VISUAL INTERPRETATION OF FACES IN THE NEIMO MULTI-MODAL HCI TEST-BED

Patrice De Marconnay , James L. Crowley

LIFIA (IMAG)
46 av. Felix Viallet, 38031 Grenoble, FRANCE.

and Daniel Salber

Laboratoire de Génie Informatique, IMAG
B.P. 53 X, 38041 Grenoble Cedex, France

Abstract

The Wizard of Oz technique (WOz) is an experimental evaluation mechanism. It allows the observation of a user operating an apparently fully functioning system whose missing services are supplemented by a hidden wizard. NEIMO is a multimodal WOz platform including four modalities: two-dimensional manual gesture (mouse), written and spoken natural language and visual interpretation of the subject's face. A sub-system, NEIMOVision, is responsible for the visual interpretation of the face. It's functionalities are to normalize the image intensity, locate the face position, orientation and size, identify the subject, detect the direction of gaze, analyse the facial expression and detect speech acts. These functionalities are mainly based on the Eigenface technique developped by Turk [Turk 91]. Demonstration pprograms based on face recognition have been implemented on a Macintosh Quadra 700 equiped with a small CCD camera. At the same time we are currently studying other technical issues related to the integration of vision as an HCI mode. We are using this system to investigate the requirements in computing power, the trade-offs between resolution and processing speed, the use of active vision, the possible extensions to hand-tracking and the role of visual processing as a source of HCI modalities.

1. Introduction

This paper presents the use of visual interpretation of faces within the NEIMO test-bed for experiments in multi-modal human-computer interaction (HCI). The paper will begin with a presentation of the NEIMO project. It will then describe the use of real time visual interpretation of faces as a basis for communication modalities. Face interpretation within the NEIMO project is based on the technique of principal components analysis developed by Turk [Turk 91]. Turk's "Eigenface" technique is reviewed. Subsections then discuss techniques for normalising the contrast of a face image, for locating the face, for classifying the face as a known user, eye tracking as a pointing mechanism, and lip motion as a trigger for speech recognition. The paper concludes with technical problems to be resolved for the use of face interpretation as a HCI communications mode.

2. The NEIMO “Wizard of Oz” test-bed

Speech and vision have both made rapid progress in the last few years. Unfortunately, these techniques have not quite matured to the point where they can provide reliable HCI modalities. While the necessary computer power is becoming available, technical and HCI problems remain about how such modalities should best be integrated into the user interface. The “Wizard of Oz” approach permits us to investigate the role for these new modalities by replace one or more components with a human wizard.

2.1 The Wizard of Oz technique

The Wizard of Oz (WOz) technique is an experimental evaluation mechanism. It allows the observation of a subject operating an apparently fully functioning system whose missing services are supplemented by a hidden operator (a wizard). The subject is not aware of the presence of the wizard and is led to believe that the computer system is fully operational. The wizard may observe the subject by any means such as a dedicated computer system connected to the observed system over a network, an internal video circuit, or a combination of both. When the subject invokes a function that is not available in the observed system, the wizard simulates the effect of the function. Through the observation of subject behaviors, designers can identify user needs for accomplishing a particular set of tasks and can evaluate the particular interface used to accomplish the tasks.

Telephone information services such as telephone directories, flight or train reservation services have previously been studied using this approach [Fraser 92]. In such systems, the wizard answered phone calls and pretends callers are talking to an automatic information system. Other case studies involve databases or advisory systems interrogation [Whittaker 89] as well as dialogues with expert systems [Polity 90].

Most of the existing WOz systems have been developed on a case-per-case basis and support the observation of a single modality. . Similarly, automated analysis tools have been limited in scope and rarely integrated into the WOz platform from the start. There has been no attempt to produce a generic, reusable WOz platform that would make possible the observation and analysis of multimodal interaction. With recent advances in interactive media, multi-modal user interfaces are becoming popular. So, we have adapted the WOz technique for the experimental evaluation of modalities for human-computer interaction [Salber 93].

2.2 The NEIMO Testbed

NEIMO (New Evaluation of Interfaces using the Wizard of Oz technique) is a generic and extensible multimodal WOz platform. It is designed and developped for evaluating communication modalities in human-computer interactions.

