Robust tracking of stochastic deformable models in long image sequences

Charles Kervrann and Fabrice Heitz

IRISA/INRIA, Campus Universitaire de Beaulieu,
35042 Rennes Cedex, France
E-mail: kervrann@irisa.fr, heitz@irisa.fr

1st IEEE International Conference on Image Processing
November 1994, Austin Texas, USA

ROBUST TRACKING OF STOCHASTIC DEFORMABLE MODELS IN LONG IMAGE SEQUENCES

Charles Kervrann and Fabrice Heitz

IRISA/INRIA, Campus Universitaire de Beaulieu,
35042 Rennes Cedex, France
E-mail: kervrann@irisa.fr, heitz@irisa.fr

ABSTRACT

In this paper, we describe a method for the temporal tracking of stochastic deformable models in long image sequences. The object representation relies on a hierarchical statistical description of the deformations applied to a template. A bayesian estimate of the deformations is obtained by maximizing a highly non-linear joint probability distribution. Time consuming global (stochastic) optimization techniques are necessary to obtain optimal solutions unless a good initial guess is available. A good initialization is provided by a recursive temporal filtering of the parameters of the deformable template, combined with a detection of abrupt changes. This procedure yields robust segmentations and enables to track reliably complex deformable structures as is demonstrated here on real-world image sequences showing hand and mouth movements.

1. INTRODUCTION

Many natural objects undergo nonrigid (deformable) movements that are subject to specific constraints. In the case of Fig. 2 for instance, deformations are the consequence of articulated motion combined with soft tissue motion. In [9], we have introduced a statistical model able to constrain the *a priori* structure and the deformations of shapes. The approach relies on the description of both global and local shape deformations. A modal decomposition based on the Karhunen Loeve (KL) transform, introduced by Cootes *et al.* in [4], provides a compact representation of global deformations. Five to ten parameters are sufficient to obtain an accurate description. Local deformations are considered as local stochastic perturbations and are modeled by a first-order Markov process [6]; they capture details that the global abstraction misses. A Maximum A Posteriori (MAP) estimate of the deformations is obtained by maximizing a highly non-linear joint probability distribution describing the interactions between observations (spatial or temporal gradients extracted from the image sequence) and the deformation process [9]. Global (stochastic) optimization techniques are necessary to obtain optimal solutions (which do not depend on the initial configuration of the model), but remain generally computationally demanding [6].

A good initialization may be provided by a temporal tracking of the model which enables to resort to fast local (deterministic) optimization procedures. Kalman filtering techniques have been widely used in computer vision for tracking various image features (points [1], edges [5], regions [11]). The case of deformable structures has been handled recently [2, 3, 13] by considering contours, characteristic points or regions as features for the Kalman filter governing equations. In this paper, we develop a Kalman filter which performs prediction and filtering of the *global deformation parameters* of the deformable model introduced in [9]. A detection of abrupt changes is combined with the temporal filtering to ensure a robust tracking of the model.

2. STOCHASTIC DEFORMABLE MODEL-BASED MOTION SEGMENTATION

The stochastic deformable model under concern has been described in [9] and [8]. The object of interest is represented by the point distribution of a "deformable template" which incorporates *a priori* knowledge on the structure of the object and on its variability. To represent global deformations of the shape, a modal analysis technique introduced by Cootes *el al.* [4] is adopted. A particular shape $\mathbf{X}$ is represented by a set of n labeled points which approximate its outline. The global deformations of shape $\mathbf{X}$ are characterized by a displacement vector $\mathbf{dX}$ with respect to a pre-computed mean shape ("template") $\mathbf{X}^\star$ [4, 9]. A KL expansion of the displacement vectors observed on a representative population allows to obtain a good approximation for the actual deformations on a low dimension space. If $\mathbf{P}$ designates the matrix of the m unit eigenvectors corresponding to the m largest eigenvalues, and if $\mathbf{b}$ denotes the vector $(m \times 1)$ corresponding to the m most significant deformation modes, the deformable template is represented by the following model:

$$\mathbf{X} \;=\; \mathbf{M}(k,\theta)\,[\mathbf{X}^\star + \mathbf{P}\mathbf{b}] \;+\; \mathbf{T}. \qquad (1)$$

