# Tracking facial features in video sequences using a deformable model-based approach

Marius Malciu* and Françoise Prêteux*

Institut National des Télécommunications
ARTEMIS Project Unit

## ABSTRACT

This paper addresses the issue of computer vision-based face motion capture as an alternative to physical sensor-based technologies. The proposed method combines deformable template-based tracking of mouth and eyes in arbitrary video sequences with a single speaking person with a global 3D head pose estimation procedure yielding robust initializations. Mathematical principles underlying deformable template matching together with definition and extraction of salient image features are presented. Specifically, interpolating cubic B-splines between the MPEG-4 Face Animation Parameters (FAPs) associated with the mouth and eyes are used as template parameterization. Modeling the template a network of springs interconnecting the mouth and eyes FAPs, the internal energy is expressed as a combination of elastic and symmetry local constraints. The external energy function, which allows to enforce interactions with image data, involves contour, texture and topography properties properly combined within robust potential functions. Template matching is achieved by applying the downhill simplex method for minimizing the global energy cost. Stability and accuracy of the results are discussed on a set of 2000 frames corresponding to 5 video sequences of speaking people.

**Keywords:** Facial motion capture, MPEG-4 face description, deformable template, 3D head pose estimation, monocular video sequences.

## 1. INTRODUCTION

Achieving non-supervised tracking of human faces and facial features in video sequences has motivated intensive research in such application areas as human-machine interaction[1-3], face and facial expression recognition[4-9] and model-based facial image coding[10-17]. Though intuitive for biological vision systems, locating faces and facial components in video sequences remains today a challenging and largely open issue in computer vision. The main encountered difficulties refer to the complexity and high variability of face morphology and head motion, and the lack of universal assumptions on scene structure which often involves arbitrary and complex background together with unknown and variable lighting conditions.

Here, we are concerned with devising a non-supervised motion capture technique dedicated to human faces, in order to enable the automated animation of virtual signing avatars for the development of deaf people oriented telecommunication services. Within this framework, computer vision is believed to provide a relevant alternative to sensor-based motion capture technologies.

The most successful approaches for face motion analysis are model-based techniques[10,12,14,17] in which a 2D/3D face model is fitted onto image data. These techniques involve (i) a global adaptation of the model onto the rigid/affine motion of the entire head, followed by (ii) a local fit to recover the local deformations of facial components. In a previous work[22], we have developed a robust object-based 3D head pose estimation method. Relying on these results, we now tackle the issue of achieving the local registration of some salient facial features by adopting a deformable template-based approach. The main steps involved are the following: 1) devising parameterized geometric models for each facial feature under consideration; 2) extracting salient image features in the regions corresponding to each facial feature; 3) selecting a non-rigid parametric template deformation model; and 4) matching the facial component models onto image features using an optimization procedure w.r.t. the deformation model parameters. Achieving an accurate template matching on arbitrary video sequence of speaking people strongly depends on the image features taken into account.

* Correspondence: Email: {Marius.Malciu,Francoise.Preteux}@int-evry.fr
WWW: www-sim.int-evry.fr/Artemis

Facial components of interest refer to elements characterizing facial expressions such as mouth and eyes. Here, facial component models are defined as cubic B-spline approximations of feature boundaries, and template deformation models consist of parameterized 2D geometric mappings. Significant image features are selected as texture, edge elements and local topography in mouth and eye inner regions, and extracted by means of gradient- and gray level morphological operators. Specifically, a self-dual connection cost operator is applied in order to extract in a reliable manner the local topography whatever the lighting conditions. The matching procedure is achieved by minimizing an energy function combining internal and external terms via the simplex optimization method.[21] The internal template energy is derived from a linear elasticity deformation model, whereas the external template energy is defined by properly combining retained image features. Initial template positioning on each frame of the sequence is performed by compensating the rigid head motion via a 3D global head pose estimation procedure.[22]

This paper is organized as follows. In Section 2, the principles underlying deformable template matching are presented together with the definition and extraction schemes of image features. Specifically, mouth and eye spline-based templates together with their internal and external energy functions are defined. Section 3 deals with template matching on image data. Internal and external template energy are combined within an optimization scheme based on the downhill simplex algorithm. In Section 4, preliminary results obtained on a set of 2000 frames corresponding to 5 video sequences of speaking people are presented and discussed.

## 2. MATCHING DEFORMABLE TEMPLATES ON IMAGE DATA

Deformable template modeling'[19] is a generic model-oriented energy minimization-based approach for solving non-rigid segmentation and matching problems in computer vision. A deformable template is a discrete parametric model that provides an archetypic description of shape properties of a specific class of objects. The ability of templates to model in a compact fashion highly variable objects with multiple parts and complex topologies makes them particularly relevant for face analysis in video sequences, including segmentation, non-rigid motion estimation[10], coding[12,14,15,23], indexing and recognition[4,5,6,9,17].

Specifying a deformable template $\boldsymbol{T}$ requires to define:

1. a discrete parameterized geometry consisting of (i) a set of nodal points and connectivity relationships that control the global shape of the model, and (ii) a family of shape functions that determine the continuity properties at any non-nodal point along the model;
2. an internal energy functional, denoted by $E_{\text{int}}$, that sets *a priori* constraints on the variability of shape properties of the template model;
3. an external energy functional, denoted by $E_{\text{ext}}$, that establishes interaction constraints in order to maintain the consistency between the template geometry and relevant image features.