The goal of the NEIMO project is to experimentally evaluate interaction modes. The test-bed includes the possibilities of interaction using speech understanding (speech wizard), gesture-recognition (mouse wizard), and visual interpretation of the face (face wizard) as well as the standard HCI tools of a mouse and graphics on a bit-mapped screen.

Figure 1 shows an example configuration involving the following modalities: mouse pointing, speech, and facial expression.

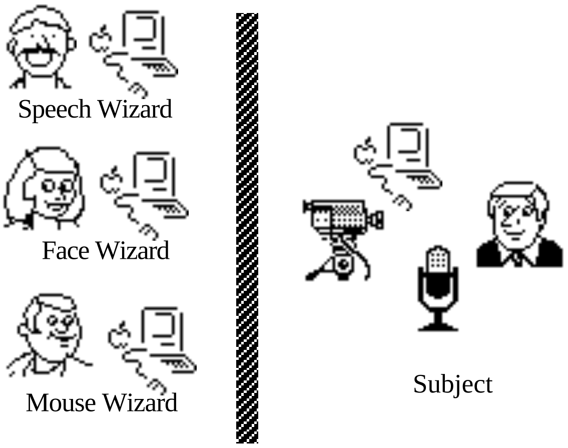


Fig. 1. An example configuration
for a Neimo experiment (adapted from [Salber 93])

In the NEIMO test-bed, a human subject uses the communication modalities provided by the system to perform a task. One or more wizards are hidden in a separate room apart from the subject. Wizards have two different functions: they interpret the actions of the subject, and they generate or supervise the machine response to the subjects actions.

NEIMO permits modalities to be provided by a wizard or by the computer system, and provides tools for the automatic observation of a subject and for the analysis of his behaviour while performing a task. The system records the actions of the subject to permit an analysis of his choice of modalities and his use of a particular interface configuration. Observation tools permit the physical actions of the subject as well as those of the wizard to be recorded. The captured actions form the subject of an automatic off-line interpretation.

The NEIMO system is implemented using AppleMacintosh Quadra computers, connected via Appletalk using Ethernet. The system currently includes the following modalities:

- two-dimensional manual gesture (mouse);
- written (typed) natural language;
- spoken natural language;
- visual interpretation of the subjects face.

In the following section we present software tools for face interpretation that have been constructed within the NEIMO platform: THe NeimoVision system.

3. NEIMOVision

3.1. Introduction

The part of the NEIMO system concerned with visual interpretation has been named "NeimoVision". NeimoVision performs three tasks within NEIMO. Firstly, NeimoVision displays the subject image on the face wizard workstation, permitting the wizard to efficiently observe the subject. Secondly, it performs some automatic processing (identification, gaze direction,...) to reduce the wizard cognitive charge. Thirdly, it provides images for data recording to make history files.

NEIMOVision has been designed to provide the following functionalities:

- Locate a face;
- Identify the subject;
- Detect the direction of gaze;
- Analyse the facial expression;
- Detect speech acts.

These functionalities are illustrated in figure 2.

From the lowest level of abstraction (i. e. an image of pixels) NeimoVision detects and locates the face, then normalizes the image gray levels. Through the four high level data processing, a subject's state model can be estimate.

Face interpretation functionalities are based on the Eigenface technique developed by Turk [Turk 91]. In this technique, an image is considered to be a vector in which each pixel is a dimension. Thus, a 128x128 image leads to a 16,384 dimensions vector. The database of images are decomposed into a smaller set of vectors using principal components analysis. This set of eigenvectors (called "Eigenfaces" because they represent images of face) defines an hyperplan: the "face-space". Classes of patterns may be defined as occupying regions of this face-space.

The Eigenface interpretation technique can be used for several purposes. Classifying an image as containing a face or not, recognizing a known-person, are two straightforward applications. Another one is to estimate face parameters by measuring the position along a trajectory through an eigen-space defined by sample images of a face or face components.

