REAL TIME FACIAL INTERACTION

Igor Sunday Pandzic, Prem Kalra, Nadia Magnenat Thalmann

MIRALab, CUI, University of Geneva
24, rue du General Dufour,
1211 Geneva, Switzreland

ABSTRACT

Human interface for computer graphics systems is taking an evolutionary turn where it would involve multimodal approach. It is now moving from typical keyboard operation to more natural modes of interaction using visual, audio and gestural means. This paper discusses real time interaction using visual input from a human face. It describes the underlying approach for recognizing and analyzing the facial movements of a real performance. The output in the form of parameters describing the facial expressions can then be used to drive one or more applications running on the same or on a remote computer. This enables the user to control the graphics system by means of facial expressions. This is being used primarily as a part of a real-time facial animation system, where the synthetic actor reproduces the animator's expression. This offers interesting possibilities for teleconferencing as the requirements on the network bandwidth are low (about 7 Kbit/s). Experiments are also done using facial movements as means of controlling a walkthrough or performing simple object manipulation.

Keywords: facial analysis, interaction, facial animation, facial communication

INTRODUCTION

Interaction in graphical systems is becoming more and more multimodal[1]. For many applications it is unnatural to use the conventional mode of 2D widget-mouse or keyboard interaction. In addition, for disabled persons who are unable to use the hand input devices, we need to explore means for them to have interactive controls of the graphical systems. In this paper we discuss real time interaction through facial input. The paper presents our method of extracting dynamic facial movements which can be hooked as controls for the desired application. For example, in performance driven facial animation, the method enables recognition of facial expressions of a real person which are appropriately mapped as controlling parameters to simulate facial expressions of a synthetic actor in real time. In other applications, the extracted parameters can provide real time estimates of positions and orientations in a virtual scene. The system requires a video camera (CCD) input and extracts motion parameters through a small set of visually tracked feature points.

Recognition of facial expressions is a very complex and interesting subject. However, there have been numerous research efforts in this area. Mase and Pentland[2] apply optical flow and principal direction analysis for lip reading. Terzopoulos and Waters[3] reported on techniques using deformable curves for estimating face muscle contraction parameters from video sequences. Waters and Terzopoulos[4] modeled and animated faces using scanned data obtained from a radial laser scanner and used muscle contraction parameters estimated from video sequences. Saji et al.[5] introduced a new method called "Lighting Switch Photometry" to extract 3D shapes from the moving face. Kato et al.[6] use isodensity maps for the description and the synthesis of facial expressions. Most of these techniques do not perform the information extraction in real time. There have been some implementations of the facial expression recognition using colored markers painted on the face and/or lipstick[7,8,9]. However, the use of markers is not practical and the methods are

needed to perform recognition without them. In another approach Azarbayejani et al.[10] use extended Kalman filter formulation to recover motion parameters of an object. However, the motion parameters include only head position and orientation. Li et al.[11] use the Candid model for 3D motion estimation for model based image coding. The size of the geometric model is limited to only 100 triangles which is rather low for characterizing the shape of a particular model.

Magnenat-Thalmann et al.[12] propose a real time recognition method based on "snakes" as introduced by Terzopoulos and Waters[3]. The main drawback of this approach, is that the method relies on the information from the previous frame in order to extract the next one. This can lead to the accumulation of error and the "snake" may completely loose the contour it is supposed to follow. To improve the robustness we adopt a different approach, where each frame can be processed independently from the previous one.

First we describe our recognition method. Then facial animation system using real performance as input is briefly discussed. Some other applications of controlling movements and manipulation are also outlined. A short note on implementation provides the communication protocol between the recognition module and the application. Finally, we discuss the current state and possible future enhancement to the system followed by the concluding remarks.

RECOGNITION METHOD

Accurate recognition and analysis of facial expressions from video sequence requires detailed measurements of facial features. Currently, it is computationally expensive to perform these measurements precisely. As our primary concern has been to extract the

features in real time, we have focused our attention on recognition and analysis of only a few facial features.

The recognition method relies on the "soft mask", which is a set of points adjusted interactively by the user on the image of the face as shown in Figure 1. Using the mask, various characteristic measures of the face are calculated at the time of initialization. Color samples of the skin, background, hair etc., are also registered. Recognition of the facial features is primarily based on color sample identification and edge detection. Based on the characteristics of human face, variations of these methods are used in order to find the optimal adaptation for the particular case of each facial feature. Special care is taken to make the recognition of one frame independent from the recognition of the previous one in order to avoid the accumulation of error. The data extracted from the previous frame is used only for the features that are relatively easy to track (e.g. the neck edges), making the risk of error accumulation low. A reliability test is performed and the data is reinitialized if necessary. This makes the recognition very robust. The method enables extraction of the following facial features.

