIMP (Windows, Icons, Menus, Pointing device). This paradigm finds its roots in the windowing and direct interaction system developed at Xerox PARC and used commercially first in the Apple Macintosh (from 1984), soon to be followed by similar implementations by Microsoft (Windows) and by MIT for Unix (X-windows) [2]. The current implementations have a remarkable degree of commonality, which has allowed for easy use of different computer systems after the basic principles of WIMP have been mastered.

[1] The term *physical selection* used here covers both *physical pointing* and physical selection based on *proximity*.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 372–383, 2003.

When small mobile terminals, such as PDAs (Personal Digital Assistant) and smart mobile phones, appeared on the market, they were largely deriving from this successful WIMP paradigm. However, the small display has too little space for windowing solutions or large menus to be displayed. Pointing by means of the touch screen is not very precise and the amount of concurrent information displayed is necessarily limited. Browsing a large amount of information, or selecting an item from a large list, are very tedious tasks with the interface provided by a mobile terminal. The lack of a keyboard has also called for alternative text-input methods like handwriting and speech. Research on finding new ways for post-WIMP interaction is broad and will probably contribute significantly to the natural interaction with mobile terminals.

There is another important observation one needs to make with respect to the use of mobile terminals in a ubiquitous computing setting. The interaction will no longer be limited to the applications within the mobile terminal. Instead, the mobile terminal will be used as a mediator between the user and services and devices in the environment. The mobile terminal will negotiate with the environment, convey the available services to the user, and facilitate the interaction with this service. As in the case of the direct interaction metaphor, interaction involves the selection of a target and the subsequent manipulation of that target. Now the targets are not just found on the screen, but also in the real-world environment of the user and his hand-held mobile terminal. Therefore extending the selection to the real-world by means of a pointing action is a natural step. Interaction with a service is initiated by means of a real-world selection action (physical selection), followed by manipulation facilitated by the mobile terminal. This “manipulation” may just mean the presentation of information, but also displaying a complete user interface for further interaction, or invoking an action in the environment.

Ideas close to physical selection have been suggested [14],[16],[5],[11]. Ulmer and Ishii [14] developed the idea of Phicons, which serve as physical icons for the containment, transport and manipulation of online media in an office environment. Their paper does not discuss the role of mobile personal terminals, such as smart phones or PDAs, but instead relies on fixed devices, such as digital whiteboards, projectors, and printers. Kindberg and co-workers study infrastructure to support "web presence" for the real world [5], their main idea being connecting physical objects with corresponding web sites. Infrared (IR) beacons, electronic tags or barcodes are suggested for creating the connection.

In this paper physical selection as a means of human computer interaction by using mobile personal terminals communicating with the ambient smart devices and services is suggested. The employment of widely used and increasingly popular mobile terminals as a tool for physical selection is estimated to have high potential with a much wider application than accessing web pages associated with physical places or objects. It can be seen as a building block for fulfilling the ubiquitous computing vision. We describe three examples of using physical selection, analyse the implementation issues, and give an implementation example. Finally, the potential of physical selection as well as the future direction of the research is discussed.

2 Using Physical Selection

In this section the possible usage of physical selection is considered. First different characteristics are identified and then three examples are presented.

2.1 Main Characteristics of Physical Selection

The main characteristics of physical selection are

- selection characteristics,
- data transfer characteristics and
- data handling and storage characteristics.

Conceptually physical selection can be based on proximity or pointing. In the case of proximity, physical selection occurs when the mobile terminal[2] ("selection device") is within a certain distance of the target object ("tag"), say closer than 5 cm. In the case of pointing, selection is initiated by the mutual alignment of the pointing device and the target object. For example, the pointing beam may need to be aimed at the target object (within its field of view) within a margin of 5 degrees. Choosing between these two concepts influences not only the usage of physical selection but also its implementation.

The data transfer characteristics determine whether the selection initiates uni- or bidirectional communication between the mobile terminal and the tag. They also cover issues like data rate, latency time and communication range between the mobile terminal and the tag. The amount of data to be transferred may vary from one-bit on/off data to files of Java code. In some cases, physical selection only initiates data transfer using some other communication channel [10], e.g. via a wireless local area network.

The data storage and handling capacity of the tag varies widely according to application needs. The data content of the tag may be fixed, such as in barcodes or passive RFID tags (Radio-Frequency Identification), or it may be dynamic as in sensors or systems with embedded processors. The processing power in the tag ranges from dumb processor-less to smart systems which act as the front ends for sophisticated applications such as heating systems or burglar alarms.

In the following subsections three use-cases exemplifying these characteristics are given.

2.2 Updating the Context Profile of a Mobile Terminal

Developing context sensitivity [3] of devices and services has been seen as one of the main goals of ubiquitous computing research. Although some successful demonstrations and applications have been presented, two key problems seem to hinder the

[2] The word *mobile terminal* is used for the selection device held by the user. Usually the terminal has means of input and output. Practical examples of a mobile terminal are a PDA and a smart phone. A "*tag*" is the target object to be pointed at. A tag can be a fixed information storage (e.g. barcode), an interface to an ambient device (e.g. smart temperature sensor), or to a system (e.g. a computer controlled heating system).

progress in this area: 1) the methods for automatically and reliably assessing the context, e.g. location, task at hand or situation, has proven to be much more difficult than expected; and 2) more fundamentally, the answer to the question "Does the user really want automatic context sensitivity?" is not always "Yes". As a solution to both of these problems in the specific case of updating the context profile of a mobile terminal we suggest physical selection. This approach seems to be highly potential: it resolves the ambiguities and uncertainties related to context reasoning and at the same gives the user a notion of control over her/his digital environment by allowing the user to update and change the context deliberately but in a fashion more natural than using keys or mouse.

The context profile of a mobile phone or a PDA should relate to the current location, task or social situation. The location specific context could be e.g. office, meeting room, car or home. Changing or updating the context profile of a mobile terminal could be done as shown in Figure 1 by physically selecting a Context Tag with the mobile terminal and accepting the new profile. A natural place for Context Tags would be near door posts of rooms. In a similar way, the task or situation context could be chosen by selecting physical symbols of each named context with the mobile terminal.

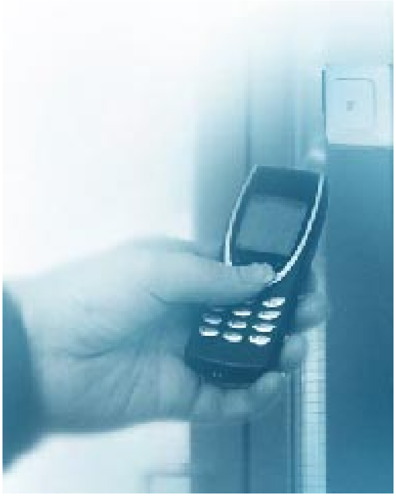


Fig. 1. Changing the context profile of a mobile terminal by physically selecting a Context Tag

Updating the context profile of a mobile terminal requires only unidirectional data transfer, namely from a Context Tag to a mobile terminal. The amount of data to be transferred is small, and thus there is no need for fast communication. The data content of a Context Tag can be either static or semi-static. The activity required by the user is easier, more subtle and faster than selecting the context profile through menus with small keys or by voice commands.

2.3 Activating a Function by Physical Selection

Physical selection can be used to activate functions, either in the tagged device or in the mobile terminal. The former case is the familiar way of using remote controllers to control TV, VCR and audio appliances or to operate car locks and garage doors. The latter case is a more novel idea, which again is associated with mobile terminals. Examples of this kind of activating functions might include activating a phone call to a person by selecting her/his picture or starting a game or an operation in the game in a mobile terminal by selecting a corresponding physical tag. It is also easy to envisage applying tags as URL-addresses [5],[10], where by the user could receive more information, such as maps, manuals, product data etc. via the Web.

Unidirectional data transfer with moderate bit rate and short latency time are typical requirements for this type of application. The data content of the tag may be fixed and there is no need for smartness in the tag. The selection activity may be based either on proximity or pointing.

2.4 UI for Devices and Services without Display and Keys

Physical selection could be used for activating - and even downloading - the User Interface for devices without display or keys. This kind of applications can of course be implemented with, for example, Bluetooth radio communication, but we assume that it can be made much easier and more natural when the UI of a device or service is activated by selecting it - or its physical icon - instead of a complicated series of menu selections. Figure 2 illustrates the principle. Potential applications include burglar alarm systems, home appliances, cars, and industrial systems.

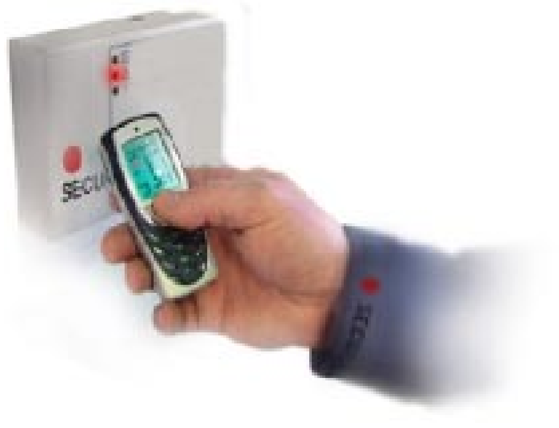


Fig. 2. Activating the UI for a device without a display and keyboard

Activating and using the UI via physical selection requires bidirectional communication, moderate or reasonably high data rate (in the case where the UI is first downloaded), and short latency time. The device to be controlled must contain some data processing capability, otherwise the whole concept is meaningless. The main advantage over more standard approaches, such as using a wired terminal or Bluetooth connection, is seen to be the easy and natural way of establishing the connection by physical selection, and consequently the speed of operation.

3 Implemetation of Physical Selection

This section analyses the practical implementation of the physical selection. In addition to the selection function, the data transfer function is also considered, because these two functions are often implemented by the same technology, even though this is not necessary. Three major alternatives for wireless physical selection exist:

- visual codes, e.g. barcode,
- electromagnetic technologies and
- infrared (IR) technologies.

Vision based solutions as used for augmented reality are excluded in this study, since they often rely on modelling the environment and use a fixed infrastructure, which are not feasible for the mobile usage we envision.

3.1 Visual Codes

The common barcode is the best known visual code. Barcode is a one-dimensional code consisting of vertical stripes and gaps, which can be read by optical laser scanners or digital cameras. Another type of visual code is a two-dimensional matrix code, typically square shaped and containing a matrix of pixels [9]. Optical Character Recognition (OCR) code consists basically of characters, which can be read by people and machines.

The introduction of mobile terminals with embedded digital cameras has made visual codes a feasible solution for physical selection. A code can be read with the camera and analysed by image recognition software.

Visual tags are naturally suitable for unidirectional communication only, as they are usually printed on a paper or other surface and the data in them can not be changed afterwards [6]. When printed on paper or adhesive tape, the tag is very thin, and it can be attached almost anywhere. The most significant differences between barcode, matrix code and OCR lay in the information density of the tag and the processing power needed to perform the image recognition. Barcodes have typically less than 20 digits or characters, while matrix tags can contain a few hundred characters. The data content of an OCR is limited by the resolution of the reading device (camera) and the available processing power needed for analysing the code. Visual codes do not have any processing capability and they do not contain any active components, thus their lifetime is very long and they are inexpensive. The reading distance ranges from contact to around 20 centimetres with hand held readers and it can be up to several meters in the case of a digital camera, depending on the size of the code and

resolution of the camera. By nature, visual codes are closer to the pointing class than the proximity class.
Barcodes are widely used for labeling physical objects everywhere. There are already myriad of barcode readers, even toys, on the market. Commercial image recognition software is also available.

3.2 Electromagnetic Technologies

The following electromagnetic technologies can be applied for physical selection especially in applications based on the proximity concept:

- RFID technologies based on inductive coupling,
- short-range communication technologies based on inductive coupling and
- short-range technologies based on RF (Radio Frequency).

RFID systems incorporate small modules called tags that communicate with a compatible module called a reader [4]. The communication is usually based on a magnetic field generated by the reader (inductive coupling), but with very short operating ranges it is also possible to apply capacitive coupling. The tags are typically passive, which means that they don't operate or produce signals independently from the reader. Instead, the passive RFID tags receive the energy needed for the operation from the magnetic field generated by the reader module, eliminating the need for a separate power supply. In addition, there are active RFID tags that incorporate a separate power supply for increasing the operating range or data processing capability. RFID technology can be applied for physical selection by integrating a tag in the ambient device and a reader in the mobile terminal or vice versa.

Typically the tags incorporate an antenna and one IC (Integrated Circuit) chip providing data transfer, storage and possibly also processing capability. Usually the data transfer is unidirectional from the tag to the reader, but also bidirectional tags exist. The operating range is typically from a few millimetres to several tens of centimetres depending on the antenna, operating frequency, modulation method, operating power and bit rate. Examples of operating frequencies typically used are 125 kHz and 13.56 MHz. The basic advantages of the inductive RFID technology compared to other electromagnetic technologies are low price, small size, operation without a power supply and good commercial availability. These advantages make the inductive RFID technology very attractive from the viewpoint of physical selection applications based on the proximity concept.

Originally the RFID tags were aimed at the electrical labelling of physical objects, replacing visual barcodes. Currently, the RFID technology has established itself in a wide range of applications, e.g. automated vehicle identification, smart cards, access systems and toys. There are several manufacturers providing RFID ICs, tags and systems.

There are also technologies particulary aimed for short-range communication that are based on magnetic induction, but contrary to the RFID technologies incorporate an active transmitter at both ends of the communication link, which also makes the tags able to operate and produce signals independently from the reader. When compared to the inductive RFID technologies, the disadvantage of these technologies is that the tags always require a separate power source for operation. The advantages are longer operating range and increased data processing capapility of the tag due to the separate power source. When compared to RF based technologies, magnetic induction has some advantages in short-range (below 3 m) wireless communication such as power consumption, interference and security [1]. There are also some commercial components available which are applicable in physical selection applications.

Longer operating ranges than by magnetic induction can be achieved by RF-based technologies such as Bluetooth, other wireless personal area network (WPAN) technologies and long-range RFID technologies. WPAN based tags are always active in the sense that they require a separate power source and can produce signals independently from the reader. However, the backscattering technology used in the long-range RFID tags makes them able to operate without a separate power source. On the other hand this means that they can operate and produce signals only when being in the operating range of the reader. The operating range of all these RF-based technologies is typically several meters, which is too long for most of the physical selection applications. However, it is possible e.g. to reduce the operating range by external shielding or to use received signal strength indication (RSSI) if available. Examples of the operating frequencies of WPANs and long-range RFID tags are 868 MHz, 915 MHz or 2.45 GHz. One possible disadvantage of Bluetooth, concerning especially ambient devices, is the high power consumption. Bluetooth, other WPAN and long-range RFID components and modules are available from several manufacturers.

3.3 Infrared Technologies

Infrared (IR) is widely used in local data transfer applications such as remote control of home appliances and communication between more sophisticated devices such as laptops and mobile phones. In the latter case, the IrDA standard is widely accepted and it has a high penetration in PC, mobile phone and PDA environments. Due to the spatial resolution inherent to the IR technology, IR is the most obvious technology for implementing physical selection applications based on the pointing concept.

