Journal of Ambient Intelligence and Smart Environments 1 (2009) 23–30
DOI 10.3233/AIS-2009-0003
IOS Press

Multimodal interfaces: Challenges and perspectives

Nicu Sebe[a,b]
[a] *University of Amsterdam, The Netherlands. E-mail: nicu@science.uva.nl*
[b] *University of Trento, Italy. E-mail: niculae.sebe@unitn.it*

Abstract. The development of interfaces has been a technology-driven process. However, the newly developed multimodal interfaces are using recognition-based technologies that must interpret human-speech, gesture, gaze, movement patterns, and other behavioral cues. As a result, the interface design requires a human-centered approach. In this paper we review the major approaches to multimodal Human Computer Interaction, giving an overview of the user and task modeling, and of the multimodal fusion. We highlight the challenges, open issues, and the future trends in multimodal interfaces research.

Keywords: Multimodal interfaces, human-centered computing, adaptability, fusion, interface design

1. Introduction

In human-human interaction, interpreting the mix of audio-visual signals is essential in communicating. Researchers in many fields recognize this, and thanks to advances in the development of unimodal techniques (in speech and audio processing, computer vision, etc.), and in hardware technologies (inexpensive cameras and sensors), there has been a significant growth in multimodal HCI research. Unlike traditional HCI applications (a single user facing a computer and interacting with it via a mouse or a keyboard), in the new applications (e.g., intelligent homes [41], remote collaboration, arts, etc.), interactions are not always explicit commands, and often involve multiple users. This is due in part to the remarkable progress in the last few years in computer processor speed, memory, and storage capabilities, matched by the availability of many new input and output devices that are making ubiquitous computing [77] a reality. Devices include phones, embedded systems, PDAs, laptops, wall size displays, and many others. The wide range of computing devices available, with differing computational power and input/output capabilities, means that the future of computing is likely to include novel ways of interaction (e.g., using gestures [59], speech [62], haptics [3], eye blinks [23], and many others). Glove mounted devices [8] and graspable user interfaces [20], for example, seem now ripe for exploration. Pointing devices with haptic feedback, eye tracking, and gaze detection [27] are also currently emerging. As in human-human communication, however, effective communication is likely to take place when different input devices are used in combination.

Multimodal interfaces have been shown to have many advantages [12]: they prevent errors, bring robustness to the interface, help the user to correct errors or recover from them more easily, bring more bandwidth to the communication, and add alternative communication methods to different situations and environments. Disambiguation of error-prone modalities using multimodal interfaces is one important motivation for the use of multiple modalities in many systems. As shown by Oviatt [48], error-prone technologies can compensate each other, rather than bring redundancy to the interface and reduce the need for error correction. It should be noted, however, that multiple modalities alone do not bring benefits to the interface: the use of multiple modalities may be ineffective or even disadvantageous. In this context, Oviatt [49] has presented the common misconceptions (myths) of multimodal interfaces, most of them related to the use of speech as an input modality.

Extensive surveys have been previously published in several related areas such as face detection [24,78], face recognition [80], facial expression analysis [19,56], vocal emotion [43,46], gesture recognition [38,59,71], human motion analysis [18,22,26,42,76], audio-visual automatic speech recognition [62], and eye tracking [13,16]. Reviews of vision-based HCI are presented in [61] and [30] with a focus on head tracking, face and facial expression recognition [58], eye tracking, and gesture recognition. Adaptive and intelligent HCI is discussed in [15] with a review of computer vision for human motion analysis, and a discussion of techniques for lower arm movement detection, face processing, and gaze analysis. Multimodal interfaces are discussed in [45,51–53,60,63,65, 69]. Real-time vision for HCI (gestures, object tracking, hand posture, gaze, face pose) is discussed in [34] and [33].

Our goal is not to present a comprehensive survey of all the related work, but mainly to discuss the most important issues, the current trends, and the perspectives in the area of multimodal HCI systems in general and multimodal interfaces in particular.

