Multi-Modal Information Processing in *Companion*-Systems — A Ticket Purchase System

Ingo Siegert[†], Felix Schüssel[‡], Miriam Schmidt[‡], Stephan Reuter[‡], Sascha Meudt[‡], Georg Layher[‡], Gerald Krell[†], Thilo Hörnle[‡], Sebastian Handrich[†], Ayoub Al-Hamadi[†], Klaus Dietmayer[‡], Heiko Neumann[‡], Günther Palm[‡], Friedhelm Schwenker[‡], and Andreas Wendemuth[†] [1]

Abstract A demonstration of a successful multimodal dynamic human-computer interaction (HCI) in which the system adapts to the current situation and the users state is provided using the scenario of purchasing a train ticket. This scenario demonstrates that Companion Systems are facing the challenge of analyzing and interpreting explicit and implicit observations obtained from sensors under changing environmental conditions. In a dedicated experimental setup, a wide range of sensors was used to capture the situative context and the user, comprising video and audio capturing devices, laser scanners, a touch screen, and a depth sensor. Explicit signals describe user's direct interaction with the system such as interaction gestures, speech and touch input. Implicit signals are not directly addressed to the system, they comprise the users situative context, his or her gesture, speech, body pose, facial expressions and prosody. Both multimodally fused explicit signals and interpreted information from implicit signals steer the application component which was kept deliberately robust. The application offers stepwise dialogs gathering the most relevant information for purchasing a train ticket, where the dialog steps are sensitive and adaptable within processing time to the interpreted signals and data. We further highlight the system's potentials of a fast-track ticket purchase when several information indicate a hurried user.

A video of the complete scenario in German language is available at: http://www.uni-ulm.de/en/in/sfb-transregio-62/pr-and-press/videos.html

1 Introduction

Companion-Systems are faced with the challenge of analyzing and interpreting observations obtained from sensors under changing environmental conditions. A

[†]Otto von Guericke University Magdeburg, D-39016 Magdeburg · [‡]Ulm University, D-89081 Ulm

[1] All authors contributed equally.

demonstration of a successful multimodal dynamic human–computer interaction in which the system adapts to the current situation by implementing multimodal information processing and the user's state is provided using the scenario of purchasing a train ticket[2]. The technical setup defines the available sensors and may change depending on the location of the user. The environmental conditions parametrize the interpretation of the observations from the sensors and constrain the reliability of the information processing. The inferred knowledge about the user's state and the context of use finally enables the system to adapt not only its functionality, but also the way of interacting with the user in terms of available in- and output modalities.

An experimental platform shown in Fig.1 was equipped with a wide range of sensors to capture the situative context and the user. The sensors comprise video and audio capturing devices, laser scanners, a touch screen, and a depth–sensing camera. The information processing is depicted in Fig.2 and was realized by multiple components. Some of them retrieve data from sensors while others are pure software components, depicting in a complex conceptual information flow. The components are categorized according to explicit and implicit user signals. Explicit signals describe commands performed by the user with the intention to interact with the system such as interaction gestures, speech and touch input. Implicit signals are not directly addressed to the system but nevertheless contain a rich set of relevant information. These signals comprise the user's situative context, his or her gesture, speech, body pose, facial expressions and prosody.

Fig. 1 A user interacting with the ticket purchase system. While the application is shown on the right screen in front of the user, the left screens visualize internal system states and signal processing details. Various sensors can be seen mounted on the rack.

While explicit user signals are directly fed into the input fusion component, sending signals to the ticket application, implicit signals are first combined and further abstracted within a dedicated data fusion component. The architecture and the communication middleware of the underlying system represents a specific instance of the generic *Companion* architecture as described in Chap. 22. The planning and dialog management tasks are realized within the application component. The same

[2] A video of the complete scenario in German language is available at:
http://www.uni-ulm.de/en/in/sfb-transregio-62/pr-and-press/videos.html

applies to the storage about the user and his or her preferences using a knowledge base component. The application offers stepwise dialogs gathering the most relevant information for purchasing a train ticket, where the dialog steps are sensitive to the interpreted signals and data. The dialog flow can be automatically adapted within processing time.

