Analysis and Synthesis of Facial Motions

R. Magnolfi, P. Nesi
Department of Systems and Informatics, Faculty of Engineering
University of Florence, Via S. Marta 3, 50139 Florence, Italy
Tel.: +39-55-4796265, Fax.:+39-55-4796363, NESI@INGFI1.ING.UNIFI.IT

Abstract

Many researchers have studied the problems related to the analysis and synthesis of human faces under motion and deformations. These techniques can be used for defining low-rate image compression algorithms (model-based image coding), cinema technologies, videophone, as well as for applications of virtual reality, etc. Such techniques need real-time performance and a strong integration between the mechanisms of motion estimation and that of 3D head/face modeling, rendering and animation. In this paper, a complete and integrated system for tracking and synthesising face motions in real-time on low-cost architectures is presented. For modeling curves associated with face features spatio-temporal B-splines have been adopted. In addition, the system proposed is capable of adapting a generic 3D wire-frame model of a head/face to the face that must be tracked, and thus the simulation of the face motions estimations can be simulated on a realistically patterned face.

1 Introduction

In the recent years many researchers have studied the problems related to the analysis and synthesis of faces under motion. These techniques can be used for defining low-rate image compression algorithms (model-based image coding), new cinema technologies as well as for applications of virtual reality, and videophone, etc. In order to be effectively used, such techniques have to integrate mechanisms for motion estimation with that of 3D head/face modeling, rendering and animation. For most of these new applications the processes of motion estimation and rendering must be mandatorily performed in real-time.

The head/face motion estimation problem can be divided into two sub-problems, that is, the estimation of head motion (global motions) and the estimation of face deformations due to changes of expressions (local motions). The first problem is also know as head tracking and can be solved with traditional techniques for motion estimation – e.g., [2], [6]. In the literature, the second problem, that is, the problem of estimation of facial deformations and motions (lips tracking, eyes tracking, etc.), has been addressed by using several techniques. These techniques can be classified in three main categories which can be distinguished for what they adopt for modeling facial features (mouth, eyes, eyebrows, nose, etc.) that have to be tracked: (i) deformable or dynamic contours (spline/snake, B-spline) – e.g., [7], [5], [16], [4]; (ii) deformable templates – e.g., [19], [18]; (iii) points or patterns (by using optical-flow- or matching-based techniques) – e.g., [9]. These features are tracked in subsequent image frames in order to estimate the face deformations – i.e., by measuring motions and deformations of features selected.

The head/face synthesis and animation is obtained by using techniques for (i) modeling the head/face 3D structure as a wire-frame object – e.g., [12], [14], [11]; (ii) smearing the face/head pattern on the corresponding 3D wire-frame model instead of using classical algorithms for shading with uniform colors (e.g., Phong, Gouraud); and for (iii) animating the non-rigid patterned head/face – e.g., [17].

Since in the animated model must be as similar to the real model as possible, a correspondence among the face features (which are used for tracking the face deformations) and the corresponding mathematical structures in the reconstructed model must be defined. The association between these two domains is defined in a phase in which a parametrized 3D wire-frame model is adjusted in accordance with the real measures of the model under analysis. The process of adjustment can be also simplified by deforming the wire-frame model in order to match a frontal image of the face shape – e.g., [13]. In some cases the structural model of the face can be defined by considering also facial muscles for a certain depth [15].

In this paper, a complete method for tracking facial features like (mouth and eyebrows) and synthesising that in quasi-real-time by using low-cost architectures is presented (see Section 2). The method for tracking features is based on dynamic contours, which in turn are mathematically modelled as spatio-temporal B-spline. The wire-frame model of the human head used has been obtained by improving the well known CANDIDE model due to [14], [8] (see Section 2.1). The 3D wire-frame model is adapted to the face under analysis by using a single frame, by means of a method derived from [13]. Moreover, an *ad hoc* algorithm for guarantee a fast smearing of the real face pattern on the model has been also defined. Experimental results are discussed in Section 3. Conclusions are drawn in Section 4.

