MULTIMODAL INFORMATION PROCESSING: THE TICKET PURCHASE – A DEMONSTRATION SCENARIO OF THE SFB/TRR-62

Ingo Siegert[1], *Stephan Reuter*[2], *Felix Schüssel*[2], *Georg Layher*[2], *Thilo Hörnle*[2], *Sascha Meudt*[2], *Andreas Wendemuth*[1,3]
[1]*Otto-von-Guericke University Magdeburg,* [2]*Ulm University*
and [3]*Center for Behavioral Brain Sciences*
ingo.siegert@ovgu.de

Abstract: The demonstration scenario of the SFB/TRR-62 shows multimodal, dynamic interactions between a human being and a technical system that are adaptive to the situation and the emotional state. It uses the example of purchasing a train ticket to demonstrate how a companion system is able to adapt its dialog with the user according to the situational context and the emotions of the user. One special feature of this scenario are the simultaneous analyses and evaluations of explicit and implicit input data. The scenario demonstrates further how background knowledge about the user can be included; for example often visited destinations, the user's timetable or the number of travelers.

1 Introduction

Technical cognitive systems recently received increasing attention. Besides making the operation of future technical systems as simple as possible, one goal is to enable a natural interaction. In this context, future systems have to be adaptive to both the actual situation and the user's emotional state [6]. Furthermore, they are faced with the challenge to analyze and interpret observations from various sensors. Systems providing this extended recognition abilities, adapting to the user's needs and the actual situation, ultimately become the user's *companions* [2].

In this paper, a demonstration of an impressive multimodal dynamic human-computer interaction conceived by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" is described[1]. As scenario, the purchasing of a train ticket is used. The presented system adapts to the current situation and the user's state by implementing multimodal information processing. The system therefore simultaneously analyses and evaluates explicit and implicit input data.

The remainder of the paper is structured as follows: the technical setup and available sensors are described in Section 2. Afterwards, Section 3 addresses in detail the explicit and implicit user signals. The course of the demonstrated ticket purchase is then described in Section 4. Section 5 concludes this paper.

2 Technical Setup

An experimental platform is equipped with a wide range of sensors to capture the situative context as well as the user. The sensors comprise video and audio capturing devices, laser scanners, a touch screen, and a stereo camera, see Figure 1.

[1]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`

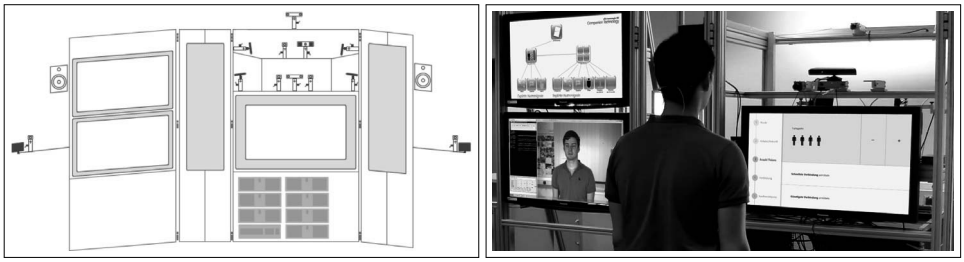

Figure 1 - The setup of the demonstration scenario (left) and a picture of a user interacting with the ticket purchase system (right). While the application is shown on the right screen in front of the user, the left screens visualize internal system states and signal processing details. Various sensors can be seen mounted on the rack.

The information processing is realized by multiple components, retrieving data directly from sensors or using software components interpreting pre-processed sensor information. As message exchange system, the Semaine API based on Apache ActiveMQ is used [10]. This allows the various information gathering modules to exchange massages about state changes as well as new user inputs in real-time. More details about the middle-ware can be found in [4].

3 Explicit / Implicit User Signals

The various components can be distinguished into explicit and implicit user signals. Explicit signals are emitted 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 relevant information, comprising the user's situative context, his non-interaction gesture, non-commanding speech, body pose, facial expressions and prosody.

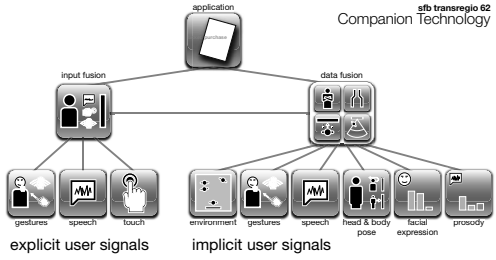

Figure 2 - Bottom-up information flow in the train ticket purchase system depicting the separation of explicit user signals, implicit signals and their particular processing.

