VISUAL INTERPRETATION OF FACES IN THE NEIMO MULTI-MODAL 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, NEIMO-Vision, is responsible for the visual interpretation of human faces. NEIMO-Vision includes procedures 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 developed by Turk [Turk 91]. Demonstration programs based on face recognition have been implemented on Macintosh Quadras equipped 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 new interaction modes for human-computer interaction.

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 begins with a presentation of the NEIMO project. It then describes the use of real time visual interpretation of faces as a basis for new modes for human-computer communication. 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 technologies have made rapid progress in the last few years. Unfortunately, these techniques have not quite matured to the point where they can provide reliable modes for human-computer interaction. While the necessary computer power is rapidly decreasing in cost, technical and HCI problems remain concerning 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 acts so that callers believe that they 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. Automated analysis tools have been limited in scope and rarely

been integrated into the WOz platform from the start. The NEIMO system has been designed as a generic re-usable WOz platform 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 developed 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.

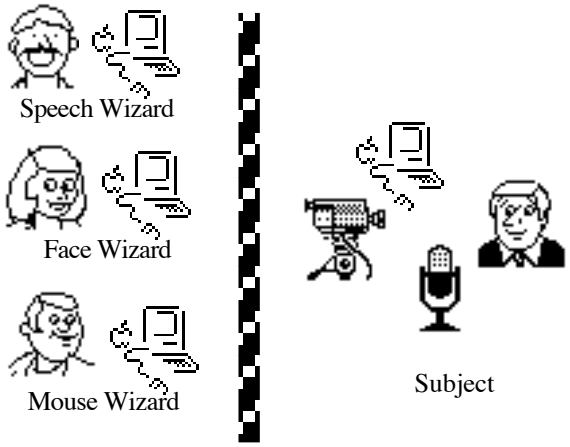

Fig. 1. An example configuration
for a Neimo experiment (adapted from [Salber 93])

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 Apple Macintosh 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.

The remainder of this paper is concerned with software tools for face interpretation that have been constructed within the NEIMO platform.

3 The NEIMO-Vision System

This section describes the techniques for face interpretation that are currently implemented in NEIMO-Vision. It describes the hardware environment of NEIMO-Vision and the presents techniques for locating a face, gray level normalisation, face recognition, determination of gaze direction, and the recognition of facial expressions.

3.1 Introduction

NEIMO-Vision performs three tasks within NEIMO.

1) The system displays the subject image on the face wizard workstation, permitting the wizard to efficiently observe the subject.
2) It performs automatic processing (identification, gaze direction,...) to reduce the wizard cognitive charge.
3) It provides images for data recording to make history files.

NEIMO-Vision 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. Images are acquired continuously by NEIMO-Vision and processed in real time. The system detects and locates a face and then normalizes the image gray levels.

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 128 by 128 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 a multidimensional hyperplane which is a sub-space of the 16 K dimensions. Turk has fancifully named this sub-space the “face-space”. Classes of patterns may be defined as occupying regions of face-space. The degree of resemblance of a face, or “face-ness”, of an image may be defined as the Euclidean distance of the image from this sub-space.

The Eigenface interpretation technique can be used for several purposes. Obvious examples include locating a face, classifying an image as containing a face or not, and recognizing a known person. It is also possible 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.

