3D model fitting for facial expression analysis under uncontrolled imaging conditions

Pierre Maurel
Odyssée, ENS
Paris, France
pmaurel@di.ens.fr

Aileen McGonigal
UMR Inserm U 751
Faculté de Médecine
Marseille, France

Renaud Keriven
CERTIS
École des ponts
Paris-Est, France

Patrick Chauvel
UMR Inserm U 751
Faculté de Médecine
Marseille, France

Abstract

This paper addresses the recovering of 3D pose and animation of the human face in a monocular single image under uncontrolled imaging conditions. Our goal is to fit a 3D animated model in a face image with possibly large variations of head pose and facial expressions. Our data were acquired from filmed epileptic seizures of patients undergoing investigation in the videotelemetry unit, La Timone hospital, Marseille, France [1].

1. Introduction

Facial expression analysis has been an active research topic for behavioral scientists and psychologists since the work of C.Darwin in 1872 [4, 7]. In 1978, Suwa et al. [11] presented a preliminary investigation on automatic facial expression analysis by tracking the motion of several identified spots on an image sequence. Since then considerable progress has been made in building computer systems that attempt to automatically analyze and recognize facial motions [8, 10].

Two principal classes of approaches have been developed: Image-based [2, 9] and Model-based approaches [5, 6]. Image-based methods extract features from images without relying on elaborate knowledge about the object of interest. Their principal quality is their quickness and their simplicity. However, if the data images are very diverse (e.g.: variation of illumination, of view, of head pose) image-based approaches can become erratic and unsatisfactory. On the other hand model-based methods use models which maintain the essential characteristics of the face (position of the eyes relative to the nose for example), but which can deform to fit a range of possible facial shapes and expressions.

[1] We thank O. Faugeras for suggesting the topic of this article and facilitating the collaboration.

In this work, we are interested in analyzing the facial expression of several patients during epileptic seizures. In fact, detailed study of such facial expressions produced during epileptic seizures could help in understanding the cerebral organization of the seizures. Because of the unsupervised nature of the data acquisition (a fixed camera in a hospital room), we chose to use a model-based approach. A large class of methods developed in the last decade was based on the Active Appearance Models [12] and more recently on 3D Morphable Model [3]. These methods construct a model from a learning set of several images of different persons showing different expressions. In our case, the expressions of the epileptic patients during their crises could be individual and complex and consequently a model built from a learning set of common expressions (typically such as anger, sadness or happiness) would not have been sufficient. This is why we chose to use the Candide [1] model in our work.

We first introduce the 3D face model we used in this paper. Then we present our method to fit this model on a facial image. The next section deals with the analysis of a facial expression. The final section shows some of the results on the real data.

2. The 3D face Model

2.1. Candide face model

In this work we use a modified version of the Candide 3D face model [1], but our method can easily be applied to any other model. Candide is a parameterized face mask specifically developed for model-based coding of human faces. The original Candide Model contains 113 vertices and 184 triangles. Fig.1 shows the mesh of the model.

To control the model, 14 shape units and 71 animation units are provided (see Fig.2 for some examples of

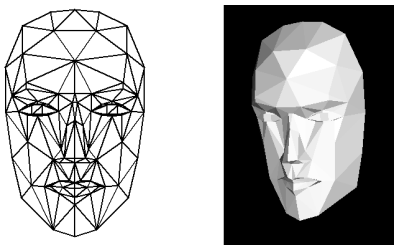

Figure 1. The Candide face model

these deformation vectors).

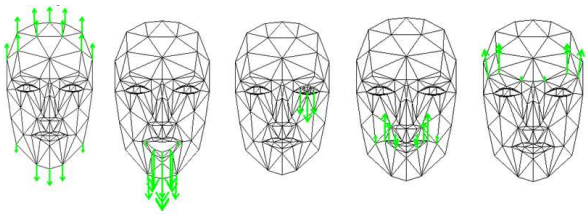

Figure 2. one example of a shape unit, and four examples of animation units

For our work we had to slightly modify the original model, principally in order to remove the top part of the head and to add some specific animation units. These modifications were motivated by the specificity of our images. The patients in fact wore EEG electrodes on the head and the forehead was often hidden. Furthermore, the expressions encountered during the crises were often asymmetric and some additional animation parameters were necessary to cover all the scope of the possible expression.

2.2. The Reference Texture

We also added a *reference texture* to the model. This texture has been computed as the average between a few number of faces on which the model has been manually placed. Fig.3 shows this *reference texture* on the model.

