Exploiting Vision and Speech in Advanced E-learning Systems

Marco Porta
Dip. di Informatica e Sistemistica
Università di Pavia
Via Ferrata, 1 – 27100 – Pavia – Italy
marco.porta@unipv.it

ABSTRACT
Technological advances offer new paradigms for training, allowing novel forms of teaching and learning to be devised; unfortunately, however, many e-learning systems are still based on complex procedures and unusable interfaces, which are regarded as intimidating, confusing or simply frustrating by the user. In this paper we consider the possibility of exploiting vision and speech as intuitive communication channels, to enhance the quality of the interaction between users and e-learning platforms. Through an analysis of current research in the field of vision-based and speech-based perceptive interfaces, we will discuss some application scenarios for e-learning, stressing the important role that such natural communication forms could play in improving the interaction process.

Categories and Subject Descriptors
H.1.2 [**Models and Principles**]: User/Machine Systems – *human factors*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *input devices and strategies, interaction styles*; K.3.1 [**Computers and Education**]: Computer Uses in Education – *distance learning*.

General Terms
Design, Human Factors.

Keywords
E-learning, multimodal interaction, natural interaction, perceptive interfaces, vision-based interfaces, speech-based interfaces, auditory interfaces, conversational animated agents.

1. INTRODUCTION
Thanks to the WIMP interaction paradigm, introduced in the eighties and based on Windows, Icons, Menus and Pointing devices, the personal computer (PC) is now a household instrument, accessible to everybody and usable for many kinds of purposes. Undoubtedly, the greater intuitiveness of the graphic interaction has simplified the approach to the computer for novices, and has also increased the productivity of those who use it professionally. Nevertheless, according to some experts the WIMP paradigm will not be able to scale properly to match all the uses of computers in the future [14].

If in the last decade one of the most recurrent keywords in the computer field has been *multimedia*, another term is now contending with it for the first place: *multimodal*. Multimodal systems usually combine natural input modes—such as speech, pen, touch, hand gestures, eye gaze, and head/body movements—with multimedia output. Sophisticated multimodal interfaces can integrate complementary modalities to get the most out of the strengths of each mode and overcome weaknesses [9].

As we move towards a world where information technology will affect almost any aspect of our life, the need arises for more intuitive ways of interacting with the computer and other electronic devices. From a technical point of view, it is already possible to implement interactions that exploit the perceptive abilities which so far have characterized human-human communication only. *Perceptive* user interfaces (also called *perceptual* user interfaces when integrated with multimedia output and other possible forms of multimodal input) try to provide the computer with perceptive capabilities, so that implicit and explicit information about the user and his or her environment can be acquired; the machine thus becomes able to "see", "hear", etc. Interface research is now moving towards several directions, and new and more natural input modalities will probably find application in graphical user interfaces (GUIs), joining and partly replacing traditional interaction paradigms based on keyboard and mouse.

In the e-learning context, the quality of the interaction is of paramount importance, as it directly influences the learning process by imposing specific communication modalities. Often, however, technologies employed are perceived by the user as unfriendly, "mysterious" and distant, lacking great part of the informal social interaction and face-to-face contact of traditional classroom training (see for example [4]); undoubtedly, this is one of the main reasons that are put forward by detractors of e-learning to support their ideas. Inexperienced computer users suffer more from such drawbacks, but even those who daily work in the informatics field may undergo their negative effects.

In this paper we consider the application of perceptive interfaces to e-learning systems, as a way to improve the quality of the interaction through more natural forms of communication. Of course, it is (and will also be in the future) practically impossible to achieve the levels of real human-human communication, but, thanks to constant developments in the fields of computer vision and speech recognition, we will get closer and closer to such goal. At the University of Pavia, in particular, we are currently focusing on vision-based interfaces, to be used for the recognition of hand/arm/head gestures and basic face expressions (see for instance [2]). Our aim is to apply these interface modules to e-

learning systems, for enhancing the interaction between the learner and the virtual teacher (possibly, also exploiting speech-based and auditory interfaces).

Since we concentrate on the communication channels per se, for the present we do not consider specific teaching subjects. However, our natural test-beds will be basic courses in the computer science area, and we strongly think that they will be very suitable to the kinds of interaction we describe in this paper. Technical topics, in fact, lend themselves to more structured organizations, with step by step explanations, examples and simulations which can greatly benefit from high levels of interactivity.