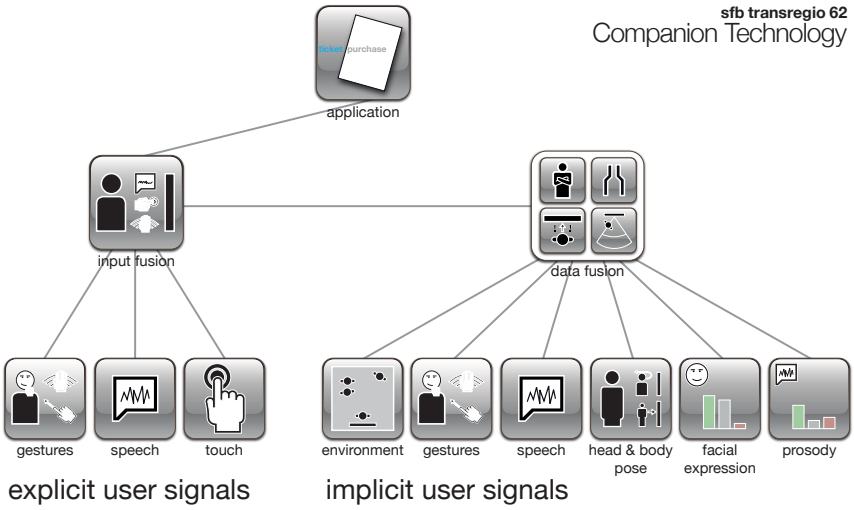

Fig. 2 Bottom-up information flow in the train ticket purchase system. Explicit user signals (i.e. gestures, speech and touch inputs) are directly combined in the input fusion and sent to the application. Implicit signals stem from various sensors. Laser scanners observe the user's environment, video capturing devices gather implicit gestures, facial expressions, as well as the head and body pose, audio devices analyze the speech and nonverbal signals. In the data fusion component, implicit signals are combined into high level information about the user's state, such as disagreement with the system behavior, attentiveness to the system, or hastiness.

The following section describes a complete run through of a normal ticket purchase with details on aspects of signal processing, information fusion, user adaption and interaction. The final section highlights the possibilities of a fast–track ticket purchase when several information indicate a hurried user.

2 User– and Situation–adaptive Ticket Purchase

The ticket purchase starts with a user approaching the system. The standard process requires the specification of the destination, travel time, number of tickets and train connection. The leaving of the device finally marks the end of a normal purchase.

Approaching the Device: The activation of the ticket purchase system is triggered based on the environment perception system and the head and body pose recogni-

tion. The environment perception system estimates the locations of all persons in the proximity using two laser–range–finders and the multi–object tracking algorithms introduced in Chap. 15. All tracked persons heading towards the ticket purchase system are considered as potential users. In order to prevent an activation of the system due to passers-by, the data fusion software component combines the state of the potential users with the results of the head and body pose estimation (cf. Chap. 17). Hence, only users which are approaching and facing the system will trigger the beginning of a ticket purchase. After a new user has been detected, the application starts using a range–dependent fade–over from the stand-by screen to the welcome screen of the purchase process. The position of the active user in the calibrated coordinate system is transferred to other software components to prevent the system from confusing the active user with non-active users.

Destination Selection: The next step in the ticket purchase process is the selection of the travel destination. Given that the user is already known to the system[1], individual contextual information is provided, e.g. the user's most frequent travel destinations are automatically suggested on a graphical map. The user selects a destination by performing a touch input on the displayed map or by specifying the destination via direct speech input.

Time Selection: The third step comprises the selection of the date and time. Again, the system displays a personalized dialog of the user's schedule given that the user is already known. The dialog allows the user to select a time slot for the trip. The *Companion*-System is aware that the interaction takes place in a public area and, therefore, observes the predefined privacy policy. Hence, the system will display only whether a specific slot is already blocked or not.
The travel time selection is conducted using both, speech and gestural input. While it is more convenient to specify the date and time using speech, e.g. "I want to travel at 8 am on Wednesday", the browsing trough the calendar is performed more naturally with the gestural input, e.g. by using the "wiping to the left" gesture to get to the next week. However, the complete functionality is provided by both modalities, e.g. a specific time slot is selected by holding a pointing gesture for a few seconds. The screen coordinate of the pointing direction is computed in two ways. If the gesture recognition system recognizes the user's arm as being outstretched, the line from the head to the hand is extended until it intersects with the screen. When the user's hand is recognized as being close to the user's body, the pointing direction is adjusted by local hand movements. Additionally, a graphical feedback is presented on the screen to indicate the location the user is pointing at. The speech and gestural inputs are recognized independently and integrated within the input fusion (see Fig. 2). The systems further allows combined and relative inputs such as pointing on a specific time and uttering the speech command "This time" to perform a fast confirmation of the selected time. In this case the input fusion does not wait, since it can take advantage of the explicit speech command.

[1] This can be elicited e.g. by authorized data transfer from the user's mobile device.

Ticket Number Selection: The system further performs an adaption to the current context. This is exemplified by the automatic pre-selection of the number of tickets. The system has initially observed with the help of the multi–object tracking algorithm whether the user detached himself from a group of other persons. The group is defined by spatial regions and common trajectories in the past. The tracking algorithm provides the group size such that the application is able to automatically suggest to buy either a single ticket or tickets for all people in the group.

