Integrating Intra and Extra Gestures into a Mobile and Multimodal Shopping Assistant

Rainer Wasinger[1], Antonio Krüger[2], and Oliver Jacobs[1]

[1] DFKI GmbH, Intelligent User Interfaces Department,
66123 Saarbrücken, Germany
{rainer.wasinger, oliver.jacobs}@dfki.de
[2] University of Münster, Institute for Geoinformatics,
48149 Münster, Germany
antonio.krueger@uni-muenster.de

Abstract. Accompanying the rise of mobile and pervasive computing, applications now need to adapt to their surrounding environments and provide users with information in the environment in an easy and natural manner. In this paper we describe a user interface that integrates multimodal input on a handheld device with external gestures performed with real world artifacts. The described approach extends reference resolution based on speech, handwriting and gesture to that of real world objects that users may hold in their hands. We discuss the varied interaction channels available to users that arise from mixing and matching input modalities on the mobile device with actions performed in the environment. We also discuss the underlying components required in handling these extended multimodal interactions and present an implementation of our ideas in a demonstrator called the Mobile ShopAssist. This demonstrator is then used as the basis for a recent usability study that we describe on user interaction within mobile contexts.

1 Introduction

Mobile computing has seen significant advancements in recent years, as applications begin to span multiple and changing contexts. Multimodal user interfaces have also emerged as an integral area of development, as users gradually break free from the stationary desktop computing paradigm and enter the realms of pervasive computing. Multimodal interfaces that provide more intuitive ways to interface the computational power of an environment will gain more and more importance if users start to interact with computationally empowered artifacts that provide no obvious clue on their computational abilities. This is for example true in a shopping scenario where products are electronically identifiable (e.g. through RFID-tags) and where users are able to interact with the products that are on sale to retrieve product information through a shopping assistant. The interaction with the products could be based on speech utterances and performed user gestures (e.g. by picking up a product), however this causes problems relating to the privacy of the request and the presentation of the results, for ex-

H.W. Gellersen et al. (Eds.): PERVASIVE 2005, LNCS 3468, pp. 297 – 314, 2005.

ample if the environment delivers the requested information through a loudspeaker in the user's vicinity. Most of these problems can be overcome by including a personal device in the scenario, which allows users to silently pose requests and receive information from the environment unnoticed by others. In this paper we will present the Mobile ShopAssist demonstrator, which aids a user in finding out product information and product comparison information while shopping in an RFID enabled store. The system accommodates for multimodal input interaction in the form of speech, handwriting, and gesture, which can be performed either on the mobile device or by directly picking up an object from the shelf.

Newcomb et. al. [9] present some interesting design guidelines for a PDA based shopping assistant in a grocery store. They state that one important aspect that has to be regarded during the design process is that shoppers often use their hands to touch the products – something we have tried to incorporate into the design of our ShopAssistant. The authors further highlight the importance to find appropriate breaks in the shopping routine to be able to provide situated assistance. This has motivated us to choose the digital camera domain where normal shoppers usually rely on the help of a shop assistant to make their choice.

To illustrate our ideas, consider the following scenario. A user is interested in buying a digital camera. In possession of a device such as a PDA, the user may enter an electronics shop and start browsing the available range of digital cameras. The store might be crowded, or noisy, and there may also be no shop assistant in sight or available to the user to ask for assistance. Bearing these environment characteristics in mind, the user instead connects to a shelf of interest and downloads the shelf's product database onto the PDA through the use of the Mobile ShopAssist. Upon synchronization, the user will be able to browse both the products that are currently available in the store as well as those that are currently out of stock. With the help of the personal device the user is able to use multimodal queries to obtain information on the different cameras, regardless of whether they are present in the shelf or not. In particular, the system allows the user to pick up a product from the shelf and to then compare it with a product that is displayed on the PDA through the use of spoken natural language (e.g. "*compare this camera to that one*"). After having processed this multimodal input the system provides a comparison chart of both products.

The goal of this paper is to investigate the technical requirements of such a system and to explore the usability issues that arise from the various forms of new modality input combinations that are introduced. In the next section we describe the many modes of interaction that permit flexible user interaction within our shopping scenario. We describe how modality types may be combined on a mobile device, and also extend the modality input combinations to account for interaction with real world environments. In section 3, we describe the key elements required for merging modalities together, the modality fusion module, and include discussion on time-frames and parameters such as confidence values that are used in this process to resolve both non-conflicting and conflicting information sources. In section 4 we describe a usability study that was recently conducted on the Mobile ShopAssist demonstrator, including results on the use of observable modalities within public spaces, modality preference for unimodal and multimodal interaction spanning the virtual and physical world, and modality intuition. Section 5 provides some final conclusions and an outlook on future activities.

2 Multimodal Interaction

Interaction may take place on a unimodal or a multimodal basis, in which unimodal interaction refers to input that is composed of a single modality such as speech or handwriting only (e.g. "*What accessories are available for the EOS 300D*"), and multimodal interaction refers to input that spans multiple modalities such as speech and gesture (e.g. "*Does this <gesture> camera have a wireless control?*"). With respect to time, we suggest that modality information provided by a user can always be categorized as either non-overlaid, overlaid and non-conflicting, or overlaid and conflicting. In more detail, non-overlaid information occurs when input provided to a system does not have any of the same information represented by multiple modalities (e.g. "*What is the price of this <gesture=PowerShot S70>?*", while overlaid and non-conflicting information occurs when the same information has been represented by multiple modalities, but does not create a conflict (e.g. "*What is the price of the PowerShot S70 <gesture=PowerShot S70>*?), and overlaid and conflicting information occurs when the same information has been provided by multiple modalities and this information conflicts (e.g. "*What is the price of the PowerShot A40 <gesture=PowerShot S70>*?). Finally, with respect to origin, (non-overlaid and overlaid) modality input may originate from devices of the same type or of a different type, in which (e.g. for overlaid information), same-type device input such as speech and speech might arise through the use of both public and private microphones, while different-type device input such as speech and gesture might arise through the use of devices capturing different types of modal input.