Global transformations from the similarity group (rotation of angle θ, scale change by a factor k and translation $\mathbf{T}$) are taken into account in this model (Equ. 1). This representation is refined by introducing additional local deformation vectors $\mathbf{t}$ which are considered as random perturbations on the location of the points belonging to the

globally deformed pattern [9]. Local deformations $\mathbf{t}$ are assumed to follow a first-order zero-mean Gauss-Markov random process with covariance matrix $\mathbf{R}$. The complete model (denoted $\mathbf{Y}$) becomes:

$$\mathbf{Y} \;=\; \mathbf{M}(k,\theta)\,[\mathbf{X}^{\star} + \mathbf{P}\mathbf{b}] \;+\; \mathbf{T} \;+ \mathbf{t}. \qquad (2)$$

This model has been used in [9] to extract moving objects from images sequences in the case where the camera is static. Let $\mathbf{O}$ designate an observation field related to spatial or temporal gradients. The Maximum A Posteriori (MAP) estimate of the deformable template is defined by:

$$\mathbf{Y}_{opt} \;\triangleq\; \arg\max_{\mathbf{Y}}\; p(\mathbf{O}|\mathbf{Y})p(\mathbf{Y}). \qquad (3)$$

The prior distribution of $\mathbf{Y}$ is a Gauss-Markov distribution according to the assumption on $\mathbf{t}$. A (Gibbs) distribution for $p(\mathbf{O}|\mathbf{Y})$ describes the statistical interactions between the observations and the deformations to estimate. In [9] for instance, binary-valued observations corresponding to thresholded temporal gradients have been used for the extraction of moving objects. The distribution $p(\mathbf{O}|\mathbf{Y})$ is specified in this case to enclose moving points within the inner region R^- of the deformable model and to reject static points belonging to the background (denoted R^+) outside the outline of the model [9], yielding the following optimal estimate:

$$\begin{aligned}\mathbf{Y}_{opt} \;=\; \arg\max_{\mathbf{Y}}\; & \frac{1}{C}\, exp[-\frac{1}{2}\,(\mathbf{Y}-\mathbf{X})^{\,T}\,\mathbf{R}^{-1}\,(\mathbf{Y}-\mathbf{X}) \\ & - \sum_{s\,\in\,R^-} |\mathbf{O}(s)-1\,| - \sum_{s\,\in\,R^+} |\mathbf{O}(s)-0\,|\;\;]. \end{aligned} \qquad (4)$$

This estimate has been considered here for the extraction of the moving hands presented in Fig. 2, 3 and 4. A different form for distribution $p(\mathbf{O}|\mathbf{Y})$ has been adopted for the segmentation of the mouth (Fig. 5). The distribution $p(\mathbf{O}|\mathbf{Y})$ specified for the mouth tends to attract the deformable template toward salient features of the image corresponding to large spatial gradients [14]. The global distribution (Equ.4) is a highly non linear function of the model parameters $\mathbf{M}(k,\theta)$, $\mathbf{T}$ and $\mathbf{b}$. When no initial guess is available for these parameters the computation of the MAP estimate requires global optimization techniques [9]. In this case there is no need for a precise initial estimate of the model (no human interaction) but this comes at great cost in computational complexity. Typically 6mn cpu time are necessary to process a 256x256 frame on a workstation. This problem may be alleviated by using the temporal coherence of the movement of the deformable structure to provide good initial estimates from one frame to the next, as explained in Section 3.