2 Modeling Facial Features as Spatio-Temporal B-Snake

The tracking of facial features is reduced to the problem of tracking curves, which model the shape of that features. Since the contour can change his form in subsequent frames a process of tracking from the old position to the next is needed. This process can be obtained by defining an energy model for deformable contours [7], [5], [16], dynamic contour [4], [10], and deformable template [19], [18] (following the classification of Blake [1]). The approaches based on deformable contours (spline) and/or templates are usually computationally too heavy to be used for real-time tracking on low-cost architectures. Moreover, deformable contours are so flexible that in many cases is very difficult to maintain under control their shape. On the contrary, deformable templates work well only when the shape of the features under tracking is quite known and its deformations are small and don't change the shape structure (e.g., inversion of curvature). Dynamic contours are based on B-splines and attempt to integrate the above aspects since they model curves as a combination of elementary templates. In addition, they use a parametrized representation of the curve which makes their estimation cheaper with respect to classical splines and templates. As in [10] we call this model for curves representation as "B-snake". Moreover, since the proposed model extends the adoption of B-snake to track curves in time it is a "spatio-temporal B-snake based model" (STB-snake).

A STB-snake is a deformable parametrized surface controlled by the temporal behavior of internal and image forces which act in each point of the surface. The internal forces, F_{int}, represent the constraints on the shape curve (regularity, elasticity, etc.), while the image forces, F_{img}, guide the contour to match certain image features (luminance, contrast, etc.). Integrating that forces along the curve $v(s,t)$ the corresponding energies are obtained and from these the total energy:

$$E_{tot} = E_{int} - E_{img} = \int_T \int_s (F_{int} - F_{img}) ds dt, \quad (1)$$

where $v(s,t)$ is the parametric description of the curve and $v(s,t) = v(x(s,t), y(s,t))$. The goal is to find the surface that minimises the total energy in time. When a minimum for E_{tot} is reached, $x(s,t)$ and $y(s,t)$ expressions define a curve which best fits the contour according to its definition in terms of E_{img}.

The *Internal Energy*, E_{int}, is defined as:

$$E_{int} = E_1 + E_2 + E_t, \quad (2)$$

where E_1 and E_2 take into account the tension and the rigidity of the curve shape (the surface at a given time instant), respectively (i.e., they impose the regularity of the curve shape). The corresponding forces are weighed with functions $\alpha(s)$ and $\beta(s)$, respectively:

$$E_1 = \int_T \int_s (\alpha(s) \mid v_s(s,t) \mid^2) ds dt, \quad (3)$$

$$E_2 = \int_T \int_s (\beta(s) \mid v_{ss}(s,t) \mid^2) ds dt, \quad (4)$$

E_t takes in consideration the spatio-temporal regularity of the surface in time:

$$E_t = \int_T \int_s (\tau(s) \mid v_t(s,t) \mid^2) ds dt. \quad (5)$$

The *Image Energy*, E_{img}, consists of two terms: E_c, that depends on the image contrast of the points belonging to the curve, and E_v, that considers the changes in contrast with time:

$$E_{img} = E_c + E_v, \quad (6)$$

$$E_c = \int_T \int_s (\rho(s) H(I(x(s,t), y(s,t), t))) ds dt, \quad (7)$$

$$E_v = \int_T \int_s (\gamma(s) I_t(x(s,t), y(s,t), t)) ds dt, \quad (8)$$

where $H()$ is a gradient operator, $I(x(s,t), y(s,t), t)$ is the value of the image brightness at time t in the point $(x(s,t), y(s,t))$, I_t is the first order partial derivatives of the image brightness with respect to time, $\rho(s)$ and $\gamma(s)$ are suitable weight functions. The operator $H()$ must be capable of identifying the shape of the curve that must be tracked in the image sequence.

At each time step, the minimization of (1) is obtained by estimating the solution of the system of equations which have been obtained by taking the derivatives of the functional with respect to the unknowns (i.e., points through which the approximating curves must pass). Thus, a system of $2(p+1)$ unknowns is defined where $p+1$ is the number of curve points. Using a curve representation based on B-spline the dimension of the system of equations is strongly reduced since:

$$x(s) = \sum_{i=0}^{m} X_i B_i(s); \qquad y(s) = \sum_{i=0}^{m} Y_i B_i(s),$$

where $B_i()$ for $i = 0,..m$ are polynomials defining the basis of the B-spline representation, and (X_i, Y_i) for $i = 0,..m$ are the control point (i.e., the knots) of the curve. Thus, with this representation the number of unknowns is reduced from $2(p+1)$ to $2(m+1)$ where $m \ll p$, and the equation set can be written as:

$$\begin{aligned} \mathbf{AX} + \mathbf{G}_x(x,y,t) + \mathbf{VX}_t + \mathbf{E}_{vx} &= 0, \\ \mathbf{AY} + \mathbf{G}_y(x,y,t) + \mathbf{VY}_t + \mathbf{E}_{vy} &= 0, \end{aligned} \quad (9)$$

where $\mathbf{A}$ is an $(m+1) \times (m+1)$ matrix and $\mathbf{G}_x$, $\mathbf{G}_y$ are $(m+1)$-dimensional vectors:

$$A_{ij} = 2 \sum_{h=0}^{p} \alpha(s_h) B_{si}(s_h) B_{sj}(s_h) + \beta(s_h) B_{ssi}(s_h) B_{ssj}(s_h),$$

$$G_{xi} = \sum_{h=0}^{p} \rho(s_h) B_i(s_h) \frac{\partial H(x(s_h,t), y(s_h,t), t)}{\partial X_i},$$

$$G_{yi} = \sum_{h=0}^{p} \rho(s_h) B_i(s_h) \frac{\partial H(x(s_h,t), y(s_h,t), t)}{\partial Y_i},$$

$$V_{ij} = 2 \sum_{h=0}^{p} \left[\tau(s_h) B_i(s_h) B_j(s_h) \right].$$

Discrete versions of the above mentioned operators are obviously adopted during the numerical computation of curve parameters.

If traditional methods for solving non-linear systems of equations are adopted then the solution of the above system of equations can be computationally very heavy. In order to solve this problem, a specific and very fast method has been defined. According to many other applications, in which splines have been used for modeling curves in vision, the first

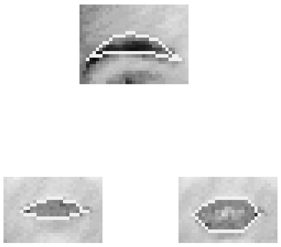

Figure 1: Modeling with STB-snake: a mouth closed, a mouth opened where can be seen the teeth, and an eyebrow on the top.

hypothesis is that the initial data is not very far from the final solution. If the deformations are supposed to be slow or the number of image per second high, the above hypothesis can be expanded to be applied to the changes between two subsequent images. The method is based on the estimation of the sign of the derivatives of total energy with respect to each variable (X_i, Y_i) for $i = 0, ..m$. Once the derivatives are estimated, the coordinates of each point (X_i, Y_i) is increased or decreased of a given amount, δ, according to the corresponding sign. This process is performed for each node for Q iterations (the stop criterion is based on a threshold on the value of the derivative of total energy with respect to the iteration number). In order to decrease the number of iterations and thus to improve the system performance, experimental results have demonstrated that the value of δ at the generic iteration q can be profitably obtained on the basis of an initial value δ_o and by using the iteration number: $\delta_q = \delta_o \sigma^q$ where $\sigma < 1$. This technique allows the estimation of the minima at each time step by using only few iterations, typically no more than 10-15 iteration, with $\delta_o = 1$ and $\sigma = 0.75$. The values have been chosen also by considering that the final goal is to reproduce the synthetic model on a screen.

Since the process is driven by the image energy (i.e., when the image energy changes the curve follows the changes in order to reach the minima) in certain conditions the curves can lose some points since they found a low energy due to the presence of more prominent image gradients, as it has been many times noted for classical splines. A typical example is the case in which a mouth opened where the teeth are visible (the appearing of the teeth changes the conformation of the energy surface). In these conditions the points are attracted from the center of the mouth. In order to solve this problem an *ad hoc* energy of repulsion has been defined among the points belonging to the upper and the lower parts of the mouth. This factor has been added to the expression of E_{int} (2) in the complete model. This constraint can also be profitably used for eyebrows since their dimensions can be considered constant in time.

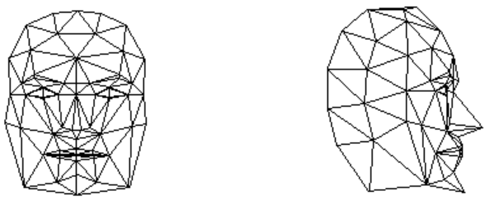

Figure 2: The generic 3D wire-frame model of the head/face: frontal and side view.