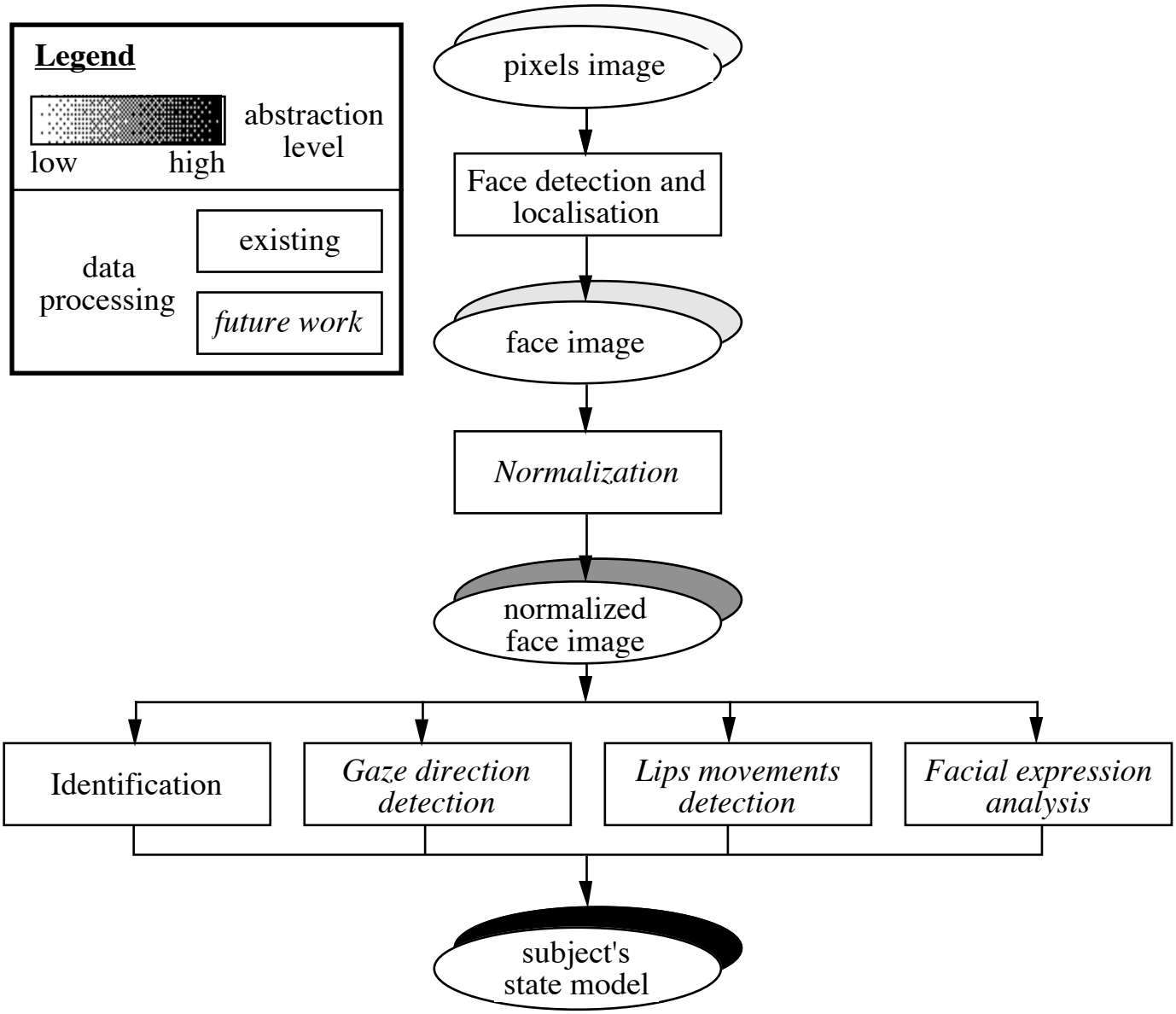

Figure 2. The NEIMO-Vision functionalities

3.2 The NEIMO-Vision hardware environment

The NEIMO-Vision 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 illustrates the NEIMO-Vision 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.