3. ROBUST TRACKING OF STOCHASTIC DEFORMABLE MODELS

3.1. Dynamic estimation using Kalman filters

Tracking the deformable structure over a long image sequence [3, 12] reduces significantly the computational cost of the segmentation method and enables to process large movements more reliably. Applying the Kalman filtering theory, we construct a recursive estimator of the stochastic model parameters. This provides valuable information about the global dynamic behaviour of the deformable structure over time which might be used for interpretation purposes. Measurements are obtained by a fast local optimization procedure which fine tunes the predicted parameters to match the observations of the current frame. This avoids the time consuming global optimization step. The dynamical model relies on a second order Taylor expansion of vector $\mathbf{X}(t)$ and $\dot{\mathbf{X}}(t)$ ($\mathbf{X}(t)$ denotes the configuration of the deformable template at time t):

$$\mathbf{X}(t+\delta t) \;=\; \mathbf{X}(t) \;+\; \delta t\,\dot{\mathbf{X}}(t) \;+\; \frac{\delta t^2}{2}\ddot{\mathbf{X}}(t) \qquad (5)$$

(and similar equations for $\dot{\mathbf{X}}(t+\delta t)$). The corresponding dynamics of the global deformation parameters are easily derived from Equ. 1 and Equ. 5:

$$\left\{\begin{array}{lcl} \mathbf{M}(t+\delta t) & = & \mathbf{M}(t) \;+\; \delta t\dot{\mathbf{M}}(t) \;+\; \frac{\delta t^2}{2}\ddot{\mathbf{M}}(t), \\ \mathbf{T}(t+\delta t) & = & \mathbf{T}(t) \;+\; \delta t\dot{\mathbf{T}}(t) \;+\; \frac{\delta t^2}{2}\ddot{\mathbf{T}}(t), \\ \mathbf{b}(t+\delta t) & = & \mathbf{b}(t) + \delta t\dot{\mathbf{b}}(t), \\ & & + \;\frac{\delta t^2}{2}\,\mathbf{\Phi}(t+\delta t)\,[\mathbf{M}(t)\mathbf{P}\ddot{\mathbf{b}}(t) - \delta t\ddot{\mathbf{M}}(t)\mathbf{P}\dot{\mathbf{b}}(t)] \end{array}\right. \qquad (6)$$

where

$$\mathbf{\Phi}(t+\delta t) = [\,(\mathbf{M}(t+\delta t)\mathbf{P})^T(\mathbf{M}(t+\delta t)\mathbf{P})\,]^{-1}(\mathbf{M}(t+\delta t)\mathbf{P})^T \qquad (7)$$

(and similar equations for the parameters derivatives $\dot{\mathbf{M}}(t)$, $\dot{\mathbf{T}}(t)$ and $\dot{\mathbf{b}}(t)$). Based on these dynamics, we can derive the Kalman filter equations associated to the global parameters $\mathbf{M}(t)$, $\mathbf{T}(t)$ and $\mathbf{b}(t)$. Two different filters (a constant velocity and an instantaneous velocity filter) are described in section 3.2 and 3.3. We first recall the general equations of the Kalman filter, that will be used in the following.

Kalman Filter

The dynamical evolution of the system is described by:

$$\mathbf{x}(t+\delta t) \;=\; \mathbf{A}(t)\mathbf{x}(t) \;+\; \mathbf{B}(t)\mathbf{u}(t) \;+\; \xi(t) \qquad (8)$$

where $\mathbf{x}(t)$ is the state vector, $\mathbf{u}(t)$ denotes a deterministic input, $\mathbf{A}(t)$ is the state transition matrix and $\xi(t)$ is a white Gaussian noise. The observations $\mathbf{y}(t)$ are a linear function of the state vector:

$$\mathbf{y}(t) \;=\; \mathbf{H}(t)\mathbf{x}(t) \;+\; \eta(t) \qquad (9)$$

with a noise term $\eta(t)$, modeled as white Gaussian noise (in our case $\mathbf{H}(t)$ is just the identity matrix $\mathbf{I}$).

