

# Gesture Analysis and Synthesis for Intelligent Tutoring Systems

<sup>1</sup>Ana Luisa Solís, <sup>2</sup>Homero Ríos, <sup>1</sup>Lourdes Guerrero, <sup>2</sup>Joaquín Peña, <sup>1</sup>Jorge Castro

<sup>1</sup>Departamento de Matemáticas, Facultad de Ciencias  
Universidad Nacional Autónoma de México

<sup>2</sup>Laboratorio Nacional de Informática Avanzada, A.C.  
Rébsamen 80, c.p. 91090, Xalapa, Veracruz, México

**Abstract.** This paper describes a new interface for intelligent tutoring systems based on recognition and synthesis of facial expressions and hand gestures. This interface senses the emotional state of the user, or his/her degree of attention, and communicates more naturally through face animation.

## 1. Introduction

This work describes a new interaction technique for intelligent tutoring systems based on the integration of results from computer graphics and computer vision. In the last few years this integration has shown important results and applications [Eisert, 98; Pentland, 96]. For example, Richard Szeliski described the use of image mosaics for virtual environments in 1996 and, in the following year for combining multiple images into a single panoramic image. H. Ohzu et al. described hyper-realistic communications for computer supported cooperative work.

Facial expression understanding is a good example of the rich middle ground between graphics and vision. Computer vision provides an excellent input device, particularly for the shapes and motions of complex changing shapes of faces when expressing emotions [Pentland, 96; Parke et al, 96].

We have been studying how to analyze efficiently video sequences for capturing gestures and emotions. Relevant expressions and their interpretations may indeed vary depending upon the chosen type of application [Rios et al., 98].

In this work, relevant expressions are assigned with a tutorial system when it is important to know the user interest on the information that is displayed or when she/he is interacting in a virtual environment.

Facial expression recognition is useful for adapting interactive feedback in a tutoring system based on the student's level of interest. The type of expressions associated with these applications are: degree of interest, degree of doubt of the information presented, boredom among other expression, or to assess the time of interest or lack of interest presented by an application.

It is also possible to communicate freely without mouse or data glove, by using hand gestures, since these are a natural part of human dialogue [Maggioni, C. and Kammerer, B., 98; Pentland, 98]. In this paper we explore hand representation and detection on digital images using trainable deformable models.

The work mentioned here also strives to capture the high resolution motion and appearance of an individual face. The goal is to use this information to animate and render synthetic faces and to have interaction with a tutorial system.

## 2. Analysis and interpretation of facial expressions

The communicative power of faces makes it a focus of attention during social interaction. Facial expressions and the related changes in facial patterns inform us on the emotional state of people and help to regulate both social interactions and spoken conversation. To fully understand the subtlety and expressive power of the face, considering the complexity of the movements involved, one must study face perception and related information processing.

For this reason, face perception and face processing have become major topics of research by cognitive scientists, sociologists and more recently by researchers in computer vision and computer graphics.

The automation of human face processing by computer will be a significant step towards developing an effective human-machine interface. We must consider the ways in which systems with this ability understand facial gestures (analysis), and the means of automating this interpretation and/or production (synthesis) to enhance human-computer interaction.

### 2.1 Facial Expressions as Emotional Signals

Facial expressions convey emotions and there are ongoing debates about their discreteness and universality. One of the most documented research efforts led by Ekman has permitted to identify six basic universal emotions: fear, anger, surprise, disgust, happiness, and sadness [Ekman and Friesen, 75]. Others like Russel prefer to think that facial expressions and labels are probably associated, but the association may vary with culture [Russel, 94].

### 2.2 Face analysis

The analysis of faces by computer is difficult since there are several factors that influence the shape and appearance of faces on images. Some of these factors are illumination, viewpoint, color, facial hair and the variability of faces. In addition, we still do not know the exact mechanisms used by humans for face processing. For instance is face processing a holistic or feature analysis process?. The brain itself has specialized structures like IT cells on the visual areas to handle face detection [Bruce & Green, 89].