2.1 On-device and Off-device Interaction

In contrast to desktop systems that generally make use of devices such as keyboards and the well-established 'Windows, Icon, Mouse, and Pointer' (WIMP) paradigm, we believe that instrumented environments require the addition of new interaction types like speech, handwriting and gesture. This need has arisen through the requirements of scenarios dealing with difficult environment contexts, and contexts in which the user is mobile and/or performing multiple tasks at a single instance in time. It is expected that users will in the future interact directly with artefacts in their surrounding, but will also most likely be aided by personal computing devices. This section outlines on-device and off-device interaction, as found on the Mobile ShopAssist demonstrator.

The Mobile ShopAssist demonstrator accepts modality input of type speech, handwriting, and gesture. Furthermore, user interaction with a set of products can take place either directly on the mobile device, or directly with artefacts in the surrounding environment, i.e. off-device. Interaction may also be part on and off device, as seen in Fig's 1A and B, where a user asks for comparison information on two products via speech, intra-, and extra- gesture.

Speech input in our system is provided by a user via the inbuilt microphone on the PocketPC. We use IBM's Embedded ViaVoice[1] speech recognizer to interpret all or

[1] IBM Embedded ViaVoice, http://www.ibm.com/software/pervasive/products/voice/vv_enterprise.shtml

part of a user's input. This is done through the use of limited-domain rule grammars that are generally between 50 and 100 words in size, and are dynamically loaded and activated by the system. We consider speech to aid both on and off-device interaction.

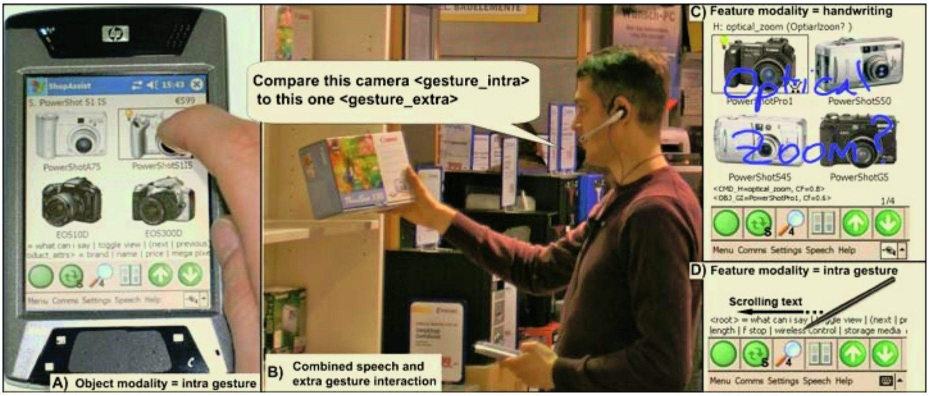

A) Object modality = intra gesture
B) Combined speech and extra gesture interaction
Compare this camera <gesture_intra> to this one <gesture_extra>
C) Feature modality = handwriting
H: optical_zoom (Optiarlzoon?)
Optical Zoom?
PowerShotPro1
PowerShotS50
PowerShotS45
PowerShotG5
<CMD_H=optical_zoom, CF=0.8>
<OBJ_G=PowerShotPro1, Cf=0.6>
1/4
Menu Comms Settings Speech Help
D) Feature modality = intra gesture
Scrolling text
<root> = what can i say | toggle view | (next | pr
length | f stop | wireless control | storage media
Menu Comms Settings Speech Help
PowerShotA75
PowerShotS1IS
EOS10D
EOS300D

Fig. 1. A and B show the combined use of speech, intra- and extra- gesture. C shows feature selection via handwriting, while D shows feature selection via an intra-gesture point on the scrolling text bar

Handwriting input is a typical on-device interaction modality, and requires a user to write on the PocketPC's display through the use of a stylus. We use Microsoft's embedded character recognizer called Transcriber[2], to interpret the user's input as a string of characters, and these characters are then mapped to a valid entry in our corresponding handwriting grammars.

Gesture is a broad term defined in common usage as "motions of the limbs or body, used as a means of expression" [Merriam-Webster dictionary]. As described in [6], gestures are used for everything from pointing at a person to draw their attention, to conveying information about space and temporal characteristics. Current research on gesture ranges from the recognition of human body motion (including facial expressions and hand movements) [13, 1], to pen and mouse based research [12] and sign language. Gesture input within our system is limited to the 'selection' of products, and may be of either type intra-gesture or extra-gesture. **Intra-gestures** are on-device interactions, and occur when a user touches objects displayed on the screen of the PocketPC. The screen coordinates are then mapped internally to the underlying data objects represented as either 2D or 3D graphics. Intra-gestures are provided in the form of stylus or finger input, and can currently be of type "point". **Extra-gestures** are off-device interactions and occur when the user interacts with the physical world around them by physically handling an object. Extra-gestures may be of type "pick-up", or "put-back". Pick-up and put-back actions are evaluated through the use of RFID technology that allows for the detection of objects being either in or out of a given space. In [8], we have experimented with integrating "extra-point" gestures

[2] Microsoft Transcriber, http://www.microsoft.com/windowsmobile/downloads/transcriber.mspx

based on a digital compass from Pointstar[3] in which we map real-world directions to known locations of physical objects, and have also augmented the intra-gesture library to include a “slide” gesture (alongside pointing). In our scenario, typical spaces include shelves that are instrumented with RFID readers and antennas, while typical objects include shopping products that are fitted with passive RFID tags. Within our scenario, intra gestures may also be combined with extra gestures, as shown by the speech utterance: “*Compare this camera <intra gesture> with this one <extra gesture>*” (see Fig. 1). It is this type of on- and off- device interaction that underlies our primary objective of providing interaction possibilities for instrumented spaces.

2.2 Representation of a Modality-Free Language, and Modality Combinations

The task of a modality fusion module is to combine different input streams into a single unambiguous and modality-free dialogue result, as defined by a modality-free language. The different input streams arise through the need for recognizers to use modality-specific grammars to deal with inputs such as speech and handwriting. A simplified version of the modality-free language used within our system is defined as follows:

$$\text{<FEATURE><OBJECT>+} \qquad (1)$$

The feature and object tags are represented in XML[4] (similar to EMMA[5]) and have a variety of attributes associated with them such as confidence and timestamp information. Valid values for the feature tag include (in reference to digital cameras) ‘optical zoom’ and ‘mega pixels’, while valid values for the object tag include ‘PowerShot S60’ and ‘CoolPix 4300’. User input is only considered well-formed or valid if it consists of one feature and at least one object. For this reason, the modality fusion module is only ever activated once a feature has been provided, which can take place via speech, handwriting and/or intra-gesture.

