FACIAL EXPRESSION ANALYSIS ROBUST TO 3D HEAD POSE MOTION

A. C. Andrés del Valle and J.-L. Dugelay
{andres, dugelay}@eurecom.fr — http:// www.eurecom.fr/~image/Clonage/vc_mainpage.html
Institut Eurécom. 2229, rte. des Crêtes 06904 Sophia Antipolis - France

ABSTRACT

Most face expression algorithms assume a front or 'near-to-front' head position. This assumption becomes an important limitation when studying input from real systems. In this article we present a new approach to robustly determine face expression independently of the head pose. Our analysis-synthesis cooperation, possible thanks to the use of a highly realistic 3D head model and the application of Kalman filtering to predict the user pose, permits to correctly track the interesting face features. Adapting 'near-to-front' analysis techniques based on the predicted pose enables us to use such algorithms with moving speakers.

1. INTRODUCTION

Face video analysis is often performed through the analysis of specific face features (eyes, eyebrows and mouth) to extract the most significant information regarding expression and speech. Many of the current feature analysis algorithms are developed to work in 'near-to-front' face position. Whether the algorithms are used to do only expression analysis [1,2] or they are oriented to perform Model Based coding [3], these systems do not allow the user to move freely.

Assuming that we control the user pose is an important restriction when doing analysis for videoconferencing purposes. Yet, most virtual telecommunication schemes [4,5] try to avoid the pose-expression coupling issue by minimizing its effects. Nevertheless, for their analysis algorithms to remain robust, they only allow the user to do slight movements.

Y. Tian et al. [6] overcome the pose limitation in their analysis by defining a "multiple state face model", where different facial component models are used for different head states (left, left-front, right, down, etc.). This approach proves to be heavy. The complexity of such a solution increases with the number of the states, which would be large if much accuracy was needed. The approach given by Y. Chang et al. [7] tracks the head pose to use their 'near-to-front'-defined feature analysis algorithms over images where the head has a different pose. They use the estimated angles to rectify the input image to an almost straight frontal face. Although they do not give results on rectifying images from extreme head poses, this system seems to work well when pose changes are minor. Furthermore, they have to perform the complete face transformation, even though they only analyze some concrete features, not optimizing computation. Other approaches [8] project and fit a 3D head mesh onto the face image to keep track of the movements and perform the expression analysis. This is a complex and computing costly process that has not proved yet giving better results than analyzing the face features individually without fitting a 3D-mesh.

Developing a video analysis framework where head pose tracking and face feature analysis are treated separately permits to design specialized image analysis algorithms adjusted to specific needs, feature characteristics, etc. For our work on virtual teleconferencing environments, we first developed a pose tracking algorithm that profits from a tight analysis-synthesis cooperation. We are able to **track and predict the pose** of the speaker frame by frame with the help of the synthesis of its realistic 3D head model (clone). In parallel, we design image analysis algorithms to **study the expression motion** from a head on a 'near-to-front' position, situation at which faces show most of their gesture information. Having already developed and positively tested an eye-state analysis algorithm [10] for heads in front position, we faced the difficulty of adapting the algorithm to make it work at any pose. The solution we propose defines the eye-feature regions to be analyzed and the parameters of the eye-state analysis on 3D, over the head model in its frontal position. The complete procedure goes as follows:

(i) We define and shape the area to be analyzed on the video frame. To do so, we project the 3D-ROI defined over the head model on the video image by using the predicted pose parameters of the synthesized clone, thus getting the 2D-ROI.

(ii) We apply the eye image analysis algorithm on this area extracting the data required.

(iii) We interpret these data from a three dimensional perspective by inverting the projection and the transformations due to the pose (data pass from 2D to 3D). At this point, we can compare the results with the eye-state analysis parameters already predefined on the neutral clone and decide which has been the eye action.

The technique used differs from other previous approaches on that we explicitly use the clone data to define the analysis algorithm on 3D. The main advantages of our solution are the complete control of the location and shape of the region of interest (ROI) and the reutilization of image analysis algorithms already tested for a head front position. We can also improve the synthesis feedback utilized for the pose tracking by updating the face expressions of the clone.