2. NATURAL INTERACTION PARADIGMS FOR E-LEARNING

In the next subsections, we will examine the two perceptive capabilities which most characterize human-human interaction, namely vision and hearing (also combined each other). After short descriptions of each technology, we will try to envision application scenarios in the e-learning context.

Besides the input side, we will briefly consider visual and audio outputs too, as, although not new, they are integral parts of the whole interaction experience.

2.1 Vision-based and Visual Interfaces

Vision-based interfaces (VBIs) are perceptive interfaces which exploit vision as a communication channel from the user to the computer [11]. Cameras are almost always non-invasive input tools and their costs are lowering at an increasing pace (often, even standard commercially-available webcams are sufficient). Typical uses of VBIs in PC applications include the following:

- *Head tracking*: the position of the head is exploited to provide some kinds of input to the computer (see Figure 1).
- *Face/facial expression recognition*: automatically identifying who is in front of the computer screen or distinguishing among different face expressions can help make the interface more "human-like" (research in the *affective computing* area has shown that emotions may greatly influence the user's behaviour in the interaction with the computer [10]).
- *Eye tracking*: detecting the user's gaze direction can be extremely useful in presence of severe disabilities which prevent normal use of the mouse or as a support to head tracking and face/expression recognition.
- *Gesture recognition*: hand/arm postures and movements can be effective ways to provide the computer with input, particularly when messages to be conveyed are inherently manipulative.

Essentially, VBIs fall into one of two possible (not mutually exclusive) categories, namely [13]:

- *Control*: when the VBI is supposed to understand user acts as explicit commands (for example, hand gestures may be used to control ordinary Windows applications).
- *Awareness*: when the VBI can draw indirect information from user actions and behaviors, which are taken as implicit input (for example, a system may understand that the user is looking elsewhere when an error situation occurs and may attract his or her attention by emitting a sound).

From the output side, the communication channel corresponding

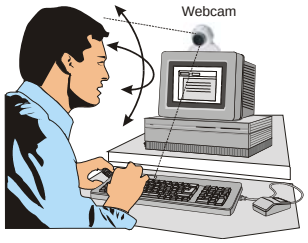

Figure 1. Head tracking as a way of providing input to the computer.

to vision perception is of course represented by visual interfaces. It is indisputable that, for almost all applications, interacting with objects placed in a two-dimensional space is extremely worthwhile. Graphic elements, in fact, have the advantage of being characterized by shape, dimension, position and possibly color, all attributes which can help better understand the meaning of what is displayed on the screen. Practically, nearly all existing GUIs are based on the WIMP paradigm and share common features which simplify their use.

Obviously, the quality of the interaction with a visual system strongly depends on its graphical interface, which should be designed according to proper usability criteria. Apart from mentioning animated agents, however, in this work we will only discuss about vision input applied to e-learning systems. Due to the vastness of the subject and its being relatively not new, in fact, we will take visual output for granted, while never forgetting its enormous importance.

2.1.1 Applications of VBIs to E-learning Systems

Vision can be conveniently exploited to improve "normal" interaction with information appliances, and, in particular, with e-learning platforms. However, VBIs are extremely useful for enhancing the *accessibility* of such systems: the use of the computer should be guaranteed to everybody, and this is all the more reason true for persons with disabilities, for whom e-learning can represent a cheaper and more practical alternative to traditional in-class teaching.

For the "control" category, the following applications of VBIs to e-learning can be identified:

1. Use of accurate head tracking techniques for mouse cursor control, object selection, choice of menu items and page scrolling (in case of users unable to utilize the hands).
2. Still for free-hand interaction, recognition of head nods and shakes, to answer yes/no dialog boxes (for example, in test sessions).
3. Use of accurate eye tracking methods for mouse cursor control, object selection and page scrolling (in case of users with severe disabilities).
4. Recognition of eye blinks to trigger specific events (e.g. to open menus or to answer yes/no questions).
5. Combination of ordinary mouse-based interaction and head/eye tracking techniques to enhance the pointing process: the mouse cursor may be automatically moved to the screen area that is currently looked at by the user (thus reducing hand shifts), and then, within that area, the mouse may be handled as usual.

Figure 2. Using the hand for page scrolling.

6. Use of intuitive hand gestures (both static and dynamic) to control basic interface functions. For instance, the hand could be simply waved upward/downward to scroll the page (Figure 2), or common commands (e.g. *Save* for documents, *Back* for browsing operations, etc.) could be triggered by specific hand postures or dynamic gestures. Also, we should not ignore the importance of sign languages for deaf-mute persons. Even if input can be provided through keyboard and mouse, a system capable of understanding basic expressions of a gestural language would be the equivalent of speech recognition for people who can speak: a more natural way to communicate with the machine (i.e. the virtual teacher).