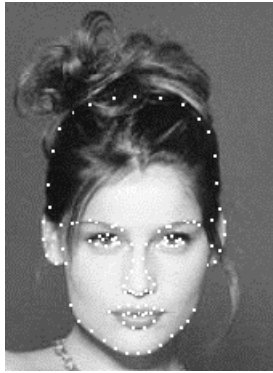
The problem of face analysis can be divided in face detection and face recognition. In the first case the goal is just to locate the general position of faces on images. On the latter, the purpose is to recognize faces using extracted features. This recognition can take place on the following conditions [Chellapa et al., 95]: a) static images, b) range images, and c) video sequences.

Since faces are non-rigid objects the best way to model them is through the use of deformable models which should have enough parameters to accommodate most of the variations in shape. Active contours have been used to detect and track facial features like head contour, lips, eyebrows and eyes [Lam & Yang, 96; Terzopoulos & Szeliski, 92]. The problem is that since “snakes” can adapt any shape, sometimes they take nonvalid shapes. One solution is the use of trainable deformable models from examples, like the point distribution model proposed by Cootes [Cootes et al. 92]. This model has been improved to learn grey and color variations, and the search of the model on images can be optimized as shown by active appearance models [Cootes et al. 98].

To train the deformable model for face detection we used 23 faces and 111 landmarks per face (Figure 1). After aligning the training faces to reduce the effect of translation, rotation and scaling, principal component analysis is used to obtain the main modes of variation [Cootes et al. 92]. Six eigenvectors account for 94% of shape variation (Figure 3). Any shape in the training set can be approximated using the mean shape and a weighted sum of the first  $t$  eigenvectors (ordered from the most significant eigenvalue to the least significant) as follows:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}$$

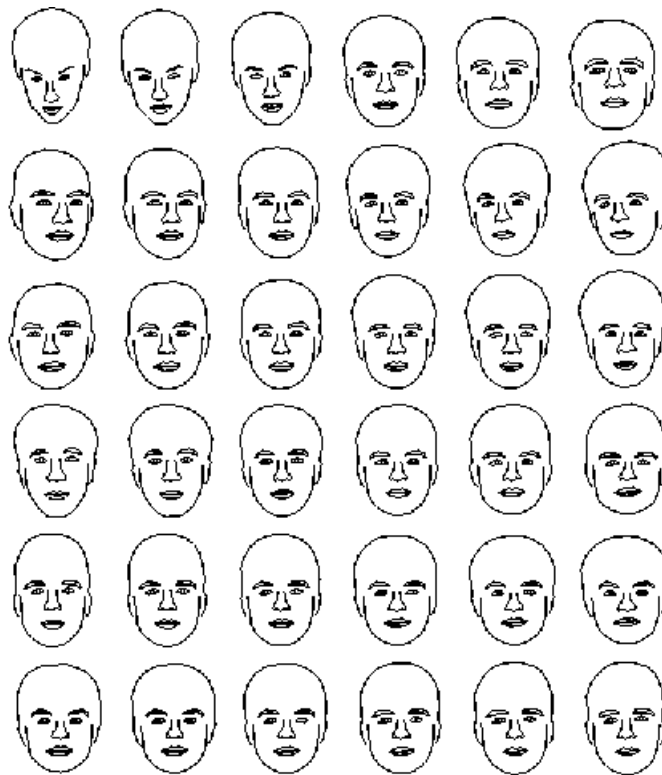
where  $\mathbf{x}$  is a shape in the training set,  $\bar{\mathbf{x}}$  is the mean,  $\mathbf{P}=(p_1, \dots, p_t)$  is the matrix of the first  $t$  eigenvectors, and  $\mathbf{b}=(b_1, b_2, \dots, b_t)^T$  is a vector of weights for each eigenvector (Figure 2 and 3)



**Figure 1.** Landmarks used for face representation.



**Figure 2.** Mean shape



**Figure 3.** From top to bottom the main modes of variation that account for 94% of shape deformation. Each row represents the mean face plus an eigenvector multiplied by a weight factor. From left to right the weight factor varies in the range  $[-3d_i, 3d_i]$ , where  $d_i$  is the square root of the  $i$ 'th eigenvalue.