Template matching is then performed by minimizing the total energy functional $E_{\text{template}}$ defined as a weighted sum of the internal and external energy functionals:

$$E_{\text{template}} = E_{\text{int}} + E_{\text{ext}} \quad .$$

The solution space on which $E_{\text{template}}$ is minimized depends on the problem at hand. Within a segmentation framework, for instance, minimization is carried out with respect to nodal point coordinates. Tracking facial features in a video sequence $(I_n)_n$ refers to a matching problem in which we search for a 2D non-rigid transformation $\hat{\tau}_{n+1}$ within some space of regular mappings that yields an optimal registration $\hat{\boldsymbol{T}}_{n+1} := \hat{\tau}_{n+1}(\boldsymbol{T})$ of the template model $\boldsymbol{T}$ from the reference frame $I_n$ into the next frame $I_{n+1}$. Hence:

$$E_{\text{template}}(\boldsymbol{T}, t, I_n, I_{n+1}) = E_{\text{int}}(\boldsymbol{T}, t) + E_{\text{ext}}(\boldsymbol{T}, t, I_n, I_{n+1}) \quad ,$$

$$\hat{\tau}_{n+1} = \arg\min_{\tau} E_{\text{template}}(\boldsymbol{T}, t, I_n, I_{n+1}) \quad .$$

### 2.1. Eye and mouth deformable templates

In this section, we specify the parameterized geometry of two deformable template models, denoted by $\boldsymbol{T}_{\text{eye}}$ and $\boldsymbol{T}_{\text{mouth}}$, and adapted to eye and mouth, respectively. The discrete geometry of these templates is designed in accordance with the new MPEG-4 Version 1 international standard that provides a complete and normalized description of face and facial components (Figure 1). MPEG-4 face description, which is primarily focused on human body animation, is however too complex for image analysis purpose, in such extent that all facial parameters cannot be associated with stable features in video sequences with arbitrary lighting conditions and camera configurations. Compared to the MPEG-4 specification, the number of degrees of freedom of $\boldsymbol{T}_{\text{eye}}$ and $\boldsymbol{T}_{\text{mouth}}$ is therefore reduced by (i) merging the corner points of inner and outer lips, and (ii) retaining only the inner eye boundaries (Figure 2). The mouth template $\boldsymbol{T}_{\text{mouth}}$ is decomposed as:

$$\boldsymbol{T}_{\text{mouth}} = L_{\text{uu}} \cup L_{\text{ul}} \cup L_{\text{lu}} \cup L_{\text{ll}} \quad ,$$

where $L_{\text{uu}}$ (resp. $L_{\text{ul}}$) denotes the upper (resp. lower) boundary of the upper lip, and $L_{\text{lu}}$ (resp. $L_{\text{ll}}$) denotes the upper (resp. lower) boundary of the lower lip. We further denote by $D_{\text{u}}$ (resp. $D_{\text{l}}$) the upper (resp. lower) lip region, and by $D_{\text{mouth}}$ the (possibly empty) region between the two lips. Similarly, the eye template $\boldsymbol{T}_{\text{eye}}$ is decomposed as:

$$\boldsymbol{T}_{\text{eye}} = E_{\text{u}} \cup E_{\text{l}} \quad ,$$

where $E_{\text{u}}$ (resp. $E_{\text{l}}$) represents the upper (resp. lower) boundary of the eye model. We also denote by $D_{\text{eye}}$ the inner region of $\boldsymbol{T}_{\text{eye}}$ (Figure 2). In the sequel, the exponent notation $X^n$ applied to some template-related entity $X$ (e.g. $L_{\text{uu}}^n$, $D_{\text{u}}^n$) will refer to the corresponding entity of the deformed template $\boldsymbol{T}^{\text{n}}$ in frame $n$.

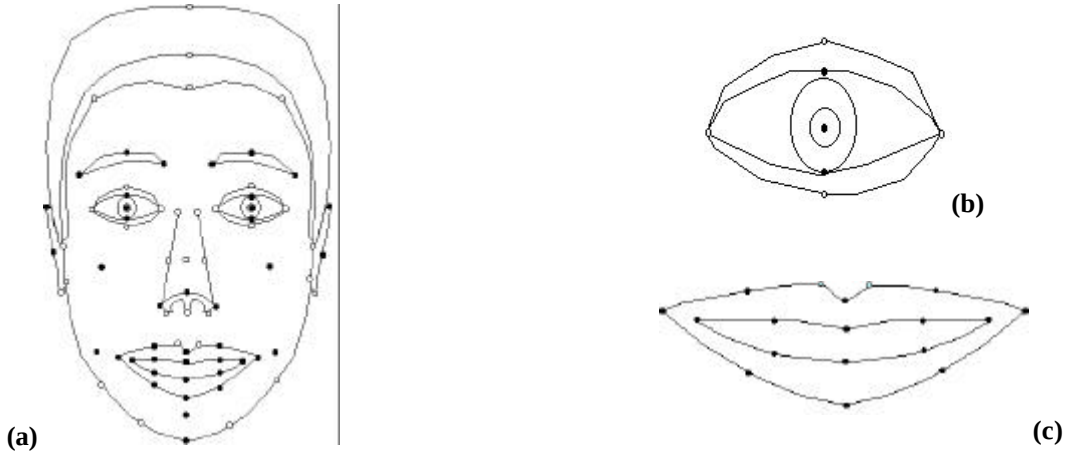



**Figure 1:** MPEG-4 Version 1 standard face description.
**(a)** Face Animation Parameters (FAPs) - **(b)** Eye animation parameters – **(c)** Mouth animation parameters.

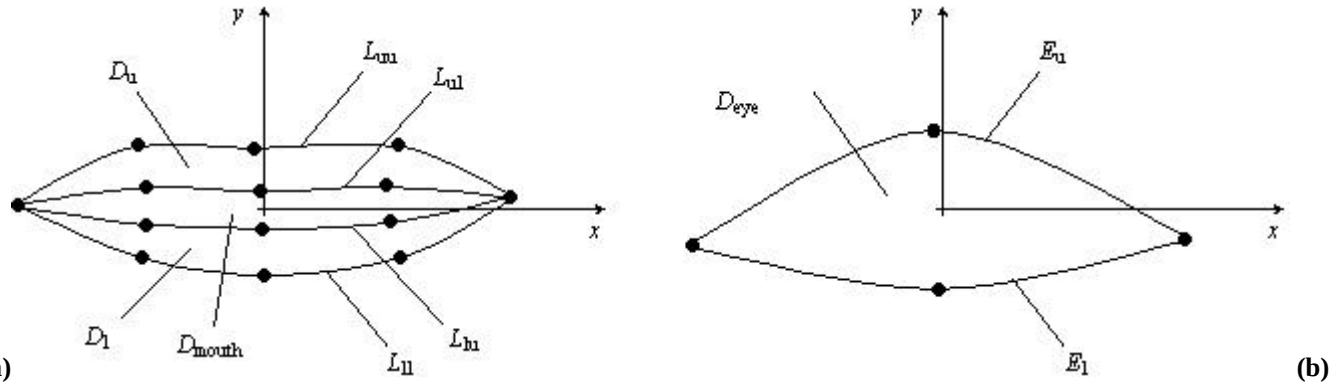



**Figure 2:** Simplified **(a)** mouth template $\boldsymbol{T}_{\text{mouth}}$ and **(b)** eye template $\boldsymbol{T}_{\text{eye}}$.