An IR tag capable of communicating with a compatible reader module in the mobile terminal would consist of a power source, an IR transceiver and a microcontroller. The size of the tag depends on the implementation and intended use, but the smallest tags could easily be attached practically anywhere. The data transfer can be unidirectional or bidirectional. The operation range can be several meters, but a free line-of-sight (LOS) is required between the mobile terminal and the ambient device. In the IrDA standard, the specified maximum data rate is 16 Mbit/s and the guaranteed operating range varies from 0.2 to 5 meters, depending on the version used. One possible problem of IrDA, concerning especially the ambient device, is its high power consumption. For reducing the mean power consumption and thus extending the lifetime of the battery, if used, the IR tags can be woken up by the signal from the reader module [7],[13]. It is also possible that the tag wakes up periodically for sending its identification signal to the mobile terminal in its operating range.

In general, IR technologies are very commonplace. Many home appliances can be controlled by their IR remote controller. Several mobile phones and laptops incorporate an IrDA port, and with suitable software they could act as tag readers. Components and modules are also available from several manufacturers.

3.4 Comparison of the Technologies

The three most potential commercial technologies for implementing physical selection are compared in Table 1. Bluetooth is included for reference since it is the best known local wireless communication technology. Obviously, exact and unambiguous values are impossible to give for many characteristics and this is why qualitative descriptions are used instead of numbers. When a cell in the table has two entries, the more typical, standard or existing one is without parenthesis, and the less typical, non-standard or emerging one is in parenthesis.

Table 1. Comparison of potential commercial technologies for physical selection (Bluetooth included as a reference)

	Visual code	**IrDA**	**RFID, inductive**	**Bluetooth**
Selection concept	proximity/ pointing	pointing	proximity	none (proximity)
Data transfer type	unidirectional	bidirectional	unidirectional (bidirect.)	bidirect.
Data rate	medium	high	medium	high
Latency	very short	medium	short	long
Operating range	short-long	medium (long)	short	medium (long)
Data storage type	fixed	dynamic	fixed (dynamic)	dynamic
Data storage capacity	limited	not limited	limited (not limited)	not limited
Data processing	none	yes	yes, limited	yes
Unit costs	very small	medium	low	medium-high
Power consumption	no	medium	no (low)	medium-high
Interference hazard	no	medium	low-medium	medium-high

In the *Updating the context profile of a mobile terminal* use-case tags are used in a variety of places, usually without easy access to a power supply. To create sufficient infrastructure, a large amount of tags is needed. This suggests that the optimal technical solutions are based on visual codes or electromagnetic tags although it is not impossible to use infrared tags.

All suggested technologies apply to the *Activating a function by physical selection* use-case. Several sub-cases of this use-case seem to be easier to use from a distance and that makes infrared as a pointing based technology more suitable than others.

The *UI for devices and services without display and keys* use-case is the most demanding of the three cases presented. Bidirectional communication, and a demand for data processing capabilities on the tag side rule out the visual code option. Of the two remaining alternatives, infrared seems to be more compelling because of the standardised bidirectional communication and the ability of the tag to act as a front-end for the device in question.

4 Example of Physical Selection in Sensor Reading Case

This section illustrates as an example a communication system with physical selection facility and its demonstration (a more detailed description can be found in [13]). The communication system is based on IR technology. In addition to demonstrating the usefulness of the physical selection paradigm, the goal was to develop a communication system for interconnections between mobile terminals and ambient devices enabling ultra-low power consumption, a very low price and a small size of the ambient devices. The requirement for ultra-low power consumption concerns especially ambient devices, because they are typically fixed installations and thus difficult to recharge, their battery difficult to replace and power supply from the mains impossible to arrange or would increase the installation costs dramatically. Whereas the batteries of mobile terminals can typically be recharged even daily, the power supply of the ambient devices must be based on a small battery lasting for several months or even years. Another possibility is to scavenge the power from ambient light, RF-field, temperature gradient, mechanical energy etc. Typically all these power supply methods call for the mean power consumption of the ambient device far below one milliwatt. [13]

The architecture of the communication system is presented in Figure 3. The communication system consists of a terminal interface unit (TIU) integrated into the mobile terminal and a micropower communication tag (TAG) integrated into the ambient device (Application object in Figure 3).

The communication protocol, architecture, electronics, and software of the TAG and the TIU were designed taking into account the requirement for ultra-low power consumption of the TAG. To evaluate the results, a macroprototype demonstrator was implemented. The mobile terminal in the demonstrator was a PDA, and the ambient device a temperature sensor readable by the PDA. According to the evaluation, the power consumption of the IR-tag can be reduced down to a few microwatts in typical applications, where the TAG is most of the time in ready-for-communication state. In spite of the reduced power consumption, the data rate during the communication can be close to the IrDA standard [13].

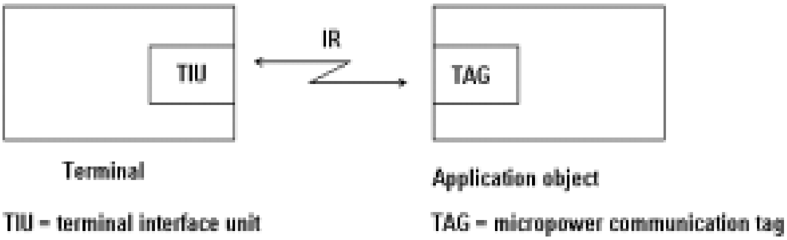

IR
TIU
TAG
Terminal
Application object
TIU = terminal interface unit
TAG = micropower communication tag

Fig. 3. Demonstrated IR based system.

5 Discussion

Physical selection offers a new way for human computer interaction in many applications of mobile and ubiquitous computing. Since smart phones and other mobile terminals, such as PDAs, are becoming very common place, it is most tempting to use them as tools of interaction between the user and the ambient intelligent devices and services. We identified some important characteristics linked to physical selection: the selection method (proximity or pointing), the data transfer characteristics, and the data handling and storage characteristics. Three use cases were analysed with respect to these characteristics. We assume, that physical selection is perceived more natural than current methods in many cases like the ones discussed here. This hypothesis remains to be tested by future research.

The three main alternatives for implementing physical selection are visual codes, electromagnetic and infrared technologies. Each of these alternatives has its benefits and drawbacks making them suitable for certain applications. For example, implementations based on infrared technology offer bidirectional communication with dynamic data content and high data rate at moderate cost. Furthermore, IR inherently supports pointing and is more privacy preserving than RF. For these reasons, we decided to build our first physical selection experiment, the reading of a temperature sensor, using IR communication and a mobile terminal.

Since the experiment showed the concept to be feasible and since IR communication is supported in hundreds of millions of existing mobile terminals by their IrDA ports, the use of IR communication technology has a good foundation and will be elaborated in our future work. The open platforms of PDAs and new Symbian® based smart phones offer the possibility for realising physical selection in mobile terminals existing today and those to come during next years. Therefore, we intend to implement physical selection with IrDA equipped mobile terminals (smart phones) and experiment with new use cases.

References

1. Bunszel, C.: Magnetic induction: a low-power wireless alternative, R.F. Design 24, 11 (Nov 2001), 78–80

2. Dettmer, R.: X-Windows - the great integrator, IEE Review 36(6), June 1990, 219–222
3. Dey A.: Understanding and Using Context. Personal and Ubiquitous Computing 5 (2001), 4–7
4. Finkenzeller, K.: RFID Handbook, Radio-Frequency Identification Fundamentals and Applications. John Wiley & Son Ltd, England, 1999
5. Kindberg, T, et. al.: People, Places, Things: Web Presence for Real World, in IEEE Workshop on Mobile Computing Systems and Applications WMCSA'00 (Monterey CA, December 2000), IEEE Press, 19–28
6. Ljungstrand, P., Holmquist, L.E.: WebStickers: Using Physical Objects as WWW Bookmarks. Extended Abstracts of ACM Computer-Human Interaction (CHI) '99, ACM Press, 1999
7. Ma, H., Paradiso, J. A.: The FindIT Flashlight: Responsive Tagging Based on Optically Triggered Microprocessor Wakeup, in UbiComp 2002, LNCS 2498, 160–167
8. Myers, B., Hudson, S., Pausch, E.: Past, Present, and Future of User Interface Software Tools, ACM Transactions on Computer-Human Interaction 7(1), March 2000, 3–28
9. Plain-Jones, C.: Data Matrix Identification, Sensor Review 15, 1 (1995), 12–15
10. Pradhan, S.: Semantic Location, Personal Technologies (2000), 4, 213–216
11. Schmidt, A., Gellersen, H., Merz, C.: Enabling implicit human computer interaction a wearable RFID-tag reader, in International Symposium on Wearable Computers, Digest of Papers, 2000, 193–194
12. Smith, D.: Designing the Star User interface. Byte April 1982, 242–282
13. Strömmer, E., Suojanen, M.: Micropower IR Tag - A New Technology for Ad-Hoc Interconnections between Hand-Held Terminals and Smart Objects, in Smart Objects Conference sOc'2003 (Grenoble, France, May 2003).
14. Ullmer, B., Ishii, H., Glas, D.: mediaBlocks: Physical Containers, Transports, and Controls for Online Media, in Proceedings of SIGGRAPH '98 (Orlando, FL, July 1998), ACM Press, 379–386
15. Want, R, Weiser, M., Mynatt, E.: Activating everyday objects in Darpa/NIST Smart Spaces Workshop (Gaithersburg, MA, July 1998), USC Information Sciences Institute, 140–143.
16. Want, R. Fishkin, P. Gujar, A., Harrison, B. L.: Bridging physical and virtual worlds with electronic tags, in Proceedings of Conference on Human Factors in Computing Systems, 1999, 370–377.

Users' Preferences for Ubiquitous Computing Applications at Home

Katja Rentto[1], Ilkka Korhonen[1], Antti Väätänen[1], Lasse Pekkarinen[1],
Timo Tuomisto[1], Luc Cluitmans[1], and Raimo Lappalainen[2]

[1]VTT Information Technology, PO Box 1206, FIN-33101 Tampere, Finland,
{Katja Rentto, Ilkka Korhonen, Antti Väätänen, Lasse Pekkarinen, Timo Tuomisto, Luc Cluitmans}@vtt.fi
[2]Department of Psychology, University of Tampere, 33014 Tampere, Finland
raimo.lappalainen@uta.fi

Abstract. We developed and evaluated a home network and ambient intelligence prototype for wellness management and home automation applications. The evaluation was based on interviews and a user trial at a simulated home environment. This paper describes users' attitudes towards ubiquitous computing technology at home, and especially what kind of applications they would prefer to use at home. We also aimed to gather qualitative information about what kind of user interfaces would be desired for using these applications. The study generated new ideas to develop the ubiquitous computing enabled home concept further.

1 Introduction

Home is a very special place for us. The importance of the home may be observed as the interest towards smart homes, or smart houses, which has been vast for many decades. Today, the development of ubiquitous computing is turning the old visions of the smart home into reality. However, still relatively little is known about the true interest and the uptake of the technology and smart home services by the ordinary customers: what kind of ubiquitous computing applications the people would truly like to use at their homes, and how they would like to use them?

Home automation and wellness management are interesting applications for smart homes. Technically, these applications have many features in common:

1. Requirement of networking of traditionally non–computational things, which have their main function in the physical rather than in the information space, such as scales, beds, white goods, etc.;
2. the need for communications bandwidth is relatively low;
3. the type of communications is characteristically rather messaging than real–time communications; and
4. the user interaction should naturally occur primarily through the physical world rather than through the information appliances i.e. through the virtual world.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 384–393, 2003.

Applications for home automation, including control of HPAC (heating, plumping, air–conditioning), fire and burglary alarms, control of lights and other electrical appliances, have already emerged to the market, and allow remote management and alarm functions through an ordinary mobile phone. However, they are usually implemented in proprietary platforms where there are little possibilities to update or install new applications, and where interaction with other systems – a central issue in ubiquitous computing [1] – is poor or non–existing. The automatic (intelligent) features are usually limited, and user interfaces do not support natural interaction and intuition. In wellness management there are similar though fever applications available, and they have similar limitations.

In the WWM (Wireless Wellness Monitor) project we aimed to develop a prototype for a home supporting ubiquitous computing applications for wellness management and home automation [2],[3],[4]. The main emphasis was on building a home network, where multiple simple household and health monitoring devices are connected, and implementing realistic applications, which may be used either with or without a specific information appliance.

The objective of the present study was to evaluate the prototype through a user trial. Especially, we wanted to gather qualitative information about the acceptability and interest to use different applications: what kind of smart home applications the users want to have, and what kind of user interfaces would be desirable for using these applications.

2 Smart Home Prototype

In WWM project we studied how simple household appliances can join the home network, and how home automation and wellness management applications may be implemented on this kind of a heterogeneous home network. The appliances and services included in the prototype were chosen so that they represent typical home automation and wellness management needs, yet being demonstratable in a home lab environment, where there are no regular inhabitants and the infrastructure sets some limitations (e.g. no access to HPAC systems). For more information about the WWM project, see [2],[3],[4].

2.1 Technical Setup

Our prototype for ubiquitous computing enabled home is based on the IP–based home network and an OSGi (Open Services Gateway Initiative, www.osgi.org) compliant home server (Figure 1). Different non–IP peripherals join the network via a device proxy, to which they may connect using any proprietary method, e.g. a RS232 or a SoapBox wireless interface [5]. The applications run on a home server (centralized intelligence model) as OSGi bundles.

The user interface (UI) appliances are essentially Java–enabled HTML–browsers, which allow the user an access to different services at home through a home portal. The UI information appliances in the prototype are a PDA (personal digital assistant,

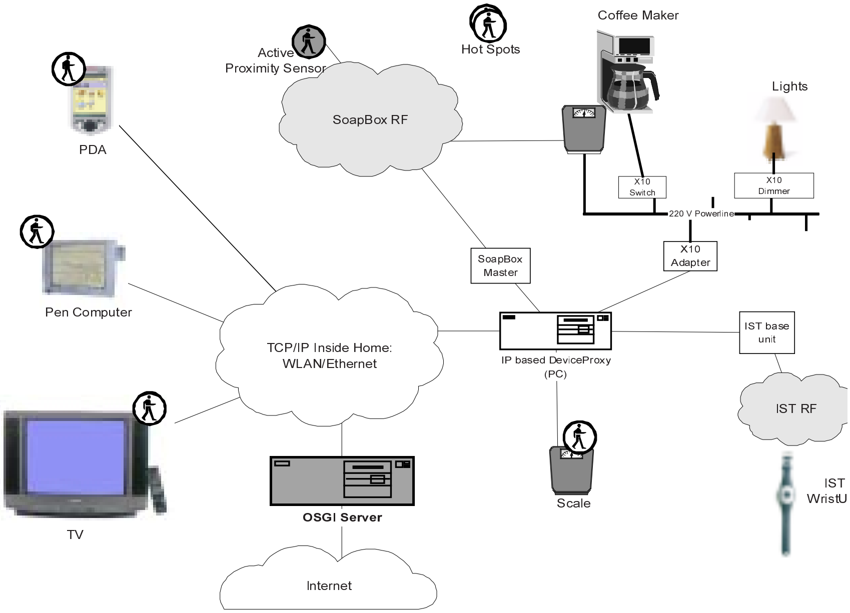


Fig. 1. Schematic presentation of the WWM prototype system used in the user trials

which is either an iPaq 3850, or a Fujitsu pen computer, both equipped with a WLAN card), or a TV with a wireless mouse.