We are working on the implementation of a search technique to locate the deformable model on images. From the extracted representation we apply an expression recognition algorithm and proceed to face animation

### 3. Facial Expression Recognition

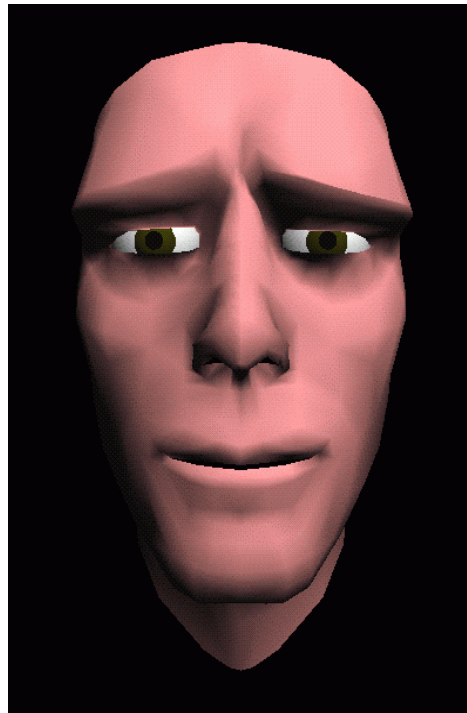
#### 3.1 Emotions and Facial Motion

Facial expression recognition can be accomplished from static images but it is also true that facial motion provides additional clues. For instance, Pentland's group use "motion energy maps" to characterize basic emotions. These maps show the pattern of motion of selected features (lips, eyebrows, etc.) from a resting state to a specific emotion [Schwartz, 95]. A general probabilistic tool used by several authors to recognize facial expressions, gestures (sign language, lip reading) and speech are Hidden Markov Models since they are well suited for time series modeling [Oliver et al., 97; Starner et. al 98; Campbell et al. 96]. Other possible technique are Bayesian Networks [Sucar and Gillies, 94].

#### 4. Face animation

Facial animation typically involves execution of a sequence of a set of basic facial actions. We use action units (AU) of the Facial Action Coding System (FACS) as atomic action units and as basic facial motion parameters.

Each AU has a corresponding set of visible movements of different parts of the face resulting from muscle contraction. Muscular activity is simulated using a parameterized facial muscle process. We can define atomic action units similar to AU and definite expressions and phonemes (Figure 4). These can be used for defining emotion and sentences for speech.



**Figure 4.** A facial expression of doubt.

For the driven facial animation the input parameters to the facial animation are the AUs. The facial expression recognition module provides a two way dialog and requires that the person on the other side to have a camera (Figure 5).



**Figure 5.** Facial Communication

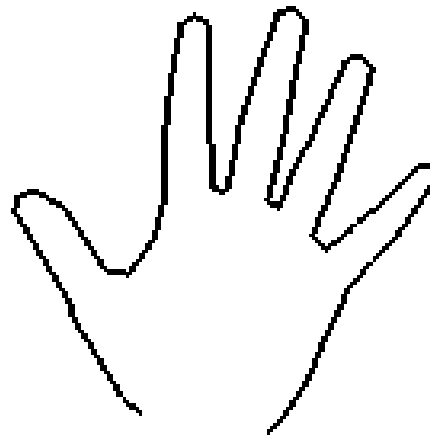
This module would perform much more task than just merely copying other's facial expression. We are developing a prototype for a virtual dialog system where a virtual actor communicates with a real person by the analysis of facial expression.