2. Overview of multimodal interaction

The term multimodal has been used in many contexts and across several disciplines (see [4] for a taxonomy of modalities). A multimodal HCI system is simply the one that responds to inputs in more than one modality or communication channel. By modality we mean a mode of communication according to human senses and computer input devices activated by humans or measuring human qualities. The human senses are sight, touch, hearing, smell, and taste. The input modalities of many computer input devices can be considered to correspond to human senses: cameras (sight), haptic sensors (touch) [3], microphones (hearing), olfactory (smell), and even taste [37]. Many other computer input devices activated by humans, however, can be considered to correspond to a combination of human senses, or to none at all: keyboard, mouse, writing tablet, motion input (e.g., the device itself is moved for interaction), galvanic skin response, and other biometric sensors.

In our definition, the word input is of great importance, as in practice most interactions with computers take place using multiple modalities. For example, as we type we touch the keys on a keyboard to input data into the computer, but some of us also use sight to read what we type or to locate the proper keys to be pressed. Therefore, it is important to keep in mind the differences between what the human is doing and what the system is actually receiving as input during interaction. For instance, a computer with a microphone could potentially understand multiple languages or only different types of sounds (e.g., using a humming interface for music retrieval). Although the term multimodal has often been used to refer to such cases (e.g., multilingual input in [5] is considered multimodal), we consider that only a system that uses any combination of different modalities (i.e., communication channels) is multimodal. For example, a system that responds only to facial expressions and hand gestures using only cameras as input is not multimodal, even if signals from various cameras are used. Using the same argument, a system with multiple keys is not multimodal, but a system with mouse and keyboard input is.

In the context of HCI, multimodal techniques can be used to construct many different types of interfaces. Of particular interest are perceptual, attentive, and enactive interfaces. Perceptual interfaces, as defined in [72], are highly interactive, multimodal interfaces that enable rich, natural, and efficient interaction with computers. Perceptual interfaces seek to leverage sensing (input) and rendering (output) technologies in order to provide interactions not feasible with standard interfaces and common I/O devices such as the keyboard, the mouse, and the monitor [72], making computer vision a central component in many cases. Attentive interfaces are context-aware interfaces that rely on a person's attention as the primary input [66] — that is, attentive interfaces [47] use gathered information to estimate the best time and approach for communicating with the user. Since attention is epitomized by eye contact [66] and gestures (although other measures such as mouse movement can be indicative), computer vision plays a major role in attentive interfaces. Enactive interfaces are those that help users communicate a form of knowledge based on the active use of the hands or body for apprehension tasks. Enactive knowledge is not simply multisensory mediated knowledge, but knowledge stored in the form of motor responses and acquired by the act of "doing". Typical examples are the competence required by tasks such as typing, driving a car, dancing, playing a musical instrument, and modeling objects from clay. All of these tasks would be difficult to describe in an iconic or symbolic form.

3. Multimodal interface design

Multimodal interface design [64] is important because the principles and techniques used in traditional GUI-based interaction do not necessarily apply in multimodal HCI systems. Issues to consider, include the design of inputs and outputs, adaptability, consistency, and error handling, among others. In addition, one must consider dependency of a person's behavior on his/her personality, cultural, and social vicinity, current mood, and the context in which the observed behavioral cues are encountered [28,31,68].

Many design decisions dictate the underlying techniques used in the interface. For example, adaptability can be addressed using machine learning: rather than using a priori rules to interpret human behavior, we can potentially learn application-, user-, and context-dependent rules by watching the user's behavior in the sensed context [60]. Well known algorithms exist to adapt the models and it is possible to use prior knowledge when learning new models. For example, a prior model of emotional expression recognition trained based on a certain user can be used as a starting point for learning a model for another user, or for the same user in a different context. Although context sensing and the time needed to learn appropriate rules are significant problems in their own right, many benefits could come from such adaptive multimodal HCI systems.

