# Gaze-X: Adaptive Affective Multimodal Interface for Single-User Office Scenarios

Ludo Maat[1] and Maja Pantic[2,3]
[1] EEMCS, Delft University of Technology, The Netherlands
[2] Computing Department, Imperial College London, UK
[3] Faculty of EEMCS, University of Twente, The Netherlands
ludomaat@zonnet.nl, m.pantic@imperial.ac.uk

## ABSTRACT
This paper describes an intelligent system that we developed to support affective multimodal human-computer interaction (AMM-HCI) where the user's actions and emotions are modeled and then used to adapt the HCI and support the user in his or her activity. The proposed system, which we named Gaze-X, is based on sensing and interpretation of the human part of the computer's context, known as W5+ (who, where, what, when, why, how). It integrates a number of natural human communicative modalities including speech, eye gaze direction, face and facial expression, and a number of standard HCI modalities like keystrokes, mouse movements, and active software identification, which, in turn, are fed into processes that provide decision making and adapt the HCI to support the user in his or her activity according to his or her preferences. To attain a system that can be educated, that can improve its knowledge and decision making through experience, we use case-based reasoning as the inference engine of Gaze-X. The utilized case base is a dynamic, incrementally self-organizing event-content-addressable memory that allows fact retrieval and evaluation of encountered events based upon the user preferences and the generalizations formed from prior input. To support concepts of concurrency, modularity/scalability, persistency, and mobility, Gaze-X has been built as an agent-based system where different agents are responsible for different parts of the processing. A usability study conducted in an office scenario with a number of users indicates that Gaze-X is perceived as effective, easy to use, useful, and affectively qualitative.

## Categories and Subject Descriptors
H1.2 [**User/Machine Systems**]: Human information processing
H.5.1 [**Multimedia Information Systems**]: Audiovisual input
I.5.4 [**Pattern Recognition Applications**]: Models, Learning

## General Terms
Performance, Design, Experimentation, Human Factors

## Keywords
Affective Computing, Facial Expressions, Multimodal Interfaces


*ICMI'06,* November 2–4, 2006, Banff, Alberta, Canada.


## 1. INTRODUCTION
We have entered an era of pervasive computing. Computers and the Internet have become so embedded in the daily fabric of our lives that we can no longer live without them [13]. We use them to work, study, communicate, shop, and entertain ourselves. With the ever-increasing diffusion of computers into society, human-computer interaction (HCI) is becoming increasingly essential to our daily lives.

Predicting the future of HCI is a difficult task, but one important source of help is the accumulated information about the preferences and limitations of humans interacting with computers. Principles can be drawn upon, which may explain why some interfaces survive and others become extinct. Rigid designs that assume that users will be explicit and fully attentive while interacting with the computer, that do not protect against errors, provide help at all times except at the right moment, and all in all make users frustrated, are likely to become quickly extinct due to their poor usability [23]. On the other hand, designs that include adequate attention to individual differences among users, support (natural) multimodal and context-sensitive interaction, expend on designs for reliability and safety, provide access to the elderly and handicapped, and properly adapt to the user level of knowledge, skills, attention, preferences, moods, and intentions, are the kind of HCI designs that are likely to become the trend in computing technology [4], [38], [32], [31], [40]. Although this list may not be complete, it points out important issues that are rather insufficiently addressed by the current initiatives [27].

Several extensive survey and position papers have been published on vision-based [33], [41], multimodal [25], [35], [16], affective [21], [14], and context-sensitive interfaces [24], [27]. Virtually all of these articles agree that approaching the naturalness of human-human interaction plays a central role in the future of HCI and that this objective can be approached by designing adaptive HCI systems that are affective, context-aware, and multimodal. However, many of these articles mention that the main application of this new technology is quickly changing from user interface models in which one user is sitting in front of the computer, to something else like ambient interface models in which multimodal multi-party interactions are observed and interpreted. We argue here that this statement is very misleading. When the main application domain in a certain field has changed from *A* to *B*, this implies usually that problems in *A* have been researched, that they have been solved, and that the research has moved on to tackle other problems. In turn, the statement in question implies that the realization of adaptive interfaces based

on affective multimodal interaction models (AMM-HCI) can be considered a solved problem for single-user office scenarios. However, an extensive research of the large body of the related literature did not confirm this. Only few works aimed at adaptive, affective, multimodal interfaces for single-user office scenarios have been reported up to date. The majority of past work in the field relate to multimodal, non-affective interaction with the user [25]. Integration of multiple natural-interaction modalities such as speech, lip reading, hand gestures, eye-tracking, and writing into human-computer interfaces has long been viewed as a means for increasing the naturalness and, in turn, ease of use [3], [42], [43]. The research in this part of the field is still very active and the majority of the current work is aimed either at support of crisis management [22], [39] or at development of personal widgets [25]. A rather large body of research can also be found in the field of human affect sensing [28], [30]. Most of these works are single modal, based either on facial or on vocal affect analysis. Recently, few works have been also proposed that combine two modalities into a single system for human affect analysis. The majority of these efforts aim mainly at combining facial and vocal expressions for affect recognition [27], although some tentative attempts on combining facial expressions and body postures have been reported as well [12], [17]. In the same way that these methods do not tackle the problem of how the sensed user's affect can be incorporated into the HCI, there is a large body of research that focuses on the role of human affect in HCI while assuming that user's affective states have been already sensed by the computer [5], [36]. Efforts that integrate these two detached research streams, and represent the related work to the one presented in this paper, are rare. Several works have been reported on developing adaptive interfaces that are based on sensing the user's affective states. These methods are usually single modal, based either on facial affect recognition [2] or on physiological affective reaction recognition [15], [34]. To our best knowledge, the only exceptions from this rule are the single-user AMM-HCI systems proposed by Lisetti & Nasoz [20] and by Duric et al. [9]. The former combines facial expression and physiological signals to recognize the user's emotion (fear, anger, sadness, frustration, and neutral) and then to adapt an animated interface agent to mirror the user's emotion. The latter is a much more elaborate system, aimed at real-world office scenarios. It combines lower arm movements, gaze direction, eyes and mouth shapes, as well as the kinematics of mouse movements to encode the user's affective and cognitive states (confusion, fatigue, stress, lapses of attention, and misunderstanding of procedures). It then applies a model of embodied cognition, which can be seen as a detailed mapping between the user's affective states and the types of interface adaptations that the system supports, to adapt the interface in a reactive or a proactive manner to the user's affective feedback. The main drawback of this system is that it is not user-profiled, while different users may (and probably will) have different preferences on how the interface should adapt. Another drawback is the employed model of embodied cognition, which is rigid and clumsy as it stores all possible combinations of inputs and outputs that are difficult to unlearn and reformat. Finally, the system has not been tested and it is not readily deployable.

