

Learning Probabilistic Deformation Models from Image Sequences

CHARLES KERVRANN
IRISA/INRIA
Campus Universitaire de Beaulieu
35042 Rennes cedex, FRANCE

submitted as a paper to SIGNAL PROCESSING
special issue on
DEFORMABLE MODELS AND TECHNIQUES FOR IMAGE AND SIGNAL PROCESSING

AUTHOR TO CONTACT:

FIRST (Personal) NAME: Charles
LAST (Family) NAME: KERVRANN
ADDRESS: INRA – Biométrie,
Domaine de Vilvert
78352 Jouy-en-Josas, FRANCE
TELEPHONE NUMBER: (+33) 1 34 65 22 35
FAX NUMBER: (+33) 1 34 65 22 28
E-MAIL ADDRESS: ck@jouy.inra.fr

List of Complementary Information

- *Number of pages:* 22
- *Number of figures:* 10
- *Number of tables:* 0
- *Key Words:* deformable models; Markov models; image sequence analysis; segmentation; tracking; learning; Principal Component Analysis.

Learning Probabilistic Deformation Models from Image Sequences¹

CHARLES KERVRANN
IRISA/INRIA
Campus Universitaire de Beaulieu
35042 Rennes cedex, FRANCE

submitted as a paper to SIGNAL PROCESSING
special issue on
DEFORMABLE MODELS AND TECHNIQUES FOR IMAGE AND SIGNAL PROCESSING

Abstract

In this paper, we present an approach for an *unsupervised* learning of probabilistic deformation modes of 2D moving objects from image sequences. The object representation relies on a statistical description of global and local deformations applied to an *a priori* prototype shape. The optimal bayesian estimate of the deformation process is obtained by maximizing a nonlinear joint probability distribution using stochastic and deterministic optimization techniques. The estimates obtained at time t are integrated in the deformation model as *a priori* knowledge for the segmentation at time $t + 1$. Deformation modes are updated on-line using a Principal Component Analysis of the distortions computed from the shapes estimated previously in the image sequence. In addition, the updated deformation modes are exploited for the segmentation of new shapes in the current image sequence. The approach yields a robust learning and is demonstrated on real-world image sequences showing the tracking of hands undergoing 2D articulated movements.

1 Introduction

Segmenting structures from image sequences and building a generic model of these structures is a difficult task in computer vision due to the complexity and variability of the objects of interest. Traditional low-level image processing techniques usually require considerable amounts of manual intervention and fail to design a consistent and coherent model of the shapes of interest. The introduction of deformable models capable to accommodate the significant variability of structures over time and across different individuals, has recently gained considerable popularity in many applications fields (meteorology, biological or biomedical images, analysis of human motion) [18, 37, 12, 39, 33, 31]. They

¹C. Kervrann was with IRISA/INRIA Rennes, FRANCE. He is now with INRA, Biométrie et Intelligence Artificielle, Domaine de Vilvert, 78352 Jouy-en-Josas, FRANCE (e-mail: ck@jouy.inra.fr).

offer henceforth a fundamental and powerful tool to represent, segment, track and match deformable structures in image analysis.

1.1 Shape and motion variability

Deformable models enable to take into account specific *a priori* knowledge on the shapes to be handled and on their variability. Their mathematical foundations are related to the geometry, physics (elasticity) theory, approximation theory or statistics. General-purpose active contours [23] or *snakes* enforce constraints controlled by elastic forces based on local structure, inflating forces and image based potentials but are not adapted to constrain the deformations for a particular object class [5] or application. On the other hand, hand-built models [9, 42] have been proposed dedicated to specific applications, but building a new model can become a quite slow and tedious task.

It seems more effective and less-time consuming to learn the shape variability from a training set. Deformation modes of an object class may be identified by matching sets of feature points [12]. Cootes *et al.* [13] consider a training set of fixed feature points specified manually to capture the variations of the shapes of interest. The Principal Component Analysis (PCA) allows to determine the main variation modes superimposed on a pre-computed mean shape. However, the specification of the the structure and the deformation modes requires an *off-line* supervised training step [12, 13, 10]. Therefore, Baumberg and Hogg have proposed a example-based representation of people walking by computing an eigenspace representation of the key points over many images from a video sequence [3]. In their approach, the spline curves fitted to the object outlines have to be normalized for training. This method uses a very large training set and therefore, yields a robust and compact representation of deformations. However, it remains sensitive to mismatches of knot points and cannot handle large rotations of objects.