The selection of an adequate family of shape functions is dictated by accuracy and compactness requirements. The most frequent choice[12,15,18] consists of second order interpolating polynomials, especially parabolic arcs. Nevertheless, quadratic representations prove to be too rigid for accurately dealing with a wide variety of facial expressions, in particular those involving large deformations of the mouth. Harmonic-based modeling[24] has been shown to provide more accurate results, but does not allow an intuitive geometric manipulation of the deformed template. Known as versatile modeling tools in computer graphics[25], spline-based representations have been successfully used for facial component modeling[26]. Following the latter approach, we will make use in this paper of interpolating cubic B-splines along each curvilinear component of the eye and mouth templates, under the following assumptions:

- boundary conditions refer to zero curvature;
- mouth/eye corners are defined as boundary points;
- spline coefficients are computed in a coordinate system with origin at the center of gravity of template points.

### 2.2. Internal energy functionals

The internal energy functionals for $\boldsymbol{T}_{\text{eye}}$ and $\boldsymbol{T}_{\text{mouth}}$ are designed to incorporate rigidity and local symmetry constraints:

- Linear elastic constraints are set by modeling the templates as systems of ideal springs connecting neighboring nodal points along each of their curvilinear components (Figure 3). In the specific case of $\boldsymbol{T}_{\text{mouth}}$, extra transversal elasticity constraints are added between upper and lower boundaries of each lip, by means of ideal strings linking analog non terminal FAPs. Denoting by $l_i$ (resp. $l_i'$) the natural (resp. current) length of spring $i$ with stiffness $k_i$, the internal elastic energy of the deformed template $\tau(\boldsymbol{T})$ is then defined as:

$$E_{\text{elastic}}(\boldsymbol{T},\tau) = \sum_{i\in\tau(\boldsymbol{T})} k_i (l_i - l_i')^2 \quad .$$

  Longitudinal (resp. transversal) springs are assumed to have the same stiffness, denoted by $k_l$ (resp. $k_t$). For the mouth template, we further assume that $k_t \approx 10\,k_l$.

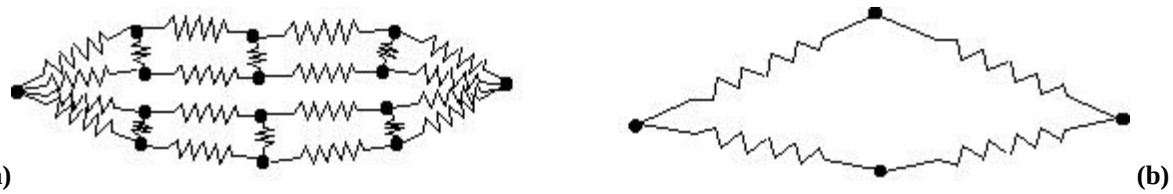
- Local symmetry constraints are added to penalize non uniform distributions of FAPs along each curvilinear component of the templates. More precisely, dealing with the mouth template, the middle points on each half lip boundary are constrained to remain roughly equidistant to their closest neighbors. For the eye template, the middle point along each eye boundary is simply constrained to remain in a central position. Using the notations defined on Figure 4, the internal geometric energy of the deformed templates are then expressed as follows:

$$E_{\text{geometric}}(\boldsymbol{T}_{\text{mouth}},\tau) = \sum_{a,b\in\{\text{u,l}\}} [(l_{ab,0}/l_{ab,1} - 1)^2 + (l_{ab,2}/l_{ab,3} - 1)^2] \quad ,$$

$$E_{\text{geometric}}(\boldsymbol{T}_{\text{eye}},\tau) = \sum_{a\in\{\text{u,l}\}} (l_{a,0}/l_{a,1} - 1)^2$$

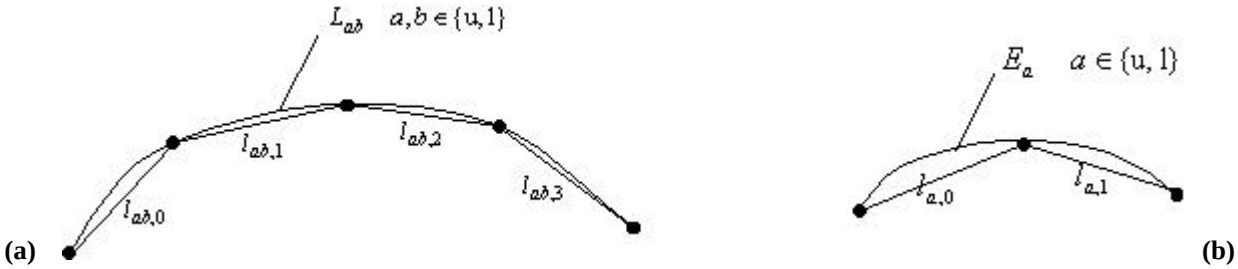
The internal energy of the template is finally defined as a weighted sum of the elastic and geometric energies:

$$E_{\text{int}}(\boldsymbol{T},\tau) = w_{\text{elastic}} E_{\text{eleastic}}(\boldsymbol{T},\tau) + w_{\text{geometric}} E_{\text{geometric}}(\boldsymbol{T},\tau) \quad .$$



**(a)** **(b)**

**Figure 3:** Elastic internal energy model for the **(a)** mouth template $\boldsymbol{T}_{\text{mouth}}$ and **(b)** eye template $\boldsymbol{T}_{\text{eye}}$.