We believe that the main reason for the lack of research on single-user AMM-HCI is a twofold. First, the misconception that the problem in question has been solved leads to the lack of interest by researchers and research sponsors. Second, it seems that the vast majority of researchers treat the problem of adaptive, affective, multimodal interfaces as a set of detached problems in human affect sensing, speech processing, and computer-human interface design. In this paper, we treat this problem as one complex problem rather than a set of detached problems and we propose a single-user AMM-HCI system similar to that proposed by Duric et al. [9]. A difference between the two systems is that ours uses a dynamic case base, an incrementally self-organizing event-content-addressable memory that reflects user preferences and allows easy reformatting each time the user wishes so. In addition, our system, which we call Gaze-X, is based on sensing and interpretation of the human part of the computer's context, known as W5+ (who, where, what, when, why, how) and, in turn, is user- and context-profiled. It is also readily deployable.

For computing technology applications, context can be defined as any information that can be used to explain the situation that is relevant to the interaction between users and the application [8]. For a single-user scenario, the six questions that summarize the key aspects of the computer's context with respect to its human user are as follows:

- *Who?* (Who the user is?)
- *Where?* (Where the user is? For single-user desktop-computer scenarios this context question is superfluous as it is known where the user is – either in front of the computer or not.)
- *What?* (What is the current task of the user?)
- *How?* (How the information is passed on? Which interactive signals / actions have been used?)
- *When?* (What is the timing of displayed interactive signals with respect to changes in the computer environment?)
- *Why?* (What may be the user's reasons to display the observed cues? Except of the user's current task, the issues to be considered include whether the user is alone and what is his or her affective state.)

The state of the art in context-aware applications ensues from two streams of research: the one on context sensing [24], [27], which focuses on sensor-signal processing (audio, visual, tactile), and the other on context modelling [8], which focuses on specifying procedures and requirements for all pieces of context information that will be followed by a context-aware application. Gaze-X integrates those two detached poles of the research. It uses a face recognition system to answer *who* the user is and to retract his or her profile (i.e., user-profiled case base). It employs an eye-tracking system and a speech recognizer in combination with event handling of standard HCI events like mouse movements, keystrokes, and active software identification to answer *what* is the current task of the user. In addition to these input modalities, a system that recognizes prototypic facial expressions of six basic emotions (anger, fear, happiness, surprise, sadness, and disgust) is used to answer the *how* context question. To answer the *when* context question, we simply keep a log of the time and the cost (in time) of HCI events associated with various input modalities. To answer the *why* context question, which is the most complex context question, we use case-based reasoning that enables evaluation of encountered events based upon the user preferences and the generalizations formed from prior input. Based upon the conducted evaluation, Gaze-X executes the most appropriate user-supportive action.

The reminder of the paper is organized as follows. Section 2 gives an overview of the Gaze-X architecture. Section 3 presents the

system's input modalities. The utilized case-based reasoning is explained in detail in section 4. Adaptive and user-supportive actions of the interface are discussed in section 5. The Graphical User Interface (GUI) of the Gaze-X is presented in section 6. The usability study that we carried out is discussed in section 7. Section 8 concludes the paper.

## 2. THE GAZE-X ARCHITECTURE

The outline of the system is illustrated in Fig. 1. The main modules are the multimodal input module, the reasoning module, and the feedback module. The user of Gaze-X experiences an adaptive interface, which changes as a function of the currently sensed context. This function is represented by the cases of the utilized dynamic case base, having the system's and the user's state as the input (represented in terms of exhibited multimodal interactive actions and cues) and the adaptive and user-supportive changes in the interaction as the output. The fact that the changes in the interaction do not have to occur in each time instance (e.g., if the user's preference is to remain undisturbed while working with a certain application), triggering of the feedback module is optional and illustrated using a dashed line.

Gaze-X has been implemented as an agent-based system. The main reason for doing so is to support concepts of *concurrency* (which allows sub-systems to operate independently and yet at the same time), *modularity/scalability* (which allows easy upgrade through inclusion of additional sub-systems), *persistency* (which ensures robust performance by saving intermediate settings), and *mobility* (which enables transport of agents to another host). We used Fleeble Agent Framework [26] to develop the Gaze-X. Fleeble can be seen as a common programming interface defining the behavior of all the agents build with the framework. The main characteristics of Fleeble and Fleeble-based multi-agent systems (MAS) can be summarized as follows (for details see [26]).