In [3], the authors have proposed to model jointly the random variations in shape arising within from the differences between objects in a given class with those occurring during object motion (due to projective effects). The distinction between the two sources of variability has been pointed out in [6, 7, 22]. In these works, the learning method is rather dynamic, using and modeling temporal image sequences. Example motions are exclusively learned by general-purpose tracker based on the assumption that the identity of the object does not change over time. The learned dynamics are then used in a Kalman filter framework which enhance tracking capability for motions similar to those in the training set [5, 40].

1.2 Learning shape variability

In this paper, we introduce a modeling framework to learn automatically deformation modes associated to the object of interest from real image data without any human interaction (such as manual point

correspondence). The method requires that the object be described as the deformations of a single prototype object [12, 2, 18, 3]. In contrast to works reported in [6, 7, 22], our approach aims at identifying the shape variability of an object class from temporal image sequences. We assume the whole estimated deformations of one single object enable to constrain the class variability of any objects belonging to the same object class. Our method differs from works described in [10, 3] in that we remove the visual rigid motion computed for each frame to determine the statistical deformation modes. In addition, the matching between sets of feature points is already known and implicit.

In our approach, the model relies on a statistical description of shape deformations, in which two deformation processes are considered [24]. The idea to combine two deformation processes (Finite Element Methods / Point Distribution Models) in a general modeling framework, has been already investigated in [11, 30]. Here, the global description of deformations relies on the modal decomposition introduced by Cootes *et al.* in [12]. Parameters describing global shape deformations include transformations from the group of similarity (translation, rotation, scale) and parameters controlling the main variation modes of the template. Local deformations are modeled as stochastic perturbations and are assumed to follow a first order Markov process [18]. It is *local* in the sense that it models deformations involving a point (and its neighbors).

The combined segmentation-learning procedure is summarized here: at the beginning of the image sequence, no training has been performed. Parameters from the group of similarity and the local deformation process only are identified. The group of similarity transformations enables a first crude registration of the shape on the input data. For each frame (except the two first frames), the Principal Component Analysis of the shapes associated to the local deformation process allows to identify and update global deformation modes as well as the shape structure over time. In addition, a Maximum A Posteriori (MAP) estimate of both local and global deformations is obtained by maximizing a highly nonlinear joint probability distribution describing the interactions between observations (temporal gradients extracted from the image in our case) and the deformation process. Global and local optimization techniques are necessary to obtain optimal solutions for the deformation process [18, 4, 16]. Global optimization algorithms saves the operator the bother of providing manual initializations for the model for the first frame. Finally, the procedure dynamically updates the shape structure as well as the deformation modes using implicit matching between sets of feature points. A completely data driven learning is then obtained.

1.3 Paper organization

The remainder of this paper is organized as follows. The motion-based segmentation framework is described in Section 2. The stochastic deformable model and the bayesian estimation of deformations

are considered in this section. In Section 3, the unsupervised method for training the probabilistic deformation models from video sequences is presented. The final section (Section 4) of the paper illustrates one possible application of our method: tracking and modeling moving hands on real image sequences.

2 Motion-based segmentation framework

Significant improvements have been obtained in image segmentation problems by introducing global statistical models such as Markov random field models (MRF) [17, 20, 26] or statistical deformable models [16, 19, 2, 18, 38] that *constrain* the segmentation process. In the following we consider the problem of extracting and tracking moving deformable objects in an image sequence. Our approach for the segmentation of deformable shapes relies on a bayesian formulation of the problem.

2.1 A stochastic deformable model

Description of global deformations To obtain a compact application-tailored description of the object of interest, the shape template and its main deformation modes are characterized here using an *on-line* training procedure. This training procedure relies on the KL expansion of the deformations observed on a training set [35]. This standard method in Pattern Recognition has been already used for the retrieval, recognition and tracking of articulated and deformable objects from grey level appearance [41, 32, 8]. The procedure, first proposed by Cootes *et al.* for modeling 2D deformations, is described in detail in [12, 10] and is briefly recalled here.