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, see Figure 2.

The planning and dialog management tasks are realized within the application component. The same applies to the storage of the user's 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.

4 User- and Situation-adaptive Ticket Purchase

Figure 3 gives a basic blockwise overview of the entire ticket purchase. A user approaching the system initiates the purchase process. The interaction further requires the specification of destination, travel time, number of tickets and train connection. The end of a purchase is indicated by the user leaving the device.

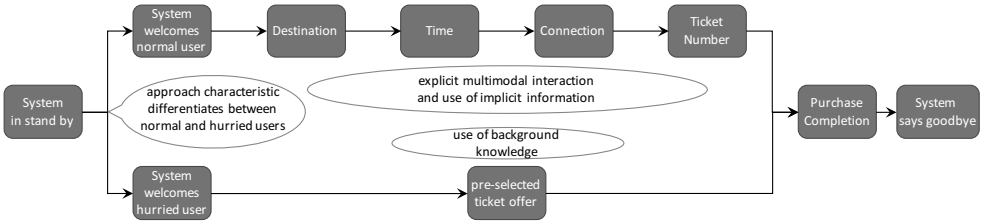

Figure 3 - Overview of the complete ticket purchase process. This comes in two flavors, depending on the characteristics of a user's approach (cf. Section 4.2). The normal purchase is shown in the top row, the quick purchase in the bottom row.

During the interaction, the user can switch freely between several explicit input modalities: touch, speech, gestures. Additionally, various sensor input states are interpreted implicitly, to distinguish hurried users, users approaching out of a group of people, recognizing user's turning away from the system or interruptions, like an incoming cellphone-call.

The scenario further demonstrates how additional background knowledge about the user can be included, e.g. frequently visited destinations or the user's timetable. The system then adapts to the user type accordingly, e.g. by presenting to the hurried user only essential and pre-selected information, applying as much system initiative and implicit verification strategy as possible.

In the following, selected parts of the interaction are presented in more detail:

4.1 Application Control by Explicit User Signals

Destination Selection:

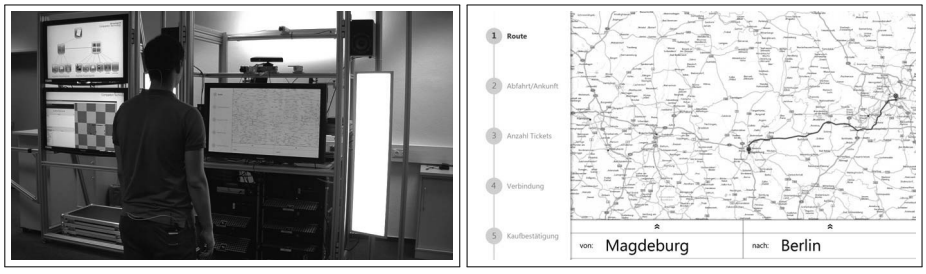

Figure 4 - Destination selection via speech input (left). The recognized destination is indicated on the screen (right).

The very first step in the ticket purchase process, after approaching the device, is the selection of the travel destination. The user selects a destination by performing a touch input on the displayed map or by specifying the destination via direct speech input. Possible input phrases

– city names, pronunciation variants – are stored in a grammar. As it can be seen in Fig. 4, the user's individual contextual information is provided, e.g. the most frequent travel destinations. For this, the user has to be already known to the system. This can be elicited by e.g. authorized data transfer from the user's mobile device.

As additional input method, the user can make use of the touch screen and select the destination either touching the city on the map, or selecting one of the user's most frequent travel destinations from a drop-down list.

Time Selection: This dialog step allows the user to select a time slot for the trip. The system displays a personalized dialog of the user's schedule given that the user is already known. 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 blocked or not.

Both, speech and gestural input can be used by the user for the travel time selection, depending on the designated task. As 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 additional gestural input, e.g. by using the "wiping to the left" gesture to change the depicted week, or by using a pointing gesture to select the time-slot, see Fig. 5.