The main difference with these UIs is that while the former is a personal and mobile device, many persons may use the latter and it is stationary.

To provide user identification and relative indoor positioning for different applications, the users wear a RF–based proximity tag, which detects other tags in the vicinity. Each ambient tag was associated with a certain location indoors, which allowed the positioning service at home server to deduce the location of the user, which, in turn, was used for detecting e.g. the user of the personal scale. In addition, the user wears a wrist unit (IST Vivago WristCare□; www.istsec.fi), which provides a signal quantifying the wearer's physical activity. In addition, the unit has a button, which may be used to control the environment (e.g. switching the lights on or off). The electronic appliances are controlled via X10 power line protocol. The peripherals include an intelligent coffee maker (capable to communicate it weight, used to determine the amount and freshness of the coffee in the pot), personal scale, lights, movement detector and a web camera.

2.2 Prototype Services

The prototype was built in our usability laboratory. The lab simulated an ordinary home environment containing a kitchen, a living room and a bathroom. The services (applications) were accessible through the home portal (Figure 2), though some

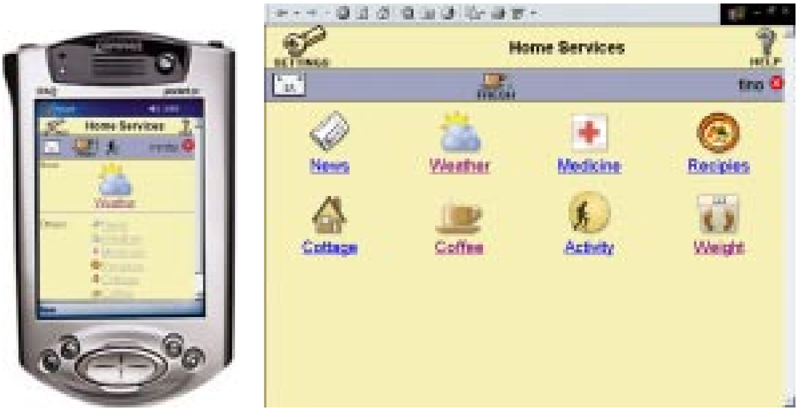


Fig. 2. Home portal UI on PDA (left) and on TV screen (right). The uppermost area of the browser window is the title bar providing access to help and settings. Below that is the status area, in which the most essential status information as well as announcements (e.g. about a new message) are shown. The lowest area is reserved for displaying the icons of the services. In the PDA user location information is used to emphasize the icons of services, which are considered relevant for that location.

services could be used without any information appliance (e.g. weighing occurred just simply by stepping on the scale, and lights could be controlled with IST Vivago manual button). The home portal was context sensitive so that while accessing it by a PDA it used the location information to emphasize the services relevant especially to that location (Figure 2).

The prototype services may be grouped into three categories:

1. Wellness management and health related services included weighing (automatic weight measurement and storage on personal database), weight history browsing (launched automatically after weighing if user has an UI open), and activity history browsing (activity was automatically logged into the personal database always when the user was wearing the wrist unit). In addition, activity status was also used to push coffee maker status information to the user if his/her activity was low and there was fresh coffee available.
2. Home automation and control related services included control of lights (by the IST wrist button), burglary alarm, intelligent coffee maker (information about coffee amount and freshness, automatic turn off the power in case of empty pot or too old coffee), and home messaging service (notifications sent to the user about different events in the house, e.g. washing machine failure).
3. Information services included access to news, weather, food recipes or medicine information, i.e. the user could access information from web through his/her UI. These services were made context aware so that their priority in the UI was dependent on user's location, e.g. news service was emphasized when the user came near the table in the living room.

Fig. 3. User using the PDA (top) and the TV UI during the user trials in the usability laboratory.

3 Evaluation Methods

Twelve users (6 men, 6 women, age 23–33 years) participated in the user trial. The chosen users were highly educated and used computers, mobile phones and Internet regularly to obtain a group of early technology adopters. This user group was chosen as it represents potential early adopters for smart homes but most importantly has a potential to understand within a given time frame the scenarios and setups they are presented during the user trials.

The user test was based on a walk-through of a pre-specified scenario, which included most of the prototype applications presented above. The test began with the introduction session in the meeting room, where the test user was given the PDA and a personal tag. After the introduction the test continued at the usability lab (Figure 3). A complete test took approximately two hours. Two evaluators carried out the tests: one as interviewer and the other as an observer. Both evaluators made notes and the test sessions were recorded on MiniDisc. We interviewed the users by e–mail afterwards. In this way we wanted to find if the users attitude towards wireless home network had changed or if they had any new ideas about the system.

The descriptions of actions in the tests are presented in the Table 1. In certain steps during the tests the users were asked to evaluate their possible motivation to use the current service or application, to provide any suggestions to improve it further, and whether they could foresee any potential problems with the use of the application in real life. Furthermore, they were asked to grade the importance of the service in scale 0 – 10 (0 is useless, 10 excellent), and what other services or features could be used in a similar manner.

4 Results

4.1 General Evaluation Results

The user trials were conducted successfully. In general, the attitude of the users towards the idea of the services and the smart home concept was positive, but not unreservedly. Privacy issues aroused suspicions and also the real need and some users questioned functionality of services in everyday life. Security and HPAC–related services and the messaging service got the most positive feedback. In fact, the users wished the home network to consist more information and control features for HPAC than our prototype did. Weight management and activity monitoring services were concerned as potential services especially for special groups such as elderly users for measuring long–term data for wellness management and to support independent living.

Advantages of the location awareness of different information services were not experienced as significant. Most test users did not experience any advantage of location awareness of different services, though some users felt that it was practical to get recipes or contents of fridge on the screen as they go to the kitchen, or weather service as they move towards the window. The users considered, however, that an optimal UI should take care of the problem of several services (i.e. if there are multitude services available simultaneously, it may be difficult to identify the relevant icon on a small PDA display, for example) and personalization could help the usage. Hence, the idea of location awareness was seen to provide some potential but obviously this was not experienced in our setup with relatively small number of available services. It was considered as important to have continuous access to all information services without a need to move to a certain location. It appears that implementing this kind of features in a proactive manner implies a larger concept for context awareness, i.e. using also other context information than just the location.

The users were not willing to continuously carry any extra tags, sensors or terminal devices with them at home. The PDA was considered as too big and heavy to carry continuously at home, and also the pointing device (pen) could be lost easily. However, for accessing the home portal the PDA was preferred to the TV. The users consider the TV to be preserved in the first instance for the entertainment and media and for the whole family. In the kitchen a flat and large display on the wall or on the fridge door was desired to access the home portal and some services such as food recipes. Home control terminal in hallway would appear as an appropriate UI concept for HPAC and security services.

4.2 Specific Comments on Applications and Technology Details

In the following, some specific comments regarding the specific applications or technology details presented to the users are summarized.

Personal Tags and Indoor Positioning. In our prototype system wearable tags were used to provide automatic user identification and indoor positioning system. However, the users had some privacy concerns related to the user identification and positioning. For example, they were concerned about who will have access to the data that the personal tag is gathering. If the tag is necessary, it should have a form factor of a jewel or alike i.e. as small unobtrusive and unnoticeable as possible. It should also be possible to switch the tag off when necessary.

Table 1. Evaluated themes and evaluation approaches used in the user trials

Issue to be evaluated	Evaluation approach
General idea of the smart home	The idea of a smart home, the smart home prototype, and the UIs are introduced to the user
Switching lamp on by IST wristband	The user controlled lights by clicking the button on IST wristband.
Messaging service	The receives different messages related to the state of the home environment and household appliances.
Weather service	A location aware information service is introduced to the user: the user moves near to a window, which causes the home portal to emphasize weather service as a relevant service.
Weighing service	User weighs him/herself by stepping on the scale. After weighing the user sees the weight result and a long term weight history automatically on his/her PDA.
Recipe service	Another location aware information service: in the kitchen containing food recipes.
Activity monitoring service	IST Vivago wristunit collects data of user's activity automatically. Information is presented as an icon and as an activity history.
Coffee machine	The users checks the status of the coffee machine remotely using the PDA
Medical chest service	Yet another location aware information service: near to the medical chest the user receiveds information of drugs for headache.
Cottage burglary alarm	The user receives an alarm about a movement at the cottage, after which the user opens a web camera picture of the cottage.
News service	Last location aware information service: the user can read the news from TV or PDA while sitting on a sofa.

UI Devices. As mentioned above, the users did not want to carry any mobile device at home. They were afraid of loosing the device. New ways of control, such as gestures and voice were also suggested. The users expressed a preference to use fixed touch screens. Almost every user wished to have a screen at the kitchen where they can see e.g. the recipe service. The screen could be integrated to the closet's door and it should be big enough so that they can e.g. read the recipes while cooking. Also a screen integrated to the fridge's door was a popular idea. Another screen could be in the hall. There it is possible to check the HPAC–status and security system and make modifications to house system, and check the lights when leaving out.

Integrated Health Service. The users expressed a wish to integrate activity monitoring and weight monitoring services as a comprehensive health service, which could also include other monitoring methods e.g. blood pressure, blood sugar, body weight index, fitness level and muscular strength. In addition, the health service should be linked e.g. to the recipe service. This integrated health service could e.g. give tips when the user should rest or what kind of exercise he should do, and also advises about a healthy diet. Especially the long–term follow–up of the health related parameters was considered interesting and valuable.

Messaging System. The idea of the messaging system was to deliver short text messages related to the status of the home appliances or some other event messages to the user's UI. When a new message became available it was notified to the user as an icon in the UI status area (Figure 2). The users mainly welcomed the messaging system. However, they also suggested that the messages should have different priority categories, e.g.: 1) urgent alarms, which demand immediate actions; 2) notifications and reminders; 3) information messages, which provide mainly feedback to some action. Most of the users wished the messaging system to resemble the one in the mobile phones: if something crucial happens, the alarm message is a loud signal and text on the UI in use, even if the inhabitant is not at home or using some UI appliance i.e. alarm messages should overrun any other actions. Notifications and reminders could be set either manually by the inhabitants or automatically by the system. The messaging system could be also used to provide monthly reports about e.g. the heating oil consumption and remaining amount of oil in the tank, i.e. the system could be used to push selected information from e.g. HPAC system.

HPAC and Security. Information related to HPAC status, humidity and air cleanness was considered very important by the users. Especially long–term data of HPAC related variables such as water, oil and electricity consumption, indoor and outdoor temperatures and humidity, etc. were considered useful. Provided that these data were certified and trusted they could also be used as a reference for the condition of the house or apartment: e.g. while selling a house the owner could provide these data for the potential buyer as a reference for expected living costs in the house. In addition, physical security issues (e.g. burglary and fire alarms) were considered essential. Many users also suggested that the system should cover not only the home but objects such as a car, summer cottage, boat, bike etc. to provide burglary and fire alarm protection for these items within the same system.

The emphasis laid on the HPAC and security issues probably reflects the fact that these systems are already available and hence the users are already familiar with their potential, but also that HPAC and security are considered truly important factors potentially providing also some reimbursement for the investment for the smart home technology.

Remote Control and Setup of Home Appliances. The home portal UI could also serve as a generic remote control UI for all home appliances connected to the home network. The users mentioned a need to control remotely e.g. the sauna stove, kitchen oven, fridge, deep–freezer, lightning, house doors and windows. For remote control, the stationary UIs (e.g. TV and an UI at the hall wall) would be preferable. Obviously, the control and the messaging system reporting about the status of different appliances should be inter-operational.

Automation. With home automation we mean that the smart home system automatically (i.e. without waiting for approval from the user) performs some functions related to the control of the physical world e.g. switches on and off some appliances or control their status otherwise. In this study, though the users were ready to accept some automation and proactive behavior from the system, they considered that predictability and feeling of control are the most important features of the system. It was also emphasized that if the user should be able to switch the system off anytime without falling into troubles. Furthermore, it must be possible to be able to track easily which features are functioning and which are not at any given time instant.

5 Discussion

There are several issues associated with the current user evaluation, which should be taken account when making conclusions of the tests.

First, the results can only be directly generalized to a group of highly educated, relatively young people who use often computers, mobile phones and net services – "early technology adapters". On the basis of these results it is very difficult to say how, for example, older people with limited experiences of net services may experience the presented system and the concept.

Furthermore, the results should be extrapolated with caution when the system and the prototype are used for a long time in the real environment – at home. For example, the long-term use of home automation may either lead the users to learn to predict the behavior of the system leading to an increased use and user satisfaction, or totally to give up using the system e.g. due to e.g. occasional unsatisfactory performance. Clearly, the critical issues in proactive functions are the frequency of unsatisfactory functioning and users' willingness to accept that at their homes in real settings – especially the latter we know relatively poorly. Another example is health monitoring: the users had very positive attitude to it but it remains unknown how frequently they would truly utilize these features in their daily life – it is well-known that the customers tend to have positive attitudes towards healthy living style (e.g. food, exercise) but poorly adopt these in their own life.

Finally, we must remember that this evaluation was based mainly on the users' verbal reports after they had received information of the concepts and tasks. People may have difficulties to report verbally their likes and dislikes associated with a complicated system. Thus, the users' reports are probably only a part of the “total experience” the users had during the test session. This evaluation difficulty is, however, hard to overcome without going into longitudinal field studies in real homes.

6 Conclusions

The study suggests that regarding the enabling technology for the ubiquitous technology at homes the proper UI devices and methods for indoor positioning and user identification are essential for successful implementation. Unlike in the office environment or outdoors, the users are reluctant to carry any extra devices with them at their homes. Hence, anything that can be done to minimize the need for that would be an advantage. Furthermore, it is essential to build the applications to meet the real needs of the users, and so that their feeling of control is not compromised.

While interpreting the results of the current study, one should bear in mind the limitations. First of all, the evaluation was based on a laboratory study and interviews. The results in an evaluation in a real home environment, with real services, in long–term, might give different results. In this kind of a study, the factors like a priori attitudes, the way and order how information and tasks are presented to the user, the actual implementation of the prototype (and compromises involved) etc. may a significant role. However, we believe that the current study still adds to our understanding how the real users would like to use their ubiquitous computing enabled home.

Acknowledgments. The study was carried out with VTT Information Technology, VTT Electronics, Nokia Research Center, IST International Security Technology Oy, and Celotron Oy. The study was financially supported by TEKES, the National Technology Agency of Finland. We would also like to thank all the user trial participants.