Figure 1 to be placed here

Following [10], a particular shape \mathbf{x}_k belonging to the training population is represented by a set of n labeled points linked by a polygonal line or a B-spline curve which approximates its outline (Fig. 1):

$$\mathbf{x}_k = (x_{k1}, y_{k1}, \dots, x_{kn}, y_{kn})^T.$$

The n labeled points correspond to the most salient points of the shape outline. In [10], these “landmarks” are extracted manually on the learning population and the shapes belonging to the learning population are normalized in scale, and aligned with respect to a common reference frame. The mean shape $\bar{\mathbf{x}}$ and the covariance matrix \mathbf{C} of shapes $\{\mathbf{x}_k\}$ are computed from this set of normalized shapes. The main deformation modes of the template model \mathbf{X} , are then described by the eigenvectors Φ of \mathbf{C} , with the largest eigenvalues [10]. The globally deformed template is defined by (Fig. 1)

$$\mathbf{X} = \mathbf{M}(k, \theta) [\bar{\mathbf{x}} + \Phi \mathbf{b}] + \mathbf{T} \tag{1}$$

where

- \mathbf{T} and $\mathbf{M}(k, \theta)$ account for rigid transformations of the template in the image plane (\mathbf{T} is a global translation vector, and $\mathbf{M}(k, \theta)$ performs a rotation by θ and a scaling by k),
- $\Phi = (\phi_1, \dots, \phi_m)$ is the matrix of the first m ($m < 2n$) eigenvectors associated to the m largest eigenvalues and $\mathbf{b} = (b_1, \dots, b_m)^T$ is a vector containing the weights for these m deformation modes.

A global configuration of the deformable template is thus described by $4 + m$ parameters corresponding to rigid transformations (four parameters) and m deformation modes b_j , $j = 1, \dots, m$. In practice, only five to seven modes are necessary to stand for more than 90% of the variability observed on the training population [10].

Description of local deformations A deformation process δ , applied to the n labeled points, is introduced to complete the global description already presented. These local deformations are considered as *random perturbations* (represented by random translations) that are superimposed to the globally deformed shape (Fig. 1). The deformation vector δ is described by a Gauss-Markov process defined on the graph corresponding to the outline of the deformable template. The Gauss-Markov distribution models the statistical interactions between the local random deformations applied to neighboring points of the template [24]. The complete model is expressed as (Fig. 1)

$$\mathbf{Y} = \mathbf{X} + \delta = \mathbf{M}(k, \theta) [\bar{\mathbf{x}} + \Phi \mathbf{b}] + \mathbf{T} + \delta. \quad (2)$$

where $\delta = (\delta_1, \dots, \delta_n)^T$ and $\delta_i = (\delta_{x_i}, \delta_{y_i})^T$. The probability distribution of δ is defined by

$$\mathbf{P}(\delta) = \frac{1}{Z_p} \exp -\frac{1}{2} \delta^T \mathbf{R}^{-1} \delta \quad (3)$$

where \mathbf{R} is the covariance matrix of δ and Z_p designates the partition function. Assuming a first-order Gauss-Markov model (*i.e.* a first-order neighborhood structure on the graph), the joint distribution of δ becomes

$$\mathbf{P}(\delta) = \frac{1}{Z_p} \exp -\frac{1}{2} \sum_{i=1}^n \left[\frac{1}{\varepsilon_i^2} \|\delta_i - \delta_{i-1}\|^2 + \frac{1}{\sigma_i^2} \|\delta_i\|^2 \right]. \quad (4)$$

Parameters σ_i^2 and ε_i^2 are interpreted as variance parameters. Parameters ε_i^2 weight the interactions between neighboring points and control the smoothness of local deformations. Parameters σ_i^2 control the amplitude of the local deformation vectors. Low values for σ_i^2 will draw the model \mathbf{Y} toward the globally deformed shape \mathbf{X} . In our experiments, these parameters were assumed to be constant: $\sigma_i^2 = \sigma^2$ and $\varepsilon_i^2 = \varepsilon^2$, $\forall i$. This is of course an approximation since σ_i^2 and ε_i^2 should depend on

the distance between the points of index $i - 1$ and i , unless the feature points are equally spaced. Adopting constant values for these parameters in our implementation has, however, proved satisfactory in practice: the goal was to favor smooth shapes. For instance, the values $\sigma = 3$ and $\varepsilon = 2$ have been adopted in our experiments.