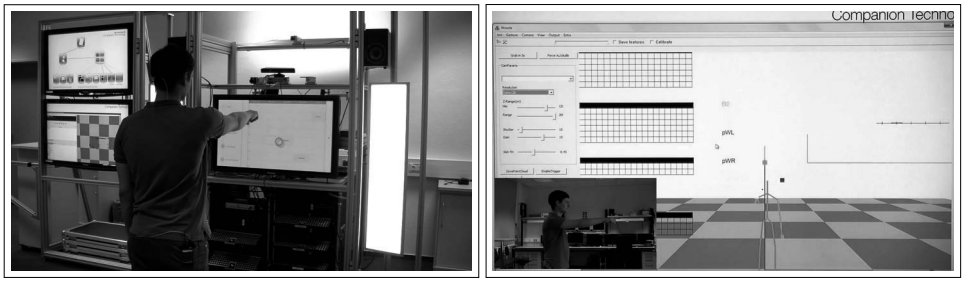

Figure 5 - User performing a pointing gesture to select a desired time slot (left). A view into the systems inner models for evaluating gestures, the skeleton pattern of the pointing gesture (right).

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.

4.2 Application Control by Implicit User Signals

Ticket Number Selection: The continuous perception of the environment facilitates the adaptation of the system behavior to the current context. The dynamic environment model is realized using a multi-object tracking algorithm [7, 8] which delivers the trajectories of all persons in the

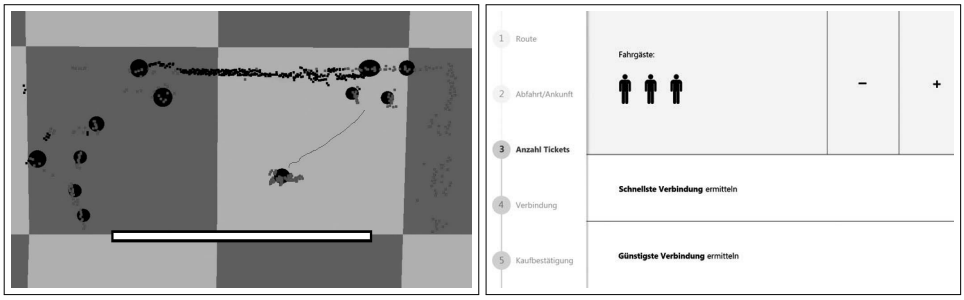

Figure 6 - Pre-selection of ticket number using environment perception. The laser tracking system observes a user approaching the system out of a group (left). This information is used to automatically include the group members into the ticket purchase process (right).

surroundings. As a first example for using the obtained information, the system automatically adapts the pre-selected number of tickets. Due to the continuous tracking of all persons in the proximity, the system is able to detect that the current user belongs to a group of persons. The affiliation of individual persons is computed based on the spatial proximity and the similarity of the trajectories. Using the estimated number of persons for the user's group, the application automatically suggests to buy a ticket for the whole group, see Fig. 6.

Interruption: In real scenarios a reliable discrimination of user commands and unrelated utterances is essential for the performance of the technical system. In speech-controlled systems, these unrelated utterances (e.g. talking with other people of the group) are often denoted as "off–talk". Since these conversations are often related to the content of the interaction with the system, a differentiation of the off–talk from system commands solely based on speech content or keywords is error-prone. The availability of additional modalities facilitates an information fusion approach which improves the reliability of the off–talk detection and the system performance itself.

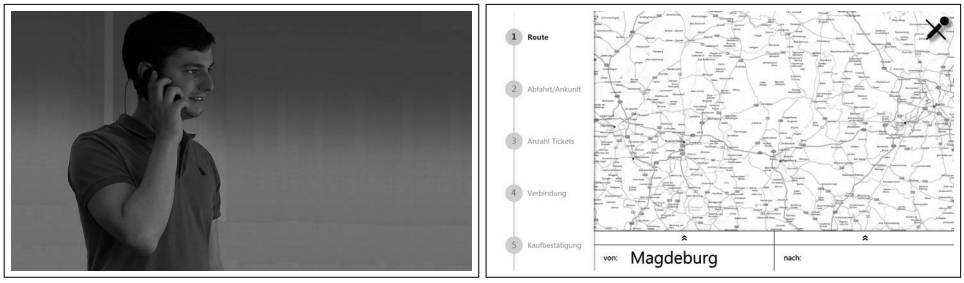

Figure 7 - The user receives a phone call (left). Due to the joint observation of phone-typical touch gestures utterances, the system disables the speech input and output modules (right).