3.1. System integration architectures

The most common infrastructure that has been adopted by the multimodal research community involves multi-agent architectures such as the Open Agent Architecture [39] and Adaptive Agent Architecture [11,35]. Multi-agent architectures provide essential infrastructure for coordinating the many complex modules needed to implement multimodal system processing and permit this to be done in a distributed manner. In a multi-agent architecture, the components needed to support the multi-modal system (e.g., speech recognition, gesture recognition, natural language processing, multimodal integration) may be written in different programming languages, on different machines, and with different operating systems. Agent communication languages are being developed that handle asynchronous delivery, triggered responses, multi-casting, and other concepts from distributed systems.

When using a multi-agent architecture, for example, speech and gestures can arrive in parallel or asynchronously via individual modality agents, with the results passed to a facilitator. These results, typically an n-best list of conjectured lexical items and related timestamp information, are then routed to appropriate agents for further language processing. Next, sets of meaning fragments derived from the speech, or other modality, arrive at the multimodal integrator which decides whether and how long to wait for recognition results from other modalities, based on the system's temporal thresholds. It fuses the meaning fragments into a semantically and temporally compatible whole interpretation before passing the results back to the facilitator. At this point, the system's final multimodal interpretation is confirmed by the interface, delivered as multimedia feedback to the user, and executed by the relevant application.

Despite the availability of high-accuracy speech recognizers and the maturing of devices such as gaze trackers, touch screens, and gesture trackers, very few applications take advantage of these technologies. One reason for this may be that the cost in time of implementing a multimodal interface is very high. If someone wants to equip an application with such an interface, he must usually start from scratch, implementing access to external sensors, developing ambiguity resolution algorithms, etc. However, when properly implemented, a large part of the code in a multimodal system can be reused. This aspect has been identified and many multimodal application frameworks (using multi-agent architectures) have recently appeared such as Jaspis framework [73,74], Rutgers CAIP Center framework [21], and the Embassi system [17].

3.2. Modeling

There have been several attempts for modeling humans in human-computer interaction literature [79]. Here we present some proposed models and we discuss their particularities and weaknesses.

One of the most commonly used models in HCI is the Model Human Processor. The model, proposed in [9] is a simplified view of the human processing involved in interacting with computer systems. This model comprises three subsystems namely, the perceptual system handling sensory stimulus from the outside world, the motor system that controls actions, and the cognitive system that provides the necessary processing to connect the two. Retaining the analogy of the user as an information processing system, the components of a multimodal HCI model include an input-output component (sensory system), a memory component (cognitive system), and a processing

component (motor system). Based on this model, the study of input-output channels (vision, hearing, touch, movement), human memory (sensory, short-term, and working or long-term memory), and processing capabilities (reasoning, problem solving, or acquisition skills) should all be considered when designing multimodal HCI systems and applications. Many studies in the literature analyze each subsystem in detail and we point the interested reader to [14] for a comprehensive analysis.

Another model proposed by Card et al. [9] is the GOMS (Goals, Operators, Methods, and Selection rules) model. GOMS is essentially a reduction of a user's interaction with a computer to its elementary actions and all existing GOMS variations [9] allow for different aspects of an interface to be accurately studied and predicted. For all of the variants, the definitions of the major concepts are the same. Goals are what the user intends to accomplish. An operator is an action performed in service of a goal. A method is a sequence of operators that accomplish a goal and if more than one method exists, then one of them is chosen by some selection rule. Selection rules are often ignored in typical GOMS analyses. There is some flexibility for the designers/analysts definition of all of these entities. For instance, one person's operator may be another's goal. The level of granularity is adjusted to capture what the particular evaluator is examining.

All of the GOMS techniques provide valuable information, but they all also have certain drawbacks. None of the techniques address user fatigue. Over time a user's performance degrades simply because the user has been performing the same task repetitively. The techniques are very explicit about basic movement operations, but are generally less rigid with basic cognitive actions. Further, all of the techniques are only applicable to expert users and the functionality of the system is ignored while only the usability is considered.

The human action cycle [44] is a psychological model which describes the steps humans take when they interact with computer systems. The model can be used to help evaluate the efficiency of a user interface (UI). Understanding the cycle requires an understanding of the user interface design principles of affordance, feedback, visibility, and tolerance. This model describes how humans may form goals and then develop a series of steps required to achieve that goal, using the computer system. The user then executes the steps, thus the model includes both cognitive and physical activities.