For the "awareness" category, the following applications of VBIs to e-learning systems can be identified:

1. Use of head/eye tracking techniques for correct interpretation of user's gaze. For instance, if the user is not looking at the screen when an important message is to be communicated to him or her, a sound alarm might be played. Recognizing gaze direction is also useful to better interact with conversational animated agents, as they can "look" at the user's eyes (see Section 2.3).
2. Expression recognition, or detection of other user-related "features", to interpret the user's "emotional status". A system able to understand whether the learner has difficulty in understanding something would be extremely useful for e-learning, as the platform could automatically adapt its "behavior" according to such implicit information. A first attempt at implementing a tutorial system of such a kind is represented by the work of Zhang, Silber and Kambhamettu [15], in which a vision-based interface is used to recognize few basic frontal-view facial expressions (then used to infer the degree of user understanding). Although finding the right connections between a perceived user signal and its mental implications is a very delicate and difficult task, a good solution to reduce the effect of wrong interpretations may be to gradually modify the behavior of the interface according to the recognized signal, without abrupt changes. For instance, if there are elements suggesting that the user is not understanding something (e.g. he or she is knitting the eyebrows), the help button of the interface might be enlarged, instead of directly displaying a help window. Besides facial expressions, additional cues for obtaining implicit information about the emotional status of the user can be found in other unconscious signs. For example, the variation of the frequency of eye-blinks during time may indicate that the learner is getting nervous; also, if he or she looks at the screen discontinuously, maybe the lesson is perceived as boring and an alternative learning path (if available) might be presented.
3. Recognition and classification of user's "activities". Even if not strictly connected to e-learning, recognizing what the user is doing could help the system to automatically adapt to the particular situation. For instance, if the user is phoning, no sound should be emitted, even if, at the same time, he or she is watching a voice-commented animated tutorial (which thus will automatically stop or show subtitles only).

2.2 Speech-based and Auditory Interfaces

Speaking is a communication skill we learn at an early age and speech recognition technology is able to free users from the constraints of the ubiquitous WIMP paradigm, towards higher levels of naturalness in human-computer interaction. Speech recognition has already been in use in several systems for a few years, and very cheap software tools are now available which give excellent results (used for example as an alternative to the keyboard to provide input to text editors).

Speech is useful as an output means as well (often referred to as *synthesized speech* or *text-to-speech* technology), as it is a practical secondary output channel, which frees our attention in "eyes-busy" tasks [6].

From voice it is also possible to draw information about the emotional status of the user. Falling, like expression recognition, in the affective computing class, interfaces of this kind are useful to "humanize" the computer. Scientific findings suggest in fact an increasingly large number of important functions of emotions, which contribute not only to irrational behavior, but also play an important role in rational decision making [10].

In addition, speech output can be used to provide animated agents (Section 2.3) with speaking capabilities, thus enhancing (often through speech input as well) their likeness with human behavior. Such kinds of applications are part of the class of *conversational agents* or *interfaces*, which try to emulate human-human communication.

2.2.1 Applications of Speech-based and Auditory Interfaces to E-learning Systems

Although, by now, speech recognition is state-of-the-art technology, a sort of resistance seems to be offered by the market and by general users towards such new input modality. The reason for this may be found in the way computers have been traditionally employed: providing input through keyboard and mouse is a relatively easy and quick task, and does not interfere with the work of possible nearby users.

However, we should always consider that the naturalness of communication is an extremely important factor in the e-learning sphere. Therefore, here are some e-learning applications that can be identified for speech-based interfaces:

1. Alternative input modality for persons who cannot easily employ the hands. Like for vision-based interfaces, speech-based systems can be useful for disabled persons, for example with limited hand/arm motion capacity. Almost any task can be accomplished by proper vocal commands, if necessary structured according to a suitable language (but advances in natural lan-

guage understanding research might simplify even more the communication process). Only applications requiring very precise mouse cursor control (e.g. drawing tools) cannot extensively exploit voice instructions. Speech input is of course also useful for blind and low-vision people.