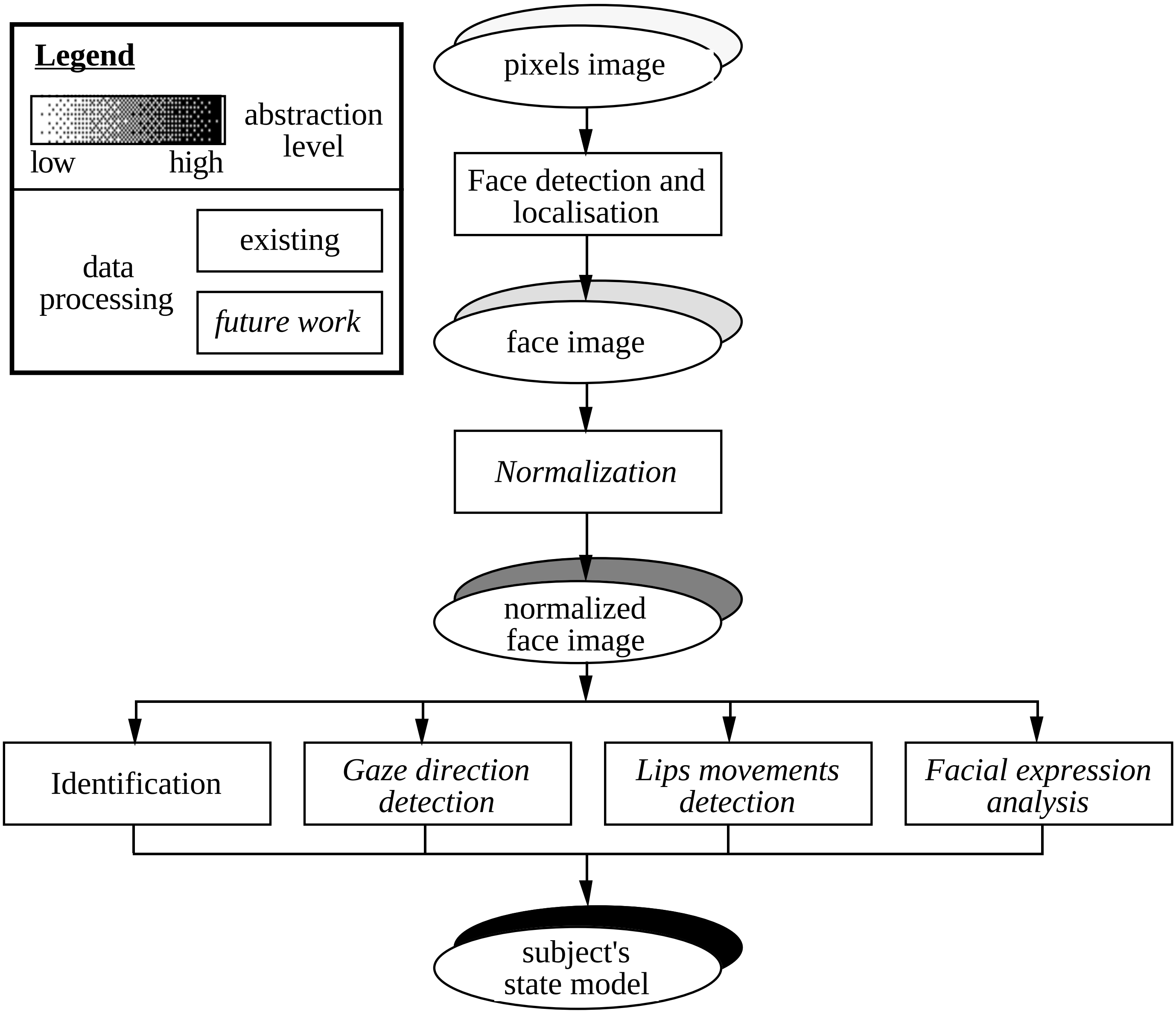


Figure 2. The NeimoVision functionalities

3. 2 The NeimoVision hardware environment

The NeimoVision hardware environment is closely dependant on the NEIMO test-bed environment. A small CCD camera is mounted at the bottom of a 16” color monitor at which the subject performs his task. This camera is connected to a video acquisition board within the workstation in a separate room at which the face wizard is seated. The camera is focused on the subject’s face when the subject is seated normally before the monitor.