• vertical head rotation (nod)

• horizontal head rotation (turn)

• head inclination (roll)

• eyes aperture

• horizontal position of the iris

• eyebrow elevation

• horizontal distance between the eyebrows (eyebrow squeezing)

• jaw rotation

• mouth aperture

• mouth stretch/squeeze

The following sections describe the initialization of the system and the details of the recognition method for each facial feature, as well as the verification of the extracted data. The recognition of the features and the data verification are presented in the order of execution, as also shown schematically in Figure 2.

Initialization

Initialization is done on a still image of the face grabbed with a neutral expression. The soft mask is placed over the image as shown in Figure 1. The points of the mask are interactively adjusted to the characteristic features of the face, such as mouth, eyes, eyebrows etc. These points determine the measures of the face with neutral expression and provide color samples of the background and the facial features. The process of setting the mask is straightforward and usually takes less than half a minute.

Head tracking

First step is to find the edges of the neck (blue circles in Figure 3, points N1 and N2 in Figure 4). During the initialization, color samples are taken at the points 1, 2 and 3 of the mask (Figure 1). Points 1 and 3 are aligned over background and skin respectively, and point 2 over the hair falling on the side of the face, if any. During recognition, a sample taken from the analyzed point of the image is compared with those three samples and identified as one of them. As each color sample consists of three values (red, green and blue), it can be regarded as a point in a three dimensional RGB space. The distance in this space between the sample being analyzed and each stored sample is calculated. The closest one is chosen to categorize the point. This method of sample identification works fine in the areas where the number of possible different colors is small and where there is sufficient difference between the colors. Next step is to find the hairline (marked with the red circle in Figure 3, point M in Figure 4). The samples of the hair and skin color are taken and edge between the two is detected. The horizontal position of the starting point is

halfway between the neck edges, and the vertical position is taken from the previous frame. At a fixed distance below the hairline the edges of the hair seen on the sides of the forehead are detected (marked with green and yellow circles in Figure 3, points L1, L2, R1, R2 in Figure 4) using the above described sample identification method.

Using the information from points L1, L2, R1, R2, N1, N2, and M (Figure 4) we estimate the head orientation for different movements. For example:

head turn = f(L1,L2,R1,R2)

head nod = f(M)

head roll = f(L1,L2,R1,R2,N1,N2)

Jaw rotation

To extract the rotation of the jaw the position of the chin has to be found. We exploit the fact that the chin casts a shadow on the neck, which gives a sharp color change on the point of the chin. Once again the sample identification is used to track this edge.

Data verification

At this point the data extracted so far is checked against the measurements of the face made during initialization. If serious discrepancies are observed the recognition of the frame is interrupted, the warning signal is issued and the data is reinitialized in order to recognize the next frame correctly. This may happen if the user partially or totally leaves the camera field of view or if he takes such a position that the recognition cannot proceed.

Eyebrows

The starting points for the eyebrow detection are above each eyebrow, sufficiently high that the eyebrows cannot be raised above them. They are adjusted interactively during

initialization (points marked 4 in Figure 1) and kept at fixed position with respect to the center of the hairline. Also during initialization, the color samples of the skin and the eyebrows are taken. The search proceeds downwards from the starting point until the color is identified as eyebrow. To avoid wrinkles on the forehead being confused with the eyebrows, the search is continued downward after a potential eyebrow is found. If that is the real eyebrow (i.e. not just a wrinkle), the next sample resembling the eyebrow will be in the eye region, i.e. too low. The points on eyebrows are marked with magenta circles in Figure 3. The relative position of each eyebrow with respect to the hairline is compared with the eyebrow position in the neutral face to determine the eyebrow-raise. The eyebrow squeeze is calculated from the distance between the left and right eyebrow.

Eyes

During initialization, a rectangle (marked as 5 in Figure 1) is placed over each eye and it's position relative to the center of the hairline is measured. During recognition the rectangles (outlined in magenta in Figure 3) are fixed with respect to the center of the hairline and stay around the eyes when the user moves.