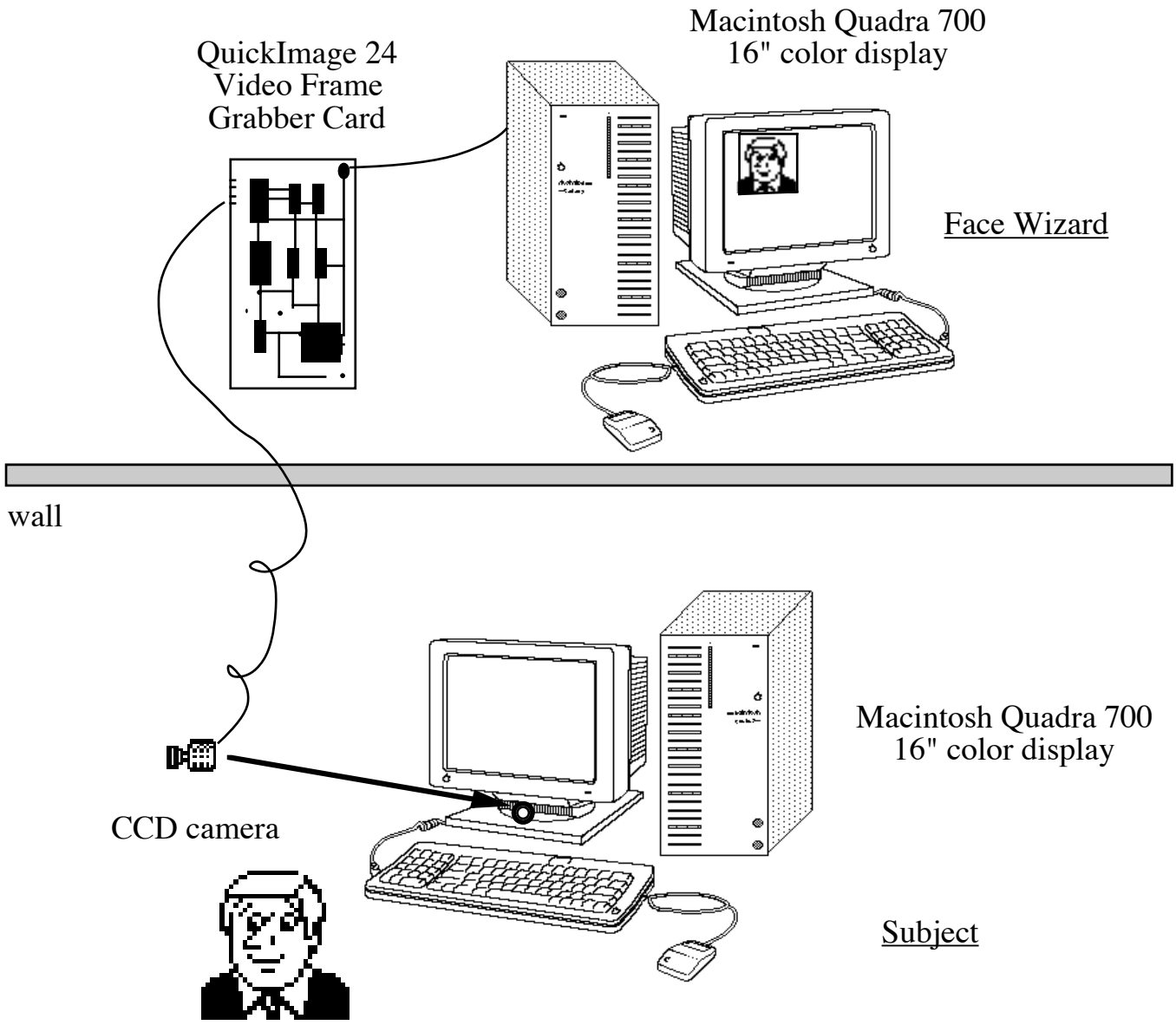

Figure 3. The NEIMO-Vision hardware configuration

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. If the system is slow, the wizard cannot compensate. Thus the wizard must have a very efficient way of observing the subject. The system must provide an effective foundation for efficiency and speed.

3.3 Locating a face

Two different techniques are being investigated 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, as well as using smaller subregions covering only the eyes and nose of the generic face image. Fast correlation techniques based on partial evaluation promise to bring this process to real time in the near future.

The second technique consists of exploiting the space-time pattern provided by blinking. Blinking is the quickest non-moving change that can be found in sequence of face images. A temporal derivative is computed by smoothing and differencing a dense 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. Unfortunately our digitizing hardware currently limits the use of this technique.

3.4 Normalization

We have found that the EigenFace approach is very sensitive to background information and gray-level intensity, as well as the position, size and orientation of the face. The normalisation step minimizes the effects of 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. The gray scale range of the resulting face image is the normalized using histogram equalisation.