3.3. Adaptability

The number of computer users (and computer-like devices we interact with) has grown at an incredible pace in the last few years. An immediate consequence of this is that there is much larger diversity in the "types" of computer users. Increasing differences in skill level, culture, language, and goals have resulted in a significant trend towards adaptive and customizable interfaces, which use modeling and reasoning about the domain, the task, and the user, in order to extract and represent the user's knowledge, skills, and goals, to better serve the users with their tasks. The goal of such systems is to adapt their interface to a specific user, give feedback about the user's knowledge, and predict the user's future behavior such as answers, goals, preferences, and actions [32]. Several studies [70] provide empirical support for the concept that user performance can be increased when the interface characteristics match the user skill level, emphasizing the importance of adaptive user interfaces.

Adaptive human-computer interaction promises to support more sophisticated and natural input and output, to enable users to perform potentially complex tasks more quickly, with greater accuracy, and to improve user satisfaction. This new class of interfaces promises knowledge or agent-based dialog, in which the interface gracefully handles errors and interruptions, and dynamically adapts to the current context and situation, the needs of the task performed, and the user model. This interactive process is believed to have great potential for improving the effectiveness of human-computer interaction [40], and therefore, is likely to play a major role in multimodal HCI. The overarching aim of intelligent interfaces is to both increase the interaction bandwidth between human and machine and, at the same time, increase interaction effectiveness and naturalness by improving the quality of interaction. Effective human machine interfaces and information services will also increase access and productivity for all users [36]. A grand challenge of adaptive interfaces is therefore to represent, reason, and exploit various models to more effectively process input, generate output, and manage the dialog and interaction between human and machine so that to maximize the efficiency, effectiveness, and naturalness of interaction [57].

One central feature of adaptive interfaces is the manner in which the system uses the learned knowledge. Some works in applied machine learning are designed to produce expert systems that are intended

to replace the human. However, works in adaptive interfaces intend to construct advisory-recommendation systems, which only make recommendations to the user. These systems suggest information or generate actions that the user can always override. Ideally, these actions should reflect the preferences of the individual users, thus providing personalized services to each one.

Every time the system suggests a choice to the user he/she accepts or rejects it, thus giving feedback to the system to update its knowledgebase either implicit or explicit [2]. The system should carry out online learning, in which the knowledgebase is updated each time an interaction with the user occurs. Since adaptive user interfaces collect data during their interaction with the user, one naturally expects them to improve during the interaction process, making them "learning" systems rather than "learned" systems. Because adaptive user interfaces must learn from observing the behavior of their users, another distinguishing characteristic of these systems is their need for rapid learning. The issue here is the number of training cases needed by the system to generate good advice. Thus, it is recommended the use of learning methods and algorithms that achieve high accuracy from small training sets. On the other hand, the speed of interface adaptation to user's needs is desirable but not essential.

Adaptive user interfaces should not be considered a panacea for all problems. The designer should seriously take under consideration if the user really needs an adaptive system. The most common concern regarding the use of adaptive interfaces is the violation of standard usability principles. In fact, there exists evidence that suggests that static interface designs sometimes promote superior performance than adaptive ones [25,67]. Nevertheless, the benefits that adaptive systems can bring are undeniable and therefore more and more research efforts are being paid towards this direction.

An important issue is how the interaction techniques should change to take this varying input and output hardware devices into account. The system might choose the appropriate interaction techniques taking into account the input and output capabilities of the devices and the user preferences. So, nowadays, many researchers are focusing on such fields as context aware interfaces, recognition-based interfaces, intelligent and adaptive interfaces, and multimodal perceptual interfaces [32,36,40,72].

Although there have been many advances in multimodal HCI, the level of adaptability in current systems is rather limited and there are many challenges left to be investigated.