In this paper, Section 2 recalls the complete tracking plus the feature analysis framework. In Section 3 we develop the novel approach for the definition of ROI for video face analysis. We describe the expression-pose analysis coupling in Section 4 and next, we give the influence of the pose prediction over the expression analysis algorithms. Section 6 shows some preliminary results and we draw our conclusions in Section 7.

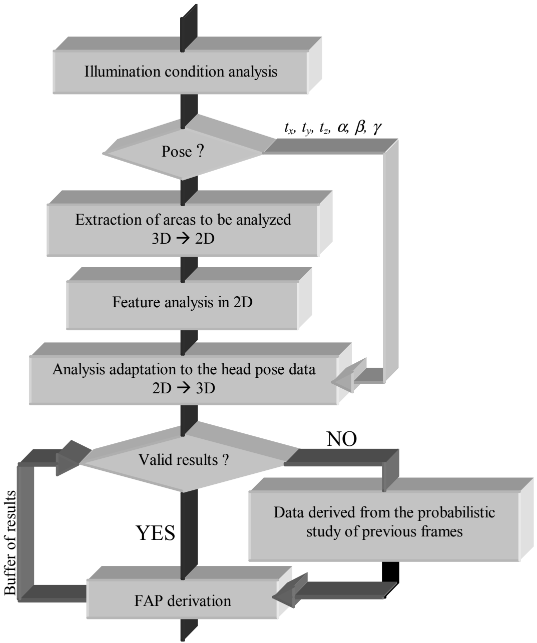

Illumination condition analysis
Pose ?
t_x, t_y, t_z, α, β, γ
Extraction of areas to be analyzed
3D → 2D
Feature analysis in 2D
Analysis adaptation to the head pose data
2D → 3D
Valid results ?
NO
YES
Data derived from the probabilistic study of previous frames
Buffer of results
FAP derivation

Figure 1. Complete diagram of the head pose and expression analysis for our teleconference system

2. COMPLETE POSE TRACKING AND FEATURE ANALYSIS

We consider face analysis from a video sequence as a function of the **general pose** of the face on the sequence, the **illumination conditions** under which the video is recorded and the face **expression** movements. To obtain animation parameters from video frames, we first study the illumination conditions of the face in the sequence; this information will enable our algorithms to work under any lighting conditions. Then, we estimate the pose of the face obtaining translation and rotation parameters. Finally, we extract some specific features from the face and we apply on them some dedicated analysis techniques to obtain face animation parameters. Synthesis cooperation can be done at different stages of the analysis chain (see Figure 1).

To obtain the global pose of the synthetic model we have developed a tracking algorithm that profits from a high analysis-synthesis cooperation. This algorithm utilizes a feedback loop. The predicted synthesized image from the clone is compared to the image of the face in the sequence to extract some 2D information. We feed a Kalman filter with this information and the filter predicts the translation and rotation parameters (t_x, t_y, t_z, α around x-axis, β around y-axis, γ around z-axis) to apply onto the synthetic clone, whose image is again compared to the following frame. This algorithm analyzes the model and the video sequence at the image level therefore we have to previously perform some light compensation on the synthetic model to adapt it to the lighting conditions of the video sequence. More details about this part can be found in [11]. The projection model used by the Kalman filter is:

$$\begin{bmatrix} x^p \\ y^p \end{bmatrix} = \frac{F}{N} \begin{bmatrix} c_\beta c_\gamma x_n - c_\beta s_\gamma y_n + s_\beta z_n + t_X \\ (s_\alpha s_\beta c_\gamma + c_\alpha s_\gamma) x_n - (s_\alpha s_\beta s_\gamma - c_\alpha c_\gamma) y_n - s_\alpha c_\beta z_n + t_Y \end{bmatrix} \text{(Eq.1)}$$

$$N = (c_\alpha s_\beta c_\gamma - s_\alpha s_\gamma) x_n + (-c_\alpha s_\beta s_\gamma - s_\alpha c_\gamma) y_n - c_\alpha c_\beta z_n - t_Z + F$$

where c_φ stands for cos(φ), s_φ for sin(φ), t_φ for tan(φ), ^{P} for projected coordinates and $_n$ for neutral 3D-coordinates.