2. Alternative input modality for everybody, to make the communication more natural: if I can "tell" directly something to the computer (e.g., answer questions in test sessions, navigate within course structures, etc.) instead of writing messages through the keyboard, the inevitable physical barrier between me and the machine is reduced. Dictating text to the computer implicitly transforms it into a sort of "personal secretary", and, especially from a psychological point of view, this can significantly improve the quality of the interaction. Some people, for example, become much more "creative" when they can directly express their "flow of thoughts", without keyboard intermediation.
3. "Dialog" with conversational agents (e.g. teaching avatars). Direct speech communication may contribute towards the perception of conversational agents as sorts of "living" entities.
4. Implicit input for simple attentive interfaces. Voice could be exploited to obtain information about the emotional status of the user, towards those applications envisaged by affective computing. In addition to the basic ability to recognize single words as they are uttered by the user, advances in natural language understanding will probably allow sophisticated interfaces to interpret relatively complex non-structured commands; for e-learning, where natural interaction is a central requirement, this would be an important breakthrough.

Besides for computer input, speech can be used to convey information to the user as well. Of course, we do not have to wait the future for making interfaces play pre-recorded or synthesized messages when necessary, but it is a fact that current systems do not exploit extensively such form of communication. Although potentially disturbing for the user when not required, voice output can be a useful (optional) communication channel in some e-learning applications, among which:

1. Synthesized text reading for blind and low-vision people. Systems of this kind are already in use today, but they will be much more widespread when the degree of penetration of information technology within all population bands will increase. Any e-learning system should offer such an alternative output modality (possibly combined with speech-input, for complete support of keyboard- and mouse-free interaction).
2. Communication of important messages. If user attention must be urgently drawn (for instance, because an anomalous event has just occurred), a spoken message may be much more incisive than a simple message box.
3. Output from conversational agents (maybe animated). Like speech input, speech output is important to liken the agent to a "social" entity with which to naturally interact.

2.3 Combination of Vision-based and Speech-based Interaction

To implement advanced forms of interaction characterized by high levels of human-like communication, vision and speech features can be combined to exploit the best from each. Input from vision- and audio-based perceptive channels is the groundwork for the implementation of the so-called *attentive user interfaces* (AUIs). Attentive interfaces fall in the "awareness" category, as they are able to catch indirect signals in the user's behavior. AUIs are part of the wider field of *perceptive presence* applications, which, for example, can share information about the number of people in an area, as well as their location, kind of activity, or where they are looking at [1]. User interfaces can be made attentive by *attentive agents*, which are systems that attend to what users do and try to anticipate what they need [8]. Actually, the distinction between attentive interfaces and attentive agents is very subtle, or even does not exist. It is usually more marked when the attentive role is clearly assigned to a certain software module or is perceived by the user as an additional, different entity with respect to the main interface (e.g. a graphically animated character which "looks" at the user).

In Section 2.2 we have already discussed about conversational agents. Real-time representation and animation of virtual living figures (e.g. humans) has been a challenging area in Computer Graphics since early eighties: however, most of these animated agents are characterized by predefined behaviors, which do not directly involve the user in the virtual environment (a very simple example is given by the "assistants" of Microsoft Office applications). What about "sensing" the user and his or her surroundings?

To imitate everyday human communication, advanced computer interfaces may combine the benefits of high visual fidelity with conversational intelligence and the ability to correctly understand users' emotions [12]. On the part of computer graphics, achieving such goals requires synthesizing digital humans characterized by photorealistic faces, capable of expressing emotions in the best possible way. On the part of multimodal user input, vision can be exploited to observe what happens in the environment and react accordingly.

Thanks to these conversational "attentive" animated agents, computer interaction may resemble natural face-to-face conversation with human-like characters [3]. At present, however, these animated entities are mostly advanced computer graphics programs combined with basic conversational agents functionalities (and, of course, one thing is simulating a conversation, another thing is carrying out an even very simple real conversation). Attentive animated agents are also an attempt at associating life-like creatures to interface agents or attentive interfaces, to engage the user in captivating interaction experiences.

2.3.1 Applications to E-learning Systems of Interfaces Based on both Vision and Speech

As stated several times in the previous sections, for users to be "at their ease" while interacting with e-learning systems it is very important that interactions be as natural as possible: proper combination of two "highly human" communication modalities such as vision and speech may greatly contribute to this purpose.

Here are possible applications to e-learning of interfaces able to receive inputs from both the vision and speech channels at the same time:

1. Joint input modalities for GUIs' elements control. In the interaction with an e-learning platform, a gesture might be used to perform a certain action (e.g. to scroll the page by waving the hand), whereas a vocal order might be exploited to trigger

a specific command (e.g. to go back to the previous learning unit by saying "Back"). The two kinds of communication could even be used at the same time (for example, if while waving upward/downward the hand the user says "Fast", then the scrolling step might be augmented). In general, redundancy allows the same commands to be (maybe optionally) specified in several ways: by traditional use of keyboard and mouse, by gestures, by speech, etc.