To determine the aperture of the eye we exploit the fact that the sides of the iris make strong vertical edges in the eye region. The points lying on vertical edges are found as the local minima of a simplified color intensity gradient function. The edges are found by searching for the groups of such points connected vertically. The largest vertical edge is a side of the iris. To find the aperture of the eye we search for the eyelid edges upwards and downwards from the extremes of the vertical edge found earlier. In Figure 3, the aperture of the eyes is marked with green lines, the vertical yellow line marking the side of the iris.

To determine the horizontal position of the iris we find the distance between the iris and the edge of the eye using simple edge detection. This distance is marked with a horizontal yellow line.

Nose and Mouth

The distance between the nose and the hairline is measured during initialization. Using this value the approximate position of the nose is determined. Edge detection is used for locating the nose. A point where the vertical color intensity gradient is above a certain threshold, is considered to lie on a horizontal edge. A 3x3 pixels gradient operator is used. The threshold value is determined during initialization by exploring the gradient values in the area. The blue line in the Figure 3 connects the edge points and the orange circle marks the nose position.

For acquisition in the mouth region we search for a horizontal edge downward the nose point to find a point on the upper lip. At the same horizontal position the search is performed from the chin in upward direction to find a point on the lower lip. This process is repeated on the next horizontal position n pixels to the right, n being 1/10 of the mouth width. The search starts in the proximity of the found vertical positions. We continue to move to the right, each time storing in memory the points on the lips edges found, until the corner of the lips is passed. This is detected when no edge is found in the area. The corner of the lips is then tracked more precisely by decreasing the step to n/2, n/4, n/8,...,1. The same process is repeated for the left side. All the points found together thus form the mouth curve. It is shown in green in Figure 3. However, due to shadows, wrinkles, beard or insufficient lip-skin color contrast, the curve is not very precise. Therefore the average height of the points in the middle third of the curve is taken for the vertical position of the lip. The bounding rectangle of the mouth is outlined in magenta. This rectangle provides

measures for the nose and chin heights, relative vertical positions of both the lips, and squeeze/stretch of the mouth etc.

FACIAL ANIMATION

Facial animation, as any other animation, typically involves execution of a sequence of a set of basic facial actions. We use what we call a Minimum Perceptible Action (MPA)[13] as a basic facial motion parameter. Each MPA has a corresponding set of visible movements of different parts of the face resulting from muscle contraction. Muscular activity is simulated using rational free form deformations[14]. MPAs also include actions like head turning and nodding. An MPA can be considered as an atomic action unit similar to AU (Action Unit) of FACS (Facial Action Coding System)[15], execution of which results in a visible and perceptible change of a face. We can aggregate a set of MPAs and define expressions and phonemes. Further these can be used for defining emotion and sentences for speech. Animation at the lowest level, however, is specified as a sequence of MPAs with their respective intensities and time of occurrence. The discrete action units defined in terms of MPAs can be used as fundamental building blocks or reference units for the development of a parametric facial process. Development of the basic motion actions is non specific to a facial topology and provides a general approach for the modeling and animation of the primary facial expressions. In our facial model the skin surface of the face is considered as a polygonal mesh. It contains 2500-3000 polygons to represent the shape of the model. Hence, the model considered has sufficient complexity and irregularity to represent a virtual face, and is not merely represented as crude mask as considered in many other systems.

For the real time performance driven facial animation the input parameters to the facial animation are the MPAs. These MPAs have normalized intensities between 0 and 1

or -1 and 1. The analysis of the recognition module is mapped appropriately to these MPAs. In most cases the mapping is straightforward. Due to the speed constraint we have concentrated on only few parameters for the motion. This reduces the degrees of freedom for the animation. However, we believe that complete range of facial motion is practically not present in any particular sequence of animation. To mimic the motion of a real performance only a set of parameters is used. Figure 5 shows some frames of the real time facial animation.

With the input from real performance we are able to reproduce individual particular feature on the synthetic actor's face (e.g. raising the eyebrows, opening the mouth etc.) in real time. However, reproducing these features together may not faithfully reproduce the overall facial emotion (e.g. smile, surprise etc.). In order to achieve this a better interpreting/analyzing layer between recognition and simulation may be included. Use of a real performance to animate a synthetic face is one kind of input accessory used for our multimodal animation system. The system can basically capture the initial template of animation from real performance with accurate temporal characteristics. This motion template then can be modified, enhanced and complemented as per the need by other accessories for the production of final animation.

OTHER APPLICATIONS

We are currently experimenting with some other applications, like walkthrough and simple object manipulation. We also discuss some potential applications that may be included in the future.