References

1. Weiser, M.: Ubiquitous Computing.http://www.ubiq.com/hypertext/weiser/UbiHome.html
2. Pärkkä, J., van Gils, M., Tuomisto, T., Lappalainen, R., Korhonen, I.: A wireless wellness monitor for personal weight management. In: Proceedings of 2000 IEEE EMBS International Conference on Information Technology Applications in Biomedicine, ITAB–ITIS 2000 (2000) 83– 88
3. Saranummi, N., Korhonen, I., van Gils, M., Kivisaari, S.: Barriers limiting the diffusion of ICT for proactive and pervasive health care. In: IFMBE Proceedings. Medicon 2001. 9th Mediterranean Conference on Medical and Biological Engineering and Computing. Part 1 (2001) 23–26
4. Van Gils, M., Pärkkä, J., Lappalainen, R., Ahonen, A., Maukonen, A., Tuomisto, T., Lötjönen, J., Cluitmans, L., Korhonen, I.: Feasibility and user acceptance of a personal weight management system based on ubiquitous computing. In: Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Istanbul, Turkey, October 25–28, 2001)
5. Tuulari E, Ylisaukko-oja A. SoapBox: A Platform for Ubiquitous Computing Research and Applications. In: F. Mattern, M. Naghshineh (Eds.): Proceesings of the Pervasive Computing. First International Conference, Pervasive 2002 (Zürich, Switzerland, August 26-28, 2002). pp. 125–138

Experiences Managing and Maintaining a Collection of Interactive Office Door Displays

Dan Fitton and Keith Cheverst

{dan.fitton, keith.chevers}@comp.lancs.ac.uk

Abstract. To date there have been very few ubicomp systems that have been deployed and evaluated 'in place' and over the longer term. Consequently, the issues related to the management and maintenance of such systems remain a very much under explored (but non-the-less critical) area. This paper discusses our experiences managing and maintaining the Hermes system, which provides asynchronous messaging services for users and runs on multiple public display appliances situated outside offices in the computing department of Lancaster University. Hermes has been running 24/7 for over 15 months, and receives regular use by a range of users.

1 Introduction

Weiser's vision of ubiquitous computing [10]; moving technology away from the focus of a 'single box' and 'into the fabric of everyday life', seems to imply the need for highly-available situated technology. Today there are many examples of mobile devices allowing users access to applications and services wherever they go, however these simply allow us to take the 'box' with us. One of the key requirements to providing a real ubiquitous computing environment is situating technology in a range places.

This paper presents experiences of managing and maintaining the Hermes system [1], a system of deployed interactive office door displays, providing asynchronous messaging services for users in the computing department of Lancaster University. The displays are small PDA based devices situated outside offices and accessible to all passers-by. One of our overriding aims has been to deploy and evolve Hermes in the longer term. This is to investigate the interactions that (do and do not) occur, observing how the system is used on a day-to-day basis. Through our approach we hope to see how situated technology can be put to use in everyday scenarios, and what the principal factors in design and development on such systems are.

Such a system deployed and used on a day-to-day basis clearly requires some sort of management element to help it run smoothly. This paper presents the additional management agent that we have developed for maintaining the 'smooth' running of the Hermes system, and in particular to address some of major technical difficulties we have encountered during the last 15 months of use.

The following section of this paper ('The Hermes System') presents an overview of the Hermes project and the third section ('Design of the Hermes System') presents

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 394–409, 2003.

some of the key design issues behind Hermes. The fourth section (‘Experiences of Long Term Deployment’) then introduces some of the motivating factors for supporting a management element within Hermes. Following this the fifth section (‘Supporting and Managing Hermes’) presents an overview of the key functionality supported by our implemented and deployed management agent. This is followed by short overview of related work and a future work section. Finally some concluding remarks are presented in the last section.

2 The Hermes System

2.1 Overview

One of our goals for developing and deploying the Hermes system was to explore whether the traditional way of leaving messages on Post-it™ notes in ‘semi-private’ places, such as office doors, could be enhanced with a digital equivalent. In order to explore this area we have designed and deployed a digital asynchronous messaging system (named Hermes after the messenger to the gods in Greek mythology) within the main computing building at Lancaster University.

The Hermes system supports remote interaction through a web portal and by allowing messages to be created using a mobile phone via SMS (Short Message Service) [3].

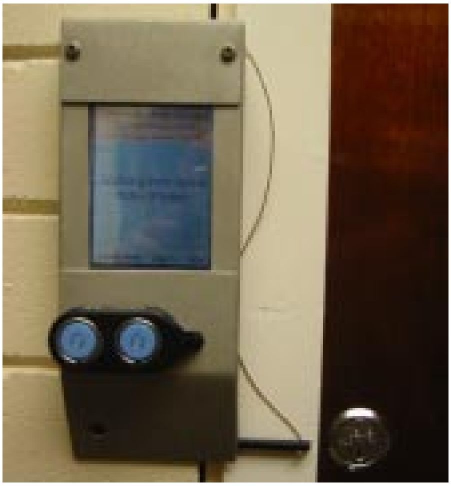


Fig. 1. Example of an early Hermes display

Devolvement work started on the Hermes system in October 2001 and the first unit was installed outside one of the offices in the computing department in March 2002. The system comprises a central server and a number of wall mounted units (referred to as Hermes displays). Figure 1 illustrates the first Hermes display to be deployed in the department. The design of Hermes displays was required to meet a number of installation requirements.

2.2 Requirements

The deployment of the Hermes system was subject to three main requirements, namely: units should comply with university health and safety regulations, units should comply with disabilities legislation, units should be both straightforward to deploy and units should be relatively secure.

Current U.K. disabilities legislation states that public facilities need to be positioned at a height that does not unduly discriminate against people using a wheelchair. For this reason, it was necessary to place the units at a fairly low height (approximately 150 cm) off the floor. However, one of the implications of this is that the device is quite awkward to use for taller people. This problem highlights an interesting variation on the theme of private ownership vs. public use.

One of our key requirements for Hermes displays was that they should be very easy to deploy. We felt that this requirement would be best met by designing the Hermes display as a self-contained unit. We also wanted the display to be relatively easy to develop and so we chose to adopt a PDA based solution. We had hoped that the use of wireless communications would mean that cabling would not be required from a Hermes display to its associated office. Unfortunately, however, the battery life of a PDA is still relatively short and so it has proved necessary to take (LV) power from offices through a small drilled hole in the door mounting.

Although access to the department is restricted during evenings and weekends the department has unrestricted access at all other times. For this reason, we needed to mount displays in such a way that opportunistic theft of the device would be difficult. The case for the displays was also required to restrict access to the buttons on the PDA device. This is to prevent malicious or accidental termination of the application.

2.3 Functionality

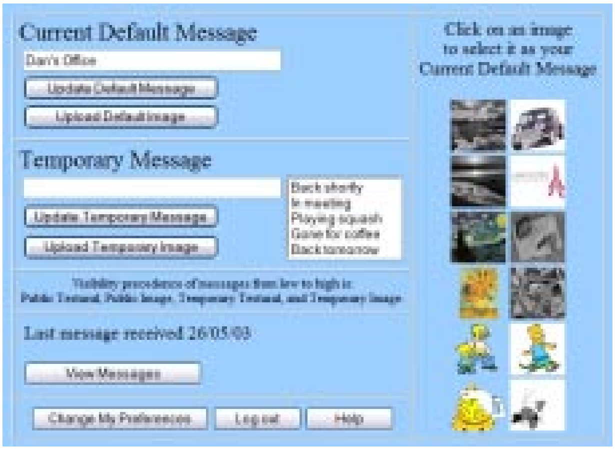
Our design approach has been implement a small number of features well, rather than attempting to provide every feature technically feasible. This enables us to provide the levels of ease of use and dependability usually found with information appliances.

The functionality supported by the system can be considered from two main perspectives: the perspective of the *owner* of the Hermes display and the perspective of a *visitor* to the Hermes display.

2.3.1 Functionality Available to the Owner

The system provides the owner of the Hermes display with two key functions: the ability to create a message to appear on the display, and, the ability to read messages left by visitors.

Typically, the owner will create a message to appear on their Hermes display by entering some appropriate text using the web interface shown in Figure 2. The web interface can also be used to upload a graphical image, such as an animated GIF.



Current Default Message

Dan's Office

Temporary Message

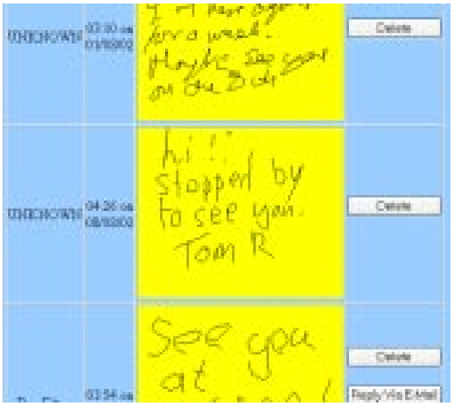
Back shortly
In meeting
Playing squash
Gone for coffee
Back tomorrow

Visibility precedence of messages from low to high is:
Public Textual, Public Image, Temporary Textual, and Temporary Image

Last message received 26/05/03

Click on an image to select it as your Current Default Message

Fig. 2. The Hermes web interface



UNKNOWN 03:30 on [illegible]

for a week! Maybe see you on the [illegible]

UNKNOWN 04:26 on [illegible]

hi!! stopped by to see you. Tom R

03:54 on

See you at

Fig. 3. Viewing messages via the Hermes web portal

Initially, only the web interface could be used by the owner of the Hermes display in order to create messages. However, after a short period of use it became clear that it is often only when closing the office door that one thinks to leave a message. For this reason we added a feature to enable the owner to create a freehand message by using an interface on the door display itself. This process does, however, require the owner to authenticate themselves with the system; this would typically be achieved by the owner docking his or her iButton [4] or entering a username/password via a simple GUI on the Hermes display. The owner can read his or her messages remotely via a web browser (see Figure 3).

2.3.2 Functionality Available to the Visitor

In the current implementation, a visitor must be co-located with a Hermes display in order to leave a message. The user simply has to tap on a 'leave note' button on the

Hermes display and then use the attached pen to ‘scribble’ a message on the touch sensitive display. Once the user taps on the ‘finished’ button the display on the unit is updated to reflect the fact that the owner has an additional message waiting to be read.

A visitor is not permitted to read the messages left by other visitors for the owner of a Hermes display, providing a privacy advantage over the traditional Post-it™ note. Authentication by a visitor allows them to remotely view any messages left on their own Hermes display.

3 Design of the Hermes System

3.1 Design of the Hermes Display

3.1.1 Hardware

Having decided to develop a system using an off-the-shelf PDA, we considered various candidates. After consideration we decided upon using the Jornada 568 for two main reasons. Firstly, the unit has a built-in compact-flash type 1 slot so does not need an additional expansion jacket. Secondly, the unit has a relatively square shape and we felt that this would simplify case design. Initially all Hermes displays were equipped with a compact flash type 1 802.11b wireless network interface.

In order to allow the units to be deployed securely, a case was designed to hold the unit. The case is made from aluminium, and can accommodate an iButton reader. This is wired to the serial port of the Jornada in order to support authentication.

3.1.2 Software

From an early stage we decided to use a Java for development. However, finding a Java virtual machine (VM) that met our requirements was a difficult process, particularly one that supported Java COMM API (to support a serial iButton reader). Eventually, we chose to adopt NSIcoms CrEme [8] version 3.2 running on the PocketPC2002 operating system as our development platform.

3.1.3 Overall System Architecture

The overall system architecture of the Hermes system is shown in Figure 4. The large oval represents the typical entities associated with a given user.

At the heart of the system is a single central server application written in Java 2 running under Linux. It provides the following key functions:

- centralized storage for messages and user profile information,
- communication with the SMS Gateway,
- hosting of the web portal.

As illustrated in Figure 4, the system utilizes both wireless (802.11b) and wired Ethernet network infrastructures. In order to support the reception of SMS messages, the central server communicates with a Wavecom DB02 GSM terminal over a serial link.

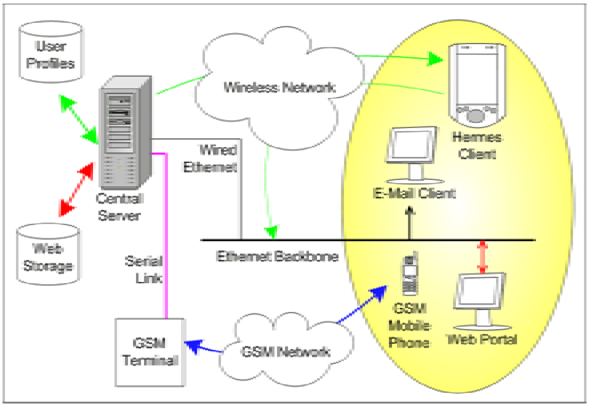


Fig. 4. The system architecture of Hermes

Messages left by visitors on Hermes displays are encoded using run-length encoding before being sent to the server, where they are re-encoded as GIF files. These file are then made available using an Apache web server.

The web portal is implemented using Java servlets running on a Jakarta Tomcat servlet runner. This enables the dynamical generation and publication of html web pages (see Figure 3) using the Java language and straightforward integration with other components using Java RMI.

3.2 Current Deployment

At the time of writing we have 10 units deployed on one floor of the Computing Department at Lancaster. The majority of units are situated outside the offices of computer scientists. However, in order to avoid including only ‘techies’ in our study, two of the units are situated outside the offices of departmental secretaries, and another outside the office of a sociologist (and certainly not a techie).

4 Experiences of Long Term Deployment

To investigate the use of Hermes over the long term obviously requires regular use during that time. This is perhaps one of the more demanding issues inherent to this system, and where it contrasts with most comparable projects, which have tended to produce few ‘end systems’ and where not deployed widely or over long periods (examples include [6],[5], [7]).

We require users to integrate the Hermes system into their current patterns of use, for this to happen there will always be an initial cost-benefit trade off for users: It generally seems to take adaptation, consideration and time to integrate new technology into a daily routine. When designing Hermes the basic functionality we attempted to provide was that of paper Post-it® notes placed on office doors & message ‘whirlers’ as shown in Figure 5.

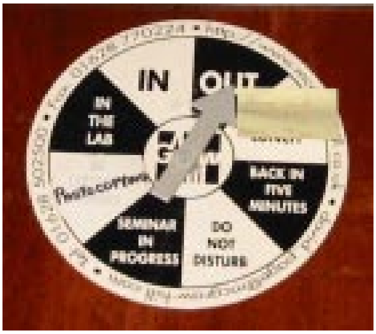


Fig. 5. Typical message 'whirler'

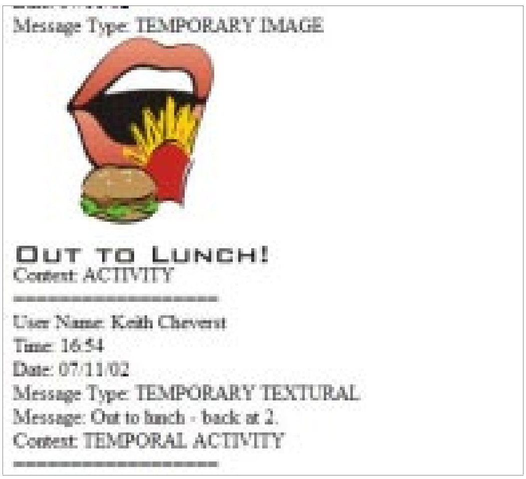
Augmenting an existing system proves advantageous as there is always a backup system for users to rely on if there is a temporary fault with Hermes. However, the problem is ensuring that people use Hermes instead of these traditional systems. One of the most disappointing scenarios we have encountered is users placing Post-it® notes next to working Hermes displays.

From our research it seems that there are two important factors to encourage and maintain use of Hermes, ease of use and trust. We found that there is a limit to how much effort a task requires before it is unfeasible as part of a daily routine. This limit proved to be lower than originally thought, and we have continually improved ease of use though the addition of various new software features.