**(a)** **(b)**

**Figure 4:** Local geometry of the **(a)** mouth and **(b)** eye template models.

### 2.3. External energy functionals

External energy functionals are meant to maintain the consistency between the template geometry and relevant image features. The key issue consists therefore in defining and accurately estimating salient and stable mouth / eyes descriptors from the video sequence, and in combining them within potential functions having a few number of local minima. We propose to combine three information sources related to the previous frame $I_n$ and current frame $I_{n+1}$:

- The edge information related to the external boundaries of the upper and lower lips, and to the eye contours prove to be stable w.r.t. varying lighting conditions and skin deformations. We therefore compute a Euclidean edge map $\|\nabla I_{n+1}\|_2$ using the Canny-Deriche operator.
- Texture information related to the lips can be taken into account provided that lighting conditions remain stationary. In this case, image luminance can be used directly. In the case of eyes, texture information cannot be exploited because of luminance fluctuation due to fast eye motions and blinks.
- Non local topographic information, expressed in terms of luminance peaks and valleys, is relevant for describing mouth and eyes regions[18]. Indeed, the luminance distribution within mouth interior consists of bright and dark zones corresponding to teeth (if visible) and lip folds, respectively. In addition, eyes appear as small dark regions (iris and pupil) partially or completely surrounded by bright areas (eye white) over a gray background (skin). Peak and valley information can be incorporated within a single potential function, denoted by VP( $I_{n+1}$), by applying the connection cost operator[20] to the original and negated image with respect to the boundaries of rectangular patches surrounding the mouth and eyes regions (see Appendix). These patches are automatically propagated from the previous frame $I_n$ into the current frame $I_{n+1}$ using a previously developed 3D head pose estimation procedure[22].

Assuming that template matching has been achieved in frame $I_n$, the external energy of the mouth template is specified as a weighted sum of edge, texture and topography energies defined as follows:

$$E_{\text{ext}}^{\text{mouth}} = w_{\text{texture}}^{\text{mouth}} E_{\text{texture}}^{\text{mouth}} + w_{\text{edge}}^{\text{mouth}} E_{\text{edge}}^{\text{mouth}} + w_{\text{topography}}^{\text{mouth}} E_{\text{topography}}^{\text{mouth}} \quad ,$$

$$E_{\text{texture}}^{\text{mouth}}(\boldsymbol{T}, t, I_n, I_{n+1}) = \sum_{a \in \{u,l\}} \frac{1}{\left|D_a^n\right|} \iint_{D_a^n} \left[I_n - I_{n+1} \circ \tau\right]^2 (\vec{x})\, dxdy \quad ,$$

$$E_{\text{edge}}^{\text{mouth}}(\boldsymbol{T}, t, I_n, I_{n+1}) = \sum_{a \in \{u,l\}} -\frac{1}{\left|\tau(L_{aa}^n)\right|} \int_{\tau(L_{aa}^n)} \left\|\nabla I_{n+1}(\vec{x})\right\| ds \quad ,$$

$$E_{\text{topography}}^{\text{mouth}}(\boldsymbol{T}, t, I_n, I_{n+1}) = -\frac{1}{\left|\tau(D_{\text{mouth}}^n)\right|} \iint_{\tau(D_{\text{mouth}}^n)} \text{VP}(I_{n+1})(\vec{x})\, dxdy \quad .$$

Here, $\left|D_a^n\right|$ (resp. $\left|L_{aa}^n\right|$) denotes the area (resp. length) of the domain $D_a^n$ (resp. the curve $L_{aa}^n$). In a similar fashion, the external energy for the eye template is defined as:

$$E_{\text{ext}}^{\text{eye}} = w_{\text{edge}}^{\text{eye}} E_{\text{edge}}^{\text{eye}} + w_{\text{topography}}^{\text{eye}} E_{\text{topography}}^{\text{eye}} \quad ,$$

where the edge and topographical external energies are related to corresponding energies for the mouth template by replacing $D_{\text{mouth}}^n$ by $D_{\text{eye}}^n$, and $L_{\text{aa}}^n$ by $E_{\text{a}}^n$.

### 2.4. Matching process

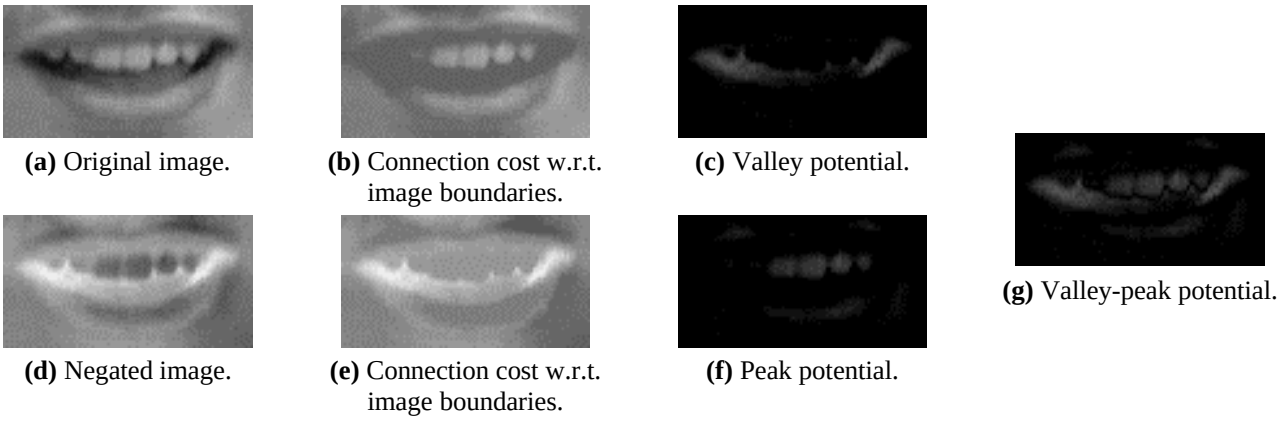
As mentioned above, template matching is achieved by searching for the mapping $t$ minimizing the template total energy $E_{\text{template}}$. In our experiments, this transform is estimated up to its second order derivatives, *i.e.* the mapping $t$ is parametrically defined as a second order polynomial:

$$t(\vec{x}) = \begin{pmatrix} a_0 + a_x x + a_y y + a_{xx} x^2 + a_{xy} xy + a_{yy} y^2 \\ b_0 + b_x x + b_y y + b_{xx} x^2 + b_{xy} xy + b_{yy} y^2 \end{pmatrix} \quad .$$