Walkthrough or navigation

Input from facial movement may be used for controlling the movement in the virtual environment. We have experimented controlling the virtual camera by the appropriate rotations of the head. At the same time user can perform a move forward in the direction of his/her view by opening the mouth. While it is very natural to turn the head in the direction we want to look at, it is relatively difficult to coordinate the head rotation around all the three axis at the same time. Consequently, we have reduced the complexity of the controls which is more natural and easy to learn. Figure 6 shows the user "walking through" a virtual environment using the face to control the movement.

Simple object manipulation

In a 3D object environment the user can control the rotation of the object around all the three axes by appropriate movements of the head. At the same time user can stretch or squash the object horizontally by stretching or squeezing the mouth, and vertically by raising or lowering the eyebrows. These controls are quite natural for the user.

Visual communication

With some extensions, our real time facial animation system can be used as a communication system. Since the recognition of the facial expression is done on the host machine, and the rendering of the synthetic actor on the remote machine, this already provides one way communication. For a two ways dialog it would be necessary for the person on the other side to have a camera and the recognition module. The bandwidth requirements on the network are about 7 Kbit/s in one way, which is not prohibiting.

High level communication with the virtual actors

The system is well suited for developing an autonomous virtual actor who may communicate and exchange the dialog with another virtual actor. This would require an

intelligent and knowledge based interpreter for the communication process. This interpreter would perform much more task than just merely copying other's facial expression. This type of communication is more futuristic and requires further research in the domain. However, the present system offers its feasibility.

Aid for disabled persons

Our experiments with the simple applications like walkthrough and object manipulation show that it is possible to control the computer by means of facial movements. Potentially, this could aid the disabled persons, enabling them to control an application by whatever limited movements they can do. Dave Warner[16] uses the bio-electric signals from eye, muscle and brain activity as input and thus enables the severely disabled persons to control their environment by means of limited movements. He reports several successful applications. We would like to explore a similar path, using the facial expression recognition as input.

IMPLEMENTATION

In our implementation, the communication protocol is designed in such a way that one or more applications can easily connect to the recognition module. The connections are schematically presented in Figure 7. The applications can run on the same machine and/or on a remote machine connected through a standard network supporting TCP/IP. The applications running on the local machine connect through shared memory. On the remote machine there is a communication process whose only task is to read the data coming from the recognition module over the network. Applications get the data from the communication process through shared memory. This type of asynchronous communication enables the application to use the data at it's own speed. This is important

when the application is slower than the recognition module -- otherwise there would be accumulating delays. The communication process allows more applications to use the same data from the shared memory.

We can use any Silicon Graphics workstation for the motion capture of the facial features with Live Video Digitizer facility. For better quality we use a professional CCD camera. The recognition speed is 10 frames per second on a SGI Indy workstation which we consider enough for all the current applications. For the animation of a virtual actor we use a faster SGI workstation (e.g. Reality Engine) to obtain real time display and match with the speed of the recognition module. The system is developed in C and uses Fifth Dimension Toolkit[17] for the interface.

DISCUSSION

We have undertaken extensive tests of the recognition system with various persons using it for real time facial animation, walkthrough and object manipulation. Our real time facial expression recognition system is capable of extracting adequate quantity of relevant information at a very satisfying speed (10 frames/s). The system is robust. If the user positions himself in such a way that the recognition cannot proceed (e.g. if he leaves the camera field of view or turns away from the camera) the system issues a warning sign, and an appropriate signal to the application(s).

The major drawback is that the recognition doesn't work equally well for all the people. In particular, the bald people cannot use our system. The users with pale blond hair and eyebrows may have problems. The people with irregular haircuts and hair falling on the forehead have to use some hairpins to straighten the hair. We are exploring better techniques to make the program more general.

For real time performance driven facial animation the number of parameters are limited by the features extracted from the recognition module. We intend to use some heuristics and rules to derive more information from the extracted features to improve the qualitative performance of the animation of the virtual actor. To add realism for the rendering we also intend to add texture information which will be captured from the real performance.

Manipulation of objects in 3D scene using facial input is found relatively difficult. This is due to the fact that we as users rarely keep our head in a real straight position with a neutral expression. Consequently, the controls are very often triggered unintentionally with noise and it is quite difficult to keep the object still. With the use of better filters for the extracted data we hope to improve the effectiveness.