If $\hat{\mathbf{x}}(t_2|\,t_1)$ denotes the minimum mean square error estimate of $\mathbf{x}(t_2)$ given the measurements up to t_1 ($t_1 \le t_2$) and $\mathbf{P}(t_2|\,t_1)$ the associated error covariance matrix, the Kalman filter correction equations are:

$$\hat{\mathbf{x}}(t|\,t) \;=\; \hat{\mathbf{x}}(t|\,t-\delta t) + \mathbf{K}(t)\,[\mathbf{y}(t) - \mathbf{H}\,\hat{\mathbf{x}}(t|\,t-\delta t)\,], \qquad (10)$$

$$\mathbf{P}(t|\,t) \;=\; [\,\mathbf{I} \;-\; \mathbf{K}(t)\,\mathbf{H}\,]\;\mathbf{P}(t|\,t-\delta t) \qquad (11)$$

where the Kalman gain is defined as:

$$\mathbf{K}(t) \;=\; \mathbf{P}(t|\,t-\delta t)\;\mathbf{H}^T\;[\,\mathbf{H}\;\mathbf{P}(t|\,t-\delta t)\;\mathbf{H}^T \;+\; \mathbf{V}(t)\,]^{-1}, \tag{12}$$

where:

$$\mathbf{V}(t) = \quad \mathbb{E}\,[\eta(t)\;\eta(t)^T] \;\;;\;\; \mathbb{E}\,[\eta(t)\,] \;=\; 0, \tag{13}$$

and $\mathbb{E}[.]$ stands for the expectation (in our case the covariance matrix $\mathbf{V}(t)$ is $\sigma_\eta^2\mathbf{I}$).

The prediction equations are:

$$\hat{\mathbf{x}}(t+\delta t|\,t) \;=\; \mathbf{A}(t)\;\hat{\mathbf{x}}(t|\,t) + \mathbf{B}(t)\mathbf{u}(t), \tag{14}$$

$$\mathbf{P}(t+\delta t|\,t) \;=\; \mathbf{A}(t)\;\mathbf{P}(t|\,t)\mathbf{A}(t)^T \;+\; \mathbf{Q}(t) \tag{15}$$

where

$$\mathbf{Q}(t) \;=\; \mathbb{E}\,[\xi(t)\;\xi(t)^T] \;\;;\;\; \mathbb{E}\,[\xi(t)\,] \;=\; 0. \tag{16}$$

The error vector $\xi(t)$ is modeled as a zero mean white Gaussian noise with known variance σ_{acc}^2.

3.2. The constant velocity filter

In the following, $\mathbf{z}(t)$ will stand for $\mathbf{T}(t)$, $\mathbf{b}(t)$ or $\mathbf{M}(t)$. The state vector is defined as $\mathbf{x}(t) = (\,\mathbf{z}(t)\;\;\dot{\mathbf{z}}(t)\,)^T$ and we assume that there is no control input $\mathbf{u}(t)$. A second order Taylor expansion of vectors $\mathbf{X}(t)$ (Equ. 5) and $\dot{\mathbf{X}}(t)$ yields the following state transition matrices, for $\mathbf{M}$ and $\mathbf{T}$:

$$\mathbf{A}(t) \;=\; \begin{pmatrix} \mathbf{I} & \delta t\;\mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \tag{17}$$

and for $\mathbf{b}$:

$$\mathbf{A}(t) = \begin{pmatrix} \mathbf{I} & \delta t\;\mathbf{I} \\ \mathbf{0} & \mathbf{I} \;+\; \delta t\;\mathbf{\Phi}(t+\delta t)\;[\,\dot{\mathbf{M}}(t) - \dot{\mathbf{M}}(t+\delta t)\,]\;\mathbf{P} \end{pmatrix}. \tag{18}$$