- *Fleeble enables easy development of agents and MAS*. The framework can instantiate and configure an agent and then start it up in a separate thread. The agent is autonym, although it is running in the framework's processing space. An agent can also instruct the framework to start up another agent. The framework keeps track of all agents and their parent agents. So, a single agent can be created which starts up the appropriate agents for each MAS. This *kickoff* agent is then the parent agent of all agents that form the MAS in question.
- *Fleeble supports simple processing of events coming from the outside world and other agents*. Fleeble offers a message distribution system for communication between agents that is based on a Publish/ Subscribe system, which is centered on the concept of a *channel*. Channels are named entities that allow a single message to be delivered to any number of agents. An agent informs Fleeble that it is interested in events pertaining to a specific channel (i.e., it subscribes to that channel). An agent can ask Fleeble to deliver a message to a channel (i.e. it publishes to the channel in question). Fleeble creates a "handler" thread for each agent that has subscribed to the channel in question. All handler threads are started at the same time and deliver the message to the subscribed agents. Hence, e.g., the user's facial cues can be simultaneously processed by both the Identity Agent and the Emotion Agent (see Fig. 1).
- *Fleeble supports the concept of concurrency needed to allow agents to operate independently and yet at the same time*. This has been achieved by starting each agent in a separate thread, allowing it to access the delivery system described above at its own convenience.
- *Fleeble supports data and state persistence*. Fleeble agents can instruct the framework to store values referenced by a key. The framework stores this (key, value) pair and allows access to it at any time, even when the execution of the framework has ceased in the meantime. State persistency allows the user to shut down a single agent (or MAS) and to restore it from the point where it was suspended later on, even when the PC has been shut down in the meantime. This makes Fleeble-based MAS very robust.
- *Fleeble supports the concepts of distribution and mobility*. It establishes socket-to-socket connections between frameworks residing on different computers and manages these connections, i.e., creates and closes them as needed. Connections can be used to transport agents to another host, allowing agents to physically move. The process of being moved to another host can be started either by the agent or by any parent agent.

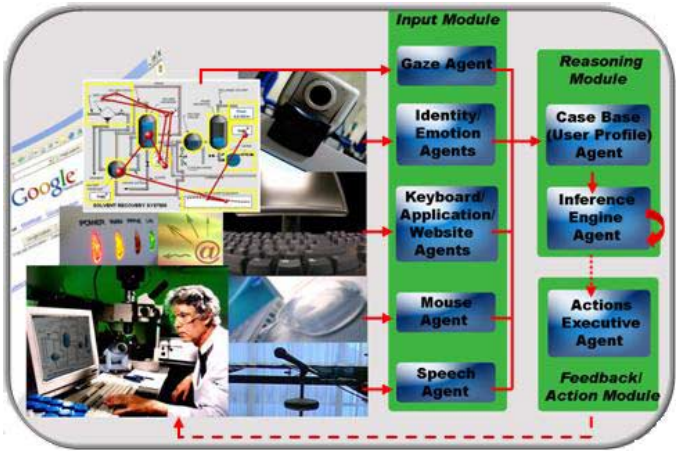


**Figure 1. Overview of the Gaze-X system including the input modalities, modules, and agents.**

## 3. INPUT MODALITIES

The front end of Gaze-X consists of the multimodal input module, which processes images of the user's face, gaze direction, speech, and actions done while interacting with the computer including the mouse movements and the keystrokes.

To process the images of the user's face acquired by a standard web-cam, we use a commercially available Face Reader system for face and facial affect recognition produced by Vicar Vision[1]. This system operates as follows. It detects candidate face regions in the input scene by comparing image regions to a number of prototype faces. These prototypes are representative of a large database of human faces. An Active Appearance Model (AAM) [6] is then fitted to the detected face region. The variations in the shape and texture of the AAM, caused by fitting the AAM to the detected face, serve as a unique identifier of the individual (identifiers for different users are stored in a database). Once the user is identified, AAM fitting is employed to detect the affective state of the user. The variations in the shape and texture of the AAM, caused by fitting the AAM to the user's face in the current frame, are fed to a neural network trained to recognize the prototypic facial expressions of six basic emotions (surprise, fear,

[1] Vicar Vision BV, 2004. http://www.vicarvision.nl.

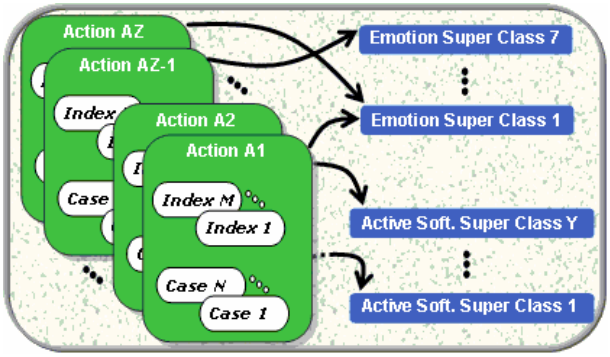


**Figure 2. Schematic organization of the case base of Gaze-X. Each user has his or her own personal case base (profile).**

anger, happiness, sadness, and disgust) [19], defined in classic psychological studies on human emotions [18]. The output layer of the neural network consists of 7 nodes (one for each emotion and one for the neutral state), each of which outputs a continuous value. This enables detection of blends of emotions (e.g., surprise and happiness in an expression of delight) and low-intensity emotions (e.g., a frown is recognized as a low intensity of anger).

To detect the user's gaze direction, we employ a commercial system for remote eye tracking produced by SMI GmbH[2]. The utilized iView X remote eye-tracking system consists of two main components, the infrared pan tilt camera and the software used for the calibration and eye tracking. To determine the direction of the gaze, the system employs the so-called red-eye effect, i.e., the difference in reflection between the cornea and the pupil. This direction determines the point on the screen, which is the focus of the user's attention. As a cursor can be displayed at this point, eye-tracking can be employed to free the user from the constraints of the mouse and, in combination with spoken commands, to allow a hands-free interaction with the computer.