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 demonstration system 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 of the image in the "face space" is 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 NEIMO-Vision demo (with a face database of three faces).

We are currently implementing a second demonstration with a very simple interface. This is a kind of login process without a password. The user is seated in front of a Macintosh Quadra equipped with a camera. If the face recognition process is able to locate the user in the face-space database, he is greeted by an audio-message and automatically logged in to his environment. If

the subject is not recognized, the system informs the user that he is not know and asks for his name and login information. This information may be registered by the subject, so that he is recognized in the future.

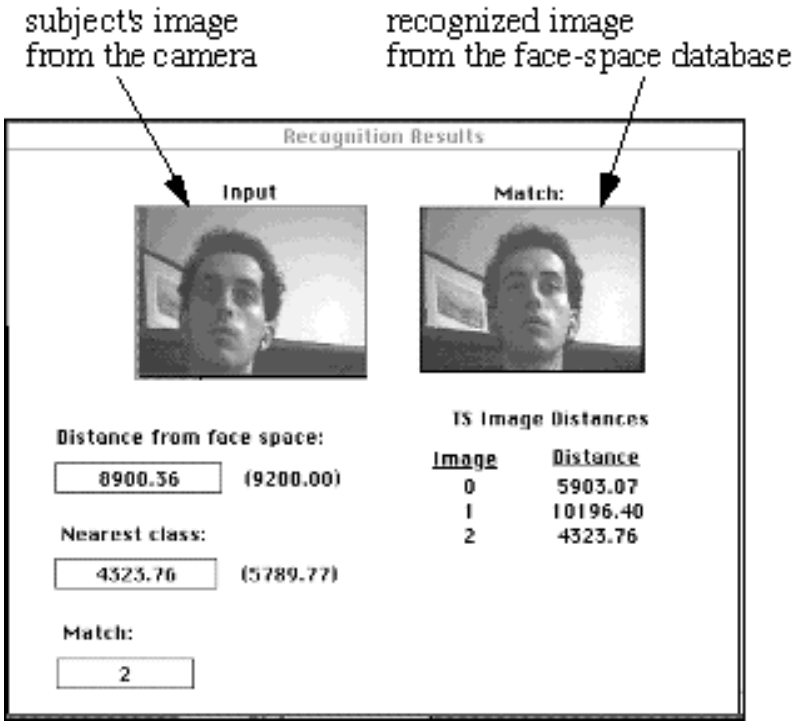

Fig. 4. The NEIMO-Vision demo: the recognition window

3.6 Measuring Gaze Direction

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 expressions

Facial expressions contains useful informations about the user's state of mind. Such information is useful in a Wizard of Oz experiment and may eventually also be used to dynamically adapt the man-machine interaction. 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 concludes with technical problems to be resolved for the use of face interpretation as a HCI communications mode. In particular we 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;

4.1 Requirements in computing power

NEIMO-Vision is designed to be used by a "Face Wizard" to observe the interaction of a human subject. The face wizard is not necessarily a computer expert. Thus the human interface of NEIMO-Vision must be relatively robust and user friendly, and must aid the wizard to efficiently perform his task. Unfortunately this raises the requirements in computational power.

The NEIMO system is implemented on Macintosh Quadra computers. Early experiments have been performed on the Macintosh Quadra 700 and 900 machines. Such machines are barely sufficient for 1 Hz acquisition and processing of small face images, and leave little processing power available for other applications. An additional factor of 2 in performance should provide a qualitative improvement in the interface.

The only specific hardware required for NEIMO-Vision is an image acquisition board. Image processing and interpretations are performed in software. With an additional investment in hardware many of these operations could be easily moved onto dedicated computer boards, freeing the central processor for data collection and analysis tasks. It is expected that that next generation of micro-processor technology, with capabilities in the 50 to 100 Million Floating point operations per second will be permit the image resolution to be substantially increased. Subsequent generations, with processing powers above 100 Million FLOPS will be more than inadequate to include such operations as a normal part of the human interface.