Our expression analysis techniques are designed for a frontal head. Generally, we are able to interpret motion more easily when the face has this pose because this way, it shows most of its expression information. Since we study specific areas of the face, these features must be tracked with precision. We use some face features during the Kalman filtering, therefore we could also utilize them to delimit the areas where we will do expression analysis. This approach is not convenient because the pose-tracking algorithm assumes the movements on the tracked features are only due to pose motion; it cannot compensate for errors from expression changes. Moreover, we need to obtain the relationship between the analysis area and the six predicted pose parameters so to know how to interpret our analysis algorithm in a 'non-frontal' position.

To overcome these limitations, we define the expression analysis features on the 3D head model and independently of the image features for the pose tracking. Next sections present how **pose prediction** becomes useful for: i) correct **ROI definition** and ii) proper **expression analysis recovery**. The procedure presented does not only allow us to study the important areas of the face on the image but also to understand the data that the analysis provides. We are able to well detect the eye, analyze it and interpret the derived information regardless of the pose of the head.

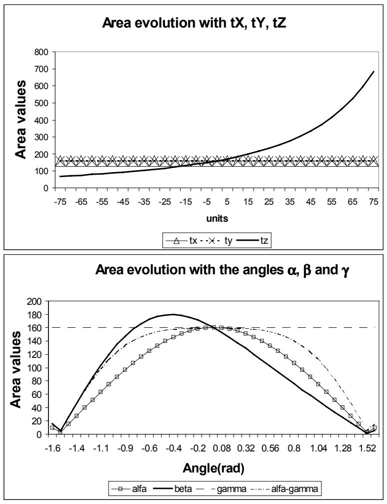
3. DEFINITION OF VIDEO ANALYSIS ROI

The control of the area being analyzed is very important for many expression analysis algorithms, more specifically for the analysis technique that we have developed for eye state tracking [10]. The definition of a well-established ROI has two purposes. On the one hand, we want to extract the maximum amount of information, optimizing the area to analyze (minimizing computation). On the other hand, we want to foresee the relevance of the information we could obtain from the feature even before having started its analysis.

To achieve these goals, we define the ROI over the 3D head model and not over the image itself. We obtain the region to analyze by projecting this 3D area on the image (Fig. 2(a-b) shows which area is chosen for the eye analysis). Projecting using the predicted pose parameters allows us to reshape the areas on the frames along with the pose and to foresee the relevance of the analysis of one feature. We define a threshold *Th*, computed as the surface of the projected ROI, below which the algorithm will not act because we consider that there will not be enough visible surface. This threshold is feature dependent. Section 6 gives details about the chosen *Th* for the eye analysis. Fig. 2(c) also shows the deformation of the area and Graphs 1-2 represent the evolution of the area depending on the pose parameters (each one independently and α-γ conjointly). The surface expression is:

$$Area = base \cdot height = \mathbf{A}^{\mathbf{P}} \cdot \mathbf{B}^{\mathbf{P}} \cdot \sin(\arccos(\frac{(x_3^P - x_1^P)(x_2^P - x_4^P) + (y_3^P - y_1^P)(y_2^P - y_4^P)}{\mathbf{A}^{\mathbf{P}} \cdot \mathbf{B}^{\mathbf{P}}}))$$

where $\mathbf{A}^P$ = dist (3^P,1^P), $\mathbf{B}^P$ = dist (2^P,4^P) and i^P=(x_i^P,y_i^P).



Graphs 1-2. Surface evolution with pose parameters independently and conjointly (α-γ) for the eye feature.