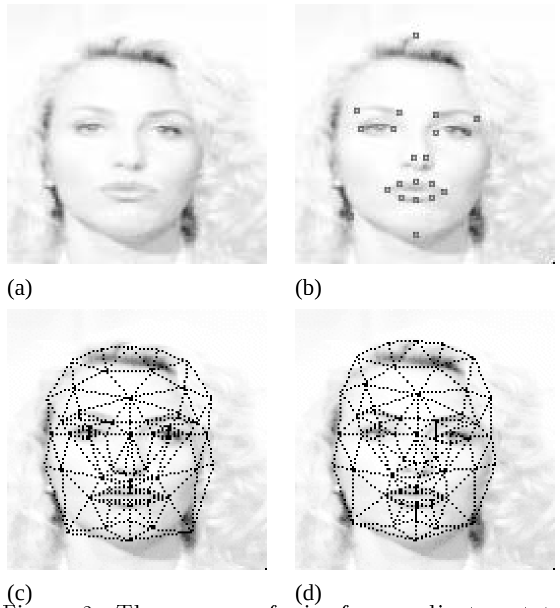

Figure 3: The process of wire-frame adjustment to the actual face dimensions and shape: (a) source image; (b) reference points on the face; (c) image with scaled wire-frame model; (d) adjusted wire-frame model superimposed on the source image.

2.1 Synthetic Model

In Fig.2, two views of the generic 3D wire-frame model consisting of 105 points are reported. It has been adopted as a generic wire-frame model and has been derived from the well known CANDIDE model (76 points) [14] by adding point around the mouth and the nose for improving realism and for providing a correspondence between the points of the model and knots of the STB-snakes.

In order to establish a true correspondence among the face under analysis and the reconstructed model during the phase of animation the generic wire-frame model must be adjusted. To this end, a procedure to adapt size and shape of the wire-frame facial model to that of the person in front of the camera has been derived and used. It is based on elastic deformations of the model, [3], and has been derived from that presented in [13]. The adjustment confers a high realism to the phase of animation of the synthetic model, even if the final face model with pattern could be improved considering also the side views of the face under analysis. The process of adjusting is summa-

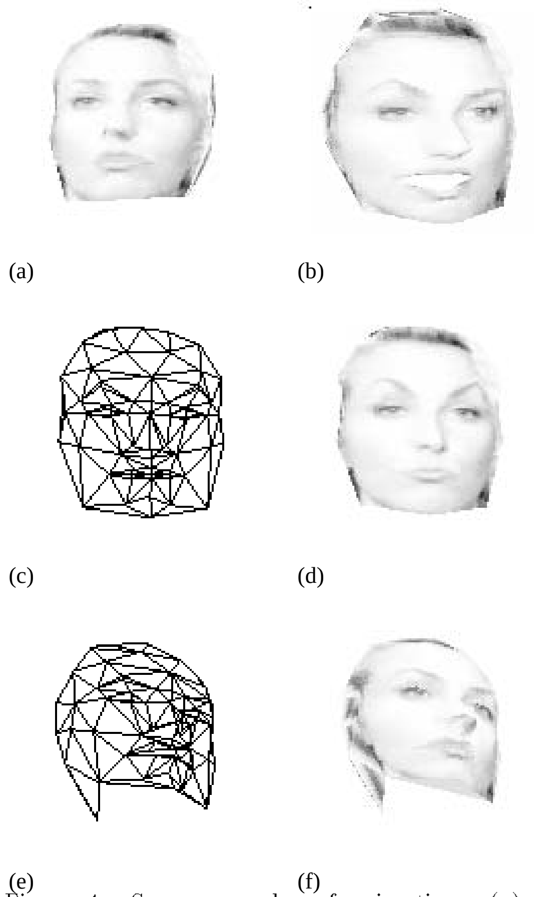

Figure 4: Some examples of animation: (a) rotate model; (b) deformed and rotated model; (c)-(d) wire-frame and patterned models rotated and deformed (see eyebrows); (e)-(f) wire-frame and patterned models rotated and deformed (see eyebrows and mouth).

rized in Fig.3 The process of adjustment proceeds as follows. Firstly some reference points (corresponding to the most important face features points and to vertices of the wire-frame model) must be marked. On the basis of these references the generic wire-frame model of the face is scaled and then, through an elastic process, the model is adjusted with respect to the frontal image. The adjustment is driven by means of an iterative process in which the marked points play the role of attractors and their forces are propagated by using a Gaussian distribution through the edges of the mesh.