Figure 3 illusttrates the NeimoVision hardware configuration for the “face wizard” and the subject. The “face wizard” can observe the subject's face digitized in real time, execute programs on the digitized images or interpret the images himself, communicate the interpretations to the other wizards to help them in their task, and store the interpretations for later analysis.

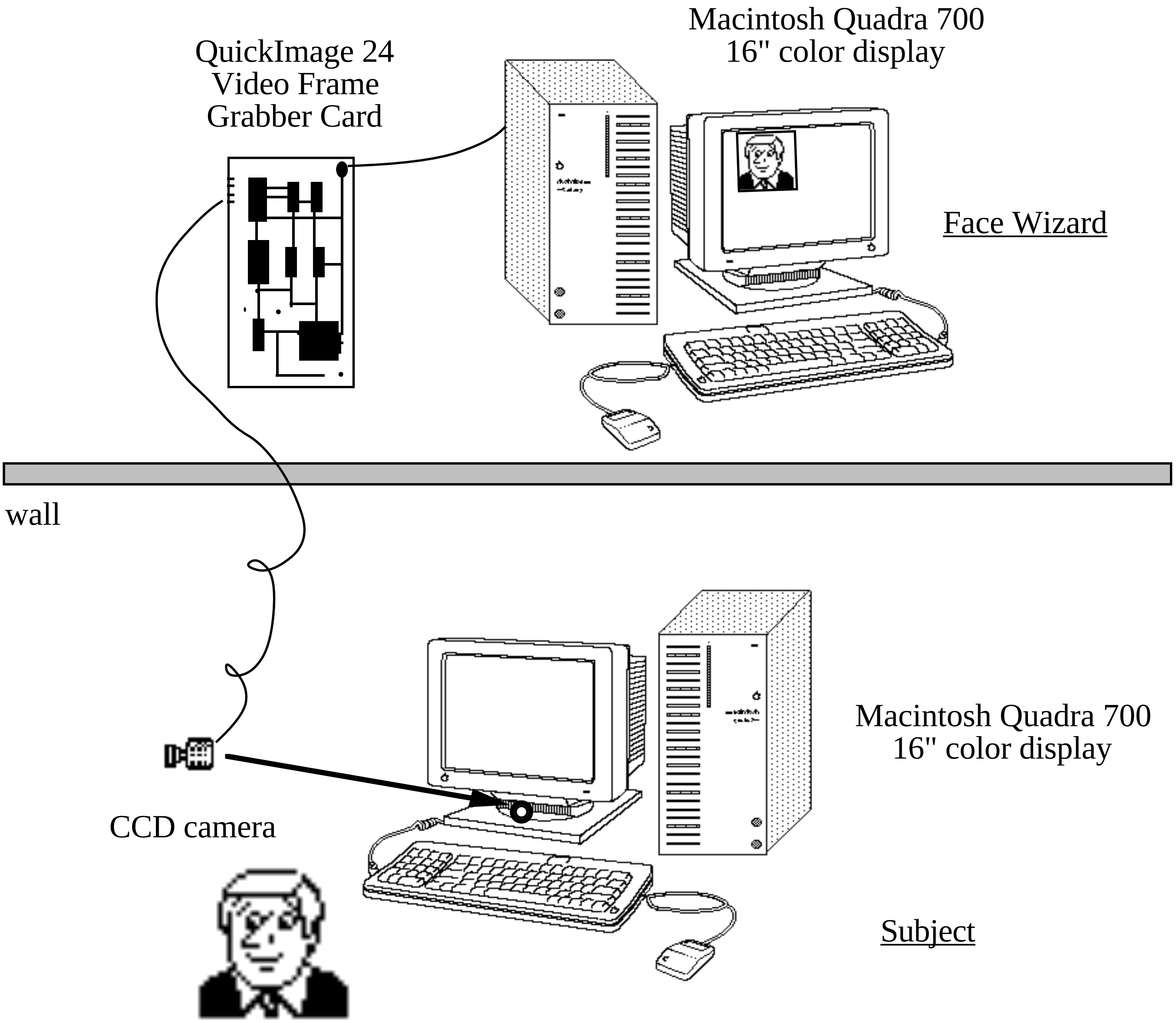


Figure 3. The NEIMOVision hardware configuration

Within an interaction where the subject can use multiple modalities concurrently, the wizard must simulate complex functions. Response time is often critical to system performance. Thus the wizard must have a very efficient way of observing the subject. . If the system is slow, the wizard cannot compensate. Thus the system has to provide the effective foundations of efficiency and quickness.