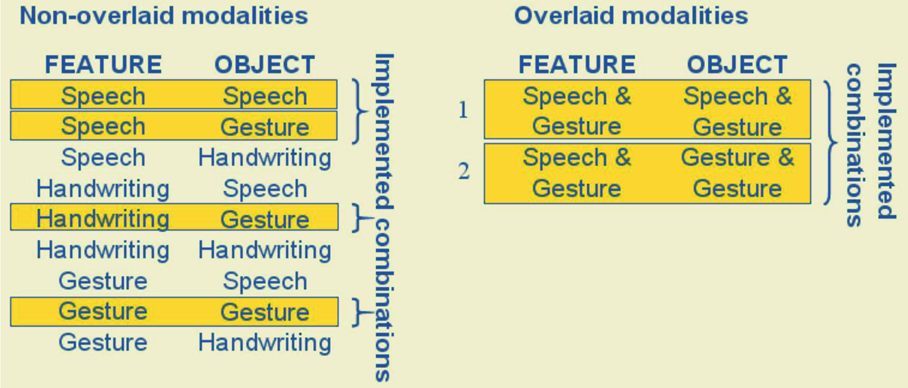
Having defined the two primary constituents of our modality-free language – i.e. FEATURE and OBJECT – it can be stated that although the system accepts speech, handwriting and gesture input, it does not accept all feature and object modality combinations. To demonstrate the extent that such a requirement would place on a system providing speech, handwriting, and gesture input, we would have 9 modality combinations[6] arising from the combination of a single feature and a single object, and 27 modality combinations for a single feature and two object references. These 27 combinations also do not consider the effect that overlaid modality information would have on the size of a full-scale implementation, as in the example: “*How many mega pixels does the PowerShot S60 <G=PowerShotS60> have?*” in which both speech and gesture are used to define the same object referent. Fig. 2 shows the modality combinations that have been implemented in our system so far.

[3] Pointstar, http://www.pointstar.dk

[4] W3C Extensible Markup Language, http://www.w3.org/XML/

[5] EMMA: Extensible MultiModal Annotation markup language, http://www.w3.org/TR/emma/

[6] Not differentiating between intra and extra gestures. Also note that extra gestures may only be used for selecting objects.



Non-overlaid modalities

FEATURE	OBJECT
Speech	Speech
Speech	Gesture
Speech	Handwriting
Handwriting	Speech
Handwriting	Gesture
Handwriting	Handwriting
Gesture	Speech
Gesture	Gesture
Gesture	Handwriting

Implemented combinations

Overlaid modalities

	FEATURE	OBJECT
1	Speech & Gesture	Speech & Gesture
2	Speech & Gesture	Gesture & Gesture

Implemented combinations

Fig. 2. Non-overlaid and overlaid input modality combinations

3 Modality Fusion

As shown in Fig. 3, our modality fusion component is based on a blackboard architecture. In contrast to the projects QuickSet [3] and SmartKom Mobile [2] in which a heavy reliance existed on distributed and client/server architectures, all of the interaction processing (except for the extra-gesture recognition which is based on RFID technology) is performed locally on the mobile PocketPC device. Indeed the blackboard itself is also located on the mobile device. Recognizable user input is defined by grammars that are associated with product types within the product database. These grammars are dynamically loaded based on the type of products contained

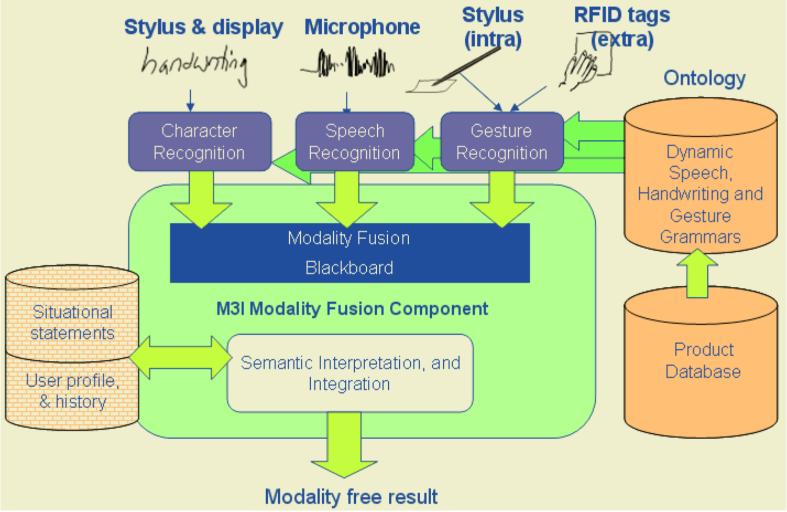


Fig. 3. Modality fusion within the Mobile ShopAssist

within the currently synchronized data container. During user interaction with the system, input is written to the central blackboard in the form of data nodes. These data nodes provide the primary information required for the modality fusion module to make informed decisions about the objects on the blackboard. As described in [14], this information includes the presumed dialog segment (i.e. feature or object), the parent modality group (i.e. speech, handwriting, intra-gesture, extra-gesture), an underlying modality type where appropriate (e.g. point, pick-up, put-back), a confidence value from 0.0 to 1.0, the start and stop times for the dialog segment, a time classification (i.e. past, or present), the raw user input, the matched user input, and the 3-best result matches including confidence scores.

The blackboard is stored on the PocketPC device itself, as too the modality fusion component. This provides the additional benefit that users can disconnect from their surroundings and continue to interact with the system offline. When browsing offline, the user has access to all modalities except extra-gesture, which is based on RFID technology and is recognized by an environment server.

Two important parameters for the modality fusion process – confidence values and timestamps – will be described in the following section. Alongside these types of parameters, we also expect statistical data from user-history log files, and context information in the form of situational statements [5] to further contribute to the modality fusion process. Situational statements refer for example to characteristics of a user (e.g. role, age, gender, walking speed, eye sight), the device (e.g. remaining battery life, working memory, speaker volume), or the surrounding environment (e.g. noisy, crowded, rainy), and are a convenient way of representing context.