## **5. Hand analysis**

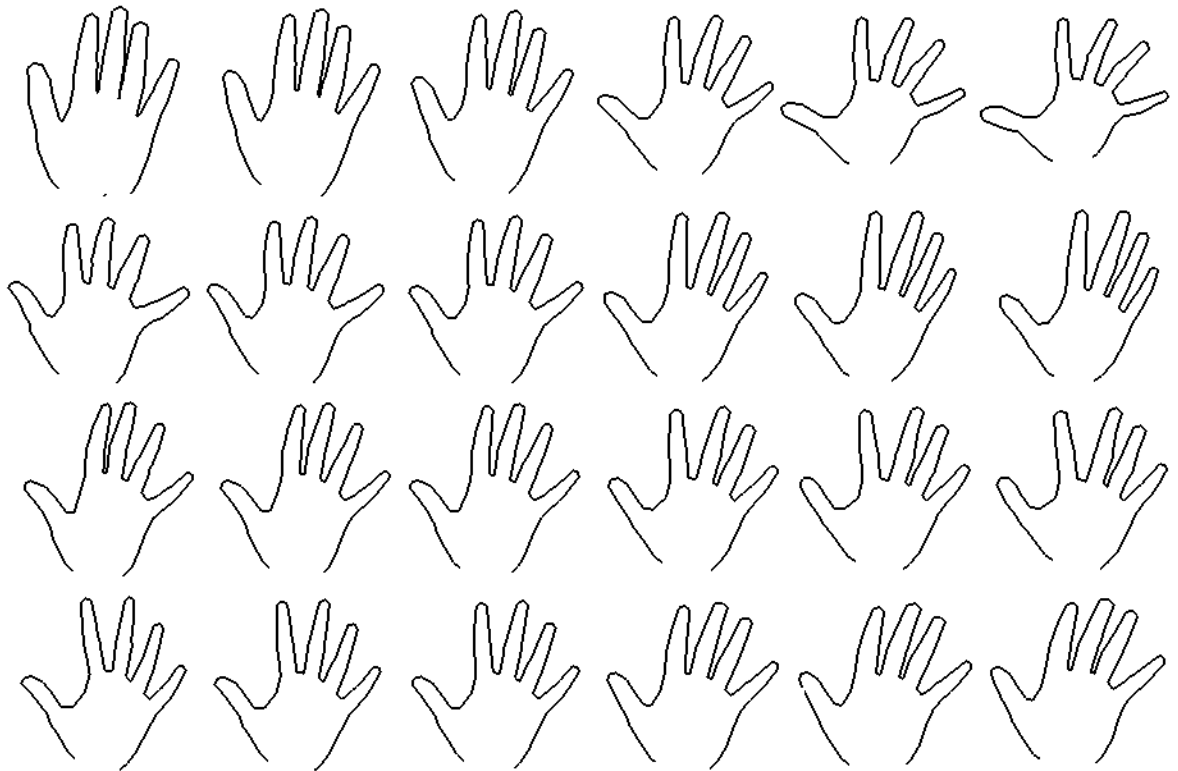
In our experiments we used 15 hands and 74 landmarks per hand to train a point distribution model [Cootes et al. 92]. Nine eigenvectors account for 96.5% of shape variation (Fig. 6-8).



**Figure 6.** Landmarks used for hand representation.



**Figure 7.** Average hand



**Figure 8.** From top to bottom the main modes of variation that account for 80% of hand deformation. Each row represents the mean hand plus an eigenvector multiplied by a weight factor. From left to right the weight factor varies in the range  $[-3d_i, 3d_i]$ , where  $d_i$  is the square root of the  $i$ 'th eigenvalue.

## 6. Integration of a Intelligent Tutoring Systems (ITS) in a Virtual Environment.

### 6.1 Face expression input for the ITS

The system is well suited for developing an autonomous intelligent tutor who may communicate and exchange a dialog with a real person through expression recognition.

We want to propose a simple architecture. Figure 9 shows the global structure of the system. The input to the analyzer is the facial expression recognition module. The result of the analysis can provide the expressions

recognition of the user. The virtual actor responds to real person, a data base is used with content words with facial expression states. These are defined in terms of constituent phonemes and facial expressions.