A variety of speech-recognition systems are now available either as commercial products or as open source. Although some novel approaches to speech recognition has been proposed recently [7], most of the existing methods utilize acoustic information of speech contained in frame-by-frame spectral envelopes which are statistically classified by Hidden Markov Models. Gaze-X utilize Sphinx 4[3] speech recognizer, which is a Java-implemented speech recognizer that recognizes a predefined set of words (vocabulary) based on acoustic features and a HMM architecture.

To monitor standard HCI events including keystrokes, currently active software and currently visited web-site, we utilize the Best Free Keylogger software[4]. It monitors website visiting/blocking, e-mail visiting, keystrokes, and application activity, and writes these data into log files. We also monitor and log the locations of the mouse cursor. In the case that the user's preference is to use eye-tracker as an alternative for the mouse, the system does not log the mouse movements.

The user's identity, his or her displayed affective state, current gaze direction, and uttered words, delimit the current user's state. The HCI events including the mouse movements, keystrokes, and the currently active software delimit the current system's state. These two states form further the input to the reasoning module.

## 4. CASE-BASED REASONING

Since the Gaze-X can have different users, each of which can be using different applications in his or her daily work with the PC, while showing emotions and a variety of interactive patterns using the standard and the natural interactive modalities that the Gaze-X supports, the mapping of the system's multimodal input onto a large number of adaptive and user-supportive changes in interface in a user-profiled manner is an extremely complex problem. To tackle this problem, one can apply either eager or lazy learning methods. Eager learning methods such as neural networks extract as much information as possible from training data and construct a general approximation of the target function. Lazy learning methods such as case-based reasoning store the presented data and generalizing beyond these data is postponed until an explicit request is made. When a query instance is encountered, similar related instances are retrieved from the memory and used to classify the new instance. Hence, lazy methods have the option of selecting a different local approximation of the target function for each presented query instance [1]. Eager methods using the same hypothesis space are more restricted because they must choose their approximation before presented queries are observed. In turn, lazy methods are usually more appropriate for complex and incomplete problem domains than eager methods, which replace the training data with abstractions obtained by generalization and which, in turn, require excessive amount of training data. Hence, we chose to implement the inference engine of the Gaze-X as the case-based reasoning about the content of a dynamic memory. The memory is dynamic in the sense that, besides generating user-supportive feedback by analogy to that provided to the user in similar situations "experienced" in the past, it is able to unlearn feedback actions that the user liked once but now tends to dislike and to learn new feedback actions according to the instructions of the user, thereby increasing its expertise in user-profiled, user-supportive, intelligent user-computer interaction.

The utilized dynamic memory of experiences is based on Schank's theory of functional organization of human memory of experiences [37]. According to this theory, for a certain event to remind one spontaneously of another, both events must be represented within the same dynamic chunking memory structure, which organizes the experienced events according to their thematic similarities. Both events must be indexed further by a similar explanatory theme that has sufficient salience in the person's experience to have merited such indexing in the past. Indexing, in fact, defines the scheme for retrieval of events from the memory. The best indexing is the one that will return events most relevant for the event just encountered.

In the case of Gaze-X memory of experiences, each event is one or more micro-events, each of which is an interactive cue (a part of the system's multimodal input) displayed by the user while interacting with the computer. Micro-events that trigger a specific user-supportive action are grouped within the same dynamic memory chunk. The indexes associated with each chunk comprise individual micro-events and their combinations that are most characteristic for the user-supportive action in question. Finally, micro-events of each dynamic memory chunk are hierarchically ordered according to their typicality: the larger the number of

[2] SensoMotoric Instruments GmbH, 2002. http://www.smi.de.
[3] Sphinx-4, 2004. http://cmusphinx.sourceforge.net/sphinx4.
[4] Best Free Keylogger, 2006. http://sourceforge.net/projects/bfk/

times the user was satisfied when the related user-supportive action was executed as the given micro-event occurred, the higher the hierarchical position of that micro-event within the given chunk. Certain user-supportive action can be preferred by the user for both different software applications that he or she is usually using and different affective states that he or she is displaying. Hence, to optimize the search through the case base, affective states and currently active software are not treated as other micro-events but are used as containers of various memory chunks. More specifically, memory chunks representing user-supportive actions that are triggered when the user is displaying a certain emotion are grouped in a super class representing that emotion. Hence, each chunk may contain a pointer to an emotion-identifying super class. Similarly, each chunk may contain a pointer to an active-software-identifying super class. A schematic representation of Gaze-X case base organization is given in Fig. 2.

To decide which user-supportive action is to be executed (if any) given an input set of interactive cues displayed by the user, the following steps are taken:

- Search the dynamic memory for similar cases based on the input set of observed interactive cues, retrieve them, and trigger the user-supportive action suggested by the retrieved cases.
- If the user is satisfied with the executed action, store the case in the dynamic memory and increase its typicality. If the user is not satisfied with the executed action, adapt the dynamic memory by decreasing the typicality of the just executed action.

The simplest form of retrieval is to apply the first nearest neighbor algorithm, that is, to match all cases of the case base and return a single best match. This method is usually to slow. A pre-selection of cases is therefore usually made based on the indexing structure of the utilized case base. Our retrieval algorithm employs a pre-selection of cases that is based on the clustered organization of the case base (super classes and memory chunks), the indexing structure of the memory, and the hierarchical organization of cases within the memory chunks according to their typicality.