3.4. Fusion

Fusion techniques are needed to integrate input from different modalities and many fusion approaches have been developed. Early multimodal interfaces were based on a specific control structure for multimodal fusion. For example, Bolt's "Put-That-There" system [7] combined pointing and speech inputs and searched for a synchronized gestural act that designates the spoken referent. To support more broadly functional multimodal systems, general processing architectures have been developed which handle a variety of multimodal integration patterns and support joint processing of modalities [6,35,39].

A typical issue of multimodal data processing is that multisensory data are typically processed separately and only combined at the end. Yet, people convey multimodal (e.g., audio and visual) communicative signals in a complementary and redundant manner (as shown experimentally by Chen [10]). Therefore, in order to accomplish a human-like multimodal analysis of multiple input signals acquired by different sensors, the signals cannot be always considered mutually independently and might not be combined in a context-free manner at the end of the intended analysis but, on the contrary, the input data might preferably be processed in a joint feature space and according to a context-dependent model. In practice, however, besides the problems of context sensing and developing context-dependent models for combining multisensory information, one should cope with the size of the required joint feature space. Problems include large dimensionality, differing feature formats, and time-alignment. A potential way to achieve multisensory data fusion is to develop context-dependent versions of a suitable method such as the Bayesian inference method proposed by Pan et al. [55].

3.5. Evaluation

Evaluation is a very important issue in the design of multimodal systems. Here, we outline the most important features that could be used as measures in the evaluation of various types of adaptive multimodal HCI systems namely, efficiency, quality, user satisfaction, and predictive accuracy.

People typically invoke computational decision aids because they expect the system will help them accomplish their tasks more rapidly and with less effort than they do on their own. This makes efficiency an important measure to use in evaluating adaptive systems. One natural measure of efficiency is the time the user takes to accomplish his task. Another facet is the effort the user must exert to make a decision or solve a problem. In this case, the measure would be the number of user actions or commands that take place during the solving of a problem.

Another main reason the users turn to multimodal HCI systems is to improve the quality of solutions of their task. As with efficiency, there are several ways in which one can define the notion of quality or accuracy of the system. For example, if there is a certain object the user wants to find then the success of finding it constitutes an objective measure of quality. However, it is clear that in some cases it is necessary to rely on a separate measure of user satisfaction to determine the quality of the system's behavior. One way to achieve this is to present each user with a questionnaire that asks about his subjective experience. Another measure of user's satisfaction involves giving the user some control over certain features of the system. If the user turns the system's advisory capability off or disables its personalization module, one can then conclude that the user has not been satisfied by his experience with these features.

Since many adaptive system user models make predictions about the user's responses, it is natural to measure the predictive accuracy to determine the success of a system. Although this measure can be a useful analytical tool for understanding the details of the system's behavior, it does not necessarily reflect the overall efficiency or quality of solutions, which should be the main concern.

4. Future directions and conclusion

Multimodal interfaces based on recognition of human speech, gaze, gesture, etc. which are currently developed represent only starting points towards intelligent interfaces capable of human-like sensory perception. The future interfaces will interpret continuous input from visual, auditory, and tactile input modes, as well as sensor-based information from the system interface and the surrounding environment. The future adaptive multimodal interfaces will support intelligent adaptation to the user, task, and the working environment. These long-term directions are expected to yield new computational functionality, greater robustness, and improved flexibility for personalization and mobility.

The most promising approach to multimodal interface design is to use a human-centered approach [29,50]. One of the major conclusions is that most researchers process each channel (visual, audio) independently, and multimodal fusion is still in its infancy. On one hand, the whole question of how much information is conveyed by "separate" channels may inevitably be misleading. There is no evidence that individuals in actual social interaction selectively attend to another person's face, body, gesture, or speech, or that the information conveyed by these channels is simply additive. The central mechanisms directing behavior cut across channels, so that, for example, certain aspects of face, body, and speech are more spontaneous and others are more closely monitored and controlled. It might well be that observers selectively attend not to a particular channel but to a particular type of information (e.g., cues to emotion, deception, or cognitive activity), which may be available within several channels. No investigator has yet explored this possibility or the possibility that different individuals may typically attend to different types of information (see [54] for a recent study on this topic).