To make the deformed areas more suitable for image analysis we enclose them in video analysis rectangles $(x_t,y_t) \rightarrow (x_b,y_b)$:

$$(x_t, y_t) = (\min(x) \in ROI, \max(y) \in ROI);$$

$$(x_b, y_b) = (\max(x) \in ROI, \min(y) \in ROI)$$

4. EXPRESSION-POSE ANALYSIS COUPLING

Expression analysis algorithms defined for a frontal position cannot be directly used over image features obtained at any given pose. Our eye analysis algorithm searches for the point of lower energy or minimum intensity of the feature area to determine the eye state (open-close, left-center-right). In [9] we show how the restriction of having the same action in both eyes allow us to define the eye states from the situation of the lower energy point. The different states are defined for a head in a frontal position. Thanks to the pose prediction we could rectify the image and then apply the algorithm. Instead, we prefer slightly adapting the algorithm. It is less computing costly and gives more accurate results. To adapt the algorithm, first, we **reformulate it to fit 3D space** by defining the analysis parameters and state measurements in 3D, over the model. Next, we **find the minimal energy point** on the projected ROI image. Then, we **inverse the projection and the pose transformation** of the point to deduce its 3D coordinates.

By inverting the pose and projection used by the Kalman filter (see Equation 1), we recover the straight line that defines the ray of possible solutions for the projected point in 3D space. To obtain its 3D coordinates, we use the information provided by the **model of the feature we are analyzing**. For the eye, we model it as the sphere — $(x_n - x_0)^2 + (y_n - y_0)^2 + (z_n - z_0)^2 = Rad^2$ — that better suits the eye on the head model (Fig. 2(a-b)). To simplify, we **linearize the model** by developing the linear approximation of the sphere on the point tangent to the pupil in its neutral position — $z_n = M = z_0 + Rad$. This plane is also used to define the 3D-ROI tracking area. The intersection of the ray with the modeled surface provides the point 3D-coordinates. These 3D-coordinates are the solution to the following system:

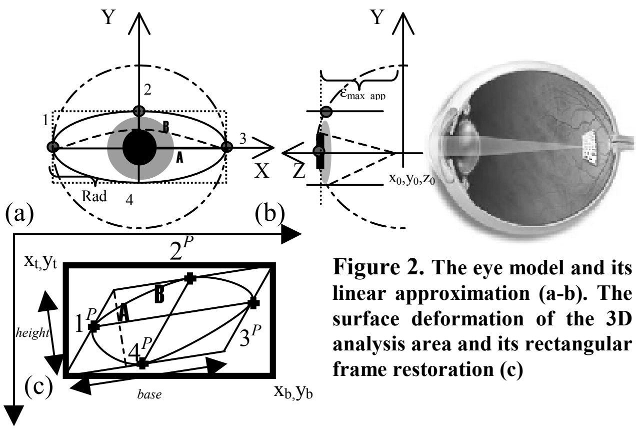


Figure 2. The eye model and its linear approximation (a-b). The surface deformation of the 3D analysis area and its rectangular frame restoration (c)

$$\begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \end{bmatrix} = \begin{bmatrix} a_4 - Ma_3 \\ b_4 - Mb_3 \end{bmatrix}$$

$$\begin{array}{ll} a_1 = -x_p(c_\alpha s_\beta c_\gamma - s_\alpha s_\gamma) + Fc_\beta c_\gamma & b_1 = -y_p(c_\alpha s_\beta c_\gamma - s_\alpha s_\gamma) + F(s_\alpha s_\beta c_\gamma + c_\alpha s_\gamma) \\ a_2 = x_p(c_\alpha s_\beta s_\gamma + s_\alpha c_\gamma) - Fc_\beta s_\gamma & b_2 = y_p(c_\alpha s_\beta s_\gamma + s_\alpha c_\gamma) + F(-s_\alpha s_\beta s_\gamma + c_\alpha c_\gamma) \\ a_3 = x_p(c_\alpha c_\beta) + Fs_\beta & b_3 = y_p(c_\alpha c_\beta) - Fs_\alpha c_\beta \\ a_4 = -x_p.(t_Z - F) - F.t_X & b_4 = -y_p(t_Z - F) - F.t_Y \end{array}$$

We consider conflictive the cases where the system does not present a solution. This always occurs for angles over $\pm\pi/4$. Therefore, coupling is reliable for angles between $+\pi/4$ and $-\pi/4$. Once the 3D position is recovered, we can compare it to the analysis parameters previously defined in 3D and deduce the eye state. The error precision due to the model linearization is known and always smaller than *Rad*, $|\varepsilon_{max\_app}| < Rad$.