The issue of trust seems to be linked directly to dependability, users will only bother to use a new system, let alone as part of their daily patterns, if they think it works. This seems a logical requirement from a user's point of view, but essentially translates to very high levels of reliability. This is difficult on a system such as Hermes, especially given the current hardware and software platforms.

4.1 Logging the Use of Hermes

All aspects of the Hermes system are logged continually. This ranges from the actions generated by UI components on Hermes displays to messages shared, sent and received by users. All information that we wish to log is sent to the central server to be recorded. We maintain this level of logging as it is hard to pre-empt what information may need to extract from the logs in the future. Recently we need to investigate the level and type of context sharing taking place through Hermes by its users. Extraction of messages shared by users from the logs proved straightforward, though we had no mechanism in place for classifying the context shared. Extracting this information involved tagging each message (of which there were over 300) by hand using the categories of activity, temporal and location (a small snippet of the log is shown in Figure 6).



```
Message Type: TEMPORARY IMAGE
OUT TO LUNCH!
Context: ACTIVITY
========================
User Name: Keith Cheverst
Time: 16:54
Date: 07/11/02
Message Type: TEMPORARY TEXTURAL
Message: Out to lunch - back at 2.
Context: TEMPORAL ACTIVITY
========================
```

Fig. 6. Information extracted from Hermes logs

Analysis of the logs showed that the vast majority (over 83%) of messages set shared some aspect of context. More details of the results of this analysis can be found in [2]. We also found similar use of sharing of context when messages were left remotely (via mobile phone), see [2] for further details.

4.2 Technical Problems Encountered

Many of the initial problems after deploying Hermes displays seemed to be attributed to the Java VM used, drivers for the various wireless network cards, and the PocketPC2002 operating system. Individually or together, running continually, these elements have proved a source of instability, occasionally crashing the Hermes displays. These crashes seem to have been reduced with updated versions of the Java VM, device drivers for the wireless network cards and PocketPC2002 updates. After initial testing is was intended to rivet the tops onto the cases for security, though to-date we have not felt confident enough in the stability of the devices, so for the moment the tops are screwed on. A current solution to this problem is to route the reset switch on the PDA device to inside the office of the owner.

A second source of problems is the experimental wireless network used by the Hermes displays. The major problem is that of signal strength. The displays can be thought of as 'thin' clients, little information is stored at them (due to their limited storage, small processing capacity and relative unreliability), and the majority of information is stored at the central server. Because of this most user interface actions (such as logging in, leaving a message etc) on the Hermes displays require communication, via Java RMI, with the server. We have found that RMI calls using our present Java VM seem very sensitive to network quality, and consume large amounts of CPU time. This means that if a user attempts to perform an action requiring communication with the server, and there is a temporary degradation in wireless signal strength (for example due to users standing in front of a Hermes display), the application may halt temporarily. During this time the application will no longer respond to user input, giving the impression that the application or the system has crashed. Unsurprisingly this tricks users into thinking that Hermes does not work properly and discourages use.

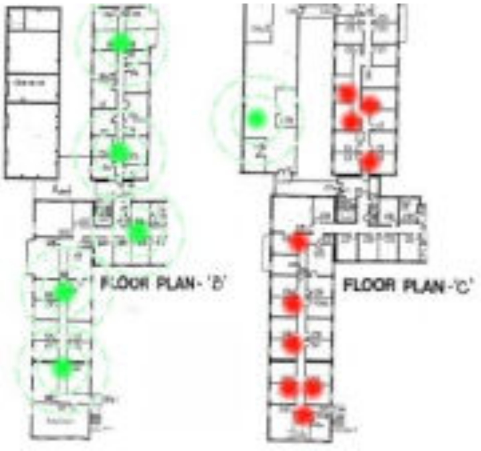


Fig. 7. Location of Hermes displays on 802.11b access points

Initially the display cases were completely enclosed aluminium boxes, this design worked well for areas with good signal strength, but in low strength areas the signal was significantly reduced or seemed to be blocked for periods. This problem is aggravated by the fact that the wireless network signal is absorbed by humans. When multiple people gather round a unit (for example to see a demonstration) the network signal can be blocked entirely. Initially it was assumed that the wireless base stations used by the devices were located in the false ceilings of the corridors, though after investigation it was found that all but one of the wireless access points (AP's) were actually located in the false ceilings of the floor below (this is shown in Figure 7, where Hermes displays are shown in red and wireless base stations in green).
This seems to lead to wireless signal 'hot spots' above the access points, though further away the signal strength seems to dictated by the structural anomalies of the building (e.g. Expansion joints between walls significantly reduce signal).

4.3 Current Solutions

After experimentation, it was found that removing sections from the top part of the case at the front and back near the protruding part of the wireless network card significantly increased the signal strength. This small modification was so successful the decision was taken to modify all the currently deployed units (shown in Figure 8).

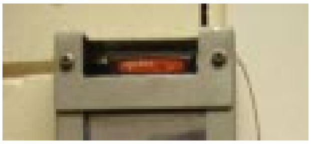


Fig. 8. Top Section of Modified Hermes display housing

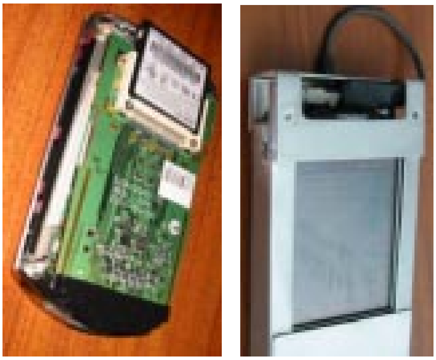


Fig. 9. Modified Compaq iPAQ

We are currently developing a prototype display using a wired ethernet card which we hope will improve users' perceptions of reliability. This has proved challenging due to hardware constraints, the PDA in use, HP Jornada 568, only accepts type I compact flash cards, and type 1 ethernet cards have proved relatively scarce. For various reasons, including the need to utilize existing hardware, our current solution is to use a modified Compaq IPAQ and compact flash expansion jacket with a type 2 wired ethernet card.

We found that an iPAQ without the rear casing (but including battery) along with a Compact Flash jacket without any of the casing (see Figure 9 - left), would fit in our housing.

The housing did require slight modifications (as can be seen in Figure 9 - right) to accommodate the slight 'bulge' at the top of the iPAQ case. Also, removing the housing from the connector 'dongle' of the ethernet card allowed it to fit in the space designed for the protruding section of a compact flash wireless network card (this can be seen in Figure 9 - right). It should be noted that all of the exposed circuit boards and electronic components are entirely insulated before the device is placed in the housing.

Presently there are no viable methods we can use to improve the current hardware and software platforms to provide the high levels of dependability we need. It is simply not possible to change or upgrade all of the Hermes display hardware. Additionally it is not always the hardware at fault; several pieces of 3rd party software (among other things) have proved themselves unreliable. The key to improving dependability and users' perception of dependability may not lie solely in preventing faults. We feel it can be improved through accepting that faults may occur, and providing measures to minimise downtime and inform users of the system status.

Previously, we relied on users notifying us when their display had crashed, or through the chance encounter of noticing when one wasn't working. Our present solution is to use a Management Agent to monitor all the displays and provide (among other features) automated notification on display failure.

5 Supporting and Managing Hermes

The main motivation for the Management Agent is to provide a system to monitor the Hermes displays, so both system administrators and users can be alerted to display failure. While designing this new system we also took the opportunity to include desirable features we felt would be useful. The following sections explore the important aspects of the Management Agent.

5.1 Notification

One of the key methods we identified to increase a users trust of their Hermes display is providing them with notification of system failure. Initially an e-mail list was set up containing the addresses of all door display owners, this is used to notify users of scheduled system downtime for upgrade work etc. This idea has been extended using the Management Agent to provide e-mail notification on individual Hermes display failure to both the owner of the Hermes display and system administrator. A similarly important feature is that of notifying user when their Hermes display is working again.

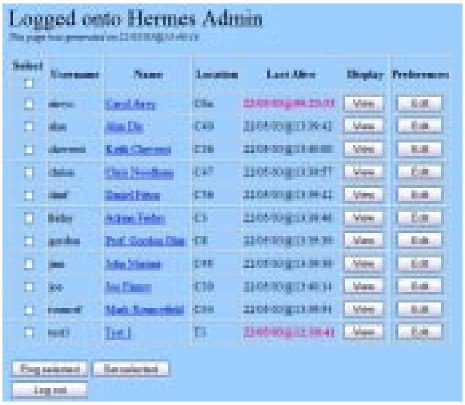
This new functionality has only recently been enabled, though it is hoped that this extra information will increase users trust in Hermes, as they will always be aware of whether their Hermes display is working or not. There is also the added advantage that the system administrator learns of a Hermes display failure at the earliest opportunity. Enabling any problems to be resolved quickly and improving overall reliability.

5.2 Management Agent Web Portal

The Management Agent also includes a web portal, accessible by Hermes user with correct privileges. This allows easy access to management features both locally and remotely. Presently there is only a single access level, allowing access to all management features. However this may be extended, allowing subsets of features to users who may need them.

5.2.1 Monitoring

The main web portal screen is shown in Figure 10, and shows the status of the Hermes displays for all users in a table, one row for each user.

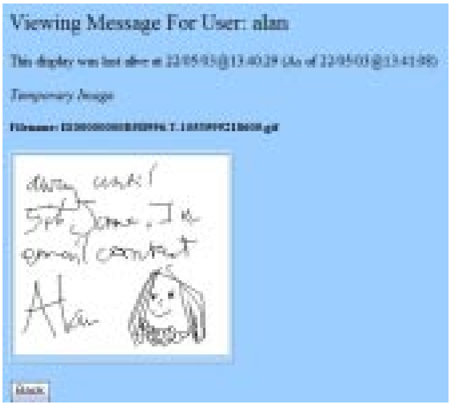


Logged onto Hermes Admin

This page was generated on [illegible]

Select □	Username	Name	Location	Last Alive	Display	Preferences
□	[illegible]	[illegible]	[illegible]	[illegible]	View	Edit
□	alan	Alan Dix	C43	22/05/03@13:39:42	View	Edit
□	[illegible]	Keith Cheverst	C38	[illegible]	View	Edit
□	[illegible]	[illegible]	C47	[illegible]	View	Edit
□	[illegible]	Daniel Fitton	C36	[illegible]	View	Edit
□	[illegible]	Adrian Friday	C3	[illegible]	View	Edit
□	gordon	Prof. Gordon Blair	C8	[illegible]	View	Edit
□	[illegible]	John Mariani	C48	[illegible]	View	Edit
□	joe	Joe Finney	C50	[illegible]	View	Edit
□	[illegible]	Mark Rouncefield	C38	[illegible]	View	Edit
□	test1	Test 1	T1	[illegible]	View	Edit

Fig. 10. Web portal main screen



Viewing Message For User: alan

This display was last alive at 22/05/03@13:40:29 (As of 22/05/03@13:41:00)

Temporary Image

Filename: [illegible].gif

away work!
5th June, I'm
email contact
Alan

Fig. 11. Viewing a users' Hermes display

It shows their username, real name (an e-mail link), office number and the last date and time their Hermes display was know to be alive.

If a user's Hermes display has not communicated with the server for a period of time (and assumed to be 'dead') the entry is highlighted in red. Each row in the table provides a 'View' button to see what is currently displayed on the user's Hermes display (Figure 11).

A second button is also provided to view and edit users personal preferences. Another important feature incorporated is that of 'pinging' a Hermes display, this feature initiates a sequence of remote calls to and from a Hermes display to verify that the Hermes application and wireless network are functioning correctly. If a 'ping' is successful the server has an updated 'last alive' time for that Hermes display, which is reflected in the web portal.

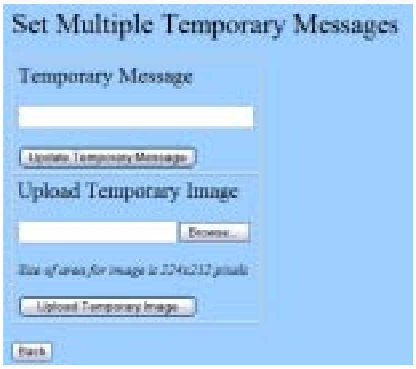
To initiate a 'ping' on a Hermes display the administrator ticks the check box on the appropriate row in the table, then selects the 'Ping Selected' button. This design allows for a simple interface and allows multiple Hermes displays to be ‘pinged’ simultaneously.

These features allow a system administrator to monitor all Hermes displays from a single point. This is useful for many reasons, such as:

- to check if a Hermes display is actually malfunctioning,
- to check where an update has been successful, and
- to monitor the Hermes system remotely.

5.2.2 Additional Administrator Features

When designing the Management Agent functionality was included allowing a Hermes administrator to change the temporary message for one or more users. When selecting the users to change the temporary message for, this feature is similar in use to pinging Hermes displays (selection using check boxes). When the administrator selects the ‘Set Selected’ button they are presented with the dialogue shown in Figure 12.



Size of area for image is 224x212 pixels

Fig. 12. Setting temporary messages for multiple users

This allows the administrator to set a textural message, or upload an image file to be used. Once the temporary messages have been changed the administrator may use the ‘View’ feature (as shown in Figure 11) to ensure that displays have been updated successfully. When a user’s temporary message has been changed from the Management Agent, they are informed via an e-mail also including reassurance that they can change or remove this message at any time.

We hope to extend this feature so users such as departmental secretaries (etc) may be able to use multiple Hermes displays to display pertinent messages and reminders. These messages may include reminders events such as a fire alarm test, notification of an important meeting, department open days etc.

5.3 Current Findings

The Management Agent has been in place for approximately 10 weeks and although initial signs are encouraging the period of time has not been long enough to justify an analysis of whether it has increased user trust of the system. All use of Hermes is logged so it will certainly be possible to see whether use of Hermes has increased, and additionally to interview users to find their opinions.

The mechanism to detect whether a Hermes display is still working (to provide notification on failure etc) relies on an RMI call back made to the server every minute by the display. A timer is used to detect a problem if the server does not receive any communication from a client for a specified period of time. During development this time was initially set to 5 minutes, though after occasional ‘false alarm’ notifications of dead displays this was increased to 15 minutes. After deployment it was found this interval was still too short, especially in one area of very low wireless network signal. The Hermes displays could loose signal for around 20 minutes then start working again normally. The timeout was further increased to take this into account, though later the decision was taken to use wired network for that display due to the low signal problem.

The greatest benefit so far from the management agent has been the automated notification of malfunctioning Hermes displays, allowing the problem to be remedied straight away. It is also very for Hermes administrators to be able to monitor the displays and use to ping feature to tell if a display is working immediately.

6 Future Work

Out immediate plans are to provide a new mechanism allowing users (such as departmental secretaries) to set temporary messages for multiple Hermes displays. The issues arsing from the devolvement of this feature will hopefully prove useful, especially as we plan to investigate a departmental navigation system to run on the Hermes displays. The whole area of appropriating a personal public display for other uses enables many appealing applications, and will be interesting to explore.

The central point of control that the Management Agent provides can certainly be exploited through extra features. Providing remote control of the Hermes displays would be very useful, perhaps allowing the Hermes application to be terminated and restarted if it crashed, and allowing updated software versions to be sent to each display. Presently these operations are done manually as all the displays are situated relatively close to each other, if they were distributed any further away this may prove a problem. Throughout the Hermes project we have applied a participatory design based approach to development, this will be carried through to the Management Agent as we refine the current design and add new features.