3. Model Fitting

3.1. Energy

For a new image of a face, we want to find the best position of the 3D model and the best values of the shape and animation parameters in the sense that the projection of the mesh in the image matches the face

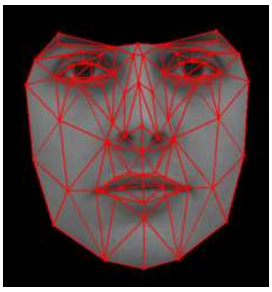

Figure 3. Reference texture, I_{ref}

(see Fig.4). Let us define an energy which measures

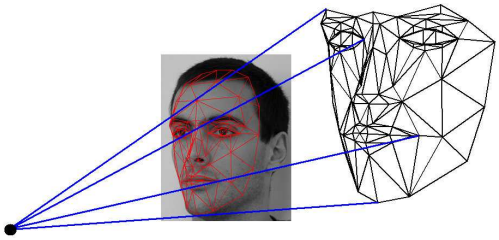

Figure 4. Projection of the 3D model in a 2D image, I

the quality of a given position of the 3D model (and its projection). We use the cross-correlation between the *reference texture*, I_{ref}, and the image in which we want to fit the model, I. The support of the cross correlation is, for each vertex P_i, the set of triangles T_j incidents to this vertex (see Fig.5).

$$\rho_i = \sum_{j|P_i \in T_j} \int_{T_j^{ref}} \left(I_{ref}(P) - \overline{I_{ref}}_{,i}\right)\left(I(\varphi(P)) - \overline{I_{\varphi}}_{,i}\right) dP$$

with

$$\overline{I_{ref}}_{,i} = \frac{1}{\sum_{j|P_i \in T_j} |T_{ref,j}|} \sum_{j|P_i \in T_j} \int_{T_{ref,j}} I_{ref}(P) dP$$

$$\overline{I_{\varphi}}_{,i} = \frac{1}{\sum_{j|P_i \in T_j} |T_{ref,j}|} \sum_{j|P_i \in T_j} \int_{T_{ref,j}} I(\varphi(P)) dP$$

Let us then define:

$$\langle I_{ref}, I \rangle_{\varphi} = \sum_{i=0}^{n} \rho_i$$

where n is the number of points in the model.

And finally the energy:

$$\rho_{I,\varphi}(M) = \frac{\langle I_{ref}, I\rangle_{\varphi}}{\sqrt{\langle I_{ref}, I_{ref}\rangle}\sqrt{\langle I, I\rangle_{\varphi}}}$$

φ is a bijection from I_{ref} to I based on the current projection of the model (see Fig.5).

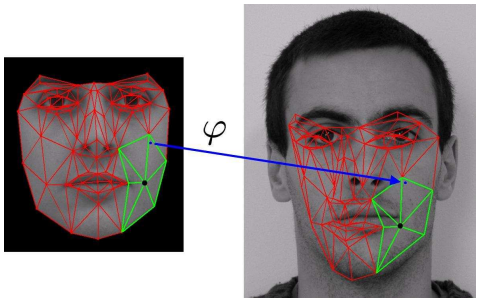

Figure 5. support of the cross correlation and φ, the bijection between the reference texture and the current projection of the model in the image

This energy has the advantage of being invariant to any affine transformation of the histogram of the image.

3.2. Energy minimization

We computed the derivatives of the energy $\rho_{I,\varphi}(M)$ with respect to the 3D global position of the model, and to the shape and animation parameters. We then perform a multi-scale gradient descent: we use a quasi-Newton method with non-linear constraints.

The multi-scale method makes the process less dependent on the initialization of the minimization and quicken the convergence. The constraints imposed in the minimization algorithm are: limits on the shape and animation parameters, some specific conditions on the mesh (e.g.: the eyelids and the lips must not pass through each other) and the fact that the projection of the mesh should not exceed the borders of the image.

4. Facial Expression

Once the model is fitted in an initial image of a subject, we define a new *reference texture* using the image and the projection of the fitted model. Given a new image of the same subject with a new facial expression, we use this new *reference texture* in order to fit the model in this image. The minimization is now done only with respect to global position and animation parameters (the shape parameters are supposed to be constant for two images of the same person). The facial expression can then be represented as the variation of the animation parameters between the neutral image of the subject and the expressive one.

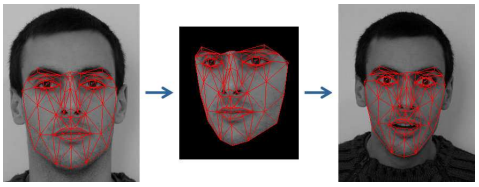

Figure 6. the two steps of the analysis of an expression. Left: the model is fitted on a neutral image. Middle: the new reference texture. Right: the model is fitted on the expressive image.

5. Results

During seizures, large movements can occur and occlusion of the image may be produced by intervention of medical personnel. Therefore we manually selected several images for each patient[2]: one of a neutral expression and some during seizures. Fig. 7 shows some of the results of the algorithm: the selected images and the final fit of the 3D model.

Once the model is fitted on the neutral image and on the expressive one, we can compute the variation of the animation parameters from the first image to the second one and for example apply that variation to another position of the model (Fig. 8).

6. Conclusion

We have proposed a method to fit a 3D animated model in a monocular single image under uncontrolled imaging conditions. Our method is based on an energy defined with a cross-correlation term and an energy minimization process. We fitted the model on real-world data obtained in a medical framework. The principal drawback of our approach is the time of computation of the optimization process which is quite far from real-time. However real-time processing was not one of our requirements, since our work was motivated by a need for an analysis tool. Further work includes investigating on the potential clinical use of this tool and on a more sophisticated model (such as including more advanced texture statistics than just an average).