Once the process of adjustment is finished the pattern of the frontal image (called source image) is smeared on the synthetic model. It should be noted that the smearing must be repeated each time the synthetic model is rotated, translated or deformed (at least for the triangles that have been changed). In addition, an algorithm for removing the hidden lines and surfaces has been defined, otherwise wrong visualisations are obtained. All these things must be done very fast if the real-time motion tracking and synthetic reconstruction is required.

3 Experimental Results

The technique proposed for motion tracking of face features and synthesis of estimated deformations on a patterned 3D model has been tested on several real image sequences produced by distinct people. The final application of our methods is the very long-term tracking of faces for video-conference, videophone and cinema.

In Fig.5 some image frames of a sequence where a man is opening his mouth are reported together with the corresponding synthetic reproductions. The snakes estimated have been superimposed on the source images. In Fig.6 some images selected from a sequence where a man is moving his eyebrows are reported. It should be noted that, as a side effect, he had opened also his eyes. On the contrary, on the right of the same figures in the images reporting the synthetic reproductions the eyes are static, since in this case only the eyebrows have been tracked. In Fig.7 some examples of synthesized images obtained by using the deformations estimated from the sequence of Fig.5. Some of these synthetic images has been obtained rotating in several directions the synthetic model and/or by assigning the deformations estimated from the sequence of Fig.5 to a different model (in particular model of a female). Therefore, in our system is also possible to assign the motion of a face to the structure of another. Moreover, the global motions and the deformations estimated can be integrated by global motions and deformations coming through keyboard or other means. This opens the way for applications of virtual reality and cinema – e.g., a synthetic actor can be animated by using the mimic of another actor.

Our experiments have demonstrated that the approach proposed for the estimation of face deformations is quite robust with respect to noise, and that the approach proposed is suitable to track face motions without limit of time. Hence, it can be profitably used in non-controlled environments to perform motion tracking in real applications of long-term motion analysis as videophone, videoconference, etc.

The system proposed for tracking facial features differs from other systems presented in the literature since it adopts a specific energy model and is computationally lighter. This is due to the STB-snake model and to the mathematical technique adopted for solving the system of non-linear equations used for estimating the minimum of the functional expressing the total energy (1).

As a result, the algorithm proposed for motion tracking is computationally very efficient. In fact, our system is capable of tracking a mouth or an eyebrow with 12 images per second (10-15 iterations per frame) on a 486 DX 33 Mhz. Also the algorithm for image reconstruction is very fast, it is capable of producing 22 images (faces) per second having a maximum resolution of 128×128 pixels, reproducing rotations, translations, zooming, and deformations on a 486 DX 33 Mhz. Therefore, a quasi real-time head/face motion tracking has been obtained with low-cost architectures.