5. INFLUENCE OF POSE PREDICTION OVER FEATURE ANALYSIS: ERROR BEHAVIOR

The coupled feature-analysis pose-tracking procedure results on a system that works under the influence of errors cumulated from two different origins.

In [12] we have developed the error expressions of the obtained 'neutral' coordinates from the analyzed projected data. Along with the analysis precision error we find errors due to the Kalman prediction. Analyzing these expressions, we check that when coupling pose prediction with feature expression analysis, the inaccuracy of pose predicting becomes the major source of error. Unlike the error introduced when analyzing the video image, the pose prediction error could hardly be minimized by any image analysis technique and therefore cannot be easily controlled. This error analysis shows how critical the accuracy of the prediction is for our method.

6. TESTS AND RESULTS

In previous articles [10,11], we showed the positive results pose and eye-state tracking algorithms had, when working separately. To study the feasibility of the pose-expression analysis coupling, first, we analyzed ROI adaptation and tracking on some synthetic sequences. Tests presented perfect adaptation of the eye ROI through the frames. Even though these preliminary tests did not show that the predicted pose parameters could recover well the minimum energy point of the tracked ROI, they showed that eye

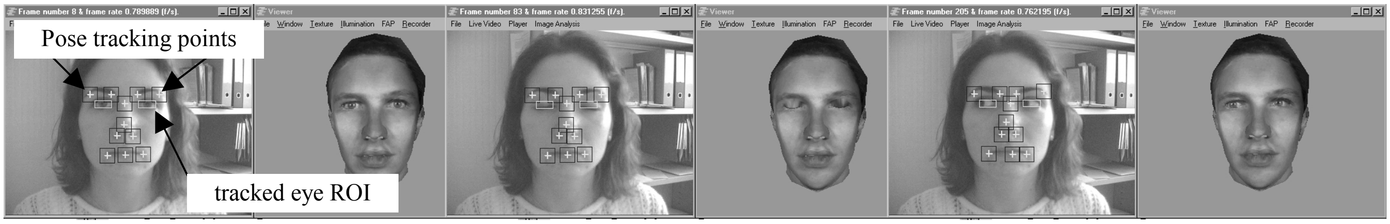

Pose tracking points
tracked eye ROI
Frame number 8 & frame rate 0.789889 (f/s).
Viewer
Frame number 83 & frame rate 0.831255 (f/s).
Viewer
Frame number 205 & frame rate 0.762195 (f/s).
Viewer
File Window Texture Illumination FAP Recorder
File Live Video Player Image Analysis

Figure 3. Evolution of the pose and feature expression analysis coupling with a non-similar 3D head model, over a real sequence.

features were correctly enclosed as long as the tracking was well performed.

Next, to test the eye-state algorithm coupled with the pose, we applied the 3D-adapted eye-tracking algorithm over the tracked ROI on real video sequences. We used a 3D model that not matched the person on the video. We are able to utilize the pose-tracking algorithm under these conditions, although it permits less freedom of movement than when used with the analysis-synthesis feedback of a realistic speaker-dependent model. In sequences where the movements were not too extreme, the adaptation worked fine. The success of the eye expression analysis algorithm, evaluated as the number of times the 3D model closed, opened its eyes, looked right, left correctly, stayed the same level as when it was not coupled with the pose tracking (80%). As expected, the system showed worse performance when the pose-analysis algorithm started losing track.