3.3. Locating a face

Two different techniques are being invetigated for the problem of locating the face in an image: correlation with a prototype generic face, and detection of blinking.

The generic face is provided by the zeroth order eigenface. The average face can be convolved with the face image to produce a “peak” at the location of a face in the image. Because convolution of an entire image with another image is computationally expensive (N^2 operations for images composed of N pixels), we are investigating methods using coarse to fine correlation within a multi-resolution pyramid. A variation of this approach involves forming a smaller mask using only the eyes which is then correlated with the image.

The second technique consists of exploiting the space-time pattern provided by blinking. Blinking is the generally the quickest non-moving change that can be found in sequence of face images. A temporal derivative is computed by smoothing and differencing the dence temporal sequence of images. A blinking motion will produce a characteristic pair of small regions of temporal change. By searching for two such regions with the proper size and separation, we can detect the location of the eyes, and thus the face by measuring the eyes separation and orientation..Digitizing hardware currently limits the use of this technique.

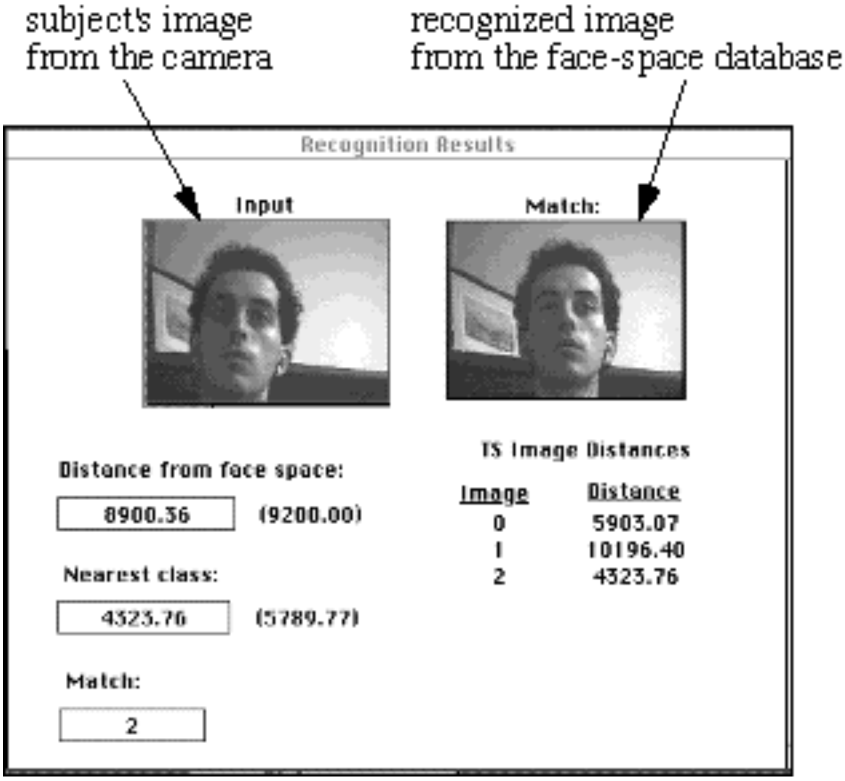
3.4. Normalization

Recognition using Eigenfaces is very sensible to the position, size and orientation of the face as well as the background pattern and gray-level intensity. In the normalisation step we try to cancel these factors.. Using the location and size of the face issued from the previous step, we transform the face to a standard position and size. We then apply a mask to eliminate the background. We have found that the EigenFace approach is very sensitive to backkground information. The gray scale range of the resulting face image is normalized using histogram equalisation. Gray-scale normalization is a difficult problem in computer vision. The efficiency of the Eigenface technique directly relies on the efficiency of the gray-scale normalization algorithm. This particular point will be precisely study.