CONCLUSION

This paper presents a tool for man-machine interaction which uses facial input. The recognition module in our system does not use any special markers or make-up and does not need "training" the system by executing the entire set of expressions. The system is adequately fast, reasonably robust and adaptable for a new user with quick initialization. We have presented several working applications of our system and discussed some potential applications. We believe that facial interaction will have it's place among the interaction techniques in the near future.

REFERENCES

1 Kalra P *An Interactive Multimodal Facial Animation System*, PhD Thesis, Swiss Federal Institute of Technology, Lausanne, 1993

2 Masse K, Pentland A 'Automatic Lipreading by Computer' *Trans. Inst. Elec. Info. and Comm. Eng.* 1990, Vol. J73-D-II, No. 6, 796-803

3 Terzopoulos D, Waters K 'Techniques for Realistic Facial Modeling and Animation' *Proc. Computer Animation* 1991, Geneva, Switzerland, Springer - Verlag, Tokyo, 59 - 74

4 Waters K, Terzopoulos D 'Modeling and Animating Faces using Scanned Data' *Journal of Visualization and Computer Animation* 1991, Vol. 2, No. 4, 123-128

5 Saji H, Hioki H, Shinagawa Y, Yoshida K, Kunii T 'Extraction of 3D Shapes from the Moving Human Face using Lighting Switch Photometry' in Magnenat Thalmann N, Thalmann D (Eds) *Creating and Animating the Virtual World*, Springer-Verlag Tokyo 1992, 69-86

6 Kato M, So I, Hishinuma Y, Nakamura O, Minami T 'Description and Synthesis of Facial Expressions based on Isodensity Maps' in Tosiyasu L (Ed) *Visual Computing*, Springer - Verlag Tokyo 1992, 39-56

7 Magno Caldognetto E, Vagges K, Borghese N A, Ferrigno G 'Automatic Analysis of Lips and Jaw Kinematics in VCV Sequences' *Proceedings of Eurospeech 89 Conference* vol. 2, 453 - 456

8 Patterson E C, Litwinowich P C, Greene N 'Facial Animation by Spatial Mapping', *Proc. Computer Animation 91*, Magnenat Thalmann N, Thalmann D (Eds.), Springer-Verlag, 31 - 44

9 Kishino F, 'Virtual Space Teleconferencing System - Real Time Detection and Reproduction of Human Images" *Proc. Imagina 94*, 109 - 118

10 Azarbayejani A, Starner T, Horowitz B, Pentland A 'Visually Controlled Graphics' *IEEE Transaction on Pattern Analysis and Machine Intelligence*, June 1993, Vol. 15, No 6, 602-605.

11 Li Haibo, Roivainen P, Forchheimer R '3-D Motion Estimation in Model Based Facial Image Coding' *IEEE Transaction on Pattern Analysis and Machine Intelligence*, June 1993, Vol. 15, No 6, 545-555.

12 Magnenat Thalmann N, Cazedevals A, Thalmann D 'Modeling Facial Communication Between an Animator and a Synthetic Actor in Real Time' *Proc Modeling in Computer Graphics*, Genova, Italy, June 1993 (Eds Falcidieno B and Kunii L), 387-396.

13 Kalra P, Mangili A, Magnenat-Thalmann N, Thalmann D 'SMILE: A Multilayered Facial Animation System' *Proc IFIPS WG 5.10*, Japan (Ed Kunii Tosiyasu L), 189-198.

14 Kalra P, Mangili A, Magnenat-Thalmann N, Thalmann D 'Simulation of Muscle Actions using Rational Free Form Deformations' *Proc Eurographics '92, Computer Graphics Forum*, Vol. 2, No 3, 59-69.

15 Ekman P, Friesen WV 'Facial Action Coding System' *Investigator's Guide Part 2*, Consulting Psychologists Press Inc.,1978.

16 Warner D 'Biologically Responsive Interactive Interface' *Proc. Imagina 94*, 52 - 59

17 Turner R, Gobbetti E, Balaguer F, Mangili A, Thalmann D, Magnenat-Thalmann N 'An Object Oriented Methodology using Dynamic Variables for Animation and Scientific Visualization' *Proc. CGI '90*, Springer Verlag, 317-328.

FIGURE CAPTIONS

Figure 1: Recognition initialization - face with the soft mask (Color Photograph)

Figure 2: Flow chart of the recognition method

Figure 3: Face with the markers from the recognition (Color Photograph)

Figure 4: Points used in head tracking

Figure 5: Images from the real time facial animation (Color Photograph)

Figure 6: Using the facial movement to control a walkthrough (Color Photograph)

Figure 7: Connections between the recognition module and the application(s)