From Equ. 18, we see that the prediction and updating of $\mathbf{b}$ depends on the result of the state of $\dot{\mathbf{M}}$. This yields a non linear filter formulation. We have resorted to an approximate linear formulation by decoupling the filters on the parameters $\mathbf{M}$, $\mathbf{T}$ and $\mathbf{b}$ and by approximiting $\dot{\mathbf{M}}(t)$ by $\dot{\mathbf{M}}(t|\,t)$ and $\dot{\mathbf{M}}(t+\delta t)$ by $\dot{\mathbf{M}}(t+\delta t|\,t)$. This approximation yields good experimental results in practice. The tracking procedure may be described as following: at time t the current estimates of the parameters are $(\,\mathbf{z}(t|\,t)\;\;\dot{\mathbf{z}}(t|\,t)\,)^T$. The prediction step (Equ. 14 and Equ. 15) define the predicted location of the deformable model in the next frame (time $t+\delta t$). This location is generally close to the optimal configuration (unless an abrupt change occurs). Relevant measurements are thus obtained by applying a fast optimization procedure that tunes the parameters toward the spatiotemporal gradients extracted at time $t+\delta t$. The updated state $(\,\mathbf{z}(t|\,t+\delta t)\;\;\dot{\mathbf{z}}(t|\,t+\delta t)\,)^T$ is finally derived from Equ. 10 and Equ. 11. This filter yields robust tracking as long as the constant velocity assumption is verified. More complex kinematical behaviours are however observed in practice. We have thus developed a second filter, based on an on-line computation of the instantaneous velocities of the parameters.

3.3. A Kalman filter using instantaneous velocity

The second model considers the state vectors $\mathbf{x}(t) = \mathbf{z}(t)$ and a control input $\mathbf{u}(t) = \dot{\mathbf{z}}(t)$ computed using *finite differences* over time. We derive the Kalman filter for the deformation parameters by considering the second order Taylor expansion of vector $\mathbf{X}(t)$ (Equ. 5). The matrices $\mathbf{A}(t)$ and $\mathbf{B}(t)$ become respectively the identity matrix $\mathbf{I}$ and the matrix $\delta t\;\mathbf{I}$; the filter evolution equations correspond to the dynamics described by Equ. 6. The modeling error vector $\xi(t)$ stands for the higher order derivatives of the global parameters. The prediction, measurement and updating scheme is the same as in section 3.2. The instantaneous estimation of $\dot{\mathbf{z}}(t)$ enables to update the deformable model even in the presence of complex (but smooth) shape motions of over time. Abrupt changes require a specific procedure, which is described in the next paragraph.

3.4. Detection of abrupt changes

In real world situations, the dynamic evolution of the system may be subject to abrupt changes. A jump detection based on Hinkley's Cumulative Sum test [7] is thus introduced on the filter innovation in order to detect abrupt changes in the movement of the deformable structure. When a jump is detected, the Kalman filtering is reinitialized and a global optimization step is performed in order to estimate new reliable model parameters. In practice, only a few re-initialization steps were necessary to obtain a robust tracking over several hundred frames using the instantaneous velocity filter.

4. EXPERIMENTAL RESULTS

In our first experiments, we have considered the segmentation and tracking of complex shapes corresponding to hands moving against a textured background with partial occlusions. The sequence presented here was composed of more than one hundred 256x256 frames. In Fig. 2 and Fig. 3, the little finger is partially occluded by a box in the foreground. Fig. 2 presents the result of the segmentation without the tracking procedure described in this paper. Fig. 2a shows the optimal MAP segmentation at time $t-1$. Fig. 2b depicts the projection of the estimate obtained at time $t-1$ in the next frame (without predicting the motion). Local optimization on the model parameters does not yield satisfactory results if this projection is taken to initialize the estimation as can be seen in Fig. 2c. Without tracking one has thus to resort to time consuming stochastic optimization (see the result Fig. 2d) [9] to obtain relevant segmentations (6mn cpu time on a Sun Sparc IPX workstation in this case). The tracking procedure enables to reduce the computation time to less than 2mn with the same qualitative results as for global optimization (Fig. 3). Fig. 3 and Fig. 4 show examples of tracking with the constant velocity Kalman filter presented in section 3.2 (similar qualitative results are obtained with the instantaneous velocity filter). Fig. 3a and Fig. 3d present results of the filtered estimate respectively at time $t-1$ and time t using a fast local optimization procedure. The model prediction at time t and the observation are presented respectively in Fig. 3b and Fig. 3c. Fig. 4 presents another sequence with a detection