Gaze-X can run in two modes, an unsupervised and a supervised mode. In the unsupervised mode, the affective state of the user is used to decide on his or her satisfaction with the executed action. If a happy or a neutral expression is displayed, Gaze-X assumes that the user is satisfied. Otherwise, if the user emotes negatively, Gaze-X assumes that he or she is dissatisfied. In the supervised mode, the user explicitly confirms that an action of his preference has been executed. If the user is not satisfied with the executed action, the dynamic memory is adapted. In the unsupervised mode, the typicality of the relevant case is decreased. In the supervised mode, the typicality of the relevant case is decreased and the user may provide further feedback on the action of his/her preference that should be executed instead.

## 5. INTERACTION ADAPTATION

The final processing step of Gaze-X is to adapt the user-computer interaction based on current needs and preferences of the user and according to adaptive and user-supportive changes suggested by the system's dynamic memory of experiences. General types of interface adaptations supported by Gaze-X include the following.

- Help provision – Examples include the following. When the open-file-dialogue is open for a long time, help can be provided by highlighting the file names that were opened in the past in combination with the currently open files. Alternatively, desktop search application can be started. If the user selected a column in a table, and (s)he is scrutinizing the menu bar for a long time, help can be given by highlighting table-related menu options. Alternatively, help-menu option can be highlighted.
- Addition/removal of automation of tasks – Examples include automatic opening of all windows of an application that the user opens each time he or she starts up the application in question, automatic error correction, automatic blockage of websites that are similar to already blocked sites, and removal of such an automation if the user disapproves of it.
- Changing information presentation – Automatic selection of most relevant features/options to be displayed given the user's current task, automatic font size increase/decrease according to the user's preferences, usage of eye tracking and speech as an alternative to mouse movements, automatic sound play, etc, are typical examples of this type of interface adaptation.

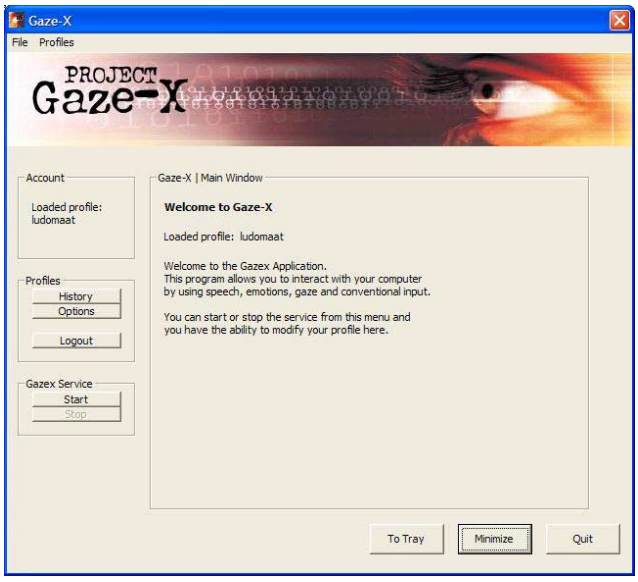


**Figure 3. The main window of the GUI of the Gaze-X.**

Gaze-X carries out interface changes in a rather conservative way. More specifically, when operating in the unsupervised mode, it executes adaptive and user-supportive actions one at the time and in a rather slow pace. The underlying philosophy is not to make an ever-changing interface a source of user's frustration in itself. In order to allow the user some time to get accustomed to the idea of a self-adaptive interface and to initialize the system using a (small) set of interface changes that the user considers helpful, Gaze-X is initially set to operate in the supervised mode. As soon as the user considers the system profiled enough, he or she can set the system to operate in the unsupervised mode.

## 6. THE GUI OF THE GAZE-X

Gaze-X has a simple, easy to understand GUI. The goal was to develop a direct-manipulation GUI in which WYSIWYG (what you see is what you get) would be the guiding principle. A simple self-explicatory GUI was developed that is easy to understand and use. As can be seen in Fig. 3, the main window of Gaze-X GUI visualizes who the current user is (i.e., whose user-profiled case base is currently used), allows loading, creation, and adaptation of

the profiles, and enables initiation of adaptive interaction mode supported by the system. Gaze-X automatically loads the user's profile for a known user based on the output of the Face Reader system as explained in section 3. However, an option to load a profile manually, using a valid username and password, also exists. Besides, since Gaze-X is developed as Fleeble-based MAS, the agents that constitute the system including the reasoning agent (representing the profile of the user) are mobile and can be moved to another host. Hence, users' profiles can be transmitted as needed to any computer where Gaze-X is installed.

When installing Gaze-X, a directory needs to be found where the main system and the supporting systems and sensors (Fleeble, the FaceReader face and facial expression detector, the iView-X remote eye tracker, a web cam, and a microphone) can be installed with proper access rights. Also a specific version of Java Runtime Environment (version 1.5 or later) needs to be properly installed before Gaze-X can run. To make this process as easy as possible for the user, a setup wizard has been implemented. It executes automatically many of the required steps to set up Gaze-X and leads the user through the rest of the required steps to ensure that the required hardware and software are properly setup.

Any new way of thinking about programming takes some getting used to, and Gaze-X is no exception to this axiom. To aid users in working out questions that they may have, Gaze-X provides a tutorial. It shows the basic functionalities and usage of Gaze-X in a step by step demonstration. The tutorial can be started by means of 'To Try' button illustrated in Fig. 3.