Considering all these aspects, multimodal context-sensitive human-computer interaction is likely to become the single most widespread research topic of the artificial intelligence research community [60]. Advances in this area could change not only how professionals practice computing, but also how mass consumers interact with technology.

Acknowledgment

This work has been partially supported by the MIAUCE European Project (IST-5-0033715-STP).

References

[1] J.K. Aggarwal and Q. Cai, "Human motion analysis: A review," CVIU, 73(3):428-440, 1999.

[2] M. Balabanovic, "Exploring versus exploiting when learning user models for text recommendations," User Modeling and User-adapted Interaction, 8:71-102, 1998.

[3] M. Benali-Khoudja, M. Hafez, J.-M. Alexandre, and A. Kheddar, "Tactile interfaces: A state-of-the-art survey," Int. Symposium on Robotics, 2004.

[4] N.O. Bernsen, "Defining a taxonomy of output modalities from an HCI perspective," Computer Standards and Interfaces, 18(6-7):537-553, 1997.

[5] N.O. Bernsen, "Multimodality in language and speech systems - From theory to design support tool," Multimodality in Language and Speech Systems, Kluwer, 2001.
[6] M. Blattner and E. Glinert, "Multimodal integration," IEEE Multimedia, 3(4):14-24, 1996.
[7] R. Bolt, "Put-That-There: Voice and gesture at the graphics interface," Computer Graphics, 14(3):262-270, 1980.
[8] C. Borst and R. Volz, "Evaluation of a haptic mixed reality system for interactions with a virtual control panel," Presence: Teleoperators and Virtual Environments, 14(6), 2005.
[9] S.K. Card, T. Moran, and A. Newell, The Psychology of Human-computer Interaction, Lawrence Erlbaum, 1983.
[10] L.S. Chen, Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-computer Interaction, PhD thesis, UIUC, 2000.
[11] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T.S. Huang, "Semi-supervised learning of classifiers: Theory, algorithms, and their applications to human-computer interaction," IEEE Trans. on PAMI, 22(12):1553-1567, 2004.
[12] P.R. Cohen and D.R. McGee, "Tangible multimodal interfaces for safety-critical applications," Communications of the ACM, 47(1):41-46, 2004.
[13] Computer Vision and Image Understanding, Special Issue on Eye Detection and Tracking, 98(1), 2005.
[14] A. Dix, J. Finlay, G. Abowd, and R. Beale, Human-computer Interaction, Prentice Hall, 2003.
[15] Z. Duric, W. Gray, R. Heishman, F. Li, A. Rosenfeld, M. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," Proc. of the IEEE, 90(7):1272-1289, 2002.
[16] A.T. Duchowski, "A breadth-first survey of eye tracking applications," Behavior Research Methods, Instruments, and Computing, 34(4):455-70, 2002.
[17] C. Elting, S. Rapp, G. Mohler, and M. Strube, "Architecture and implementation of multimodal plug and play," ICMI, 2003.
[18] C. Fagiani, M. Betke, and J. Gips, "Evaluation of tracking methods for human-computer interaction," IEEE Workshop on Applications in Computer Vision, 2002.
[19] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," Pattern Recognition, 36:259–275, 2003.
[20] G. Fitzmaurice, H. Ishii, and W. Buxton, "Bricks: Laying the foundations for graspable user interfaces," ACM CHI, 1(442-449), 1995.
[21] F. Flippo, A. Krebs, and I. Marsic, "A framework for rapid development of multimodal interfaces," ICMI, 2003.
[22] D.M. Gavrila, "The visual analysis of human movement: A survey," CVIU, 73(1):82-98, 1999.
[23] K. Grauman, M. Betke, J. Lombardi, J. Gips, and G. Bradski, "Communication via eye blinks and eyebrow raises: Video-based human-computer interfaces," Universal Access in the Information Society, 2(4):359-373, 2003.