of abrupt changes. The prediction (Fig. 4b) and the filtered estimate at time t (Fig. 4c) obtained from the filtered estimate at time $t-1$ (Fig. 4a), are disturbed in this case because of an abrupt motion variation of the deformable structure. The fast optimization is not able to provide satisfactory results in this case (Fig. 4c) and is supplied here by global optimization (Fig. 4d). Fig. 1 shows plots of the prediction and filtered estimate for the the first deformation mode b_0 on a synthetic hand image sequence where the ground truth is known (detection of abrupt changes are represented by vertical bars).

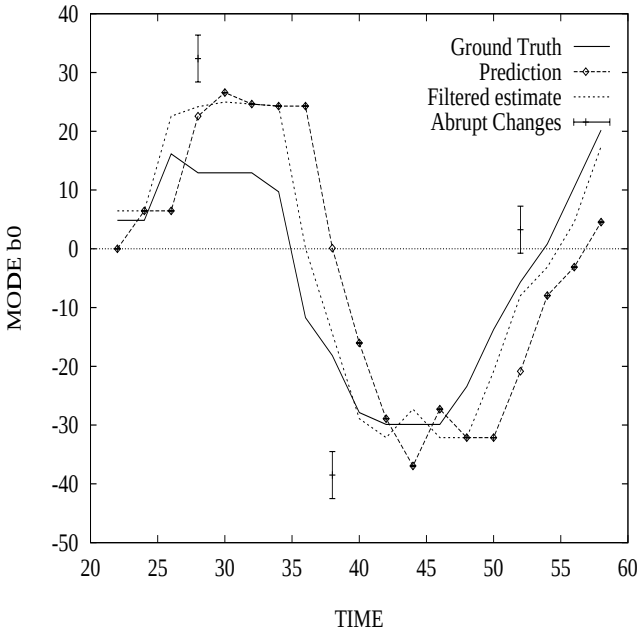

Figure 1: *Tracking of the first deformation mode* b_0

In all experiments the procedure was able to track reliably the moving hand over more than one hundred frames without failure. In a second experiment (Fig. 5), we have considered the tracking of a mouth. The algorithm yielded similar qualitative in this case.

5. CONCLUSION

In this paper, we have presented a general framework for tracking a deformable stochastic model using Kalman filtering. The technique relies on the recursive estimation of the global model parameters. A constant velocity filter and an instantaneous velocity filter have been successively considered to provide reliable tracking. A jump detection procedure has been introduced to test the validity of the prediction. The robustness of this algorithm has been tested on long sequences with partial occlusions. This method yields promising future prospects as far as the characterization and the interpretation of the dynamic behavior of complex deformable objects is concerned.

6. REFERENCES

[1] A. AZARBAYEJANI, B. HOROWITZ and A. PENTLAND. – Recursive estimation of structure and motion using relative orientation constraints. – In *IEEE Conf. Comput. Vis. Pat. Rec.*, pages 294–299 , New York City, USA, June 1993. –

[2] B. BASCLE, P. BOUTHEMY, N. DERICHE and F. MEYER. – Tracking complex primitives in an image sequence. – *Int. Conf. on Pattern Recognition, ICPR'94*, Jerusalem, Israel, Oct. 1994. –

[3] A. BLAKE, R. CURWEN and A. ZISSERMAN. – A framework for spatiotemporal control in the tracking of visual contours. – *Int. J. Computer Vision*, Vol. 11, No 2: pp. 127–145, Oct. 1993. –

[4] T.F. COOTES, C.J. TAYLOR, D.H. COOPER and J. GRAHAM. – Training models of shape from sets of examples. – In *British, Machine Vision Conf.*, pages 9–18, Leeds, UK, Sept. 1992. –

[5] R. DERICHE and O. FAUGERAS. – Tracking Line Segments. – In *First European Conference on Computer Vision*, pages 259–268, Antibes, France, April 1990. –

[6] U. GRENANDER, Y. CHOW and D.M. KEENAN. – Hands. A Pattern Theoretic Study of Biological Shapes. – Springer, New-York, 1991. –

[7] D. V. HINKLEY. – Inference about the change - point from cumulative sum - tests. – *Biometrika*, Vol. 58, No 3: pp. 509–523, 1971.