2.2 Bayesian estimation of deformations

Our approach for the segmentation of deformable shapes relies on a bayesian formulation of the problem. The hierarchical model defined in the previous section (Eq. (2)) is considered as an *a priori* statistical model describing the configurations of the shape of interest. Besides, one or more specialized modules extract from the image sequence low-level features (spatial or spatio-temporal gradients) that will be used as observations in the bayesian estimation process.

Let $\mathbf{d} = \{d(s), s \in S\}$ designates an observation field defined on a rectangular lattice S . The observation field \mathbf{d} , extracted from the image sequence, is related here to the spatio-temporal variations of the intensity function. The segmentation problem is formulated as the Maximum A Posteriori (MAP) estimation of the (hidden) random process \mathbf{Y} from the observation field \mathbf{d} :

$$\mathbf{Y}^* = \arg \max_{\mathbf{Y}} \mathbf{P}(\mathbf{d} | \mathbf{Y}) \mathbf{P}(\mathbf{Y}) = \arg \max_{\mathbf{Y}} \mathbf{P}(\mathbf{Y}, \mathbf{d}). \quad (5)$$

According to the assumption on the statistics of δ (Eq. (3)), \mathbf{Y} follows a first-order Gauss-Markov process:

$$\mathbf{P}(\mathbf{Y}) = \frac{1}{Z_p} \exp -\frac{1}{2} (\mathbf{Y} - \mathbf{X}(\Theta))^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{X}(\Theta)). \quad (6)$$

where \mathbf{R} is the covariance matrix of process δ (see Eqs. (3) and (4)) and the mean $\mathbf{X}(\Theta)$ corresponds to (see Eq. (1))

$$\mathbf{X}(\Theta) = \mathbf{M}(k, \theta) [\bar{\mathbf{x}} + \Phi \mathbf{b}] + \mathbf{T}. \quad (7)$$

$\Theta = (\mathbf{M}(k, \theta), \mathbf{T}, \mathbf{b})$ denotes here the (deterministic) set of *hyperparameters* of this probabilistic model [14].

The distribution $\mathbf{P}(\mathbf{d} | \mathbf{Y})$ is the likelihood of the observation field given the deformation process. This distribution depends on the image attributes used in the segmentation and on the observations at hand. In our case, this likelihood is specified as a Gibbs distribution [17] which incorporates specific knowledge on the application

$$\mathbf{P}(\mathbf{d} | \mathbf{Y}) = \frac{1}{Z_d} \exp -E_d(\mathbf{Y}, \mathbf{d}), \quad (8)$$

where $E_d(\mathbf{Y}, \mathbf{d})$ is an energy function and Z_d is the partition function ($E_d(\mathbf{Y}, \mathbf{d})$ is specified in the next section).

The joint distribution appearing in Eq. (5) is thus also a Gibbs distribution:

$$\mathbf{P}(\mathbf{Y}, \mathbf{d}) = \frac{1}{Z_p Z_d} \exp - E_{\Theta}(\mathbf{Y}, \mathbf{d}), \quad (9)$$

where

$$E_{\Theta}(\mathbf{Y}, \mathbf{d}) = E_d(\mathbf{Y}, \mathbf{d}) + \frac{1}{2} (\mathbf{Y} - \mathbf{X}(\Theta))^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{X}(\Theta)). \quad (10)$$