# 7. USABILITY STUDY

To make a preliminary assessment of effectiveness, usability, usefulness, affective quality, and ethical issues relevant to Gaze-X, we conducted a small evaluation study with the help of six participants. The participants were 18 to 61 years old, 33% female, 50% Caucasian, 33% Asian, 17% African, 17% expert, 50% intermediate, and 33% novel computer users. We asked them to install Gaze-X first and then to use it as demonstrated by the tutorial integrated into the system. For each session with another user, we used either a Linux or a Windows machine from which the Gaze-X was removed. We did not require the users to use the Gaze-X for a certain period of time, each of them was engaged in exploring Gaze-X for as long as he or she wanted. We also did not require from the participants to work with specific software applications. Note, however, that the software installed on the machines we used for the experiments had an Internet browser, an e-mail handler, a text editor, Adobe Reader, and a number of multimedia handlers such as music and movie players.

We used a custom-made questionnaire to elicit users' attitudes towards the Gaze-X-supported interaction with the computer. The questionnaire includes questions soliciting users' attitudes toward:

- the effectiveness of the HCI design (i.e., whether the interaction with computer is more natural than it is the case with standard HCI designs and whether it is robust enough, [35]),
- the usability of Gaze-X (i.e., whether technological variety and user diversity are supported and whether gaps in the user's computer knowledge play an important role, [38]),
- the usefulness of the system (i.e., whether the utility of system's functionalities and the utility of interface adaptation supported by the system are obvious to the users, whether they will choose to use the system if it was publicly available),
- the ethical issues related to the HCI design (i.e., do users feel uncomfortable under the scrutiny of machines that monitor their affective states, work- and interaction patterns, will users adapt soon to emote just to make the computer do something, [21]),
- the affective quality of the HCI design (i.e., whether the GUI of the Gaze-X is aesthetically qualitative in terms of orderly, clear, and creative design, [44]).

The utilized questionnaire also invites participants' suggestions on how to improve the system in any of the aspects mentioned above. It employs a 5-point Likert scale ranging from strongly disagree (1), via neutral (3), to strongly agree (5). 'I do not know' is also a possible answer. The main points of the obtained survey results are listed in Table 1.

**Table 1.** Users' satisfaction with effectiveness, usability, usefulness, affective quality, and ethical issues relevant to Gaze-X. The percentages in the table show the percentage of *agree* and *strongly agree* answers. EU stands for expert users (total: 17%) and NU stands for intermediate and novel users (total: 83%). × indicates an 'I do not know' answer.

| Survey question | EU | NU |
|---|---|---|
| The interaction is more natural than in standard HCI | 0% | 100% |
| The interaction is robust enough | 0% | 66% |
| Gaze-X supports technological variety | × | 100% |
| Gaze-X supports users of different age, skills, culture… | 100% | 100% |
| Gaze-X is easy to use even if users lack IT knowledge | 100% | 100% |
| Face identification is a useful functionality | 100% | 100% |
| Monitoring affective states is a useful functionality | 0% | 66% |
| Having mobile user profiles is useful | 100% | 100% |
| Having multimodal interaction is useful | 0% | 66% |
| Adaptive and user-supportive interface is useful | 0% | 100% |
| Gaze-X makes the interaction with computer easier | 0% | 100% |
| I do not mind to be observed by the camera | 100% | 100% |
| I like the aesthetics of the Gaze-X GUI design | 100% | 100% |
| I would use Gaze-X if it was publicly available | 0% | 66% |

All participants seem to agree on the usability, affective quality, and ethical issues relevant to Gaze-X. The perceived usability is directly related to the following properties of Gaze-X: it runs on various platforms like Linux, Windows, and Mac Os X (as a result of being Java-implemented), it accommodates users of different age, gender, culture, computing skills and knowledge, and it bridges the gap between what the user knows about the system and HCI in general and what he or she needs to know (it provides a setup wizard and a tutorial that shows how to use Gaze-X in a step by step demonstration). The perceived affective quality of Gaze-X is directly tied to aesthetics qualities of the Gaze-X GUI: (i) it has an orderly and clear design in accordance to the rules advocated by usability experts, and (ii) it reveals designers' creativity, originality, and the ability to break design conventions manifested by, for example, multimodal design, affective design, tutorial demonstration, etc. These findings are consistent with research finding of Zhang and Li [44], who suggested that the perceived affective quality of a software product is directly tied to aesthetics qualities of that product and who argued that affectively qualitative products have significantly larger chance to be widely accepted technology. The perceived ethical issues relevant to Gaze-X were somewhat surprising as it seems that the users have no problem with being continuously observed by a web cam. Standard concern that the user's behavioral and affective patterns could be used to mind-read and manipulate him or her was not mentioned a single time. However, all participants in our study stressed the importance of privacy

and asked about measures taken to prevent hacking and intrusion. Ultimately, if users have control about whether, when, and with whom they will share their private information such as the observations of their behavioral patterns while interacting with the PC (as stored in their personal profile), fears from big-brother-is-watching-you scenarios vanish.

However, as can be seen from Table 2, the question that remains is whether or not an individual user values the capability of Gaze-X to be aware of his or her behavioral and affective patterns. Our study suggests that there might be a very large difference between the expert and non expert computer users when it comes to this question. While all users agreed that having a face-identification-based access to the system is a very useful functionality, expert users found other functionalities of Gaze-X less appealing and in some instances irritating (e.g., popup widows used in supervised operating mode to ask the feedback about the user's preferences). On the other hand, less experienced computer users perceived Gaze-X as very useful since it provides support when needed, in a way needed and preferred, making the interaction with the computer more efficient and easier (e.g., less time was spent on searching various functionalities and undoing erroneous actions). In turn, it seems that Gaze-X is very suitable for novel and intermediate computer users but is much less so for experienced users. As suggested by the expert computer user who participated in our study, much more sophisticated user support should be provided to an experienced user than it is currently the case. For example, support should be provided only when a new application is installed and used for the first couple of times, no support should be provided for long installed software applications. Note, however, that only one expert computer user has participated in the present usability study. A much more elaborate survey must be conducted with experienced users if some firm conclusions are to be made about the ways to make Gaze-X useful and appealing to experienced computer users.