3.1 Confidence Scoring

Confidence scoring is the ability to attach a probability to a recognition result in order to measure how confident a recognizer was in matching a result with what was actually inputted into the system. For each of the modalities within our system (speech, handwriting and gesture), we generate an N-best list of results and assign confidence values between 0.0 and 1.0 to each result. This occurs each time a user interacts with the system. "N" in the case of the ShopAssist is equal to 3 and means that the 3 most likely results are returned for a given modality (instead of just one). N-best result lists and their associated confidence values play an essential role in the disambiguation of multimodal input [11]. As an example, user input may be overlapped and conflicting (destructive), or it may be overlapped and non-conflicting (constructive). By keeping a hold of the N-best lists, we are able to store information that a single recognizer such as a speech recognizer might ultimately have thrown away, and are thus able to compare this with results from other recognizers at a later stage (i.e. in the modality fusion module). Confidence values within mobile multimodal systems are also important because the methods used in calculating these values are specific to the individual modalities, and are likely to be affected differently by surrounding environment characteristics. As an example, speech is likely to be affected by background noise differently to gesture, and handwriting will be affected by motion (i.e. while a user is walking) differently to speech.

Confidence values when processing **speech** are generated by matching a user's spoken utterance to a sequence of word hypothesis based on a given language model.

Our language model consists of word-phoneme mappings and grammar files written in a format similar to the Backus-Naur Form (BNF). A current mobile device limitation of our system is that each utterance in the N-best list receives only one confidence value. Speech engines with greater disposable resources are capable of returning confidence values for the individual words in a recognized utterance's word lattice [4]. This limitation means that if a feature and an object are both provided in a single speech utterance, our system will tag both values with the same confidence value. These values may however still be different from one another within the returned N-best list.

The process of generating confidence values for **handwriting** within our system is a two stage process. Written input (see Fig. 1C) is first sent to a character recognizer that converts a user's handwriting input to block characters. The second stage is to then match the recognized input to our own handwriting grammars, which consist of keywords such as "optical zoom", and "mega pixels". This is based on a simple character matching algorithm, which disregards case, punctuation and white space, and performs a sliding character match on the given user input and all entries in the grammar. Grammar entries that are either too long or too short (i.e. +-3 characters in length) are discounted, and grammar entries that start with the same character as the given input are given an additional bias (Val=Val+0.1). The top 3 results are then returned to the modality fusion blackboard.

Intra-gestures are used for the resolution of features and objects on the display of the PocketPC. Objects refer to camera products such as the "EOS 300D" and are provided in the form of graphical pictures, while features refer to keywords such as "price" and are provided as scrolling text on the user's PocketPC display (see Fig.'s 1A and 1D respectively). Regarding object resolution, nine, four, two, or one object rectangles may be displayed on the PocketPC's display at any one time depending on the current mode that the user is in. We generate confidence values by drawing a rectangle around the user's "point" coordinates, equal in size to that of the other rectangles on the screen. The intersection between this Active Area (AA) and each of the Image Rectangles (IR) is then calculated and used to generate the confidence value (AA/IR), which generates a value between 0.25 and 1.0. If the user points to an image rectangle at perfect centre, the rectangles line up and the score is 1.0. If the rectangles only half line up (side by side), the score is 0.5, and if the user points to a corner, the score is 0.25. The 3 best results are then mapped onto the range 0.0 and 1.0 through the use of exponents, for compatibility with the other modalities. For feature resolution, which is based on a user pointing out a keyword from a scrolling text bar at the bottom of the PocketPC's display, we currently allocate the static values 0.8, 0.4, and 0.1 to the 3-best results. 0.8 is assigned to the keyword the user clicked on, while 0.4 and 0.1 are assigned to the keyword to the left and the keyword two to the left of that which the user clicked on respectively. We do not currently consider point coordinates above or below the scrolling text bar when resolving features. We intend in the future, to allow the user to increase and decrease the speed of the scrolling text, and to have access to the font size and direction in which the text flows. These changes would undoubtedly also need to be considered in future confidence scoring techniques for this modality.

All **extra-gestures** are currently given a confidence weighting of 1.0, and no N-best list is returned for this modality in the current scenario. In the case that multiple objects are taken from the shelf (and no "compare" command was given), timestamps are used to resolve the most recent object. The confidence weighting limitation is due to the fact that we do not keep track of the actual physical location of the individual products on the shelf. When compared to a real shopping scenario where each product has a defined position on a shelf (and labeled name and price information), accommodating for a finer resolution would not be difficult to program into our application, and would perhaps allow a modality fusion component to consider products to the left and to the right of the product selected. Information on similar looking product boxes might also be useful for extra-gesture resolution, and thus also confidence scoring.

It should be noted that the generation of adequate confidence values for use by a multimodal system is still an area of ongoing research. Although speech recognizers for example are nowadays built on top of a great wealth of statistical data that has arisen through decades of experience, there is still limited statistical data for determining how best to rate the confidence of modalities such as gesture in different mobile environment contexts. Kumar et al [7] have presented a study that investigates the performance of a multimodal (speech and intra-gesture based) system under field conditions. For this purpose subjects were involved in strenuous activities while performing map-based tasks. The results show that although performance of the single recognizer gets worse with a rising degree of exertion, the overall multimodal recognition results remain stable. A further concern when comparing confidence values between same-type and different-type recognizers is that the confidence weightings may never have been "fair" to begin with, as would result from designers using different statistical models to train their recognizers. As such, an application using multiple recognizers may have to incorporate a penalty-reward system, in which accurate and inaccurate results are used to balance out discrepancies between the different recognizers.

3.2 Time-Frames and Input Synchronization

Dealing with the temporal order in which modal input occurs - often referred to as *multimodal synchronization* [10] - is one aspect of particular importance to multimodal systems. Different input modalities may occur at different times in a dialog act, and each of these possibilities must be correctly accommodated for. As an example, the speech input "*What is the optical zoom of this camera?*" might have a gesture input accompanying it either before, during, or after the actual utterance, and each of these three possibilities would be correct. In our scenario, we use the issuing of a feature to determine our timeframe markers. We distinguish between the period before (a timeframe of up to several minutes), during (a timeframe typically 4 to 5 seconds long) and after (a timeframe of either 0.5, 1.5 or 3.5 seconds long, depending on a user's familiarity with the system) a feature has been issued. If the timeframe values are too small the chances of 'valid' user input being disregarded will be high, and if the values are too large the chances of old or 'invalid' user input being accepted will be high.