7 Related Work

Work has previously been conducted on this specific area at Georgia Tech. However, there work on dynamic door displays [6] appears to have stopped before any significant deployment or reasonable evaluation of the system was able to take place.

McCarthy developed the ‘OutCast’ service to provide “a personal yet shared display on the outside of an individuals office” [5]. Although OutCast supports a range of features, including a message box feature, it does not support remote interaction and does not allow users to utilise the displays of other office workers. The system is also reported as having only one unit deployed.

Research on utilizing doors as interruption gateways and aesthetic displays is being conducted at CMU [7]. The most similar aspect of this work is its concern for supporting the personalization of information on office doors which it supports in a very rich way through the use of projection systems.

Perhaps the most striking similarity to the motivation and ideas of the Hermes system is the concept of ‘infoDoors’ described by Ben Schneiderman [9]. In common with our work, Schneiderman envisions buildings with displays outside every office and supporting a variety of uses such as displaying personal messages.

8 Concluding Remarks

This paper has described our experiences of managing and maintaining the Hermes system, a system of deployed office door displays which provide asynchronous messaging services for users in the computing department of Lancaster University. We believe that, in general, the management and maintenance of deployed ubicomp systems is a critically under researched area and we hope that some of the issues and anecdotes presented in this paper and based 15 months experience deploying and maintaining the Hermes system will help to contribute some useful knowledge in this area.

Acknowledgments. The work described in this paper has been conducted through the auspices of the CASCO project which is funded by the UK Engineering and Physical Sciences Research Council. We would also like to acknowledge the support of members of Lancaster University’s computing department.

References

1. Cheverst, K., Fitton, D., Dix A. “Exploring The Evolution Of Office Door Displays”, In K. O’Hara, M. Perry, E. Churchill, D. Russell (ed) Public and Situated Displays: Social and Interactional aspects of shared display technologies. To appear
2. Cheverst, K., Fitton, D. Dix, A., Rouncefield, M. “’Out To Lunch’: Exploring the sharing of Personal Context through Office Door Displays” (OZCHI 2003) To appear
3. Cheverst, K., Fitton, D., Dix, A., Rouncefield, M. “Exploring the use of Remote Messaging and Situated Displays”, In Proc of Fifth International Symposium on Human Computer Interaction with Mobile Devices and Services (MobileHCI ‘03). To Appear
4. iButton home page: www.ibutton.com
5. McCarthy, J., Costa, T., Liongosari, E. “UniCast, OutCast & GroupCast: An Exploration of New Interaction Paradigms for Ubiquitous, Peripheral Displays”, Workshop on Distributed and Disappearing User Interfaces in Ubiquitous Computing at CHI 2001 (2001)
6. Nguyen, D., Tullio, J., Drewes, T., Mynatt, E. “Dynamic Door Displays.” Unpublished, Georgia Tech: www.cc.gatech.edu/fce/ ecl/projects/drewes/DynDoorDisplays.pdf

7. Nichols, J., Wobbrock, J., Gergle, D., Forlizzi, J. “Mediator and Medium: Doors as Interruption Gateways and Aesthetic Displays” In Proc. of DIS'2002, London, UK. June 25–28. pp. 379–386. (2002)
8. NSiCom CrEme home page: www.nsicom.com
9. Schneiderman, B. “Leonardo’s laptop: Human Needs and the New Computing Technologies”, MIT Press, ISBN: 0-262-19476-7 (2002)
10. Weiser, M. “The Computer for the 21st Centuary”, Scientific American, Vol. 265 No. 3, pp. 66-75. (1991)

Applications of Vision-Based Attention-Guided Perceptive Devices to Aware Environments

Bogdan Raducanu and Panos Markopoulos

Faculty of Industrial Design, Technical University of Eindhoven
Den Dolech 2, PO Box 513, 5600MB Eindhoven, The Netherlands
{b.m.raducanu, p.markopoulos}@tue.nl

Abstract. This paper discusses a computer vision based approach for enhancing a physical environment with machine perception. Using techniques for assessing the distance and orientation of a target from the camera, we detect user presence and estimate whether an object of interest is the focus of attention of the user. Our solution uses low-cost cameras and is designed to be robust to lighting variations typical of home and work environments. We argue why this approach is a useful component for the incremental construction of aware environments and discuss some practical applications of using such a system.

1 Introduction

1.1 Overview

A key component of the Ambient Intelligence vision [1] is the capability to interact with a computational environment in natural and personalised ways. One way to achieve naturalness is by means of *implicit input* [2], Implicit input entails that our natural interactions with the physical environment provide sufficient input to a variety of non-standard devices without any further user intervention. Such automatically captured input contrasts the current model of interaction where the user has to perform several secondary tasks relating to the operation of the computing device in order to achieve their primary task. Attaining this capability involves in endowing everyday objects with machine perception capabilities. By machine perception we refer to the whole class of sensing and pattern recognition techniques that can be deployed to sense and interpret aspects of activities taking place within the physical environment of interest.

Potentially, machine perception offers significant advantages to users. It can help disambiguate input, e.g., knowing who's talking to whom, which display is attended to so as to route system output, etc. Machine perception can be achieved with a variety of technologies (pressure sensors, video cameras, radio-frequency tags, fingerprint readers), which are integrated in objects commonly found in our environment (chairs, tables, displays). Each of these technologies has their respective advantages

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 410–418, 2003.

and disadvantages concerning robustness, complexity, costs and the type of inferences that can be made about user activity. This paper discusses some applications of computer vision techniques to support implicit input in ubiquitous computing environments. In particular, we discuss the detection of person proximity to an object of interest and whether this person faces towards the object. This helps to estimate whether that objects becomes user's focus of attention. The intuitive idea that people will face objects they are interested in is supported by some empirical research at Microsoft [4]. Computer vision offers several advantages over competing technologies (e.g., movement sensors, ultrasound) to answer this question.

1.2 Indoor Person Presence Detection

The problem of indoor person presence detection can be addressed at different scales and varying resolutions [3]. At a building level, one might be interested to know which room people are in, e.g. for supporting an Intercom type of application [12]. At a room level, we might wish to broadly detect which part of the room a person is at (near the window, door, table). At sub-room level we might want to know more precisely the coordinates of a person, or whether this person is directing his attention to a particular object of interest. Depending on the type of detection we address, several ranges of error are tolerable. For example, an error of a few meters is considered a good approximation for presence detection at building level, while for other tasks, e.g., whether a user is attending to a small display, the error cannot be greater than a few centimetres.

Current solutions for building level detection are based on wireless communication devices. A classical demonstration of location awareness is the Active Badge system [15] developed at the Olivetti research labs, based on infrared emitting badges. In [3], they present a system that relies upon RF identifiers placed in the shoes of users and floor-mat antennas. Based on a history of information recorded by the receiver the system can estimate whether the person is in a room or not. A more recent method based on ultrasound active badges is presented in [10].

For room-level presence detection, in [13] they created a “smart floor”, whose purpose is to identify and track the people stepping on it. The floor contains force measuring load cells. In [3], [7] and [9] they use computer vision in order to track several persons in real-time. The persons' location in the room is established based on a combination of knowledge of the cameras' relative location, fields of view of the cameras and heuristics on the movements of people. Compared to computer vision based solutions, such technologies are robust to varying lighting conditions but do not help with verifying focus of attention (where a person is facing to).

Several computer vision based techniques have been proposed for sub-room level detection. These rely either on localizing a badge carried by the user [8] or on face localization [5], [11] and [14].

This paper discusses a method for estimating the attention of a person towards an object, by detecting whether a special badge carried by the person is facing towards the camera and whether the distance between the badge and the camera is below a certain threshold. It is assumed that the camera is placed in such a position in order to optimise the interaction between the user and the device that is attached to. We want to avoid situations in which, for objects of very large dimensions (like wall-mounted displays, for instance), when the user is facing towards the object the badge cannot be visible by the camera.

The paper is structured as follows. Section 2 describes the system developed. In section 3 we discuss some potential applications that are currently under way. Finally, in section 4, we will present our conclusions and draw some guidelines for future work.

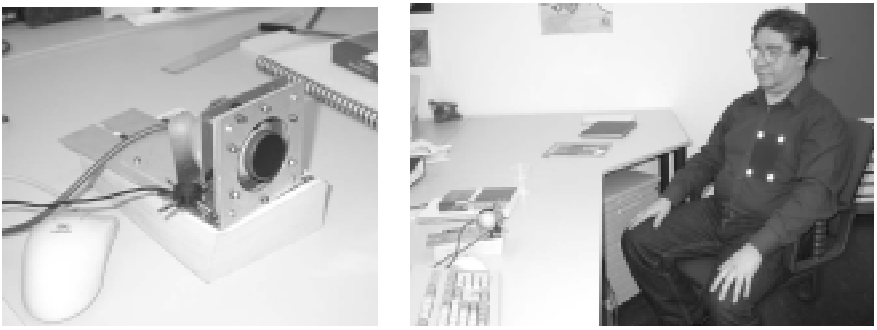
2 System Description

2.1 Overview

Our system consists of two components (see Figure 1):

- a perceiving component, represented by a Logitech™ webcam with an infra-red filter and an array of infra-red LEDs placed around the camera in form of a ring;
- a passive component, represented by a badge, which has reflector tape patches attached on it.

The role of the filter is to let pass only those frequencies of the light that are close to the infrared rays spectrum. By using this kind of filter, the camera will perceive mostly the light that is reflected from these reflector patches. In consequence, background information is discarded from the beginning, making the image analysis process much simpler.



(a) (b)

Fig. 1. Experimental hardware setup: (a) a Logitech webcam provided with an IR-light source and IR filter and (b) the first author wearing the target with IR reflector material and facing the camera

The badge is a rectangle made of rigid paper and has attached patches of reflector tape (in shape of discs), on each of its four corners. A very important property of this material is that it reflects the infrared light back, on the same direction it came from the source (infrared LEDs). The size of the badge is 10x15 centimetres and the discs have a diameter of 2 centimetres.

2.2 Description of the Method

The standard approach for estimating the 3D coordinates of a point in space, a stereo vision system is needed. Thus, the 3D coordinates are estimated based on the pixel disparity, i.e. the difference in object pixels' location in the two images. One of the disadvantages of using a stereovision system (besides its higher cost and necessity for specific hardware) is that accidental changes in the orientation of one of the cameras require a recalibration of the whole system. This makes stereo-vision based solution insufficiently robust for dynamic environments like the home or the office.

Here we explore an alternative, i.e. to estimate the 3D coordinates based on the information about the points and lines whose perspective projection we observe. Such relations with the perspective geometry constraints can often provide enough information to uniquely determine the 3D coordinates of the object. Such knowledge can come about when we have a model of the object being viewed in the perspective projection. The technique to inference the 3D point coordinates when knowing its 2D coordinates on the image plane of the camera is called "inverse perspective projection". In our case, we use a technique that is described in [6], which allows the 3D reconstruction based on the observed perspective projection of two parallel line segments (the lines connecting the centres of the two patches situated along the vertical edge of the badge, for instance). This method does not use any information regarding camera's absolute orientation in the scene. We always express the relative position of the person with respect to the camera. In consequence, small modifications in the camera position will not affect system's performance.

2.3 System's Performance

Since we looked for a low-cost solution to our problem, we tested the algorithm on a PC of modest performances with 128 MB of RAM and a processor of 730 KHz. We set the frame rate of the webcam at 10 frames/sec, which is acceptable for real-time tracking.

The experiments performed were intended to assess the accuracy of the distance measured by the proposed algorithm. In the case of infrared technology, the only factor that can affect the performances of the algorithm described above is the amount of infrared radiation presented in the environment. Empirical experiments demonstrated the robustness of our system in case of diffused natural light, considered during different moments of the day, and also artificial light. In our view, these are the most likely scenarios in an office or home environment. We found that only direct sunlight, presented in the scene covered by the camera, can affect system's performance.

We estimated the accuracy of the distance measured when the target was positioned at orientations of 0, 30 and 45 degrees with respect to the camera. The distance range was set between 40 centimetres and 2 meters. The error in distance estimation, at 2 meters (which is the maximum distance perceived by the system in its current configuration), was about 4%.

For most applications inside a home or a small office space, it can reasonably be expected that the threshold distance relating to when a person is paying attention to a specific device, should fall within the mentioned range. The low-end of the distance range would corresponds when the user is in front of a PC, while the high-end would correspond for the case of "wall mounted display".

3 Applications

In what follows, we sketch out some applications of this system that we are currently developing to demonstrate and to test the concept of attention detection.

3.1 Magnifying Map

The first demo of our system is envisioned as an application for info-kiosks, which may be found at tourist offices, for instance, and offer practical information for the newcomers. When there is no person in front of the camera, the system will display a neutral screen. Once the system notices the presence of a person (for a couple of seconds), it assumes that as an "attention-getting" event and displays a map on the screen. The level of details shown on the map is depending on the distance of the person with respect to the camera. If the person is situated afar (about 2 meters), then the system displays a general map of the city. As the person approaches, the level of details changes accordingly, so that at a closer distance (half meter), the current neighborhood is highlighted on the map. When the person moves away, then the system returns to its stand-by mode, waiting for the next visitor. This behavior can help overcome the limited resolution of large-scale electronic displays when compared to printed paper (for which more detail can be attained simply by walking up to the map).

3.2 "Follow-Me" Displaying Message System

This application is intended to illustrate how the contextual information can be used in order to have the incoming messages displayed at the most convenient location. We install our system on several workstations (WS). These WSs, on their turn, are connected to a message delivery server. In Figure 2 we depicted a very simplified sketch of the configuration described.

Each time a WS notices the presence of a person nearby, it sends an event to the server (the WS can send any ID, maybe the most easy way is to send its own network address). The server keeps track of the last WS where the person "has been seen", so that when a message destined to that person arrives to the server, it knows which is the terminal closest to him/her. This way, the person doesn't have to go to a specific location in the building in order to check for new messages. This is a realistic scenario, taking into account that is very common that a person can be present, throughout the day in several locations, not only in his/her office. As an example, we can refer to the university environment, where the researchers, besides their office, often has to go to a lab to do some experiments or have to attend a discussion session in a meeting room.

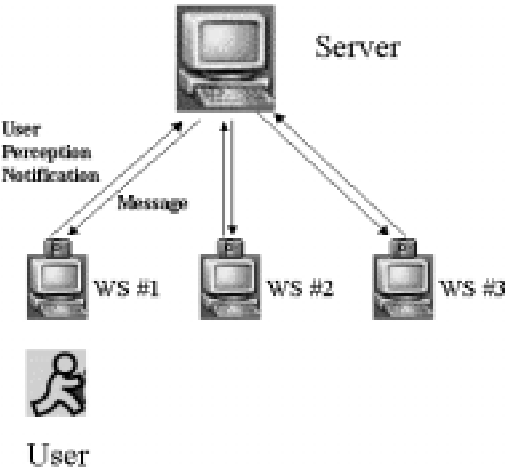


Fig. 2. A sketch of the system architecture used for the distributed messaging application

On the other hand, this application shows that the creation of an aware environment can be addressed incrementally, starting with one perceiving device and dynamically add others, as they become available.