Finally, all participants remarked that the robustness of the system can be improved. Two issues are of importance here. The first is the sensitivity of the Face Reader system, used for face and facial expression analysis, to changes in lighting conditions. The second is the sensitivity of the iView-X remote eye tracker to changes in the user's position. More specifically, the user is expected to remain in front of the computer and is not allowed to shift his or her position in any direction for more than 30 cm. Otherwise, the iView-X system should be recalibrated before it can be used again. These findings indicate that in the future more robust systems for facial expression analysis and eye tracking should be considered for inclusion in the Gaze-X.

From other remarks mentioned by the participants in the present study, arguably the most important one relates to the choice of the affective states to be tagged by Gaze-X. Most of participants said that they may experience confusion, frustration, understanding, tiredness, and satisfaction while interacting with the computer. However, the currently employed Face Reader system recognizes only facial expressions of six basic emotions including disgust, fear, and sadness, for which the participants in our study said that they are not likely to be experienced in a HCI setting like office scenarios. This indicates that in the future we should employ either automatic analyzers of attitudinal and non-basic affective states like attentiveness [10] and fatigue [11], or systems for user-profiled interpretation of facial expressions [29].

# 8. CONCLUSSIONS

In this paper we proposed one of the first systems for adaptive, affective, multimodal human-computer interaction in standard office scenarios. Our system, Gaze-X, is based on sensing and interpretation of the human part of the computer's context, known as W5+ (who, where, what, when, why, how) and, in turn, is user- and context-profiled. The user of Gaze-X experiences an adaptive interface, which changes as a function of the currently sensed context. This function is represented by the cases of the utilized dynamic case base, having the system's and the user's state as the input (represented in terms of exhibited multimodal interactive actions and cues) and the adaptive and user-supportive changes in the interaction as the output. A usability study conducted in an office scenario with the help of six users indicates that Gaze-X is perceived as effective, easy to use, and useful by novice and less experienced users and as usable and affectively qualitative by all participants in the present study. In turn, Gaze-X seems to be very suitable for novel and intermediate computer users but is much less so for experienced users. As only one experienced user has participated in the present usability study, a much more elaborate survey must be conducted with experienced users if some firm conclusions are to be made about the ways to make Gaze-X useful and appealing to this group of users. Ultimately, as the majority of current software products are still designed for experienced frequent users and designing for a broad audience of unskilled users is still seen as a far greater challenge [38], we are very glad and proud that Gaze-X was perceived as useful and easy to use by novice and less experienced users, who in general still experience computing technology as too difficult to use. Except of a more elaborate usability study with a large number of various users, our future efforts will also be aimed at enabling the system to tag attitudinal and non-basic affective states like confusion, stress, fatigue and satisfaction.

# 9. ACKNOWLEDGMENTS

The authors would like to thank Marten den Uyl of Vicar Vision BV for providing the Face Reader system. The work has been conducted at Delft University of Technology. The work of M. Pantic has been supported by the Netherlands Organization for Scientific Research (NWO) Grant EW-639.021.202. The cooperation with Vicar Vision BV has been carried out in the scope of the Dutch BSIK-MultimediaN-N2 project on Interaction.

# 10. REFERENCES

[1] Bartsch-Sporl, B., Lez, M. and Hubner, A. Case-based reasoning – survey and future directions. *Lecture Notes in Artificial Intelligence, 1570*, 67-89, 1999.

[2] Bianchi-Berthouze, N. and Lisetti, C.L. Modeling multimodal expression of user's affective subjective experience. *User Modeling and User-Adapted Interaction, 12*, 1 (Feb. 2002), 49-84.

[3] Bolt, R.A. Put-that-there: Voice and gesture at the graphics interface. *Computer Graphics, 14*, 3 (July 1980), 262-270 (Proc. ACM SIGGRAPH'80).

[4] Browne, D., Norman, M. and Riches, D. Why Build Adaptive Interfaces? In *Adaptive User Interfaces*. Browne, D., Totterdell, P. and Norman, M., Eds. Academic Press, London, UK, 1990, 15-57.

[5] Conati, C. Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence, 16*, 7-8 (Aug. 2002), 555-575.

[6] Cootes, T.F., Edwards, G.J. and Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence, 23*, 6 (June 2001), 681-685.

[7] Deng, B.L. and Huang, X. Challenges in adopting speech recognition. *Communications of the ACM, 47*, 1 (Jan. 2004), 69-75.

[8] Dey, A.K., Abowd, G.D. and Salber, D. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *J. Human-Computer Interaction, 16*, 2-4 (Dec. 2001), 97-166.

[9] Duric, Z., Gray, W.D., Heishman, R., Li, F., Rosenfeld, A., Schoelles, M.J., Schunn, C. and Wechsler, H. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceedings of the IEEE, 90*, 7 (July 2002), 1272-1289.

[10] El Kaliouby, R. and Robinson, P. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *Proc. Int'l Conf. Computer Vision & Pattern Recognition, 3*, 154, 2004.

[11] Gu, H. and Ji, Q. An automated face reader for fatigue detection. In *Proc. Int'l Conf. Face & Gesture Recognition*, 111-116, 2004.

[12] Gunes, H. and Piccardi, M. Affect Recognition from Face and Body: Early Fusion vs. Late Fusion, In *Proc. Int'l Conf. Systems, Man and Cybernetics*, 3437- 3443, 2005.