Assuming that a registered template $\hat{\boldsymbol{T}}_n = \hat{t}_n(\boldsymbol{T})$ is available in frame $I_n$, template matching in frame $I_{n+1}$ is then performed as follows:

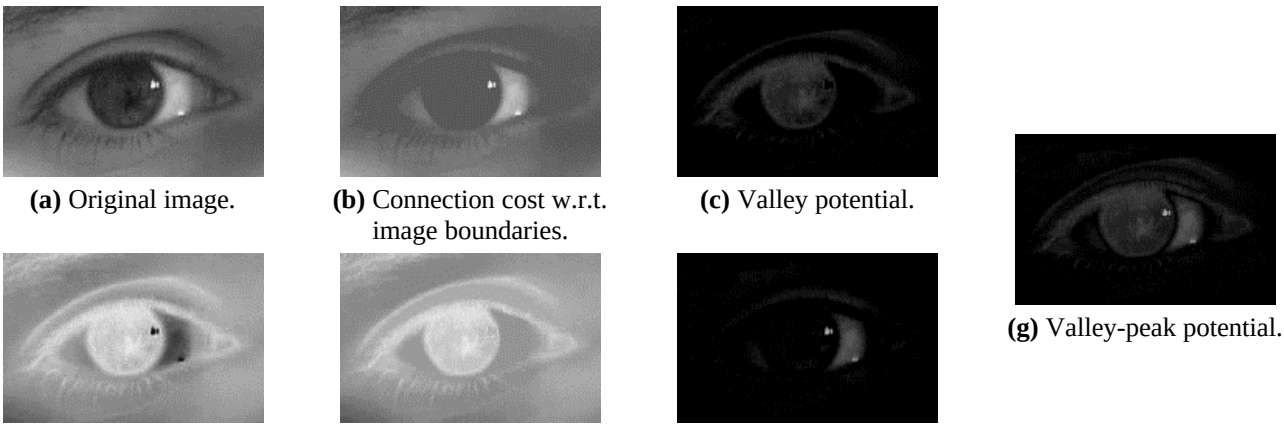
- The affine component of head motion is compensated using a robust 3D global head pose estimation procedure[22] (Figure 7). The template $\boldsymbol{T}$ is initialized in frame $I_{n+1}$ by applying the estimated 3D affine transform to $\hat{\boldsymbol{T}}_n$;
- The optimal template deformation $\hat{t}_{n+1}$ is estimated by minimizing $E_{\text{template}}$ w.r.t. the parameters of the polynomial transform using the downhill simplex method[21]. Simplex iterations are initialized with $a_x = b_y = 1$ and zero for all other coefficients;
- The deformed template $\hat{\boldsymbol{T}}_{n+1} = \hat{t}_{n+1}(\boldsymbol{T})$ is generated by transforming the spline nodal points and computing spline updating accordingly;
- Relaxing the elastic deformation model, by resetting the spring network rest position as $\hat{\boldsymbol{T}}_{n+1}$.

The template nodal points are set interactively in the first frame of the sequence.



**(a)** Original image. **(b)** Connection cost w.r.t. image boundaries. **(c)** Valley potential. **(d)** Negated image. **(e)** Connection cost w.r.t. image boundaries. **(f)** Peak potential. **(g)** Valley-peak potential.

**Figure 5:** Valley-peak potential generation for the mouth template using the double connection cost operator.



**(a)** Original image. **(b)** Connection cost w.r.t. image boundaries. **(c)** Valley potential. **(g)** Valley-peak potential.

**(d)** Negated image. **(e)** Connection cost w.r.t. image boundaries. **(f)** Peak potential.

**Figure 6:** Valley-peak potential generation for the eye template using the double connection cost operator.

## 3. RESULTS

The proposed algorithm has been applied to a small data set consisting of five monocular and non calibrated video sequences, each comprising approximately 500 frames, showing a speaking person with variable facial expressions and head motions. In addition, a training sequence, showing exaggerated facial expressions, has been used for calibrating the template hyper-parameters: spring stiffness coefficients and energy potential weights (Figure 8). The purpose of such a calibration procedure is to set up under-regularized elastic models capable of undergoing large deformations and to empirically establish the relative saliency of the image potentials. Table 1 shows the final settings that have been used in our subsequent experiments.

| Hyper-parameter | Mouth template | Eye template |
|---|---|---|
| $k_l$ | 0.15 | 1.5 |
| $k_t$ | 1,35 | |
| $w_{elastic}$ | 0.9 | 0.5 |
| $w_{geometric}$ | 0.1 | 0.5 |
| $w_{edge}$ | 0.6 | 0.8 |
| $w_{texture}$ | 0.2 | |
| $w_{topographic}$ | 0.2 | 0.2 |

**Table 1:** Experimental hyper-parameter values.

These settings have proven to yield satisfying results (Figure 8) for most of the frames. Several failures, however, have occurred when dealing with degenerated template configurations (*e.g.* temporarily lips vanishing) and for very large template deformations.

On the test sequences (more than 1800 frames), the obtained results demonstrate accurate tracking performances (Figure 9) despite some incidental local attraction defects on some isolated frames overcome by the template on subsequent frames.

**(a)** **(b)** **(c)**

**Figure 7:** Template initialization via model-based global 3D head pose estimation.
**(a)** Video sequence frame and 3D generic head mesh model – **(b)** Affine model registration – **(c)** Rough location of eye and mouth templates.