3.5. Face Recognition

One of the simplest applications of the eigenfaces method [Turk 91] is the recognition of the subject. We have prepared a simple demo which works as follows. At the beginning of a session, the system classifies the subjects face in order to determine if the subject is known. Classification is performed by multiplying the normalized face by each of the principle component images in order to obtain a vector. The vector positions the image in the “face space” defined by the current Eigenfaces. If the face is near a position of this space which corresponds to a known subject, then the subject’s image from the face-space database is displayed. If the vector is not near a known subject, the subject is classified as unknown and no face is displayed. Using the Eigenface technique, our Quadra 700 with no additional hardware can digitize and classify a face within a 108 by 120 image for a database of 12 images at about 1 frame per second.

The figure 4 shows the recognition window of the NeimoVision demo (with a face database of three faces).



subject's image from the camera

recognized image from the face-space database

Distance from face space:

8900.36 (9200.00)

Nearest class:

4323.76 (5789.77)

Match:

2

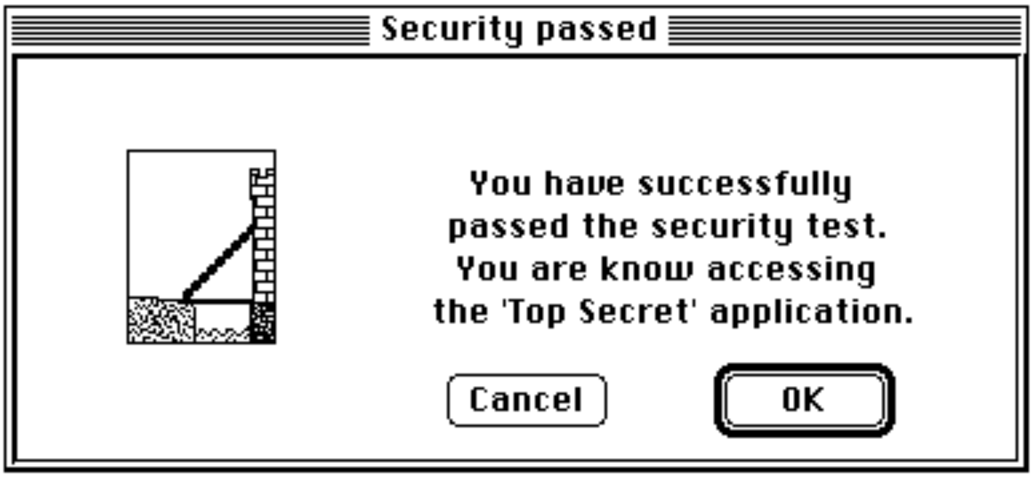
TS Image Distances

Image	Distance
0	5903.07
1	10196.40
2	4323.76

Fig. 4. The NEIMOVision demo: the recognition window

We are currently implementing another demo with a very simple interface. This is a kind of login process without password. The user is asked to sit in front of the camera. The system takes a snapshot of him and try to recognize the user. If the user is in the face-space database, he has successfully passed the security test and is able to use a "Top Secret" application. If he is not recognized, the system informs the user that he cannot access the "Top Secret" application because he isn't registered.Then, the user can make a demand to be registered.

The window displayed when the user has successfully passed the security test is shown in figure 5.



You have successfully passed the security test. You are know accessing the 'Top Secret' application.

Fig. 5. The new demo:
the Security Passed Window

The location, orientation and position of the subjects face, as well as his identity are useful informations. If we are able to determine where he's currently looking, we can begin to elaborate a model of his internal state of mind.

3.6. Determining Direction of gaze

It is possible to use the Eigenface technique to measure parameters. One example of this is for eye-tracking. We train a set of images of the subject looking in different directions and use these images to form an Eigen-Space. During execution of a task, a high-resolution window is placed over the subjects eyes, and the position in the Eigen-Space is computed. The nearest principle components are used to interpolate the current horizontal and vertical direction.

We are experimenting with this technique to determine the trade-off between resolution of the windows on the eyes, the number of eigen-images needed, and the precision which we can obtain in eye tracking. The goal is to be able to drive a pointing device, such as a mouse with such eye tracking.