The implemented system is able to detect two typical types of off-talk: 1) turning away from the system and 2) cell phone calls. The case of turning away from the system is based on the pose estimation for the active user. As soon as the user starts to turn away from the system, the pose estimation module informs the system about the possibility of off–talk. In combination

with the information of the multi-object tracking about other persons in the surroundings, the system infers the likelihood of an off–talk event and reduces the reliability of the speech input. The detection of cell phone calls is based on the combination of speech recognition and gesture classification. The beginning of a phone call is characterized by moving the phone to one of the ears as well as uttering typical phrases, see Fig 7. The movement of the hand towards the ear is recognized and classified by the gesture module based on depth images. Obviously, this gesture could also be part of an action like smoothing one's hair back. Hence, a reliable detection of the phone call additionally requires the detection of typical greetings at the beginning of the conversation using a speech recognizer. The data fusion component combines the information of both modules in a probabilistic way and delivers the probability of an off-talk event.

In case of a detected off-talk, the system disables the speech recognizer and additionally stops to use the speech synthesizer as output modality in order to prevent interruptions of the user's off–talk, see Fig 7. The presented off-talk scenarios are intuitive examples for implicit user signals where the system imitates the behavior of a human conversation partner. Although the system does not receive an explicit command from the user, it infers the intended behavior using the implicit signals.

Connection Selection: When the system has gathered all required information, the train connections which suit the known user's preferences are seeked 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 assume that no suitable train connection exists and the system can only approximately match the known user preferences. We demonstrate a case where either reservations are possible or a low number of changes between trains can be achieved. Thus, the system shows connections in which the user can make a reservation, but unfortunately has to change trains very often.

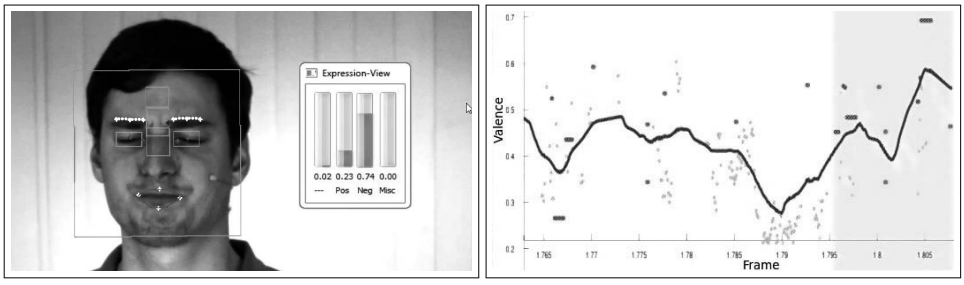

Figure 8 - Screenshot of the facial recognition module (left). Observing lip and eye brow movement, a negative expression is recognized. Screenshot of the fusion module (right), integrating acoustic (filled dots) and visual information (small dots).

In this phase of the interaction, face and voice data of the user are analyzed to capture changing emotional states which will serve as implicit input [9, 11]. The goal is to recognize whether the user shows a facial expression or performs an utterance indicating dissatisfaction with the offered 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 analyses 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 Fig. 8. The recognition using the audio channel starts by extracting mel frequency cepstral coefficients which are then classified using a probabilistic support vector machine. 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. [5]).
If a negative reaction is recognized, the input fusion module triggers the application component 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.

Approaching and Leaving the Device: In standby, the system combines the estimated head and body pose with the distance of subjects nearby to detect whether a potential user is approaching the system. Once a subject is close enough and turned towards the system, he is marked as the active user and his position, head and body pose are tracked and monitored by the system. The dialog now is initiated and the ticket purchase starts. After the purchase process is completed, 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 monitored distance and the head and body pose of the active user, 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.

Hurried User: For demonstrating the adaptivity, the ticket purchase system offers the user two different interaction modes: a "normal" mode and one for quicker purchases for hurried users. In the mode for hurried users the system limits the possible choices for the user. By applying an adaptive preselection the purchase process is speeded up. The adaptivity is achieved using the data from the system's knowledge base and the recognition modules of the various sensors. The system decides to switch to the quicker interaction mode when the user approaches the system more quickly than with a predefined threshold. To secure the selection of the quicker mode, the system asks for confirmation to prevent unnecessary limitations of the user's interaction possibilities.
With the knowledge about the current time and the time table of the current train station, the system preselects the departure time – more specifically, in the hurried mode trains that are leaving in a short period of time are preselected. The user only selects his destination and the system automatically selects the next available train to this destination, allowing a very fast interaction. At the end, the system shows the user an overview about the complete purchase for confirmation. The interaction is finished and the user can move on to his train.