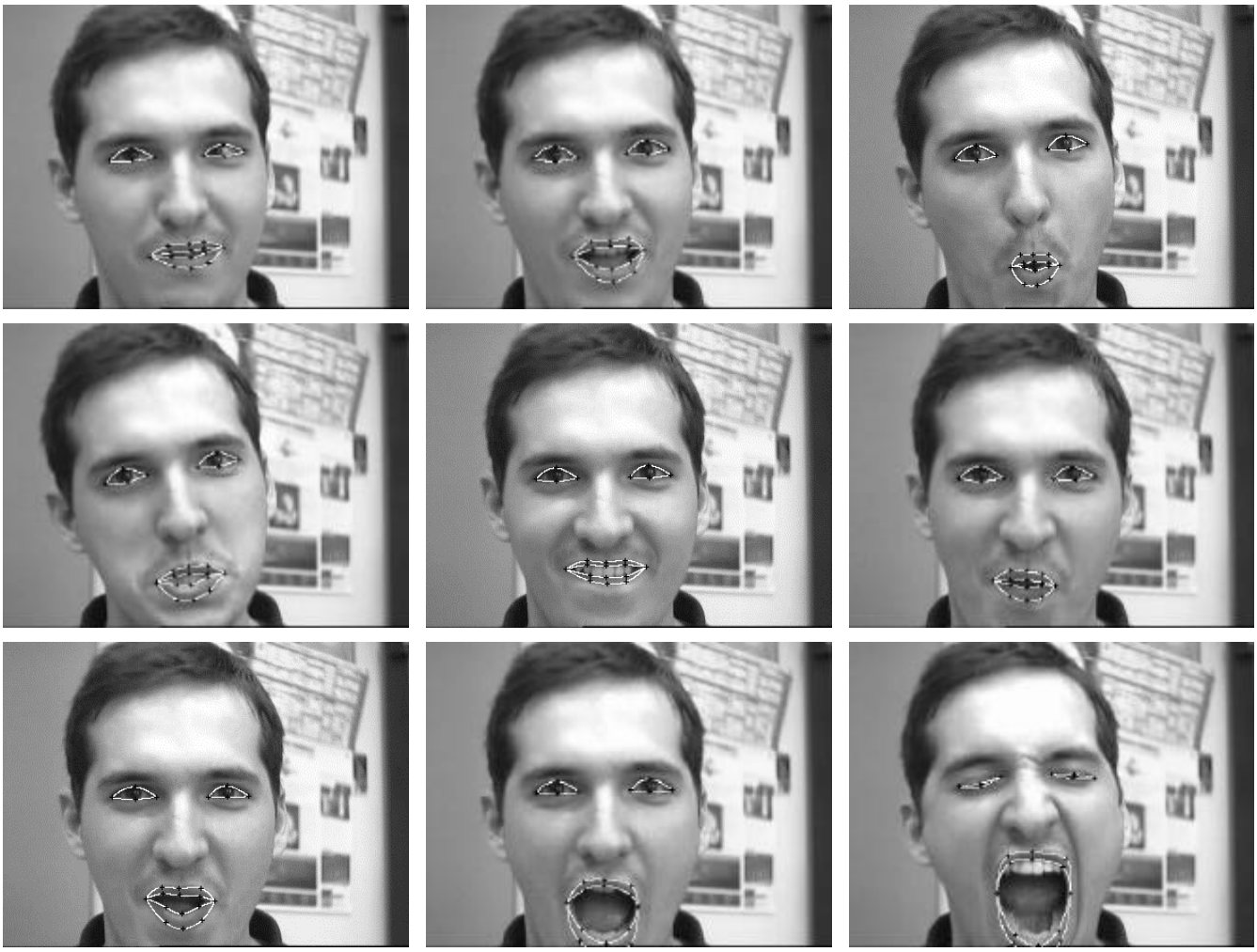


**Figure 8**: Matching of the mouth and eye templates for various facial expressions.

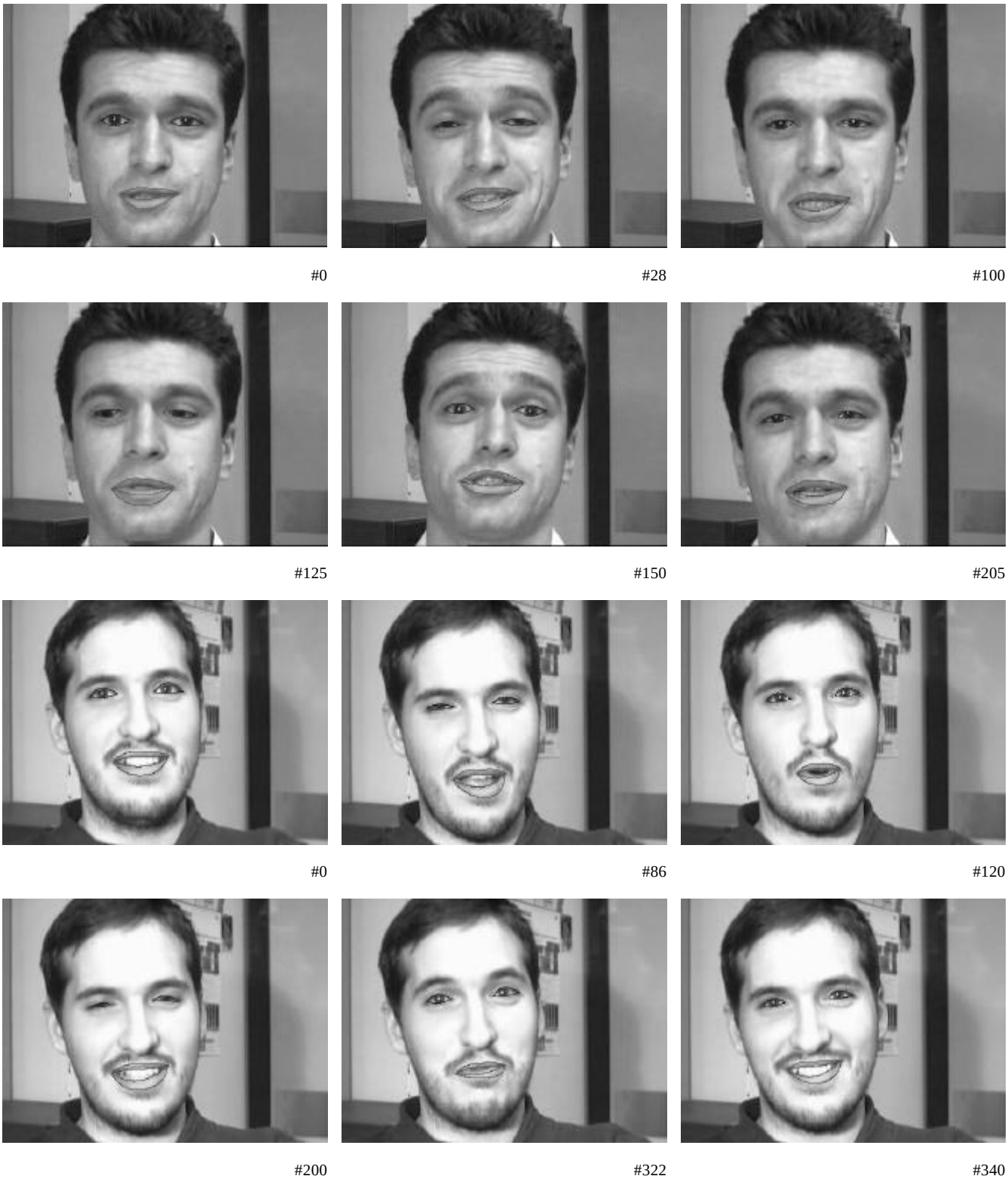



**Figure 9:** Some results of the deformable template matching for eyes and mouth.

## 4. CONCLUSION AND FUTURE WORK

In this paper, we have presented a deformable template-based method for mouth and eyes tracking in non-calibrated monocular video sequences. The external energy of the templates combines texture, contour and topography information. The internal energy is defined as a linear spring network. Template matching is achieved using the downhill simplex method for total energy minimization. Initialization is performed by directly estimating the 3D head pose via a robust technique. Experiments demonstrate that the proposed method is stable and accurate for a variety of facial image sequences.

The future work will deal with vision-based face motion capture and synthesis within an MPEG-4 compliant framework. The target application concerns the automated animation of a realistic signing avatar from video sequences showing signing deaf people.


## ACKNOWLEDGMENTS

The authors wish to thank Catalin Fetita and Tudor Murgan for their help in data set acquisition and Dr Nicolas Rougon for proof reading the manuscript.


## APPENDIX

**Connection cost definition**

Let $l \in \Re$ be a real value and $f : \Re^n \to \Re$ a function of connected support and upper bounded on any bounded subset of its support, denoted by *supp(f)*. The transform:

$$(f, l) \to X_{f,l} = \{x \in \text{supp}(f),\ f(x) \le l\},$$

denotes the *threshold* $X_{f,l}$ of *f* at level λ. The connection cost of two points $x, y \in \Re^n$ with respect to *f* is defined as:

$$\forall (x, y) \in \Re^n \times \Re^n,\ x \neq y,\ C_f(x, y) = \inf\left\{ l \in \Re / \tilde{d}_{f,l}(x, y) < \infty \right\},$$

where $\tilde{d}_{f,l}$ denotes the geodesic distance with respect to $X_{f,l}$. By convention, $C_f(x, x) = -\infty$.

Similarly, the connection cost of a point $x \in \Re^n$ to a non-empty subset $Y \subset \Re^n$ is given by:

$$\forall x \in \Re^n,\ C_f(x, Y) = \begin{cases} \inf\left\{ l \in \Re / \tilde{d}_{f,l}(x, Y) < \infty \right\} & \text{if } x \notin Y \\ -\infty & \text{otherwise} \end{cases}.$$

Intuitively, the connection cost $C_f(., Y)$ results in "filling in" all local valleys of *f*, excepting those marked by Y.

**Topographic potential generation**

In order to extract the topographic information underlying the topographic potential VP, the following procedure based on the connection cost operator is applied to an image patch $P$ including the facial element under consideration:

- The connection cost of $P$ with respect to its borders, denoted by $C_P$, is computed (Figure 10(c)). The difference $C_P - P$ provides a map of the valleys of $P$ (high values are associated with initial local low ones - Figure 10(e));
- the same procedure applied to the negative of $P$, denoted by $\hat{P}$ and defined as $\hat{P} = \left[ \max_{(x,y) \in \text{Supp}P} P(x, y) \right] - P$, where *x* and *y* denote the spatial coordinates, leads to the map of the image peaks (high values are associated with initial local high ones - Figure 10(f)).

Assuming that the mouth template $\boldsymbol{T}_{\text{mouth}}$ has been matched on frame $I_n$, and knowing the global head motion parameters between frames $I_n$ and $I_{n+1}$, a rough position of $\boldsymbol{T}_{\text{mouth}}$ in $I_{n+1}$ is computed. A small image patch $P_{\text{mouth}}^{n+1}$ of frame $I_{n+1}$ containing the new mouth region is then defined, and image valley map $C_{P_{\text{mouth}}^{n+1}} - P_{\text{mouth}}^{n+1}$ and peak map $C_{\hat{P}_{\text{mouth}}^{n+1}} - \hat{P}_{\text{mouth}}^{n+1}$ are computed. The sum of these two measures, yields a potential map, denoted by $\text{VP}_{\text{mouth}}^{n+1}$ and adapted to the description of the

mouth interior location (Figure 5). The same morphological-based processing is applied to the image patch including the eye in the frame $I_{n+1}$, to get the local valley-peak map in the eye region, $VP_{\text{eye}}^{n+1}$ (Figure 6).