3.7. Recognizing facial expression

Facial expression contains useful informations about the user's state of mind. In the Neimo experiment, the user's state of mind is a very interesting information. The Eigenfaces idea can be easily extended to classifying the users facial expression.

A set of facial expressions are obtained by the face wizard as the subject performs his task. These facial expressions are then used to form an Eigenface. At each instant, the system determines the face expression class which most closely corresponds to the users current expression. In this way, we can experiment with anticipating the users "mood" based on facial expression.

4. Technical issues related to the integration of vision as an HCI mode

The paper will conclude with technical problems to be resolved for the use of face interpretation as a HCI communications mode. In particular we will consider:

1) Requirements in computing power;
2) Trade-offs between resolution and processing speed;
3) The use of active control of camera parameters such as pan, tilt, focus and zoom;
4) Extensions to hand-tracking;
5) The role of visual processing as a source of HCI modalities.

4.1. Requirements in computing power

The user of NeimoVision is the Face Wizard. He is not inevitably a computer expert. Therefore the human interface of NeimoVision has to be very complete; not only a researcher's application ! Moreover the quality and accuracy of the interface is an important element of the wizard efficiency. So NeimoVision has to be designed with a very accurate interface. Such interfaces are known to be computationnaly high.

The NEIMO project is implemented on Macintosh computers. The Macintosh Quadra 700 runs on a 25 MHz Motorola 68040 processor. It is a RISC and CISC processor, equiped with an arithmetic coprocessor. The next generation (Quadra 800) seems to be a little quicker. Nevertheless we cannot plan yet to compute the same operations on the Quadra as on a Unix Workstation.

For now, NeimoVision is completely software implemented; no specific hardware is used except the video acquisition board. Particularly mage multiplications and additions are done by software. Such operations are computationally high. Specific cards may be used to manage these operations very usual in computer vision. Thus we can save CPU time that can be used for other usefull processings.

4.2 Trade-offs between resolution and processing speed

With our video acquisition board, we can work on digitized image with resolution up to 768 by 576 (442,368 pixels). Actually the images used in NeimoVision are 108 by 144 images (only 15,552 pixels!). This restriction is due to the fact that half a million of pixels cannot be used by our system in real time.

The image resolution is one important information to evaluate the processing time of usual computer vision algorithms. Typically the processing time grows along with the square of the pixel numbers. But precision in detection involves high resolution images. In our case, we probably need a high resolution window around the eyes to detect the glaze direction reliably. We have the same problem for detecting speech acts. A balance between resolution (i.e. precision in visual detection) and processing time (i.e. efficiency in the NEIMO experiment) is likely to be found.

One of the solution is to actively control the camera parameters.

4.3. Use of active control of camera parameters

4.4. Extensions to hand-tracking

4.5. Role of visual processing as a source of human-computer interaction modalities

5. Conclusion

References

[Ambone 92] G. Ambone and J. Coutaz.Projet NEIMO : Cahier des charges. IMAG-LGI, Univ J. Fourier, Grenoble,Feb. 1992.

[Ducret 92] V. Ducret, Projet NeimoVision : Rapport de stage, IMAG-LGI, Univ J. Fourier, Grenoble, Sep. 1992.

[Fraser 92] N. Fraser, N. Gilbert and C. McDermid: "The Value of Simulation Data", Third Conference on Applied Natural Language Processing, Trento, Italy, 31 march—3 April 1992.

[Polity 90] Y. Polity, J.-M. Francony, R. Palermiti, P. Falzon, S. Kazma: "Recueil de dialogues homme-machine en langue naturelle écrite", Les Cahiers du CRISS, n° 17, 1990.

[Salber 93] Daniel Salber & Joelle Coutaz: "Applying the Wizard of Oz Technique to the Study of Multimodal Systems", to be published in Proceedings of the East-West HCI Conference '93, Moscow, Russia, 3-6 August 1993.

[Turk 91] M. Turk and A. Pentland. “Eigenfaces for Recognition” *Journal ofCognitive Neuroscience*, 3(1):71-86, 1991.

[Whittaker 89] S. Whittaker and P. Stenton: "User Studies and the Design of Natural Language Systems", Fourth Conference of the European Chapter of the ACL, proceedings 291-8 1989.