3.3 Discussion

The presented applications are limited to a single (anonymous) user. Other applications that can be envisioned (using this limitation) are related with a museum environment. Detecting a gallery visitor, who wears a badge, in front of a painting, may then trigger the playback of recorded audio information related with it.

The fact that all users of the system wear a unique badge and are indistinguishable offers advantages and disadvantages. In some cases, anonymity may be preferred by users, for instance in the museum example. In other cases, correct allocation of a display would benefit from user identification, to show a personalized message on the monitor he is facing to. The current limitation can be overcome, by extending the current badge with a new one, having encoded (through a number of dots) the identity of the person who carries it. This way, we could enhance the contextual information, by adding person's identity. This thing would be very useful for the second application mentioned in this section ("Follow-Me"), because the system could know which person actually stays in front of it. By delivering only personalized messages, user's privacy can also be protected.

In order to experiment with an 'aware display' that will be used in the context of messaging applications, we have constructed a prototype screen enhanced with the camera as shown in Figure 3. This device, which is under construction, will enclose a single-board PC [16] and will offer PC functionality from its touch-screen. This packaging of our system is crucial to enable field testing of awareness applications, so that they will be acceptable to install temporarily in people's home. It also serves as a crude prototype of the devices we expect to furnish an aware home.

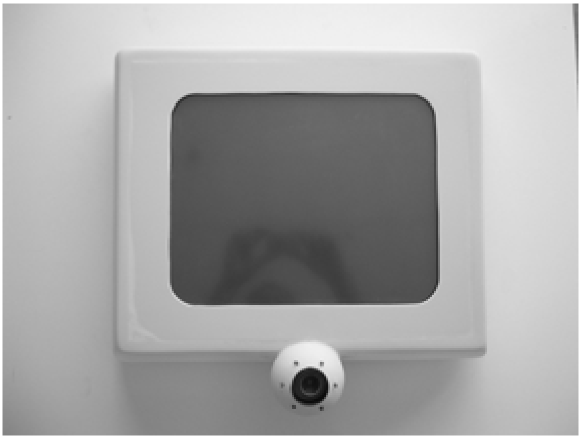


Fig. 3. A "transparent" representation for an ubiquitous computing component: a touchscreen enhanced with perceptive capability due to the embedded webcam

4 Conclusions and Future Work

In this paper we proposed a new, low-cost solution to detect user's presence at sub-room level resolutions. This approach presents a high robustness against varying lightning conditions. The detection range is between 40cm and 2m, which makes it suitable for a large variety of applications. In consequence, this will allow a redefinition of the term "near", depending on the context the application will be developed for. Applications that are currently under development have been discussed and also other potential ones have been proposed.

While the presented method gave some very encouraging results, it obligates the person to wear this target attached to his close. In everyday context this can be an onerous obligation for the user. On the other hand, it provides a direct mechanism to the user to control when his activities are monitored and responded to: the user can simply remove the badge.

There are several alternative mechanisms to detect user proximity, e.g., using RFID tags, or using ultrasound signals. These approaches can work very accurately in domestic environments and can be very robust when there is no interference with other electronic devices. However, computer vision is better suited for the specific problem of detecting the direction the user is facing in. In our next step we shall investigate the feasibility of detecting user's attention without the need for reflecting badges, by directly detecting the human face and head pose in the scene.

Acknowledgments. The authors want to thank to Martin Boschman for the design of the infrared circuitry and to Charles Mignot for making the "aware display".

References

1. Aarts, E., Harwig, R., Schuurmans, M. Ambient Intelligence. In: Denning P.J. (ed.) The Invisible Future. McGraw Hill New York (2001) 235–250
2. Abowd, G.D., Mynatt, E.D.: Charting Past, Present and Future Research in Ubiquitous Computing. ACM Transactions on Computer-Human Interaction (2000), 7(1):29–58
3. Aware Home Research Initiative. Georgia Institute of Technology, http://www.cc.gatech.edu/fce/ahri/
4. Brumitt, B., Cadiz, J.J.: Let There Be Light: Comparing Interfaces for Homes of the Future. Proceedings of Interact'01, Japan (2001) 375–382
5. Darell, T., Tollmar, K., Bentley, F., Checka, N., Morency, L.-P., Rahimi, A., Oh, A.: Face-Responsive Interfaces: From Direct Manipulation to Perceptive Presence. Proceedings of Ubicomp, Sweden (2002) 135–151
6. Haralick, R.M. and Shapiro, L.G.: Computer and Robot Vision. Addison-Wesley, New York (1993)
7. Intille, S.S., Davis, J.W., Bobick, A.F.: Real-Time Closed-World Tracking. MIT Media LabTech Report TR-403 (1996)
8. De Ipina, D.L., Mendonca, P.R.S., Hopper, A.: TRIP: A Low-Cost Vision-Based Location System for Ubiquitous Computing. Personal and Ubiquitous Computing Journal, Springer (2002) 6(3):206–219
9. Krumm, J., Harris S., Meyers B., Brumitt, B., Hale, M., Shafer, S.: Multi-Camera Multi-Person Tracking for EasyLiving. Proceedings of 3rd IEEE International Workshop on Visual Surveillance, Dublin, Ireland (2000) 3–10
10. Krumm, J., Wiliams L., Smith, G.: SmartMoveX on a Graph – An Inexpensive Active Badge Tracker. Microsoft Research, Technical Report MSR-TR-2002-70 (2002)
11. Nakanishi, Y., Fujii, T., Kiatjima, K., Sato, Y., Koike, H.: Vision-Based Face Tracking System for Large Displays. Proceedings of Ubicomp, Sweden (2002) 152–159
12. Nagel, K., Kidd, C.D., O'Connell, T., Dey, A.K., Abowd, G.D.: The Family Intercom: Developing a Context-Aware Audio Communication System. Proceedings of Ubicomp, Atlanta, USA (2001) 176–183

13. Orr, R.J., Abowd, G.D.: The Smart Floor: A Mechanism for Natural User Identification and Tracking. GVU Technical Report GIT-GVU-00-02, Georgia Institute of Technology (2000)
14. PosterCam Project: Compaq Research:
http://crl.research.compaq.com/vision/interfaces/ppostercam/default
15. Want, R., Hopper, A., Falcao, A., Gibbons, J.: The Active Badge Location System. ACM Transactions on Information Systems (1992) 10(1):91–102
16. Workbox Computer: http://www.zerez.com/producten/workboxp3/techspecs.htm

Addressing Interpersonal Communication Needs through Ubiquitous Connectivity: Home and Away

Natalia Romero[1], Joy van Baren[1], Panos Markopoulos[1], Boris de Ruyter[2], and Wijnand IJsselsteijn[1]

[1]Technische Universiteit Eindhoven, Den Dolech 2, 5600 MB Eindhoven, The Netherlands
{N,.A.Romero, J.K.v.Baren, P.Markopoulos,W.A.IJsselsteijn}@tue.nl
[2]Philips Research Eindhoven, Prof. Holstlaan, 45656 AA Eindhoven, The Netherlands
boris.de.ruyter@philips.com

Abstract. This paper describes a user study regarding the human need to stay in touch with closely related people. The study was a combination of interviews and diaries. This user needs analysis has informed the design of a novel end-to-end communication system for helping closely related people, who are spread geographically, to stay in touch. The design concept is described in brief, followed by a summary of ongoing implementation and assessment work.

1 Introduction

Imagine you start the coffee machine at home and some digital picture-frame nearby displays a photograph of your elderly father - a poor telephone user- meeting one of his best friends, some time earlier in the day at the other side of Europe. Or, it would prompt you to choose among 4-5 pictures of today's activities, for one to display this evening on the frame on his mantelpiece. It is reasonable to hypothesize that the next phone-call will come sooner and be more lively and satisfying for both.

This scenario provides an example of an *awareness system* supporting existing and valued social relationships. Communication in this example does not serve a specific task and is effected without the explicit expression of intent by the sending or the receiving party. Such communication can potentially address affective needs of people living alone or may strengthen emotional ties between remote friends or family members. Further, the awareness system is seen as complementary to current communication media rather than a substitute for phone conversations or face to face encounters. This paper discusses some research that investigates whether this indeed will be the case and an effort to identify the requirements for the informal social use of this class of systems.

Awareness systems are a class of computer mediated communication (CMC) systems that support individuals to maintain a peripheral awareness of each other's activity, similar to that which we have of our next-door neighbours at home or our co-worker across the corridor at the office. In this paper we focus on the use of an awareness system to help people stay in touch with remote friends and family.

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 419–429, 2003.

Traditionally this need has been served by letter writing, telephony and, more recently, e-mail, SMS and Internet chatting.

Experimental awareness systems have been developed and tested within working contexts, as for example the early work in Xerox PARC [2]. The interest in the domestic or leisure use of awareness systems is more recent. Several design concepts have been discussed by the Casablanca project [6] and the INTERLIVING project [3]. To this point little is known about how to design such systems for domestic use, whether they do serve actual user needs and whether the anticipated benefits may be delivered to the user. The design of awareness systems must address requirements relating to control of such systems. For instance, should the communication be effected automatically or by an explicit user action? Should systems be allowed to automatically capture information about user activities? How can the user stay in control of private information without having an excessive workload?

Research at the TU/e and collaborative efforts with Philips Research aim to provide some answers to these questions. An exploration of interaction techniques to support impromptu messaging between family members is discussed in [11]. A recent study on communication needs of elderly [8] showed that communication with remote family members is a high priority for elderly people who live alone. Video based communication at the home is, however, not accepted for social communication. One main obstacle for this acceptance is that video captures private information (e.g., clothes, presence of others in the room, untidiness, etc.) in a way that may be threatening or annoying. Consequently, we focus on CMC that does not use video based communication but, rather, aims to help people stay in touch by selective, low effort and discrete capture and display of awareness information.

From a business perspective it is crucial to identify actual user needs and to establish whether the proposed technologies can address them. We can contrast the sometimes surprising penetration of technologies supporting socializing, e.g., SMS, e-mail, mobile telephony, with technologies that provided access to content that was not valued by consumers, e.g., WAP service for mobile telephones. Ambient Intelligence promises seamless access of content coupled with situation awareness technologies that can make a home computing platform be 'aware' of the activities inside its physical confines. In this respect, the users / residents of an Ambient Intelligence environment provide or even become the content themselves and consume it for the purposes of maintaining interpersonal relations. For example, De Ruyter et al [1] showed how peripheral awareness of friends watching the same soccer match while in different locations was found to increase their feeling of group attraction. Awareness of close friends and family seems to be a valuable benefit that Ambient Intelligence can offer to people. This paper is part of a research program that aims to assess and document this claim and to develop concepts that will support this peripheral awareness in a manner pleasurable and acceptable for end-users.

This paper discusses an extensive qualitative user study regarding the interpersonal needs that can be served by awareness systems and the communication patterns that currently address these needs. The requirements study has informed the design of a concept of an awareness system, which is currently in the implementation and testing phase. In the following we summarise the process followed during the study and its results. We then outline the To-Tell system concept for supporting the

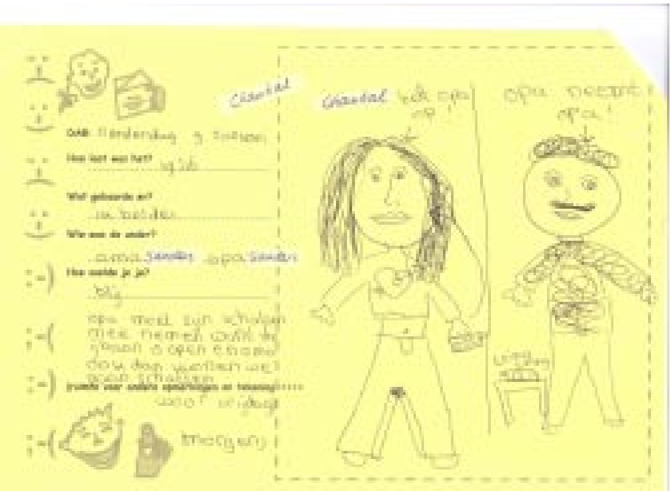


Fig. 1. In this diary page the child illustrated a contact with her grandfather.

communication activity envisaged and we discuss the planned steps for testing the concept.

2 Process of the Study

The requirements study aimed to answer the following questions:

- Why are informal social communications valued?
- What communication media currently address the need to stay connected and aware of close family?
- How are these media experienced in use?
- Which user needs should be considered for future communication systems?

A two-tier study was followed, comprising of an interview study and a diary study. This combination of techniques targeted informal, serendipitous communications that are interlaced with daily activities. The diary method offers the advantages of a prolonged data collection, that is close temporally to the phenomenon of interest (avoiding problems of recollection) and that is conducted in the context of the actual experience studied, see [10]. On the other hand, diaries can easily degrade to dry factual collections of data, so they have been combined with interviews, to obtain a deeper insight into the emotional aspects of the communication studied.

Semi-structured interviews were conducted with 17 participants: 3 family clusters and 4 additional individuals. On average, an interview lasted about 1 hour. The same interviewer conducted all interviews. A memo-recorder was used to capture the interviews and a digital camera to take pictures of the interviewees and their communication tools (see Figure 1)

The interview consisted of the following parts:

1. Communication with Family. The first part consisted of questions about communication with family members in general, so with all possible means. First, participants gave a short description of their family and were asked to estimate which percentage of all their contacts was with family members. Also, participants were asked if communication with their family was important to them and why. They also had to remember one recent contact (with a family member) that they liked very much and one contact that they did not like (critical incident technique).
2. Tech-Tour. In this part of the interview, participants showed which means they use to communicate with their family. About each medium, they were asked how often they use it, for what reasons they use it, how they feel when they use it, if they feel connected to the other person, how long this feeling lasts, what are main advantages and disadvantages of the medium and what they would like to change about it.
3. Guided Speculation. In the final part, the interviewer explained the concept of asynchronous communication and gave several examples. Participants were asked to react on the concept. They were also asked what kind of information they would like to send and for what purposes they would use an asynchronous system. They were also asked about the participation structure and whether the system should maintain a record of communication exchanges. Finally, they were asked whether they would like to send and/or receive automatically captured information.

The diary study aimed to get a closer look at when, how and why family members communicate with each other. Each of the 13 participants that constitute the 3 family clusters that were studied, filled in a diary for 1 week. By family clusters we mean individuals related with family ties, but who live in more than one household (in order to study 'staying in touch' over a distance). This reflects our ambition to focus on the potential of awareness technologies to address current social problems: elderly living alone, young adults emigrating for carrier reasons, etc. Participants were selected to include different age groups, varying levels of affinity with technology and different relationships like parent-child, grandparent-grandchild and brother-sister.

We created diaries using simple, informal language and inspiring drawings. All questions were open-ended to elicit personal, qualitative information. The participants for the diary study were asked to write down each contact with family members and answer questions about the time, duration, communication means, location, the other person, initiative, reason, effect on themselves (feelings, connectedness and after-effect) and effect on the other person. For the children, we created a separate and simpler version of the diary. They had to write down the time, describe what happened, who the other person was and how they felt. Also, they were asked to make a drawing of each contact. Figure 1 shows an example of a completed page of the diary given to children.

3 Results of the User Study

A wealth of qualitative data has resulted from the study. We summarise some of the most important points and include some quotes from the participants that illustrate these points (Note, that numerous confirmatory remarks and data are not presented for reasons of space). All quotes are translated from Dutch.