The correct behavior of the pose-tracking algorithm ensures good analysis coupling, therefore we expect better results using the analysis-synthesis feedback with high realistic models. The threshold, *Th*, used to judge the relevance of the feature analysis is dependent of the ROI analysis technique. In our tests, we used *Th*=3**AreaMinEnergySearch,* because our algorithm is able to detect up to three possible sight states (left-center-right).

The robustness of our approach is best appreciated by visually comparing video input with the synthetic results obtained from interpreting the analysis done over it. Video sequences showing video input analysis and its synthesis representation can be found on the web site of our project:

http://www.eurecom.fr/~image/Clonage/demos.html

We are currently able to perform the coupled analysis over real-time video input at a rate of 1.5 f/s (PC: Pentium III –BiPro at 700MHz). These results show the possibility of future online performance with better equipment and implementation.

7. CONCLUSION AND FUTURE WORK

Proper ROI definition is critical to accurately do face expression analysis. We have shown that controlling and predicting ROI evolution becomes very useful in this context. From our analysis, we have concluded that the Kalman filter based pose-feature analysis coupling is completely stable to head translations and becomes unstable for those head angles beyond $\pm\pi/4$. Nevertheless our tracking system finds its limitation beyond these angles because the tracked data start to fall onto the same projected points.

Parallel to the eye analysis, we have developed algorithms to study the eyebrow motion. We intend to perform our pose-feature analysis coupling over these algorithms as well.

8. ACKNOWLEDGEMENTS

Our work is partially supported by FRANCE TELECOM Research.

9. REFERENCES

[1] F. Piat, and M. Tsapatsoulis, "Exploring the Time Course of Facial Expressions with a Fuzzy System," *ICME 2000*, Sydney - Australia, 27-30 Aug. 2000.

[2] M. Pantic, and L. J.M. Rothkrantz, "Automatic Analysis of facial Expressions: The State of the Art," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, pp. 1424-1445, Dec. 2000.

[3] E. Cosatto, G. Potamianos, and H.P. Graf, "Audio-Visual Selection for the Synthesis of Photo-Realistic Talking-Heads," *ICME 2000*, Sydney - Australia, 27-30 Aug. 2000.

[4] T. Goto, S. Kshirsagar, and N. Magnenat-Thalmann, "Automatic Face Cloning and Animation," *IEEE Signal Processing Magazine*, Vol. 18, No. 3, pp. 17-25, May 2001.

[5] P. Eisert, and B. Girod, "Analyzing Facial Expressions for Virtual Conferencing," *IEEE Computer Graphics & Applications*, pp. 70-78, Sep. 1998.

[6] Y. Tian, T. Kanade, and J. Cohn, "Recognizing Lower Face Action Units for Facial Expression Analysis," *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, Mar. 2000, pp. 484 - 490.

[7] Y. Chang, et al., "Vitual Talk: a Model-Based Virtual Phone Using a Layered Audio-Visual Integration," *ICME 2000*, Sydney - Australia, 27-30 Aug. 2000.

[8] S. Ogata, et al., "Model-Based Lip Synchronization with Automatically Translated Synthetic Voice toward a Multi-Modal Translation System," *ICME 2001*, Tokyo - Japan, 22-25 Aug. 2001.

[9] J.-L. Dugelay, and A. C. Andrés del Valle, "Analysis-Synthesis Cooperation for MPEG-4 Realistic Clone Animation," *Euroimage ICAV3D*, Mykonos - Greece, 30 May-1 June 2001.

[10] A. C. Andrés del Valle, and J.-L. Dugelay, "Eye State Tracking for Face Cloning," *ICIP 2001*, Thessaloniki - Greece, 7-10 Oct. 2001.

[11] S. Valente and J.-L. Dugelay, "Face Tracking and Realistic Animations for Telecommunicant Clones," *IEEE Multimedia Magazine*, pp. 34-43, Feb. 2000.

[12] A. C. Andrés del Valle, and J.-L. Dugelay. "Pose Coupling with Eye Movement Tracking," *Eurecom Research Report RR-01-058*, Feb. 2002.