5 Conclusion

This scenario demonstrated how several key components of companion technology can seamlessly be integrated into a prototypical ticket purchase system. Besides the possibility of using several input modalities, the system adapts its behavior to the current situation and interprets implicit input data.
By demonstrating a proper interruption handling, correct selection of ticket numbers for groups, and the automatic switching between standard and hurried mode, the adaptation to the current situation is demonstrated. Further 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 another demonstration scenario of the SFB — "The Advanced User Assistance for Setting Up a Home Theater", see for example [1, 3]. The main focus of this demonstrator was to elaborate the multimodal signal processing capabilities developed for Companion–Systems.

Acknowledgments

The work presented in this paper was done within the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” (www.sfb-trr-62.de) funded by the German Research Foundation (DFG).

References

[1] BERCHER, P., F. RICHTER, T. HÖRNLE, T. GEIER, D. HÖLLER, G. BEHNKE, F. NOTHDURFT, F. HONOLD, W. MINKER, M. WEBER and S. BIUNDO: *A planning-based assistance system for setting up a home theater*. In *Proc. of the 29th Nat. Conf. on AI (AAAI)*, pp. 4264–4265, Feb 2015.

[2] BIUNDO, S. and A. WENDEMUTH: Companion-*Technology for Cognitive Technical Systems*. Künstliche Intelligenz, 2016.

[3] HONOLD, F., P. BERCHER, F. RICHTER, F. NOTHDURFT, T. GEIER, R. BARTH, T. HÖRNLE, F. SCHÜSSEL, S. REUTER, M. RAU, G. BERTRAND, B. SEEGEBARTH, P. KURZOK, B. SCHATTENBERG, W. MINKER, M. WEBER and S. BIUNDO: *Companion-Technology: Towards User- and Situation-Adaptive Functionality of Technical Systems*. In *10th IEEE Int. Conf. on Intelligent Environments*, pp. 378–381, Feb 2014.

[4] HÖRNLE, T., M. TORNOW, F. HONOLD, R. SCHWEGLER, R. HEINEMANN, S. BIUNDO, and A. WENDEMUTH: *Companion Systems: A Reference Architecture*. In BIUNDO, S., and A. WENDEMUTH (eds.): *Companion Technology – A Paradigm Shift in Human-Technology Interaction*. Springer, Berlin, Heidelberg, New York, 2016. in press.

[5] KRELL, G., M. GLODEK, A. PANNING, I. SIEGERT, B. MICHAELIS, A. WENDEMUTH and F. SCHWENKER: *Fusion of Fragmentary Classifier Decisions for Affective State Recognition*. In SCHWENKER, F., S. SCHERER and L.-P. MORENCY (eds.): *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, vol. 7742 of *LNAI*, pp. 116–130. Springer, Berlin, Heidelberg, Germany, 2013.

[6] PICARD, R. W.: *Affective Computing*. MIT Press, Cambridge, USA, 1997.

[7] REUTER, S. and K. DIETMAYER: *Pedestrian Tracking Using Random Finite Sets*. In *Proceedings of the 14th International Conference on Information Fusion*, pp. 1–8, 2011.

[8] REUTER, S., B. WILKING, J. WIEST, M. MUNZ and K. DIETMAYER: *Real-Time Multi-Object Tracking using Random Finite Sets*. IEEE Transactions on Aerospace and Electronic Systems, 49(4):2666–2678, 2013.

[9] SAEED, A., A. AL-HAMADI, R. NIESE and M. ELZOBI: *Frame-Based Facial Expression Recognition Using Geometrical Features*. Advances in Human-Computer Interaction, 2014.

[10] SCHRÖDER, M.: *The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems*. Advances in Human-Computer Interaction, 2010, 2010.

[11] SIEGERT, I., D. PHILIPPOU-HÜBNER, K. HARTMANN, R. BÖCK and A. WENDEMUTH: *Investigation of Speaker Group-Dependent Modelling for Recognition of Affective States from Speech*. Cognitive Computation, 6(4):892–913, 2014.