Interruption: Another issue a speech–controlled technical system has to deal with is to distinguish between user commands intended to control the system and other unrelated utterances. Unrelated utterances can often be denoted as "off–talk" in human–computer interaction (cf. Chap. 20). As long as the content of this off–talk is different from system commands this differentiation can be purely based on the speech content in the speech recognizer itself. But in situations where the off-talk contains the same phrases, e.g. the user conducts clarifying dialogs with his or her co-passengers or agrees upon the journey via mobile phone, this assumption cannot be taken. In this case the decision whether the user utterances are intended to control the system or not can only be decided using additional modalities. Two types of off–talk events are recognized by the system: (1) turning away from the system and (2) talking to somebody over the phone.
In the first case, the system recognizes the pose of the active user. As long as the pose is not directed towards the system, the output of the speech recognizer will be discarded until the user's turns again towards the system.
In the second case, the system interprets both, the recognized speech and gesture. It is assumed that in order to make a phone call the mobile phone will be moved to the user's ear and typical phrases of receiving or initiating a call are uttered. The self–touch of the user's ear is detected by the gesture recognizer, while the speech recognizer detects the greeting of the beginning phone conversation. Both events are recognized independently in their respective software components and passed over to the data fusion component. If both events occur within the same short period of time, they are detected as off–talk. The off–talk disables the speech recognizer and the speech synthesizer as long as the gesture of the user does not change. Once the phone conversation ends, the system enables both the speech recognizer and speech synthesizer again.
The off–talk is a welcome example of an implicit user signal. The *Companion*-System has to make an active decision which interferes with the ongoing interaction although no direct command is given.

Connection Selection: After the required information is gathered, the system seeks train connections which suit the user's preferences known from the knowledge base. Ideally, a suitable connection can be provided and the user can successfully complete the ticket purchasing process. However, in our demonstration we will assume that no suitable train connection exists and the system can only approximately match the known user preferences, i.e. reservations are possible and a low number of changes between trains. The system shows connections in which the user can make a reservation, but unfortunately has to change trains very often.

In this phase of interaction, the video and audio data are analyzed to capture the user's emotional states which will serve as an implicit input. The goal is to recognize whether the user shows a facial expression or performs an utterance which indicates that he or she is not satisfied with the pre-selection. The emotional state, i.e. positive and negative valence, is recognized by software components for each modality independently. The recognition using the video data analyzes the facial expressions on the basis of features derived from geometric distances measured in the face of the active user (e.g. mouth width/height, eye–brow distance), see Chap. 18. The recognition using the audio channel starts by extracting mel frequency cepstral coefficients which are then classified using a probabilistic support vector machine (cf. Chap. 20). The outputs of the audio and video based recognitions are then combined in the data fusion component using a Markov fusion network which is able to deal with temporally fragmented intermediate classification inputs (cf. Chap. 19).
In case a negative reaction is recognized, this information is sent to the input fusion module (for the connections see Fig. 2) which triggers the application component in order to ask the user if the pre-selection should be adapted. The application then expands the list of train connections such that the user is able to choose a connection which is the most acceptable and to continue by paying the tickets.

Leaving the Device: After the purchase process, the system remains active as long as the user does not turn away from the system. The end of the interaction is triggered by the environment perception system and the head and body pose, i.e. the system only returns to stand-by mode if the distance of the user exceeds a certain threshold and the user is no longer facing the system.

3 Hurried User

The ticket purchase system supports two basic interaction modes: the normal ticket purchase mode and a quick purchase for users in a hurry.

The selection of the actual interaction mode (normal vs. quick purchase) is done automatically based on an analysis of the approaching speed of the user to the ticket purchase system. In case the approach speed exceeds a given threshold, the system decides for the quick purchase mode, it double–checks via asking the user at the beginning of the interaction whether he or she is really in a hurry to avoid misinterpretations of the approaching speed.

In quick purchase mode, the user has less choices during the purchase to achieve a shorter overall interaction time. The system automatically makes choices by considering the available information in order to omit the corresponding queries. For instance, the system sets the number of tickets to be purchased equal to the size of the group of people the active user has detached from. This number is derived based on the data of the laser range sensors. Since the system knows that the user is in a hurry, it proposes to take a train that leaves close to the current time. The current train station is set for departure such that only the destination needs to be selected, e.g. via speech command. If there is an ambiguity, the system resolves it by present-

ing additional user queries. Finally, the user has to confirm the pre-selected train connection to complete the purchase.

4 Conclusion

In this chapter, we showed how several key components of *Companion* technology have been exemplarily integrated into a prototypical ticket purchase system. Besides the possibility of using several input modalities, the proposed system adapts its behavior to the current situation and interprets implicit input data. The adaptation to the current situation is demonstrated by the interruption handling, the ticket number selection and the automatic switching between standard and hurried mode. Examples for implicit input data are the interpretation of the facial expressions and voice. Deliberately, the potentials of planning, data interpretation, and dialog components have been kept low in this system. They are demonstrated in Chap. 24 for example. The main focus of this demonstrator was to elaborate the multimodal signal processing capabilities developed for *Companion*-System.

Acknowledgements The authors thank the following colleagues for their invaluable support in realizing the scenario (in alphabetical order): Michael Glodek, Ralph Heinemann, Peter Kurzok, Sebastian Mechelke, Andreas Meinecke, Axel Panning, and Johannes Voss.

This work was done within the Transregional Collaborative Research Centre SFB/TRR 62 "*Companion*-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).