4.2 Trade-offs between resolution and processing speed

The cost of operations on EigenFaces is directly proportional to the number of pixels that are processed. Thus, one of the ways in which we have obtained real time performance with existing micro-processor technology has been to reduce size of the processed images until algorithms run at a reasonable speed. This raises the question of what resolution is required for each of the face interpretation processes. This issue is explored in this section.

Our video acquisition board will provide digitized image with a size of up to 768 by 576 (442,368) pixels. At the current time, we process images which are 108 by 144 (only 15,552!) pixels. This number was obtained by systematically reducing the size of the processed images until the processing time reached a desired maximum value of 1 frame per second.

Different face interpretation algorithms have different requirements in resolution. Our current efforts involve measuring the resolution requirements for each of these processes. Detection of the existence of a face seems to work reliably at relatively small images, down to sizes 32 by 32. Recognition of the user requires a slightly higher precision, in the range of 128 by 128 or 64 by 64. Detection of speech acts from lip movements requires that a processing window of about 32 by 64 pixels be placed over the lips. Making hypotheses about what has been said from lip reading is thought to also require a temporally sampling density on the order of video rates 25 frames per second.

The resolution required for eye tracking is proportional to the precision with which eye movements must be measured. A low resolution image (say 64 by 64) of the eye is sufficient for determining the region of the current screen window which is being watched by the user. Use of eye tracking to replace the mouse or track-ball will likely require substantially higher resolution, such as, covering each eye with a region on the order of the size of the precision of the screen. Since the user is constantly moving, such precisions are not available with a conventional static camera. Possibilities include using very high resolution cameras (a computationally expensive solution), mounting cameras directly on the head of the subject (a cumbersome solution), or equipping the cameras with automatically controlled zoom lens and pan and tilt axes. This last possibility would permit recent results from active vision to be applied to man-machine communications, and provide a good compromise between high resolution sampling and efficient computation.

4.3 Use of active control of camera parameters

Active vision is the use of controlled camera motion and control of processing to increase the robustness and decrease the processing speed in computer vision. It appears to us that techniques from active vision will be necessary in order to obtain robust and real time vision for man-machine communications.

A simple example of the need for active vision occurs because of the resolution requirements. As we have seen above, task such as lip motion detection and eye tracking require placing relatively large regions of the image over the eyes and mouth of the subject. For example, we require a resolution in which the image of the face is 256 pixels wide. Yet the subject is free to move his face side by side by a distance of roughly 4 times the width of his face. Thus obtaining a face width of 256 pixels would require images of 1024 pixels, a factor of 10 more than currently used. We have a choice of using hardware with 10 times the processing power (beyond our limited budget!), pinning the subjects head to prevent it moving (not ergonomically appealing!) or mounting the camera on a small pan-tilt platform. Lighting conditions can equally change. Automatic gain controls generally work on the entire image, and are thus poorly suited for

improving face images. We require a technique which will control the aperture so as to maintain a maximum of contrast on the part of the image that contains the face. Techniques for control of pan, tilt, aperture and focus have been developed [Crowley 93] and make possible an improvement in robustness and processing speed.

5 Conclusion

Vision represents a powerful new source of communication channels between man and machine. In order to properly design such channels, it is necessary to test such forms of communications. The NEIMO system is a multi-modal test bed for experiments in man machine communication. NEIMO Vision is a vision component which permits us to test the role and requirements for computer vision in the interface between man and machine.

NEIMO vision has shown us that coarse face recognition is possible with relatively inexpensive hardware using the EigenFaces techniques. The system is currently being used to explore the trade-offs necessary for face detection, eye-tracking, lip motion detection and estimating the users mood. Future plans include the addition of a visual interface for hand tracking. Such an interface could permit the definition of a small virtual world where the actions of human hands are directly perceived and interpreted.

References

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

[Crowley 93] J. Crowley and H. Christensen, Vision as Process, Springer Verlag Basic Research Series, (to appear) 1993.

[Ducret 92] V. Ducret, Projet NEIMO-Vision : 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.