Fig. 4 depicts the timeframes within a typical user interaction. Each time a user interacts with an object on the display, it is recorded onto the modality fusion blackboard. In this fashion, a user's interaction may span a timeframe of up to several minutes. However, once the system is aware that the user has started to issue a feature, which is checked each time speech, handwriting or a special subclass of intra-gesture input is provided (as shown in Fig. 1D), the user is given a limited period of time before the modality fusion process begins. This time period depends on whether or not objects with *TimeType* equal to 'present' or 'past' have been selected in the current user-turn, and whether or not the user is familiar or unfamiliar with the system (defined in a user property file). If an object has been selected within the current user-turn, the object will contain a *TimeType* value equal to 'present' and the module will conclude it's processing within 500ms. If however only 'past' objects exist on the blackboard, the user is either referring to an object selected in a previous user-turn, or the user has not yet selected an object. Familiar users are provided with an additional second in this case, while users less familiar with the system are provided with an additional 3 seconds to complete their current dialogue act. The trade-off for extending the timeframe in this manner is that the system appears more sluggish in the case that the user is indeed referring to a past object.

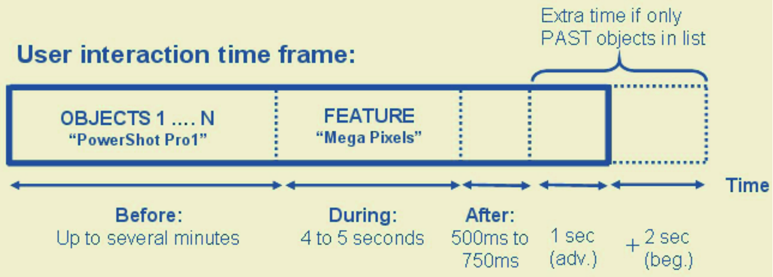


Fig. 4. Input synchronization: 'Before', 'during', and 'after' durations, as used by our system

Initial studies on our system have shown that it generally takes a trained user around 4 to 5 seconds to carry out a complete dialogue act. As an example, these 4 to 5 seconds refer to the time it takes a user to press the talk button once to start talking and once again after having spoken to stop talking for a speech-speech interaction, or the time it takes a user to select an object and to then start and stop writing a keyword on the display of the PocketPC for a gesture-handwriting interaction.

In the case that a feature is found on the blackboard but no object exists, the user is briefly informed and provided with an additional 4 seconds to complete their interaction, after which time the feature is removed from the blackboard. Overlaid and conflicting input can occur for both the feature and object values, however feature conflicts are less frequent due to the restrictive timeframes imposed on the user once a feature input has been initiated. Our system resolves object conflicts by first removing older nodes with similar modality types. We then remove nodes that fall outside a given timeframe (based on the most recent object's timestamp plus a given time-margin), and then recalculate confidence values based on a matching algorithm that considers each remaining node's N-best list (where N=3). Nodes with object values that appear in the N-best lists of multiple nodes have their confidence value increased in this way, and the node with the best overall confidence value is then finally selected.

An initial study on input synchronization in our mobile scenario has been inline with experiments conducted by Oviatt et.al. [10], who shows for example that pen onset usually precedes speech onset, and pointing gestures are integrated with parts of a speech utterance in a natural manner. Regarding speech recognition, we have observed similar to [15] that word-error rates vary directly with speaking style such that the more natural the speech delivery, the higher the recognition system's word error rate. This is more so in our mobile scenario, as we often only provide a single way for users to speak out an utterance that could normally be communicated in one of many ways, e.g. "*what is the price of* ..." and "*how much does ... cost*". Regarding setting up the RFID technology, we observed that it was sometimes required to tag product boxes multiple times in order for the system to recognize them correctly. In general however, we have found that users were quite content with the performance and accuracy of the system, and it was also observed that the learn-in time (at least for non-overlapped modalities) was acceptable.

4 Usability Study

In this section we describe the results of an exploratory user study that we recently conducted on the Mobile ShopAssist. The primary goal of the usability study was to measure the modality preference of users when interacting with the mobile system. A total of 23 different modality combinations were tested, and these were derived from the three elementary modalities speech, handwriting and gesture. For our modality-free language, <FEATURE><OBJECT>+, the combinations ranged from unimodal to multimodal interaction, and from non-overlaid to overlaid input. Aside from modality preference, we also studied how intuitive the individual modality combinations were to use, and asked users what effect being immersed in a public or private space would have on their use of the three base modality types. A total of 440 user interactions were logged by the system throughout the study, averaging 31 interactions per person. Although we kept a user history of interactions, this study does not delve into aspects of system accuracy or system learnability.

4.1 Method

Our usability study was conducted at the University of Saarland in one of the department's computer terminal rooms. We believe this laboratory setting to differ from a real-world environment in two ways. Firstly, there were few if any other people in the terminal room during the times we conducted the testing (aside from the instructor and the user), and secondly, background noises were kept to a minimum. We conducted the study on a total of 14 people who were either a little familiar or unfamiliar with the system. The study was conducted in English with users that could speak fluent English. 10 of our users were students and lecturers from the computer science department aged between 25 and 37 years, while the remaining 4 users were not from the computer science department and were unfamiliar with the system.

Each user was given a PocketPC device and a headset connected to the PocketPC's audio jack through which the user could speak into and listen to the output from. They were asked to stand in front of an instrumented shelf containing real-world camera boxes. As described in Section 2.2, the users were allowed to mix-and-match modality input combinations when creating their FEATURE-OBECT dialogue inputs, and were also allowed to overlap modalities when communicating with the system. A total of 12 non-overlapped modality combinations and 11 overlapped modality combinations were tested (as shown in Fig. 5). Each test session generally required between 45 and 60 minutes to complete. Users were explained the base modalities that could be used when building feature and object dialogue interactions. They were told that the order of the inputs was irrelevant, i.e. feature then object, or object then feature, and that system errors were to be expected but should not bias their answer as not all modality combinations had been implemented. There were a total of 10 different objects and 13 different features available to the subjects.