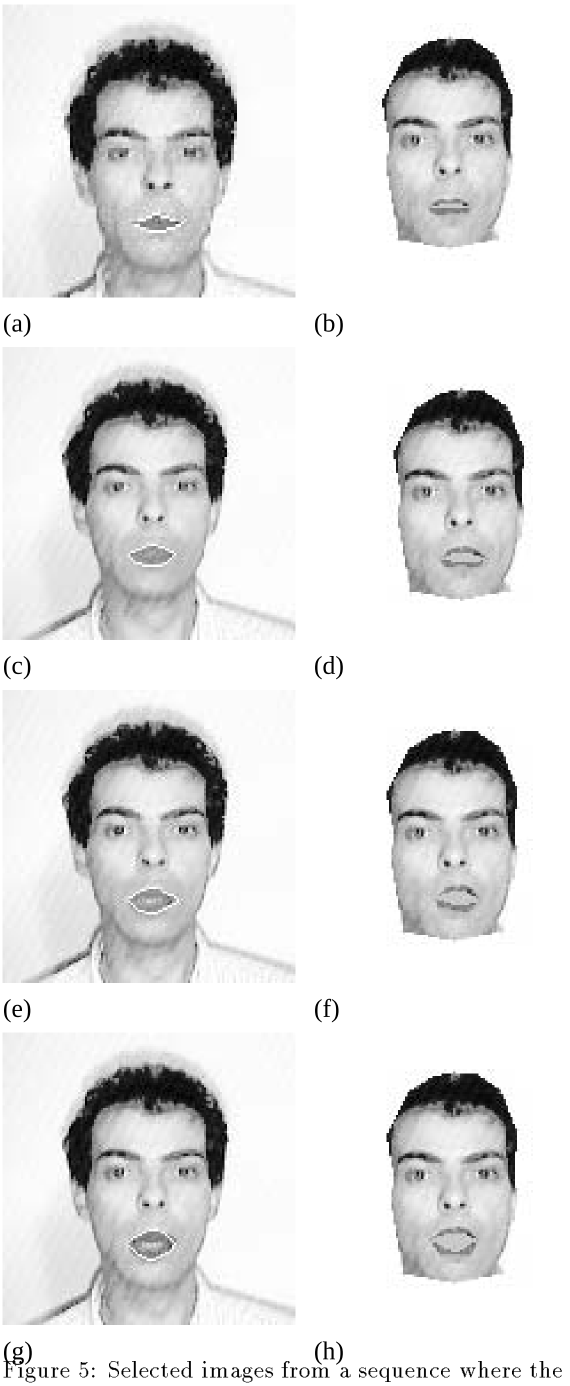

Figure 5: Selected images from a sequence where the face under analysis is opening the mouth: (a), (c), (e), (g) original images; (b), (d), (f), (h) faces synthesised by using the patterned wire-frame model with estimated deformations.

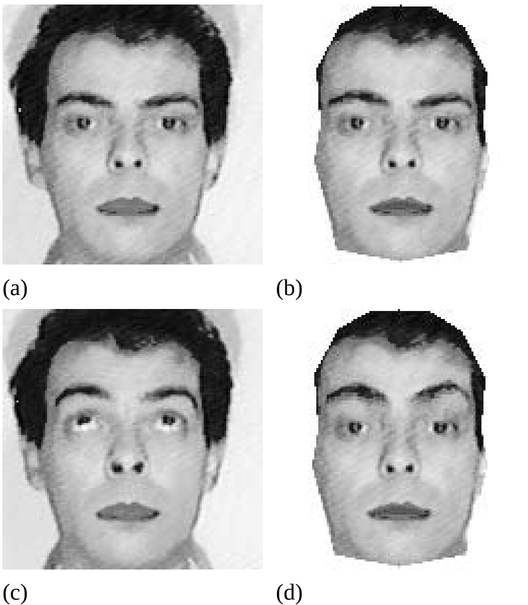

Figure 6: Selected images from a sequence where the face under analysis is moving the eyebrows: (a), (c) original images; (b), (d) the faces synthesized by using the patterned wire-frame model with estimated deformations. Note that in the reconstructed images the eyes are stationary.

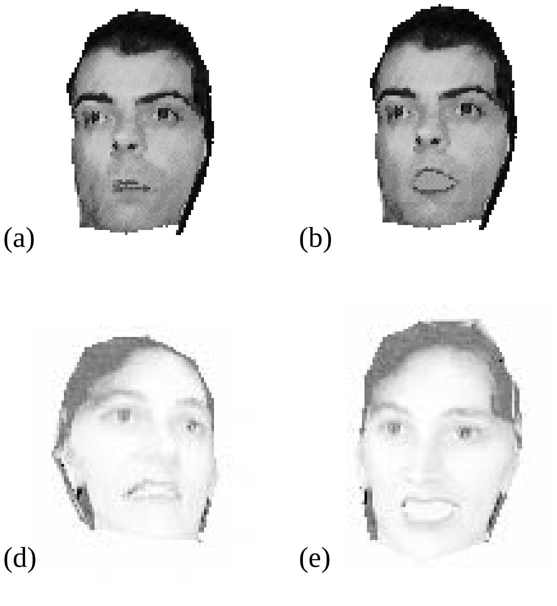

Figure 7: Synthesized images by using the image sequence reported in the previous figure: (a) rotation of the synthetic model of Fig.5(d); (b) rotation of the synthetic model of Fig.5(h); (c) synthetic model obtained by using a different wire-frame model and pattern, and the deformations estimated on image Fig.5(a); (d) synthetic model obtained by using a different wire-frame model and pattern and the deformations estimated on image in Fig.5(e).