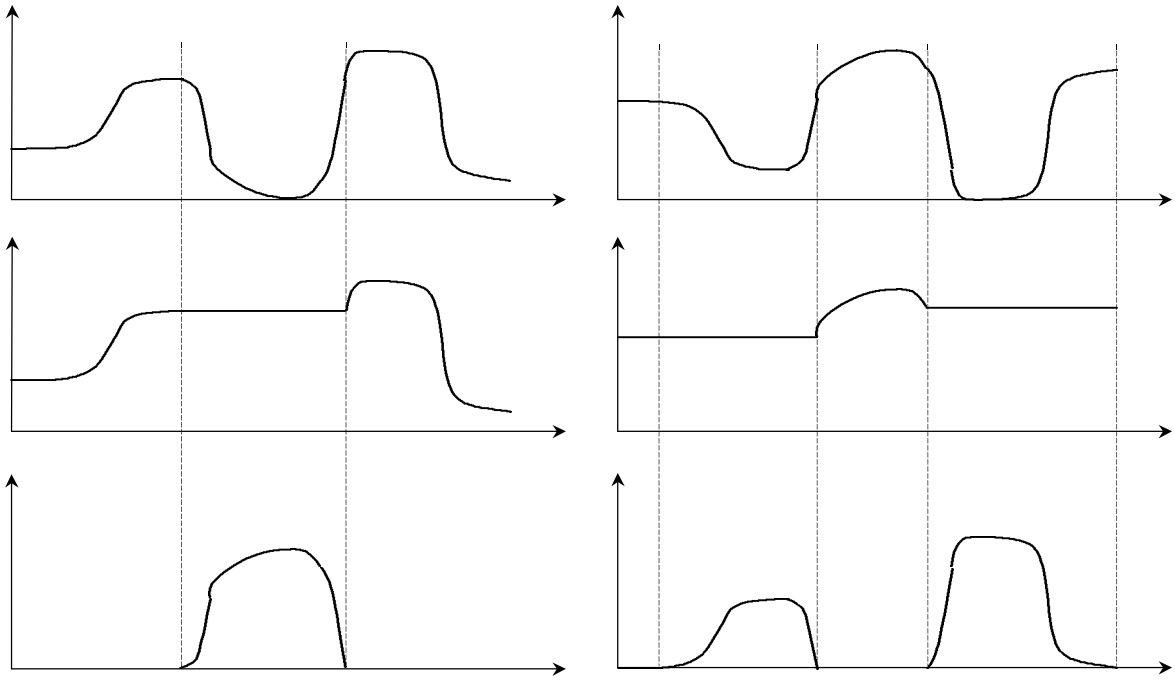


**Figure 10:** Valley / peak extraction.
**(a)** Initial image - **(b)** Negated image of (a) - **(c)** Connection cost of (a) w.r.t. patch boundaries- **(d)** Connection cost of (b) - **(e)** Valley map of (a) - **(f)** Peak map of (a).

## REFERENCES


1. A. Zelinsky, J. Heinzmann, "Human-robot interaction using facial gesture recognition", *Proceedings 5th IEEE International Workshop on Robot and Human Communication (RO-MAN'96)*, pp. 256-61, 1996.
2. H. Kobayashi, F. Hara, "Facial interaction between animated 3D face robot and human beings", *Proceedings IEEE International Conference on Systems, Man and Cybernetics. Computational Cybernetics and Simulation*, pp. 3732-3737, 1997.
3. J. L. Crowley, J. Coutaz, "Vision for man machine interaction", *Proceedings IFIP Working Conference on Engineering for Human-Computer Interaction*, pp. 28-45, 1996.
4. R. Brunelli, T. Poggio, "Face recognition: features vs. templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, **15**(10), pp. 1042-1052.
5. R. Chellappa, C. L. Wilson, S. Sirohey, "Human and machine recognition of faces: A survey", *Proceedings of the IEEE*, **83**(5), pp. 705-740.
6. C. W. Chen, C. L. Huang, "Human face recognition from a single front view", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6,** pp. 571.
7. J. Daugman, "Face and gesture recognition: Overview", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(7), pp. 675-676.

8. C. L. Huang, C. W. Chen, "Human facial feature extraction for face interpretation and recognition", *Pattern Recognition*, **25**(12), pp. 1435-1444.
9. I. A. Essa, A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(7), pp. 757-763, 1997.
10. H. Li, P. Roivainen, R. Forchheimer, "3-D motion estimation in model-based facial image coding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**(6), pp. 545-555, 1993.
11. K. Aizawa, T. S. Huang, "Model-based image coding: advanced video coding techniques for very low bit-rate applications", *Proceedings of the IEEE*, **83**(2), pp. 259-271, 1995.
12. L. Zhang, "Automatic adaptation of a face model using action units for semantic coding of videophone sequences", *IEEE Transactions on Circuits and Systems for Video Technology*, **8**(6), pp. 781-95, 1998.
13. R. Koch, "Dynamic 3-D scene analysis through synthesis Feedback control", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**(6), pp. 556-567, 1993.
14. C. S. Choi, K. Aizawa, H. Harashima, T. Takebe, "Analysis and synthesis of facial image sequences in model-based image coding", *IEEE Transactions on Circuits and Systems for Video Technology*, **4**(3), pp. 257-275, 1994.
15. M. J. T. Reinders, P. J. L. van Beek, B. Sankur, J. C. A. van der Lubbe, "Facial feature localization and adaptation of a generic face model for model-based coding", *Signal Processing: Image Communication*, **7**, pp. 57-74, 1995.
16. G. Bozdagi, A. M. Tekalp, L. Onural, "3D Motion estimation and wireframe adaption including photometric effects for model-based coding of facial image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, **4**(3), pp. 246-256, 1994.
17. P. Eisert, B. Girod, "Analysing facial expression for virtual conferincing", *IEEE Computer Graphics & Applications: Special Issue: Computer Animation for Virtual Humans*, **18**(5), pp. 70-78, 1998.
18. A. L. Yuille, P. W. Hallinan, D. S. Cohen, “Feature extraction from faces using deformable templates”, *International Journal of Computer Vision*, **8**(2), pp. 99-111, 1992.
19. A. Yuille, P. Hallinan, "Deformable templates" in A. Blake, A. Yuille A. Eds., *Active vision*, MIT Press, Cambridge, MA, pp. 21-38, 1992.
20. F. Prêteux, “On a distance function approach for grey-level mathematical morphology”, in E. R. Dougherty Eds., *Mathematical Morphology in Image Processing*, M. Dekker, 1992.
21. W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1998.
22. M. Malciu, F. Preteux, “A robust model-based approach for 3D head tracking in video sequences *Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG’2000), Grenoble, France*, pp. 169-174, 2000.
23. K. N. Ngan , R. L. Rudianto, “Automatic face location detection and tracking for model-based video coding”, *Proceedings Third International Conference on Signal Processing*, pp. 1098-1101, 1996.
24. B. Leroy, I. L. Herlin, “Un modèle déformable paramétrique pour la reconnaissance de visages et le suivi du , pp. 701-704, 1995.
25. R. H. Bartels, J. C. Beatty, *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*, Los Altos, CA, M. Kaufmann, 1987.
26. Y. Moses, D. Reynard, A. Blake, “Determining facial expression in real time”, *International Workshop on Automatic Face and Gesture Recognition (FG’95)*, pp. 332-337, 1995.
27. *Audio and video object coding*, MPEG-4 ISO/IEC 14496-1.