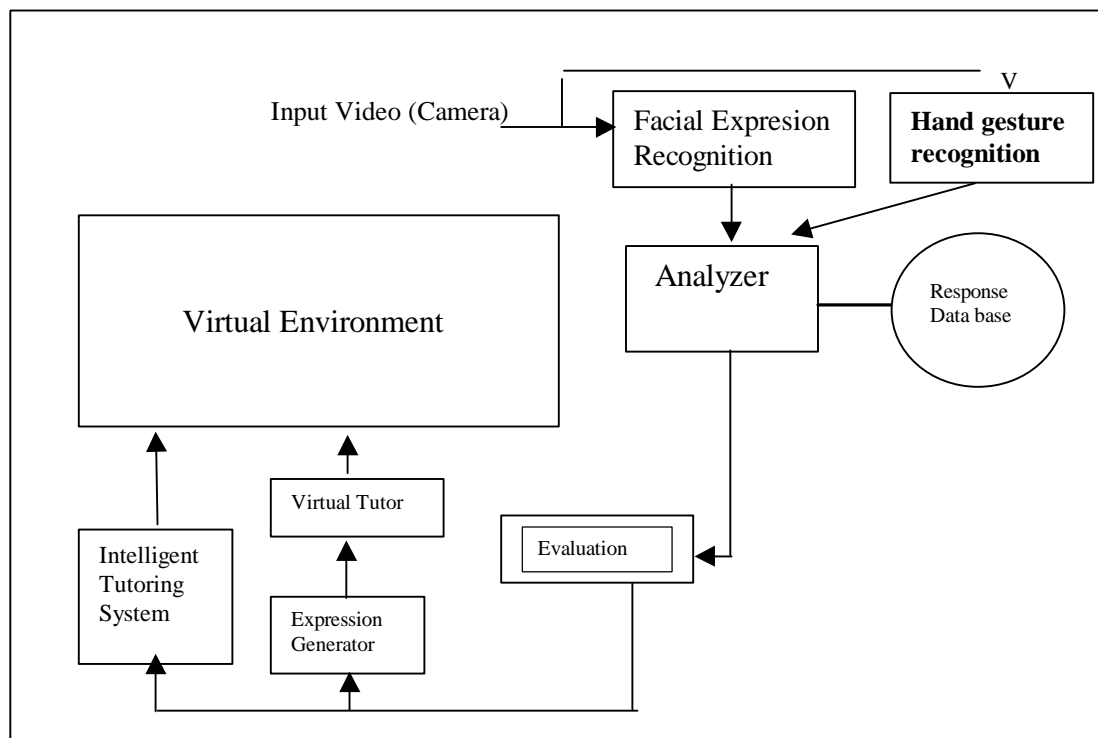
The series of timed phonemes and expressions are then compiled and decomposed into the basic action units to produce the visual changes to be done on the virtual tutor's face. The words to be spoken are transmitted to the vocal synthesizer module.

## 6.2 Interacting with the ITS through face expressions

Input from facial expression may be used to interact with the application in the virtual environment.. Relevant expressions are assigned with a virtual tutorial system when it is important to know the user interest on the information that is displayed or when the user is interacting in a virtual environment.

## 6.3 Interacting with hand movements

Some commands can be provided through hand movements allowing the user a more natural interaction. It is possible to define a special gesture language to define the meaning of hand movements or use sign language [Starner et al. 98].



**Figure 9.** Global structure of the System

## Conclusion

In this paper we have described a human-computer interface based on face analysis and synthesis, and hand gesture understanding, that enhances the communicative power of intelligent tutoring systems. The analysis allows facial expression recognition, while synthesis renders a realistic virtual tutor which could be complemented with synthetic speech.

## Acknowledgments

This work has been funded by the Mexican National Council of Science and Technology (CONACYT) as project ref. C098-A and C100-A, "Gesture recognition interfaces and intelligent agents for virtual environments". We also gratefully acknowledge the support of the National Science Foundation (NSF) through Grant 9710940.