The usability study had two parts, the first being an *observation* of each user interacting with the system, and the second being a *written questionnaire*. Within the observation, each user was free to choose their own modality combinations while interacting with the system. Most users managed 4 or 5 different modality combinations within this part before needing to be reminded of the remaining modality combinations. At this point, users were specifically told the order in which they should use the remaining modalities. After each interaction, the user was asked to rate the modality combination by answering the question "*Would you use this modality combination?*". The rating scale used was a set of preferences that we later mapped onto a scale from 0.0 to 3.0, in which "0=prefer not, 1=maybe not, 2=maybe yes, and 3=prefer yes". Following the practical component, the users were asked to complete a written questionnaire that again asked them to repeat their preference for each individual modality combination, and to also state whether or not they thought the modality combinations were intuitive. Several other questions relevant to mobile and multimodal interaction were then also asked, and the survey ended with the user stating their favourite input modality combination.

4.2 Usability Results

For simplicity, we refer to the individual modality combinations via their abbreviations – speech (S), handwriting (H), intra-gesture (GI), and extra-gesture (GE). As an example, the interaction: <FEATURE modality=speech><OBJECT modality=speech> is analogous to the modality combination SS.

4.2.1 Preferred Modality Combinations

Fig. 5 shows the modality combinations categorized into the groups non-overlapped and overlapped. From the averages shown in the figure, it can be seen that users generally prefer non-overlapped modality combinations (A_v=1.58) to overlapped modality combinations (A_v=0.60). Using a Mann-Whitney U test, this was also shown to be statistically significant in 8 out of 14 subjects: $U(12,11)<35$, $p<0.05$, and only 3 subjects had a $p>0.12$. The non-overlapped combinations have been further grouped according to their start modality, from which it can be seen that the use of speech for the feature (A_v=2.09) is preferred to the use of intra-gesture (A_v=1.39) or handwriting (A_v=1.25) for the feature. Also interesting to note is that within each subgroup of start modalities, the use of the same modality (unimodal) for both the feature and the object referents received the highest or near highest scores (see darkly shaded modality combinations in Fig. 5). Similar to the non-overlapped sub-groups, we categorized the overlapped combinations by their overlapping segment types – feature, object, or both feature and object. It can be seen that users preferred overlapped object information most (A_v=0.99) out of all of the overlapped modality combinations. This rating increases to A_v=1.33 when speech is set as one of the overlapped modalities and increases to A_v=1.57 when handwriting is excluded from the possibilities.

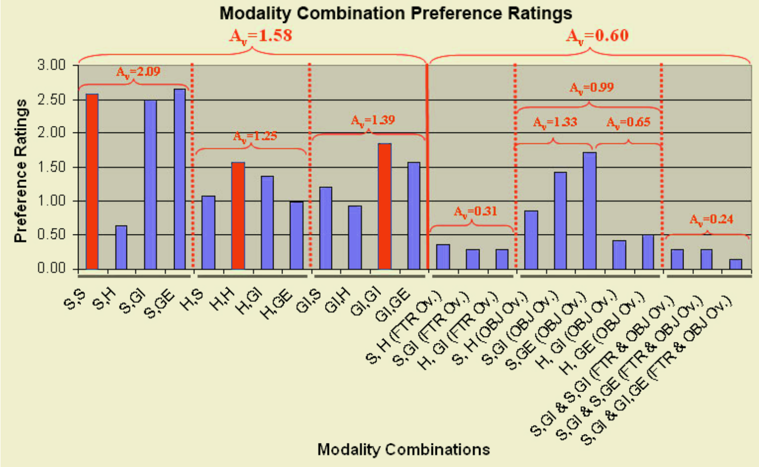


Fig. 5. The 23 modality combinations categorized into the groups overlaid and non-overlaid, and rated according to preference (0=prefer not, 1=maybe not, 2=maybe yes, 3=prefer yes). The darker shaded modalities represent unimodal combinations

Fig. 6 shows all of the modality combinations ranked in order of user preference. It can be seen that SGE is the most preferred modality combination, and that this is very closely followed by SS and SGI. Using the Mann-Whitney U test, the preference for these three modalities when compared to all other modality combinations was significant in 12 out of 14 subjects: U(3,20)<9, $p<0.05$. The benefit of allowing a user to provide deictic input can also be seen in that 2 of the top 3 modality combinations, and 7 of the top 9 modality combinations used gesture to identify the object. The successful wedding of the modalities speech and gesture is further exemplified by the overlapped object combinations SGE (rating=1.71) and SGI (rating=1.43), whose preference was shown to be significant in 6 from 14 subjects when compared to the other overlapped modality combinations (U-test, $p<0.36$). As shown in Fig. 6, the modalities have been grouped according to rating point falls between the individual modalities, where the first drop of 0.64 borders on significance (Wilcoxon, z=-1.807, p=0.071). The first set of combinations are preferred by users (A_v=2.57), while the second set of modality combinations (A_v=1.58) lie within the category "maybe no" and "maybe yes". The third set of modality combinations has a ranking value directly equivalent to "maybe no", and the fourth set of modality combinations (A_v=0.36) is least preferred. The highlighted columns represent those modalities that were not implemented in our system, and although most of these modality combinations exist on the lower side of the ranking scale, the incorporation of the modality combination HH will now be considered for future versions of the demonstrator.