[13] Hoffman, D.L., Novak, T.P., and Venkatesh, A. Has the Internet become indispensable? *Communications of the ACM, 47*, 7 (July 2004), 37-42.

[14] Hudlicka, E. To feel or not to feel: The role of affect in human-computer interaction. *Int'l J. Human-Computer Studies, 59*, 1-2 (July 2003), 1-32.

[15] Hudlicka, E. and McNeese, M.D. Assessment of user affective/belief states for interface adaptation. *User Modeling & User-Adapted Interaction, 12*, 1 (Feb. 2002), 1-47.

[16] Jaimes, A. and Sebe, N. Multimodal human computer interaction: A survey. In *Proc. Int'l Workshop on HCI in conjunction with Int'l Conf. Computer Vision*, 2005.

[17] Kapoor, A. and Picard, R.W. Multimodal affect recognition in learning environments. In *Proc. ACM Int'l Conf. Multimedia*, 677-682, 2005.

[18] Keltner, D. and Ekman, P. Facial expression of emotion. In *Handbook of Emotions*, Lewis, M., and Haviland-Jones, J.M. Eds. The Guilford Press, New York, 2000, pp. 236-249.

[19] van Kuilenburg, H., Wiering, M. and den Uyl, M. A model-based method for automatic facial expression recognition. *Lecture Notes in Artificial Intelligence, 3720*, 194-205, 2005.

[20] Lisetti, C.L., Nasoz, F. MAUI: A multimodal affective user interface. In *Proc. Int'l Conf. Multimedia*, 161-170, 2002.

[21] Lisetti, C.L. and Schiano, D.J. Automatic facial expression interpretation: Where human-computer interaction, AI and cognitive science intersect. *Pragmatics and Cognition, 8*, 1 (Jan. 2000), 185-235.

[22] Marsic, I., Medl, A. and Flanagan, J. Natural communication with information systems. *Proceedings of the IEEE, 88*, 8 (Aug. 2000), 1354-1366.

[23] Nielsen, J. *Multimedia and hypertext: The Internet and beyond*. Academic Press, Cambridge, USA, 1995.

[24] Nock, H.J., Iyengar, G. and Neti, C. Multimodal processing by finding common cause. *Communications of the ACM, 47*, 1 (Jan. 2004), 51-56.

[25] Oviatt, S. User-centred modelling and evaluation of multimodal interfaces. *Proceedings of the IEEE, 91*, 9 (Sep. 2003), 1457-1468.

[26] Pantic, M., Grootjans, R.J. and Zwitserloot, R. Teaching Ad-hoc Networks using a Simple Agent Framework. In *Proc. Int'l Conf. Information Technology Based Higher Education and Training*, pp. 6-11, 2005.

[27] Pantic, M., Pentland, A., Nijholt, A. and Huang, T.S. Front-end of human computing: Machine analysis of human behaviour. In *Proc. Int'l Conf. Multimodal Interfaces*, 2006.

[28] Pantic, M. and Rothkrantz, L.J.M. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE, 91*, 9 (Sep. 2003), 1370-1390.

[29] Pantic, M. and Rothkrantz, L.J.M. Case-based reasoning for user-profiled recognition of emotions from face images. In *Proc. Int'l Conf. Multimedia and Expo*, 2004, 391-394.

[30] Pantic, M., Sebe, N., Cohn, J.F. and Huang, T.S. Affective multimodal human-computer interaction. In *Proc. ACM Int'l Conf. Multimedia*, 669-676, 2005.

[31] Pentland, A. Perceptual intelligence. *Communications of the ACM, 43*, 3 (Mar. 2000), 35-44.

[32] Picard, R.W. *Affective Computing*. The MIT Press, Cambridge, USA, 1997.

[33] Porta, M. Vision-based user interfaces: methods and applications. *Int'l J. Human-Computer Studies, 57*, 1 (July 2002), 27-73.

[34] Prendinger, H. and Ishizuka, M. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence, 19*, 3-4 (Mar. 2005), 267-285.

[35] Reeves, L.M., Lai, J., Larson, J.A., and Oviatt, S. Guidelines for multimodal user interface design. *Communications of the ACM, 47*, 1 (Jan. 2004), 57-59.

[36] Ruttkay, Z. and Pelachaud, C. Eds. *From brows to trust: Evaluating embodied conversational agents*. Kluwer Academic Publishers, Norwell, USA, 2004.

[37] Schank, R.C. *Memory based expert systems*. AFOSR.TR. 84-0814, Comp. Science Dept., Yale University, 1984.

[38] Schneiderman, B. Universal usability. *Communications of the ACM, 43*, 5 (May 2000), 85-91.

[39] Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Maceachren, A.M. and Sengupta, K. Speech-gesture driven multimodal interfaces for crisis management. *Proceedings of the IEEE, 91*, 9 (Sep. 2003), 1327-1354.

[40] Tennenhouse, D. Proactive computing. *Communications of the ACM, 43*, 5 (May 2000), 43-50.

[41] Turk, M. Computer vision in the interface. *Communications of the ACM, 47*, 1 (Jan. 2004), 61-67.

[42] Vo, M.T. and Waibel, A. Multimodal human-computer interaction. In *Proc. Int'l Symposium on Spoken Dialogue*, 1993.

[43] Waibel, A., Vo, M.T., Duchnowski, P. and Manke, S. Multimodal Interfaces. *Artificial Intelligence Review, 10*, 3-4 (Aug. 1995), 299-319.

[44] Zhang, P., and Li, N. The importance of affective quality. *Communications of the ACM, 48*, 9 (Sep. 2005), 105-108.