[2]Written informed consent was obtained from all patients for the use of their video recordings, including for publication.

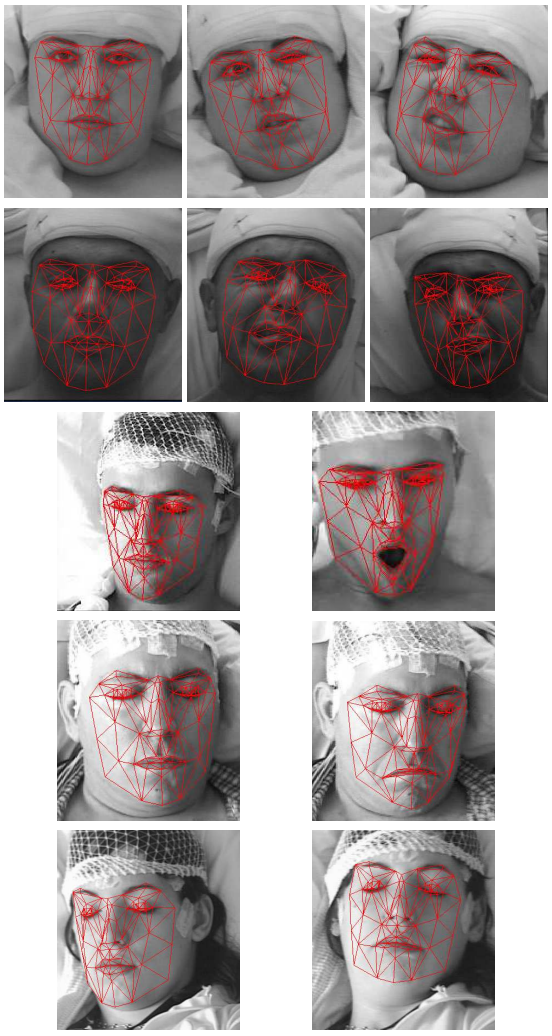

Figure 7. Some results: left column shows a neutral view and other columns are images taken during the seizure

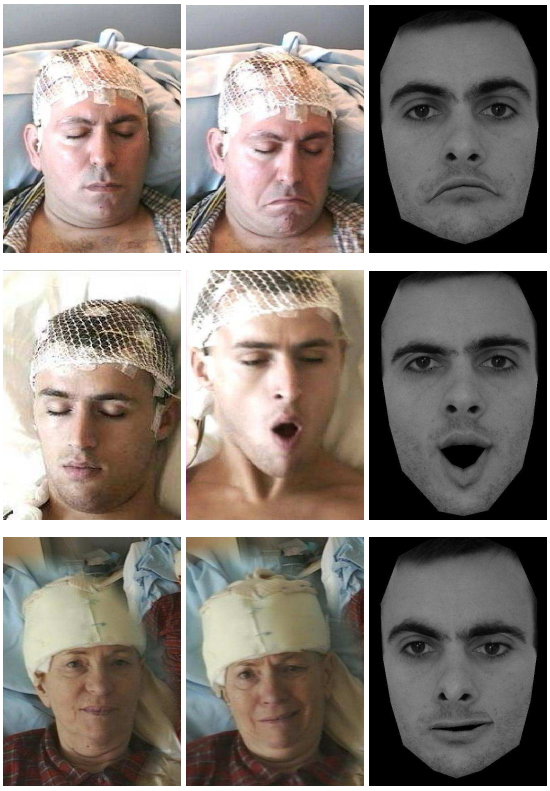

Figure 8. first column: a neutral expression, second column: image of the facial expression during an epileptic seizure and last column: the expression mapped on a new image

References

[1] J. Ahlberg. Candide-3 - an updated parameterised face.

[2] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms, 1998.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In A. Rockwood, editor, *Siggraph 1999*, pages 187–194, Los Angeles, 1999. Addison Wesley Longman.

[4] C. Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, London, 1872.

[5] M. Dimitrijevic, S. Ilic, and P. Fua. Accurate face models from uncalibrated and ill-lit video sequences. In *Conference on Computer Vision and Pattern Recognition, Washington, DC*, June 2004.

[6] F. Dornaika and J. Ahlberg. Fast and reliable active appearance model search for 3d face tracking.

[7] P. Ekman and W. Friesen. The facial action coding system. *Consulting Psychologists Press*, 1978.

[8] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36(1):259–275, 2003. IDIAP-RR 99-19.

[9] S. J. McKenna, S. Gong, R. P. Würtz, J. Tanner, and D. Banin. Tracking facial feature points with Gabor wavelets and shape models. In *Int. Conference on Audio- and Video-based Biometric Person Authentication*, 1997.

[10] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[11] M. Suwa, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. *4th International Joint Conference on Pattern Recognition*, pages 408–410, 1978.

[12] C. Taylor, G. Edwards, and T. Cootes. Active appearance models. In *ECCV98*, page II: 484, 1998.