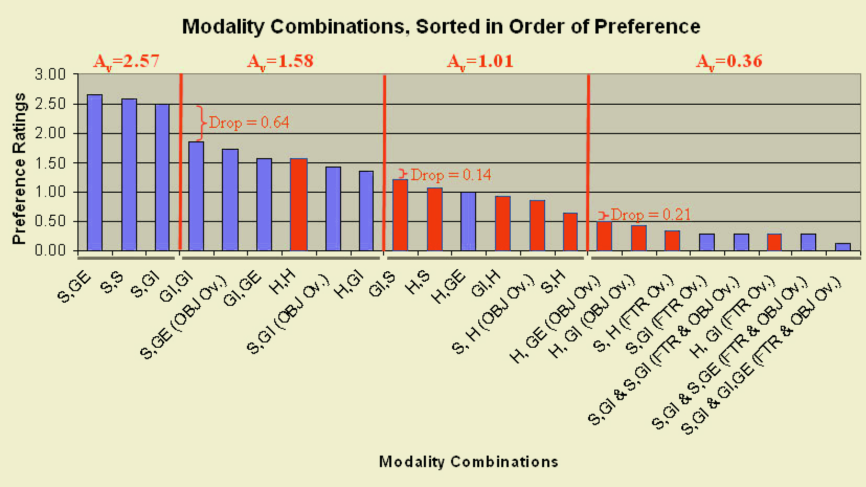


Fig. 6. The 23 modality combinations ranked in order of preference. The darker shaded modalities represent those that have not been implemented in the Mobile ShopAssist

4.2.2 Modality Intuition

We measure each modality's intuitiveness in two separate tests, one conducted during the written component (Fig. 7A), while the other conducted during the practical component (Fig. 7B). Fig. 7A shows the results provided by our subjects to the question: "*Do you feel that this modality combination was intuitive to use?*" ("no", "yes"), while Fig. 7B shows the first 4 modality combinations used by our subjects during the practical component. The modalities in Fig. 7B are weighted exponentially, such that a modality chosen 1st receives a weighting of 1000, while modalities chosen 2nd, 3rd and 4th receive the values 100, 10, and 1 respectively. The resulting weights for the individual modalities are shown in the bottom right of Fig. 7B.

The written component showed that 5 of the 12 non-overlapped modality combinations (SS, SGI, SGE, GIGI, and HH) were rated significantly *intuitive* by our subjects: $\mathrm{Chi}^2(1,N=14)>10.286$, $p<0.001$. In comparison, 6 of the 11 overlapped modality combinations were rated significantly *non-intuitive* by our subjects: $\mathrm{Chi}^2(1,N=14)>4.57$, $p<0.33$.

When correlated with the lower graph one can see that the modality combinations SGI, SGE, SS, and GIGI were mirrored as being intuitive. The modality combination HH was however never selected for use by any of our users within their 1st four interactions, despite 13 out of 14 users rating the modality as being intuitive during the written component. The overlapped modality combinations, SGI (overlapped object) and SGE (overlapped object) were also never used within the 1st four modality combinations. Many people commented that handwriting was too slow to use, and perhaps this was a reason why HH was never selected by our users within the practical component. The overlapped modality combinations may also have simply been overlooked by users due to the already wide range of non-overlapped modality combinations to choose from.

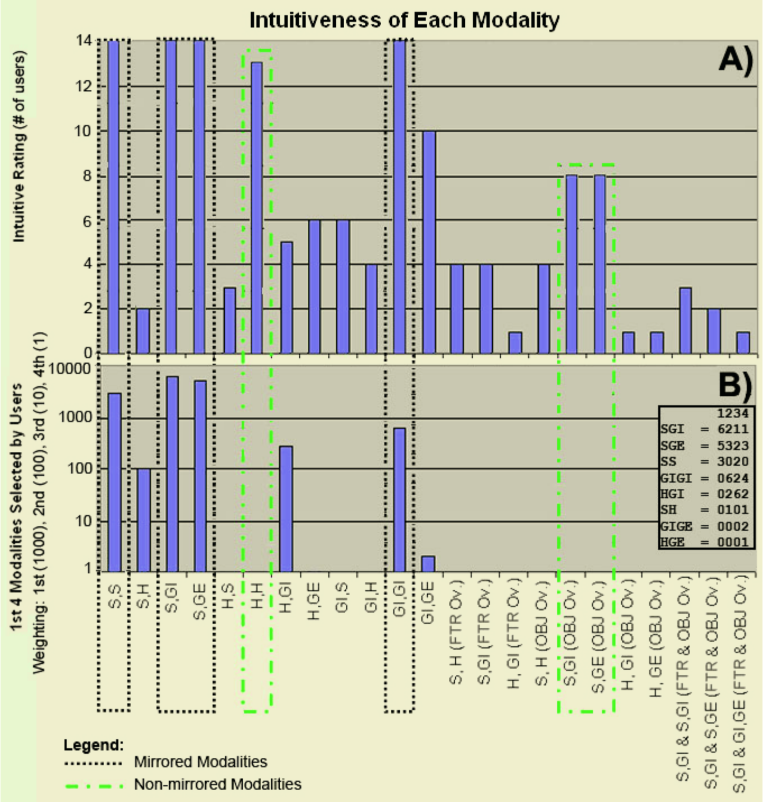


Fig. 7. Modality intuition. Graph A) shows the modality intuition results provided by users during the written component, while graph B) shows the 1st four modality combinations selected by users in the practical component, and their weightings in the bottom right

4.2.3 Public and Private Spaces, and Observable Modality Combinations

One of the questions within our written questionnaire was with respect to how the user would feel using the modalities speech, handwriting, intra gesture, and extra gesture while in a public space (e.g. a shopping mall), and while in a private space (e.g. at home). The choices given to our subjects were “embarrassed”, “hesitant”, and “comfortable”. Chi-square tests show that our subjects would feel comfortable using intra-gesture, extra-gesture, and handwriting within a public environment:

$Chi^2(2,N=14)>8.714$, $p<0.013$. Within a private environment, our users would feel comfortable using all base modalities: $Chi^2(2,N=14)>8.714$, $p<0.013$.

We also compared the group of modalities that are entirely observable by surrounding people (SGE, SS) with those that are entirely non observable by surrounding people (GIGI, HH, HGI, GIH). Excluding modality combinations that are only partly observable (e.g. SGI), it can be seen that within this usability study, users preferred the extrusive modalities ($A_v=2.61$) over the non-extrusive modalities ($A_v=1.43$). SGE (2.64) and SS (2.57) were ranked highest from the entirely observable modalities, while GIGI (1.86), HH (1.57) and HGI (1.36) were ranked highest from the entirely non-observable modalities. This implies that at least within a laboratory setting and for the product type “digital cameras”, the feelings of “embarrassment” and “hesitation” had little effect on modality preference. We are now also evaluating usability results from a second round of tests conducted on 28 users at a local electronics store[7], to see what differences might exist between a laboratory and a real-world setting. An average of 13.8 people could be seen from the shelf’s location during each of the tests, and our hypothesis that the results will see a modality preference shift towards non-observable modalities appear to be correct.