[8] C. KERVRANN and F. HEITZ. – A hierarchical statistical framework for the segmentation of deformable objects in image sequences. – Technical Report INRIA No 2133, Dec. 1993. –

[9] C. KERVRANN and F. HEITZ. – A hierarchical Statistical Framework for the Segmentation of Deformable Objects in Image Sequences. – In *IEEE Conf. Comput. Vis. Pat. Rec.*, pages 724–728 , Seattle, USA, June 1994. –

[10] D. METAXAS and D. TERZOPOULOS. – Shape and nonrigid motion estimation throught physics-based synthesis. – *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 15, No 6: pp. 580–591, June 1993.

[11] F. MEYER and P. BOUTHEMY. – Region-based tracking using affine motion models in long image sequences. – *CVGIP: Image Understanding*, to appear, Sept. 1994.

[12] C. NASTAR and N. AYACHE. – Fast segmentation, tracking and analysis of deformable objects. – *In Proc. 4th Int. Conf. Comp. Vis.*, pages 275-279, Berlin, Germany, May, 1993. –

[13] A. PENTLAND and B. HOROWITZ. – Recovery of nonrigid motion and structure. – *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 13, No 7: pp. 730–742, July 1991.

[14] A. YUILLE, P. HALLINAN and D. COHEN. – Feature extraction from faces using deformable templates. – *Int. Journal of Comput. Vis.*, Vol. 8, No 2: pp. 99–111, 1991.

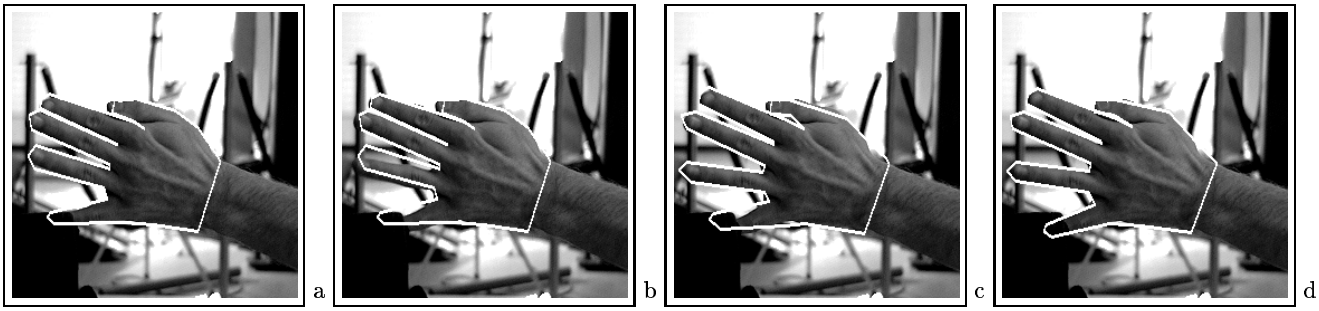

Figure 2: *Motion-based segmentation of a hand - without the tracking procedure (see text).*