Participants unanimously rated communication with their family as very important. On average, they estimated that 52% of all their contacts are with family members. The main reasons for these contacts are staying up-to-date with events in the other's life, telling about their own life, exchanging practical information, showing interest, reinforcing the relationship and giving or receiving emotional support.

From the interviews and the diaries we can distinguish four different types of contacts: social, emotional, practical and for a special occasion (e.g. a birthday). As can be seen in Figure 2 most contacts that participants felt very positive about were of a social or practical nature. We were surprised by the fact that people felt so positive about practical contacts, so we looked in more detail at these diary entries. It appears that they were mainly contacts in which people planned a real meeting.

Participants indicated that they appreciated it if a medium (e.g. the phone) allowed them to contact others in a quick and easy way and disliked it if many actions were necessary to make contact (as is currently the case with e-mail). We note how this relates to the research of Melenhorst [9] who studied how elderly adults perceive the value of communication media. While such costs (in terms of effort to use) are important to elderly, they affect their decision to use a communication medium less than the perceived benefits. In other words, the priority of for designers should not be usability, but on delivering value in terms of the expected emotional benefits.

Effort put by the sender is appreciated and is valued, but only when it is meaningful with respect to the communication. For example, the effort taken to start up a PC, log in and send a message is not valued compared to the message of choosing a postcard to fit the personal taste of the receiver. Note, that the respective costs of the communication were not valued by the receiver. However, the effort one grandmother put into the content of her e-mail message made it very much appreciated by the grandchild who described the situation during an interview:

"Grandma sent me an e-mail that she had cleaned up the garden. She made a poem of it, about slugs".

The timing of the communication is crucial. Participants indicated that they would like to share events of their everyday life with their relatives, right at the moment when they happen. At the same time, timing for receiving a communication should be negotiated with the receivers so as not to disrupt them. For example, one informant tells:

"I was very tired and did not have the energy to talk for a long time".

Almost all participants said that abstract information (e.g., symbolic icons or text labels) regarding the availability, status or activities of their family members would not be sufficient for them to create a feeling of connectedness. (N.B. Abstract visualizations of awareness information were used in concepts such as the Family Portrait of Georgia Tech or the Home Radio of Philips Research [4]). Therefore, the system should enable users to share concrete information with each other, e.g., pictures or messages. An awareness system should provide concrete and visual information. Communication with images was required by participants, perhaps as an aid to phone conversations. Further, participants mentioned that they would value storing favourite items (messages, pictures).

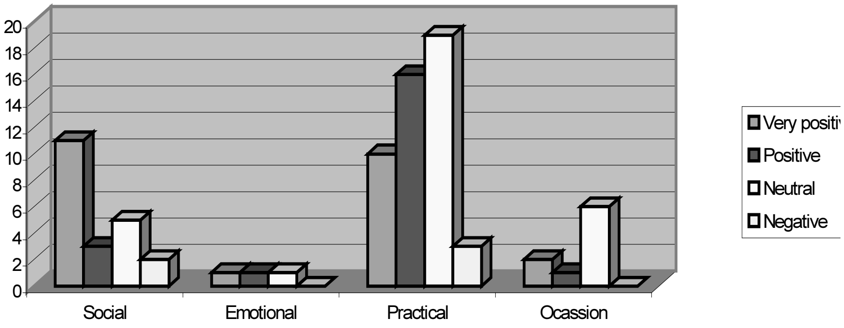


Fig. 2. Type of contact and resulting feelings.

The surprise factor is valued while routine communications are not; this observation pertains both to the timing as well as the content. E.g., one participant said:

"On New Year's eve, we came home and there was a voicemail from Laura, wishing us a happy New Year. It was a nice idea that she thought of us. I had not expected it, because she was at a party".

The opposite effect has been noted by another informant:

"My mum almost always calls me in the weekend. Most of the times it's nice, but sometimes I can tell that she only calls because she feels obliged to and then he conversation is a bit forced."

Communications for social purposes create obligations but also are valued less when they seem to be obligatory. There is reticence by participants to adopt media that will create new obligations.

Using an awareness service at home should not be disruptive to relaxation. Participants want to keep strict control of their reachability for different people. On the same token, they get frustrated when the person they try to communicate with does not respond to their efforts.

Feeling connected relates to social presence and intimacy, but can be seen as a feeling that lingers on after the communication. One participant is quoted:

"If the other person makes me feel better, I keep that feeling with me for a long time. We live very isolated, you know."

An important research aim is to seek ways to assess this lingering feeling and help create it. It also relates to some notion of social presence, i.e., focusing on the communication itself in its duration.

Participants saw value in enhancing the existing media rather than replacing them: talking about an item/message of an asynchronous service during a phone conversation, or using it as a reminder for calling someone.

An awareness system was considered by participants as a kind of a family information channel. This idea has been explored in the Home Radio project at Philips Research, see Eggen et. al. [4] and the Message Board concept of the INTERLIVING project [3].

Use in a comfortable and relaxing environment is valued (this was mentioned as a benefit of a cordless telephone or a mobile phone that you can take to a place of your choice to phone from). Participants indicated that social or emotional contacts should preferably be conducted in the home.

Primary beneficiaries for awareness systems are geographically distributed families, for cross-generational communication (parent to child and grandparent to grandchild). Interviews reported on average 52% of their communication to be cross-generational, while 45% of diary entries concerned this category.

If awareness systems do succeed in encouraging communication acts, elderly males are also important beneficiaries. Consistent with the study by Melenhorst [9], our diary study has shown that males engage in communication only with the pretext of some practical matter. Most emotional contacts were carried out by females, these were perceived most positive when one of the communicators initiated the communication to "tell' something rather than ask for advice or other information. Further all emotional contacts were between parents and children. The feelings associated with general social communications were, as expected, more positive than those associated with practical purposes.

4 To-Tell: A Privacy Respecting Way to Maintain Awareness

The concept proposed with this project is that an awareness system can act as a supplement to existing communication media. During the day users may opportunistically capture thoughts or moments they wish to share. These become a shared 'To-Tell' list, of topics for discussion over the phone or through email. The To-Tell list may be accessed or amended on the move or at home. At home it will mostly have the status of a peripheral display that stays discretely in the background when unattended. The To-Tell list should be possible to access readily during another communication (e.g., while talking on the phone or writing an email). Further, the interface to this list provides the possibility of initiating communications through the range of possible communication services and devices that are available to the user.

We hypothesize that the To-Tell list will help family members to stay in touch with each other in several ways. By providing an easy way to capture and share events without disturbing their environment, people can share more of their personal experiences with their family members. These messages can improve the quality of contacts with other media or face-to-face meetings by providing content. Finally, contacts will also be triggered by the messages and by the information about how family members can be reached.

At the time of writing this article, the service and client applications for the To-Tell list are under development, with the aim to be field tested in June 2003. The design process has proceeded in an iterative fashion, involving informal user tests. Two different mid-fidelity prototypes were constructed using Microsoft Powerpoint and Macromedia Flash and have been used for informal user tests to choose among alternative designs. An improved graphical design that has been adopted for the implementation is illustrated in Figure 3. Entries to the list are ordered in time, and are mapped upon the main spiral structure shown. To address limitations of space only a small segment of time is shown, and others are compressed (at both the outer side and the centre of the spiral). To the right of the screen, family members are illustrated along with an indication of reachability, i.e. whether they are reachable by phone, email, SMS, cellphone. Communication can be enacted by these interactive objects. Selecting persons from the thumbnails in the right of the screen, filters the items of the ToTell List to only communications to and from this person.

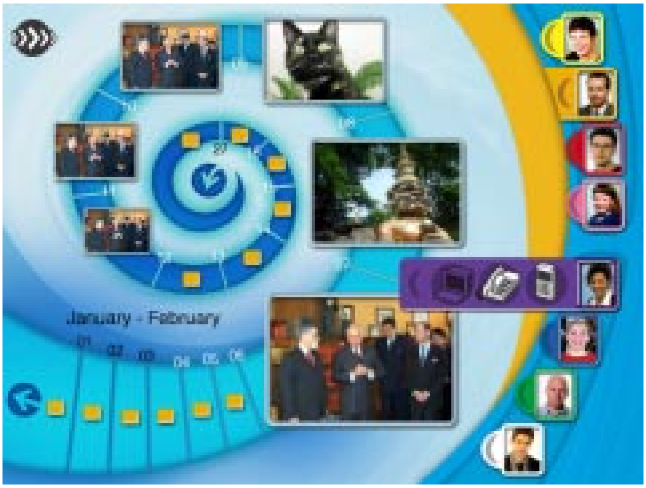


Fig. 3. Graphical design for the interface to the To-Tell list.

The To-Tell list will be maintained by a server application, to which access will be provided through Internet. The graphical design of Figure 3 is one of the screens of the iPronto client, that will also be displayed on a larger screen during background operation. A mobile client application for a Sony P-800 is under construction. This phone is GPRS enabled and offers the possibility of constructing and sending MMS messages. Through this client photographs and handwritten notes can be posted on the server. Thus, the designed functionality is in line with the user requirements of easy access, enabling meaningful effort, and providing concrete, visual information, as were found in the user study. Eventually, a special purpose application will provide instant text messages for the home user and see real time status information for other users.

5 Discussion

The constructed prototype will be deployed and assessed in user trials in order to gain a better understanding of the concepts of staying in touch and its relation to social presence. The study will have an exploratory nature aiming to generate knowledge regarding the user experience created and the usage patterns emerging with asynchronous awareness services as well as the technical requirements for the underlying technology.

The study will have two parts. The first and more controlled study will be conducted within the HomeLab facility, in the Eindhoven Campus of Philips Research in the

Netherlands. HomeLab is a future home-simulation, a test laboratory that looks like a normal house and thus provides us with a 'natural situation' to test the behaviour of the participants in the different conditions. Further, the HomeLab is a realistic test-bed for novel ambient technologies by Philips.

For each test in the HomeLab we shall involve families or groups of close friends. Two of the family members or friends will be sent on a nice outing (the Historisch Openluchtmuseum Eindhoven). They take a mobile device and will be asked to send at least 2 messages per hour. The other participant(s) stay(s) in the HomeLab. First, they will participate in a usability test in which they will be asked to perform a number of fixed tasks with the homebound device. After that, they will be encouraged to do what they like, e.g. read a book, play a game or watch a movie. At least twice every hour, they will receive messages from their family members or friends outside. They will also be allowed (but not obliged) to call each other. This part of the test will last 3-4 hours. All participants will be asked to fill out questionnaires afterwards. We will also conduct a focus-group meeting with the whole group about how they experienced using the system and how they would use it in real life. This part of the test will take 1-2 hours.

A second more extensive and realistic field-testing, will involve comparing the use of communication media by a family, one week without the device and one week using the homebound and mobile devices lent to them.

In both the laboratory and the field studies we shall be collecting qualitative data through observations and interviews and quantitative data by administering questionnaires. We shall use the IPO-SPQ questionnaire to measure the level of social presence experienced when using the system. This measurement will be repeated for different sessions, to gauge how repeated use of the system affects the level of social presence experienced. Social presence may not be the best or at least the only concept for characterising the user experience of awareness systems, like the To-Tell list, especially because communication is to a large degree asynchronous. A very important benefit that we anticipate for our users is the feeling of connectedness lasting through the day, i.e., feeling (emotionally) close to the friends and family. Currently we are developing a Connectedness Questionnaire that will complement the IPO-SPQ developed at the TU/e [5] in that it will focus on anticipated affective benefits, e.g., having company, a stronger group attraction, a feeling of staying in touch, sharing, belonging, etc.

6 Conclusion

Following Mark Weiser's call for research into ubiquitous computing applications, much of this research field has been characterised by the constant search for "killer-apps". Though very promising, it is yet unclear whether awareness systems can carry this title. The developments of market as well as research results do point to social uses of communication technologies as the 'killer activity' that is valued by increasing numbers of people.

Our earlier research and the research reported hereby suggest that people value peripheral awareness of friends and family. Intra family communication seems to be the primary domain where awareness systems could provide affective benefits. We have proposed the To-Tell concept, as one way of providing such benefits to individuals and families, in a manner that will integrate with existing communication technologies. During the assessment of this concept in the laboratory and the field tests summarised above, we plan to extend both our scientific understanding of people's needs and activities relating to staying in touch as well as the requirements for acceptance of awareness technologies at home (e.g., relating to privacy, situation awareness and staying in control).

To conclude, we underline how in our research theory building and technological developments are tightly coupled. Both contributions are essential. Pursuing practical advances in this field without a theoretical understanding of the social practices and personal needs we support can lead to expensive failures both in research and in development. The technological developments, on the other hand, are necessary to avoid basing theory and concept development on the discussion of hypothetical scenarios without empirical evidence.

Acknowledgements. The ASTRA project IST-2001-39270 is funded through the European Commission FET (Future and Emerging Technologies) Open Scheme. For more information see http://www.presence-research.org/ASTRA. Anastasia Gritsenko is thanked for her artwork in the diaries.

References

1. De Ruyter, B., Huijnen, C., Markopoulos, P., & IJsselsteijn, W., (in press) Creating social presence through peripheral awareness. Accepted for publication in HCI International, to be held in June 2003 in Greece
2. Dourish, P & Bly, S. (1992). Portholes: Supporting awareness in a distributed work group. Proceedings of ACM CHI'92, 541–547
3. Beaudouin-Lafon, M., Bederson, B., Conversey, S. Eiderbäck, B. & Hutchinson, H. (2002). Technology probes for families. Retrieved July 24, 2002, from http://interliving.kth.se/papers.html
4. Eggen, B., Rozendaal, M., & Schimmel, O.: Home Radio: Extending the home experience beyond the physical boundaries of the house. HOIT 2003, University of California, Irvine, (2003)
5. de Greef, P., & IJsselsteijn, W.: Social presence in a home tele-application. CyberPsychology and Behaviour 4: (2001) 307–315
6. Hindus, D., Mainwaring, S.D., Leduc, N., Hagström, A.E., & Bayley, O.: Casablanca: Designing social communication devices for the home. Proceedings of ACM CHI'01, (2001) 325–332
7. Beaudouin-Lafon, M., Bederson, B., Conversey, S. Eiderbäck, B. & Hutchinson, H. Technology probes for families. Retrieved July 24, 2002, from http://interliving.kth.se/papers.html (2002)
8. Markopoulos, P., IJsselsteijn, W., Huijnen, C., Romijn, O., & Philopoulos, A.. Supporting social presence through asynchronous awareness systems. In G. Riva, F. Davide, W.A. IJsselsteijn (Eds.) Being There: Concepts, Effects and Measurements of User Presence in Synthetic Environments, Emerging Communication Series, 5, IOS Press, Amsterdam, The Netherlands, (2003) 261–278

9. Melenhorst, A.S., Rogers, W.A. & Caylor, E. The use of communication technologies by older adults: exploring the benefits from the users' perspective. Proceedings HFES, (2001)
10. Rieman, J. The diary study: A workplace-oriented research tool to guide laboratory efforts. Proceedings of ACM CHI'93, (1993) 321–326
11. Vroubel, M., Markopoulos, P. & Bekker, M.M., FRIDGE: exploring intuitive interaction styles for home information appliances, In Jacko, J., and Sears, A., (Eds.) Extended Abstracts ACM CHI'01, (2001) 207–209

Author Index