4.2.4 Subjective Results Obtained from the Study

Several interesting points were raised by our subjects during the study that may serve as a future set of guidelines for interface designers. Our subjects said for example that their preference for a particular modality would change depending on the type of task at hand and the type of products that they were shopping for. Users mentioned that speech and handwriting would for example be a good alternative to gesture if an object was not accessible, not present, or difficult to find on the Pocket PC display or shelf. They noted that in comparison to the extrusive modalities like speech, non-extrusive modalities like intra-gesture would be better suited when dealing with sensitive objects like contraception. Some users also preferred the simplicity of consistent modalities (i.e. unimodal combinations) over mixed modalities, and a further distinction was made between modality combinations that were part loud and part silent such as SH.

Regarding speech, some users found the camera names such as PowerShot S1IS and FinePix A202 non-intuitive to pronounce, despite the system correctly understanding the user and despite the user being told to disregard system failures. Regarding handwriting, users often commented that the feature and object names like “mega pixels” and “PowerShot Pro1” took too long to write. Users also stated that intra-gesture for feature input (i.e. the visual WCIS) would be better if all options could be seen at the one time, rather than needing to wait for the text to scroll into focus, which they said would be problematic if they were stressed for time. With respect to extra-gesture, several users mentioned that pointing to the objects would be an improvement, especially for heavier product types, or for people that had their hands already full with shopping, winter-jackets and the Pocket PC device. These people did however like the ability to touch products as well, and many stated that they liked being able to physically touch a product before purchasing it.

[7] Conrad Electronic, Saarbrücken. http://www.conrad.de

5 Conclusions and Future Work

In this paper we have presented a multimodal mobile shopping assistant that integrates interactions on a mobile device and interactions with real world shopping products (i.e. digital cameras). For this purpose, we have incorporated three types of input modalities on a PDA: spoken language, handwriting and intra-gestures, which we have combined with extra-gestures performed with real world artifacts in the user's environment. We have discussed the various multimodal input combinations out of which we have implemented and tested 13 in our prototype. We have discussed how multimodal requests can be resolved by the use of an N-best list and provided some empirical values for an appropriate timeframe to ensure the correct input synchronization of the different modalities.

The results returned by the usability study have highlighted several important facts about mobile multimodal interaction. Most importantly, the study has shown that from the 23 modality combinations offered to our users within the mobile shopping scenario, speech and extra-gesture (SGE) were the preferred choice, closely followed by speech and speech (SS) and speech and intra-gesture (SGI). Indeed, the success of these three modalities is further iterated in that these modes are directly representative of how people interact with other people, and in particular with sales assistants. For future work, we now plan to evaluate a second round of usability testing that has recently been conducted at a local electronics store. We also plan to implement the unimodal combination of handwriting and handwriting (HH), which was ranked higher than expected by our users, and plan to make use of additional context information (e.g. user preferences and user habits) to improve the resolution of multimodal user requests. Such information could be either retrieved from an external user model or by allowing the user to correct the system's false positives through an error recovery procedure.

Acknowledgements

This work was partially funded by the German Federal Ministry for Education and Research (BMBF) under the contract no. 01 IN C02, as part of COLLATE II.

References

1. Baudel, T., Beaudouin-Lafon, M., CHARADE: Remote Control of Objects using Free-Hand Gestures, ACM Journal, Vol. 36, no. 7, 1993, pp. 28-35.
2. Bühler, D., Minker, W., Häußler, J., Krüger, S., Flexible Multimodal Human-Machine Interaction in Mobile Environments, In ICSLP, 2002, pp. 169-172.
3. Cohen, P., Johnston, M., McGee, M., Oviatt, S., Pittman, J., Simith, I., Chen, L., Clow, J., Quickset: Multimodal interaction for distributed applications, Proc. of ACM International Multimedia Conference, 1997, pp. 31-40.
4. Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen A., Zue, V., Survey of the State of the Art in Human Language Technology, Book, Chp1, 1995.

5. Heckmann, D., "Introducing Situational Statements as an integrating Data Structure for User Modeling, Context-Awareness and Resource-Adaptive Computing", ABIS Workshop on adaptivity and user modelling in interactive software systems, 2003.
6. Kendon, A., Conducting Interaction: Patterns of behavior in focused encounters, Book: Cambridge University Press, 1990.
7. Kumar, S., Cohen, P.R., Coulston, R., Multimodal Interaction under Exerted Conditions in a Natural Field Setting. In Proc. of the Sixth International Conference on Multimodal Interfaces (ICMI), 2004, pp. 227-234.
8. Krüger, A., Butz, A., Müller, C., Stahl, C., Wasinger, R., Steinberg, K.E., Dirschl, A., The Connected User Interface: Realizing a Personal Situated Navigation Service, Proc. of the 9th International Conference on Intelligent User Interfaces, 2004, pp. 161-168.
9. Newcomb, E., Pashley, T., Stasko, J., Mobile Computing in the Retail Arena, Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2003, pp. 337-344.
10. Oviatt, S., DeAngeli, A., Kuhn, K., Integration and synchronization of input modes during multimodal human-computer interaction, In Proc. of CHI, 1997, pp. 415-422.
11. Oviatt, S. Mutual disambiguation of recognition errors in a multimodal architecture. In Proceedings of the Conference on Human Factors in Computing Systems, 1999, pp. 576–583.
12. Pastel, R., Skalsky, N., Demonstrating Information in Simple Gestures, In Proc. of the 9th international conference on Intelligent User Interfaces, 2004, pp. 360-361.
13. Wahlster, W., Towards Symmetric Multimodality: Fusion and Fission of Speech, Gesture, and Facial Expression, In. Proc. of the 26th German Conference on Artificial Intelligence, 2003, pp. 1-18.
14. Wasinger, R., Krüger, A., Multi-modal Interaction with Mobile Navigation Systems, In: W. Wahlster (ed.): Special Journal Issue "Conversational User Interfaces", it - Information Technology 46 (2004) 6, München: Oldenbourg Wissenschaftsverlag (ISSN 1611-2776), 2004, pp. 322-331.
15. Weintraub, M., Taussig, K., Hunicke, K., and Snodgrass, A. Effect of speaking style on LVCSR performance. In Proc. of the International Conference on Spoken Language Processing, 1996, pp. 16–19.