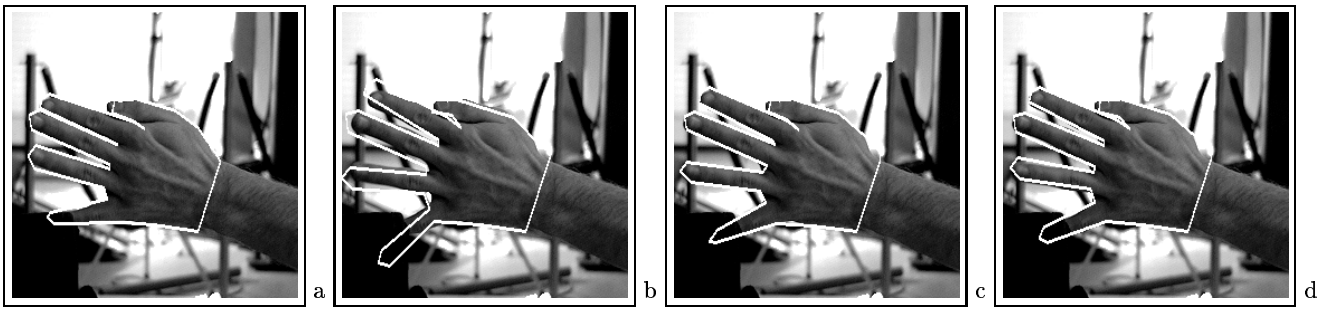

Figure 3: *Motion-based segmentation of a hand - with the tracking procedure (see text).*

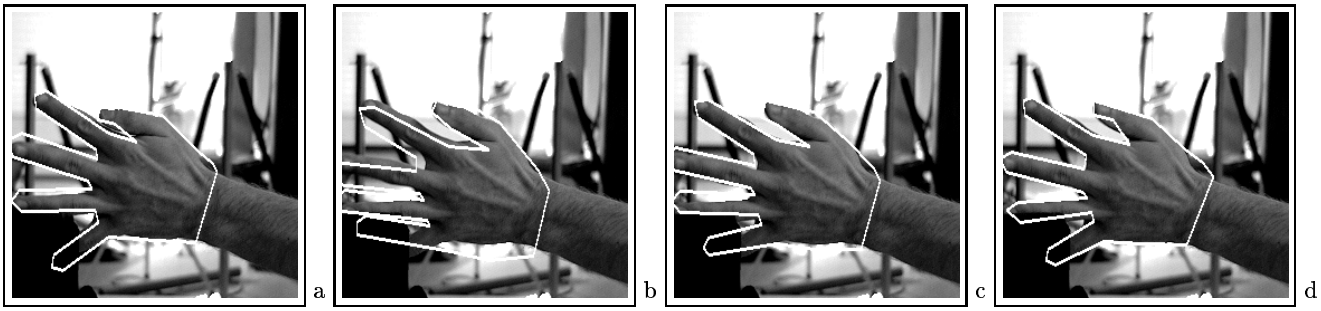

Figure 4: *Motion-based segmentation of a hand - Detection of abrupt changes (see text).*

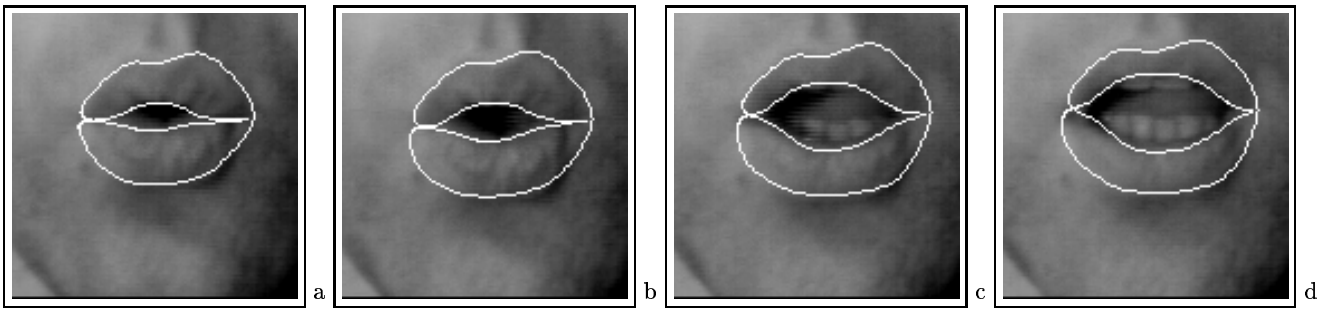

Figure 5: *Tracking of mouth movement.*