2. The combination of vision and speech within an e-learning environment can be exploited for the implementation of attentive interfaces as well. Actually, each input modality can be taken as an additional cue to confirm or better define the information acquired through the other input channel. For instance, if the user's gaze is recognized as being directed towards an animated conversational agent on the screen and the microphone is detecting speech sound, it is very likely that the user is talking to the virtual entity, and not to possible neighbors. In extremely advanced interaction systems, certain hand gestures while speaking could be even exploited to disambiguate the meaning of what the user is saying, thus helping in the difficult task of natural language understanding.
3. As regards conversational animated agents, e-learning systems, where the computer may be seen as a sort of substitute for a human being, could benefit particularly from this engaging technology. From a psychological (and possibly subliminal) point of view, the user/student may perceive an animated agent as the missing teacher in the flesh with whom to interact, thus potentially improving the learning process. More generally, an animated agent could be useful as a "mentor" in advanced tutorial systems, as well as in help modules within ordinary applications.

3. A SAMPLE APPLICATION SCENARIO

After the review presented in the previous sections, we now describe a hypothetical application scenario for vision and speech technologies. Of course, several other situations could be considered in the e-learning context, and our description is only one of the many that could be identified; however, it should be sufficient to provide an idea of what kinds of interactions we may expect from future "perceptive" e-learning systems. Even though our description refers to a "usual" interaction session, we stress again the importance that perceptive interfaces can have for disabled people, freeing them (totally or partially) from the need for assistance to accomplish many tasks.

Consider a student, who we will call John, who is attending an e-learning course of "Basic computer science". An interaction sequence might go as follows:

a. John logons onto the e-learning platform (in the usual way, by entering login and password, or through biometric systems exploiting data about face, iris, fingerprints, signature, voiceprint, etc.; see for example [5]).
b. John is presented with the course structure (learning units, lessons, topics, subtopics, etc.), possibly displayed according to effective graphical representations highlighting already-visited, partially-visited and yet-to-be-visited sections; *personalized adaptive interfaces* (see for instance [7]) could be used to tailor the course organization to the specific user. John chooses "Unit 2, Lesson 3, The Von Neumann machine", through ordinary input tools (mouse and keyboard) or by simply saying the name of the topic.
c. John starts reading the content of the first elementary information unit used by the platform to present topics (we will simply call it "page"). While he is on the third page, the vision perceptive subsystem infers, from some facial expressions and/or other behavioural signs, that John may be having difficulty in understanding the subject. Therefore, in the right lower corner of the window the link "Detailed description" appears (or also an animated agent with that link within a balloon). The text of the link, initially very small, progressively increases its size as time passes and John stays on the same page.
d. John clicks the link (or says "detailed description"), and an in-depth description of the Fetch, Decode and Execute stages is presented, which also includes an animated simulation. In the meanwhile, the other topics of the lesson might be reorganized by the adaptive module, so as to already include step-by-step explanations (probably, this lesson is rather difficult for John).
e. Since at the end of the simulation there are still signals of potential non-understanding from John, a dialog box (or an animated agent) appears which explicitly asks him whether he wants to watch the simulation once more. John nods (or says "yes", or simply presses the Ok button), and the simulation starts again. If no particular face expressions or other body signs were detected in John, the possibility to play the simulation another time would have been simply given by a link somewhere in the page.
f. While the animated simulation, which consists also of a voice description, is playing, John's mobile phone rings, and he answers. The vision subsystem recognizes the state "phoning", and decreases the audio volume; since after thirty seconds the phonecall is still afoot, the simulation is stopped. When the phonecall ends, the simulation automatically restarts from some seconds before the time when the "phoning" state was detected.
g. John continues watching the simulation, but, after a while, maybe thinking of the just received phonecall or because distracted by something, he seems not to be looking at the screen any more. After five seconds from the detection of such "inattentive" state, the REW button of the interface used to control the simulation is enlarged. After thirty seconds, the simulation is stopped, and a dialog box appears which asks whether to restart it from where the "looking away" state was detected. John nods, and the simulation is played from that point.
h. Since a part of the simulation he is viewing for the second time is already clear, John skips it by waving his left hand rightward: the play speed is augmented. When the portion to be passed over is finished, John says "normal speed", to play the simulation normally.
i. After the last page about the Von Neumann machine, John starts a test session, in which he is presented with a list of questions with multiple answers. A conversational animated agent reads the text of each query, and John chooses the right response by saying "a", "b", etc. If the answer is wrong, the agent explains why it is so, and suggests which subjects John should deepen to fill his gaps.

j. At a certain moment, John stands up and goes away; since he is not perceived in front of the screen for more than thirty seconds, he is automatically logged out. When John comes back, he has to logon again (step 'a'), and the system resumes from the exact point where it was left.