## References

- Bruce, V. and Green, P. (1989), *Visual Perception: Physiology, Psychology and Ecology*. Lawrence Erlbaum Associates, London.
- Campbell, L.W., Becker, D.A., Azarbayejani, A., Bobick, A., and Pentland, A. (1996). Invariant features for 3-D gesture recognition. *MIT Media Laboratory Perceptual Computing Section*, Technical Report no. 379.
- Cassell, J., Pelachaud, C. Badler, N. et al. (1994) Animated Conversation: Rule-based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. *Computer Graphics Proceedings*, pp.413-420.
- Chellapa, R., Wilson, C.L. and Sirohey, S. (1995). Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, Vol. 83, No.5, pp. 705-740
- Cohen, M. and Massaro, D. (1993) Modeling Coarticulation in Synthetic Visual Speech, *Models and Techniques in Computer Animation*, Springer Verlag, pp. 139-156.
- Cootes, T.F., et al. (1992). Training models of shape from sets of examples. *Proceedings of the British Machine Vision Conference*.
- Cootes, T.F., Edwards, G.J. and Taylor, C.J. (1998). *Proceedings of the European Conference on Computer Vision*, Burkhardt, H. and Neumann, B. (Eds.), Vol. 2, pp. 484-498, Springer-Verlag.
- Eisert, P. and Girod, B. (1998). Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics and Applications*, Vol. 18, No. 5, pp. 70-78.
- Ekman, P. and Friesen, W.V. (1975). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*, Englewood Cliffs, New Jersey; Prentice Hall, Inc.
- Ekman, P. and Friesen, W.V. (1978), *Manual for the Facial Action Coding System*. Consulting Psychologists Press, Inc. Palo Alto, CA.
- Lam, K.M. and Yang, H. (1996). "Locating and extracting the eye in human face images". *Pattern Recognition*, Vol. 29, No. 5, pp. 771-779.
- Ohzu, H. and Habara, K. (1996). Behind the scenes of virtual reality: vision and motion. *Proceedings of the IEEE*, Vol. 84, No. 5, pp. 782-798.
- Oliver, N., Pentland, A. and Berard, F. (1997). Lafter: lips and face real time tracker. *MIT, Media Laboratory Perceptual Computing Section*, Technical report no. 396.
- Maggioni, C. and Kammerer, B. (1998). GestureComputer - History, Design and Applications. *In Computer Vision for Human-Machine Interaction*, Cipolla, R. and Pentland, A. (Eds.), pp. 23-52, Cambridge University Press.
- Parke, F.I. and Waters, K. (1996). *Computer Facial Animation*. A K Peters.
- Pentland, A.P. (1996). Smart rooms. *Scientific American*. April 1996, pp. 54-62.
- Pentland, A.P. (1998). Smart rooms: Machine understanding of human behavior. *In Computer Vision for Human-Machine Interaction*, Cipolla, R. and Pentland, A. (Eds.), pp. 3-22, Cambridge University Press.
- Rios, H.V. and Peña, J. (1998). Computer Vision interaction for Virtual Reality, In *Progress in Artificial Intelligence*, Helder Coelho (Ed.), *Lecture Notes in Artificial Intelligence, Subseries of Lecture Notes in Computer Science*, No. 1484, pp. 262-273. Springer-Verlag.
- Russel, J. (1994). Is There Universal Recognition of Emotion From Facial Expression?. *A Review of Cross-Cultural Studies Psychological Bulletin* 115(1) 102-141.
- Schwartz, E.I. (1995). A face of one's own. *Discover the world of Science*. Vol. 16, No. 2, pp. 78-87.
- Starner, T., Weaver, J., Pentland, A. (1998). Real-time American sign language recognition using desk and wearable computer based video. *MIT, Media Laboratory Perceptual Computing Section*, Technical report no. 466.
- Sucar, L.E. and Gillies, D.F. (1994). Probabilistic reasoning in high-level vision. *Image and Vision Computing*, Vol. 12, No. 1, pp.42-60.
- Terzopoulos, D. And Szeliski, R. (1992). Tracking with Kalman Snakes. In *Active Vision*, Blake, A. and Yuille, A. (Eds.), MIT Press.

Waters, K. (1987). A Muscle Model for Animating Three Dimensional Facial Expression. *Computer Graphics Proceedings* Vol. 21, No 4 , pp.17-23.

Waters, K. and Levergood, T.(1994) An Automatic Lip-Synchronization Algorithm for Synthetic Faces. *Proceedings of Multimedia 94, ACM*, pp149-156.