[24] E. Hjelmas and B. K. Low, "Face detection: A survey," CVIU, 83:236–274, 2001.
[25] K. Hook, "Designing and evaluating intelligent user interfaces," Int. Conf. on Intelligent User Interfaces, 1999.
[26] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors", IEEE Trans. on Systems, Man, and Cybernetics, 34(3), 2004.
[27] R. Jacob, "The use of eye movements in human-computer interactions techniques: What you look at is what you get," ACM Trans. Information Systems, 9(3):152-169, 1991.
[28] A. Jaimes, "Human-centered multimedia: Culture, deployment, and access", IEEE Multimedia, 13(1):12–19, 2006.
[29] A. Jaimes, D. Gatica-Perez, N. Sebe, and T.S. Huang, "Human-centered Computing: Towards a Human Revolution," IEEE Computer, 40(5):30-34, 2007.
[30] A. Jaimes and N. Sebe, "Multimodal human computer interaction: A survey," CVIU,108(1-2):116-134, 2007.
[31] R. Jain, "Folk computing," Comm. of the ACM, 46(4):27-29, 2003.
[32] A. Jameson, R. Schafer, T. Weis, A, Berthold, and T. Weyrath, "Making systems sensitive to the user's time and working memory constraints," IUI, 1997.
[33] Q. Ji and X. Yang, "Real-time eye, gaze, and face pose tracking for monitoring driver vigilance," Real-Time Imaging, 8:357-377, 2002.
[34] B. Kisacanin, V. Pavlovic, and T.S.Huang, eds., Real-Time Vision for Human-Computer Interaction, Springe, 2005.
[35] S. Kumar and P. Cohen, "Towards a fault-tolerant multi-agent system architecture," Int. Conf. on Autonomous Agents, 2000.
[36] P. Langley, "User modeling in adaptive interfaces," Int. Conf. on User Modeling, 1998.
[37] A. Legin, A. Rudnitskaya, B. Seleznev, and Y. Vlasov, "Electronic tongue for quality assessment of ethanol, vodka and eau-de-vie," Analytica Chimica Acta, 534:129-135, 2005.
[38] S. Marcel, "Gestures for multimodal interfaces: A Review," Technical Report IDIAP-RR 02-34, 2002.
[39] D. Martin, A. Cheyer, and D. Moran, "The open agent architecture: A framework for building distributed software systems," Applied Artificial Intelligence, 13:91-128, 1999.
[40] M. Maybury, Intelligent Multimedia Interfaces, AAAI/MIT Press, 1993.
[41] S. Meyer and A. Rakotonirainy, "A Survey of research on context-aware homes," Australasian Information Security Workshop Conference on ACSW Frontiers, 2003.
[42] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," CVIU, 81(3):231-258, 2001.
[43] I.R. Murray and J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion," J. of the Acoustic Society of America, 93(2):1097–1108, 1993.
[44] D.A. Norman, The Design of Everyday Things, Doubleday, 1988.
[45] Z. Obrenovic and D. Starcevic, "Modeling multimodal human-computer interaction," IEEE Computer, pp. 65-72, September, 2004.
[46] P.Y. Oudeyer, "The production and recognition of emotions in speech: Features and algorithms," Int. J. of Human-Computer Studies, 59(1-2):157–183, 2003.
[47] A. Oulasvirta and A. Salovaara, "A cognitive meta-analysis of design approaches to interruptions in intelligent environments," ACM CHI, 2004.
[48] S.L. Oviatt, "Mutual disambiguation of recognition errors in a multimodal architecture," ACM CHI, 1999.
[49] S.L. Oviatt, "Ten myths of multimodal interaction," Comm. of the ACM, 42(11):74-81, 1999.
[50] S.L. Oviatt, "User-centered Modeling and Evaluation of Multimodal Interfaces," Proc. IEEE, 91(9):1457-1468, 2004.
[51] S.L. Oviatt, T. Darrell, and M. Flickner, eds. Comm. of the ACM, Special Issue on Multimodal interfaces that flex, adapt, and persist, 47(1), 2004.
[52] S.L. Oviatt and P. Cohen, "Multimodal interfaces that process what comes naturally," Comm. of the ACM, 43(3):45-48, 2000.
[53] S.L. Oviatt, "Multimodal interfaces," Human-Computer

Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, chap.14, 286-304, 2003.
[54] S.L. Oviatt, R. Lunsford, and R. Coulston, "Individual differences in multimodal integration patterns: What are they and why do they exist?" ACM CHI, 2005.
[55] H. Pan, Z.P. Liang, T.J. Anastasio, and T.S. Huang. "Exploiting the dependencies in information fusion," CVPR, vol. 2:407–412, 1999.
[56] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," IEEE Trans. on PAMI, 22(12):1424–1445, 2000.
[57] M. Pantic, N. Sebe, J. Cohn, and T.S. Huang, "Affective multimodal human-computer interaction," ACM Multimedia, 2005.
[58] T. Partala, V. Surakka, and T. Vanhala, "Real-time estimation of emotional experiences from facial expressions," Interacting with Computers, 18(2):208-226, 2006.
[59] V.I. Pavlovic, R. Sharma and T.S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review", IEEE Trans. on PAMI, 19(7):677-695, 1997.
[60] A. Pentland, "Looking at people," Comm. of the ACM, 43(3):35-44, 2000.
[61] M. Porta, "Vision-based user interfaces: methods and applications," Int. J. Human-Computer Studies, 57(1):27-73, 2002.
[62] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," Issues in Visual and Audio-Visual Speech Processing, MIT Press, 2004.
[63] Proceedings of the IEEE, Special Issue on Multimodal Human Computer Interface. August, 2003.
[64] L.M. Reeves, J-C. Martin, M. McTear, T. Raman, K. Stanney, H. Su, Q. Wang, J. Lai, J. Larson, S. Oviatt, T. Balaji, S. Buisine, P. Collings, P. Cohen, and B. Kraal, "Guidelines for multimodal user interface design," Comm. of the ACM, 47(1):57-69, 2004.
[65] E. Schapira and R. Sharma, "Experimental evaluation of vision and speech based multimodal interfaces," Workshop on Perceptive User Interfaces, pp. 1-9, 2001.
[66] T. Selker, "Visual attentive interfaces," BT Technology Journal, 22(4):146-150, 2004.
[67] B. Shneiderman, "Direct manipulation for comprehensible, predictable, and controllable user interface," IUI, 1997.
[68] B. Shneiderman, Leonardo's Laptop: Human Needs and the New Computing Technologies, MIT Press, 2002.
[69] O. Stock, and M. Zancanaro, (eds.). Multimodal Intelligent Information Presentation. Series Text, Speech and Language Technology. Kluwer, pp. 325-340, 2005.
[70] J. Trumbly, K. Arnett, and P. Johnson, "Productivity gains via an adaptive user interface," J. of Human-computer Studies, 40:63-81, 1994.
[71] M. Turk, "Gesture recognition," Handbook of Virtual Environment Technology, 2001.
[72] M. Turk and M. Kölsch, "Perceptual interfaces," G. Medioni and S.B. Kang, eds., Emerging Topics in Computer Vision, Prentice Hall, 2004.
[73] M. Turunen and J. Hakulinen, "Jaspis2 – An architecture for supporting distributed spoken dialogs," Eurospeech, 2003.
[74] M. Turunen, J. Hakulinen, K.-J. Räihä, E.-P. Salonen, A. Kainulainen, P. Prusi, "An architecture and applications for speech-based accessibility systems," IBM Systems, 44(3):485-504, 2005.
[75] R. Vertegaal, ed., "Attentive user interfaces: Special issue," Comm. of the ACM, 46(3), 2003.
[76] J.J.L. Wang and S. Singh, "Video analysis of human dynamics – A survey," Real-Time Imaging, 9(5):321-346, 2003.
[77] M. Weiser, "Some computer science issues in ubiquitous computing," Comm. of the ACM, 36(7):74-83, 1993.
[78] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," IEEE Trans. on PAMI, 24(1):34–58, 2002.
[79] H. Yoshikawa, "Modeling humans in human-computer interaction," in Human-computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp. 118-146, 2002.
[80] W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips, "Face recognition: A literature survey," ACM Computing Surveys, 12:399–458, 2003.