4 Conclusions

A complete and integrated system for tracking face deformations and corresponding synthetical reproduction has been presented. The motion estimation process has been based on spatio-temporal B-spline for modeling curves associated with the face features that must be tracked. In addition, an algorithm capable for adapting the generic 3D wire-frame face model to the face under analysis has been used. This has conferred a high realism to the simulations of face motions on the reconstructed faces. Experiments have demonstrated that the approach proposed is robust with respect to noise. The system proposed differs from others presented in the literature since it adopts a specific energy model for avoiding spline collapsing and it is computationally lighter being based on STB-snake and an *ad hoc* numerical method. Therefore, it can be profitably used in non-controlled environments where robust and the fast computations are mandatory, such as for videophone, video-conference, etc.

Acknowledgments

The authors would like to thank Prof. G. Bucci for his suggestions, and Prof. R. Forchheimer for the CANDIDE model.

References

[1] A. Blake and A. Yuille. *Active Vision, Proc. of Rank Prize Workshop Grasmere, England, 1991.* MIT Press, Cambridge, MA, 1992.

[2] A. Borri, G. Bucci, and P. Nesi. A robust tracking of 3D motion. In *Proc. of the Europ. Conf. on Comp. Vision*, pp. 181–188, Stockholm, Sweden, 1994.

[3] D. J. Burr. Elastic matching of line drawings. *IEEE Trans. on Patt. Ana. and Mac. Intel.*, 3(6), 1981.

[4] R. Curwen and A. Blake. Dynamic contours: Real-time active splines. In A. Blake and A. Yuille, editors, *Active Vision, Proc. of Rank Prize Workshop Grasmere, England, 1991*, Cambridge, MA, 1992. MIT Press.

[5] K. Fujimura, N. Yokoya, and K. Yamamoto. Motion tracking of deformable objects based on energy minimization using multiscale dynamic programming. In *Proc. of 11th IEEE Int. Conf. on Pat. Recog.*, pp.83–86, 1992.

[6] T. Fukuhara, A. Umahashi, and T. Murakami. 3-d motion estimation for model-based image coding. In *Proc. of the 4th IEE Int. Conf. on Image Proc. and its App.*, pp. 69–72, Maastricht, The Netherlands, 1992.

[7] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. Journal of Computer Vision*, 1(4):321–331.

[8] H. Li, P. Roivainen, and R. Forchheimer. 3-d motion estimation in model-based facial image coding. *IEEE Trans. on Pat. Ana. and Mac. Intel.*, 15(6):545–555, 1993.

[9] K. Mase. An application of optical flow - extraction of facial expression. In *Proc. of IAPR Workshop on Machine Vision App. Tokyo*, pp. 195–198, 1990.

[10] S. Menet, P. Sant-Marc, and G. Medioni. B-snakes: Implementation and application to stereo. In *Proc. of Image Understanding Workshop*, pp. 720–726. Morgan Kaufmann, 1990.

[11] S. Morishima and H. Harashima. Image synthesis and editing system for a multi-media human interface with speaking head. In *Proc. of the 4th IEE Int. Conf. on Image Proc. and its applications*, pp. 270–273, Maastricht, The Netherlands, 1992.

[12] F. I. Parke. Parametrized models for facial animation. *IEEE CG & A*, pp. 61–68, 1982.

[13] M. J. T. Reinders, B. Sankur, and J. C. A. van der Lubbe. Transformation of a general 3D facial model to an actual scene face. In *Proc. of 11th IEEE Int. Conf. on Pat. Rec.*, pp. 75–78, 1992.

[14] M. Rydfalk. Candide, a parametrised face. Technical report, Department of Electrical Engineering, Linköping University, LiTH-ISY-I-0866, Sweden, 1987.

[15] D. Terzopoulos and K. Waters. Analysis of facial images using physical and anatomical models. In *Proc. of 3rd IEEE Int. Conf. on Computer Vision, Osaka, Japan*, pp. 727–732, 1990.

[16] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. on Pat. Ana. and Mac. Intel.*, 15(6):569–579, 1993.

[17] M.-L. Viaud and H. Yahia. Facial animation with muscle and wrinkle simulation. In *Proc. of 2nd Int. Conf. Dedicated to Image Comm.*, pp. 117–121, Bordeaux, France, 1993.

[18] A. Yuille and P. Hallinan. Deformable templates. In A. Blake and A. Yuille, editors, *Active Vision, Proc. of Rank Prize Workshop Grasmere, England, 1991*, Cambridge, MA, 1992.

[19] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *Int. Journal of Computer Vision*, 8(2):99–111, 1992.