4. CONCLUSIONS

Typical human sensing capabilities, such as sight and hearing, can now be incorporated into traditional GUIs and enhance user communication with the computer. From the mere input/output side, the "dream" of natural communication with the machine is becoming a reality: speaking to a PC, for example, can now be considered as natural as typing input commands through a keyboard. Analogously, having the computer robustly recognize simple gestures is a feasible task with current technology, even with very cheap devices. If in the past such forms of interaction were a prerogative of science fiction, nowadays they are at hand, although not widespread in normal interfaces.

In this paper we have discussed about the possible use of vision and speech in advanced e-learning systems, to achieve better levels of human-computer interaction. Even if, except for very rare cases, such systems have so far not been considered in the virtual teaching realm, now the time is probably ripe for them to be taken into serious account. Beyond their effectiveness in improving the interaction with the computer, we think that their forte lies in the ability to be felt as more natural communication ways by the user.

Of course, our analysis has not been aimed at providing an exhaustive description of the field; instead, we have wanted to give an overview of perceptive technologies for e-learning, with special emphasis on potential application scenarios and interaction modalities.

5. ACKNOWLEDGMENTS

Our work is partially supported by funds from the Italian FIRB project "Software and Communication Platforms for High-Performance Collaborative Grid" (grant RBIN043TKY).

6. REFERENCES

[1] Bentley, F., Tollmar, K., Demirdjian, D., Koile, K., and Darrel, T. Perceptive Presence. *IEEE Computer Graphics and Applications*, September/October 2003, pp. 26-36.

[2] Borghi, F., Lombardi, L., and Porta, M. Basic Hand Gesture Recognition for Human-Computer Communication. *Proceedings of the 11th International Conference on Human-Computer Interaction*, Las Vegas, Nevada, USA, 22-27 July 2005, LEA Publ.

[3] Cole, R., Van Vuurel, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., and Yan, J. Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human–Computer Interaction. *Proceedings of the IEEE*, Vol. 91, No. 9, September 2003, pp. 1391-1405.

[4] Hedge, N., and MacDonald, D. Affective Considerations in Distance e-Learning. *Proceedings of the IASTED Conference on Education and Technology (ICET 2005)*, Calgary, Alberta, Canada, July 4-6, 2005.

[5] Jain, A. K., and Ross, A. Multibiometric Systems. *Communications of the ACM*, Vol. 47, No. 1, Jan. 2004, pp. 34-40.

[6] Lai, J. Conversational Interfaces. *Communications of the ACM*, Vol. 43, No. 9, 2000, pp. 24-27.

[7] Langley, P. User Modeling in Adaptive Interfaces. *Proceedings of the 7th International Conference on User Modeling*, June 20-24, 1999, Banff, Canada.

[8] Maglio, P. P., and Campbell, C. S. Attentive Agents. *Communications of the ACM*, Vol. 46, No. 3, 2003, pp. 47-51.

[9] Oviatt, S. Ten Myths of Multimodal Interaction. *Communications of the ACM*, Vol. 42, No. 11, 1999, pp. 74-81.

[10] Picard, R. W., and Klein, J. Computers that Recognize and Respond to User Emotions: Theoretical and Practical Implications. *MIT Media Lab Tech Report* 538, 2002.

[11] Porta, M. Vision-based user interfaces: methods and applications. *International Journal of Human-Computer Studies*, No. 57, 2002, pp. 27-73.

[12] Takács, B., and Kiss, B. The Virtual Human Interface: A Photorealistic Digital Human. *IEEE Computer Graphics and Applications*, September/October 2003, pp. 38-45.

[13] Turk, M. Moving from GUIs to PUIs. *Technical Report* MSR-TR-98-69 (Microsoft Research), 1998.

[14] Turk, M., and Robertson, G. Perceptual User Interfaces. *Communications of the ACM*, Vol. 43, No. 3, 2000, pp. 33-34.

[15] Zhang , Y., Silber, G., Kambhamettu, C. Facial expression driven tutorial system. *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics*, Orlando, Florida, July, 2002, Vol.IX, pp.287 –292.