Let us notice that $Z = Z_p Z_d$ does not depend on Θ since Θ only appears in the mean of the Gaussian distribution (see Eq. (6)). On the other hand, the normalizing constant $Z_d = \int \exp -E_d(\mathbf{Y}, \mathbf{d}) d\mathbf{d}$ and hence Z generally depends on \mathbf{Y} . However, we show in [28] that for the segmentation model considered in Section 2.3, $Z_d = (1 + e^{-1})^M$ is constant if M denotes the total number of sites in the image. As a result, $Z = Z_p Z_d$ does not depend on \mathbf{Y} for the aimed application and the MAP estimation of the deformable template comes to the minimization of global energy function $E_{\Theta}(\mathbf{Y}, \mathbf{d})$. The final configuration of the template is thus a compromise between prior information on the deformable structure and image-derived information (see Eq. (10)). Strong prior information about the template deformations is embedded in the model, thanks to the modal expansion. The modeling of $E_d(\mathbf{Y}, \mathbf{d})$ is considered in the next section.

2.3 Measurement of video features

In the segmentation process, the global energy function $E_d(\mathbf{Y}, \mathbf{d})$ stands for the interactions between data \mathbf{d} and the deformable template \mathbf{Y} . The modeling of $E_d(\mathbf{Y}, \mathbf{d})$ is clearly problem dependent. In this section we present a model for $E_d(\mathbf{Y}, \mathbf{d})$ which aims at extracting objects of interest from an image sequence. This model yields a motion-based segmentation of the scene: objects are characterized by their motion with respect to the background. The background may be static or may itself be in motion. This model relies on temporal gradient data (Fig. 2).

Figure 2 to be placed here

Let $I(t, s)$, $s \in S$ denote the intensity function, where $s = (x, y)$ designates the 2-D spatial image coordinates and t the time axis. We first assume that the camera is static. In order to extract moving objects from the image sequence, temporal variations (Fig. 2) are estimated using two complementary methods. The first one measures variations $d_1(s)$ on three successive images ; the second one estimates the changes $d_2(s)$ between the current image and a reference image which is created and updated on line:

$$\begin{aligned} d_1(s) &= \min (|I_t(s) - I_{t-\Delta t}(s)|, |I_{t+\Delta t}(s) - I_t(s)|), \\ d_2(s) &= |I_{ref}(s) - I_t(s)|. \end{aligned} \quad (11)$$

where Δt designates the time step between two successive frames. As can be seen, three frames are necessary to obtain a segmentation: $I_{t-\Delta t}$, I_t and $I_{t+\Delta t}$. The tracking procedure is applied, once the first segmentation has been obtained, *i.e.*, after the third frame. The reference image $I_{ref}(s)$ is constructed using a linear estimator of the background described in [15]. Observations $d_1(s)$ present high values for points belonging to a moving object and low values for background points. Temporal gradients such as $d_1(s)$ are known to yield poor observations in homogeneous (*i.e.* nontextured) regions or in the presence of self overlapping of an object mask during the displacement. The second observation $d_2(s)$ based on a reference image of the background is less sensitive to this problem and is used as complementary information.

The observation field (data) is thus defined as

$$d(s) = \max (s_\alpha(d_1(s)), s_\beta(d_2(s))) \quad (12)$$

$$\begin{aligned} \text{where: } s_\eta(y) &= 1 \quad \text{if } y > \eta \\ s_\eta(y) &= 0 \quad \text{otherwise,} \end{aligned} \quad (13)$$

where α and β are two thresholds for the detection of significant motion. Energy $E_d(\mathbf{Y}, \mathbf{d})$ then describes the statistical interaction between the thresholded temporal gradients $d(s)$ and the configuration of the deformable model. For a given configuration of the template, the image can be partitioned into two regions: the inside of the template Γ_Y^I corresponding to the object of interest and the outside of the template Γ_Y^O corresponding to the background. Energy $E_d(\mathbf{Y}, \mathbf{d})$ tends to enclose moving points inside the deformable model and to reject static points, belonging to the background, outside the outline of the model

$$E_d(\mathbf{Y}, \mathbf{d}) = \sum_{s \in \Gamma_Y^I} |d(s) - 1| + \sum_{s \in \Gamma_Y^O} |d(s) - 0|. \quad (14)$$

This model can also be generalized to situations in which the camera is itself moving (inducing a global motion on the background) by using a pre-processing step to compensate for the apparent motion of the background [28, 34].

2.4 Estimation of the model hyperparameters and optimization

In this section, we present a simple, noniterative procedure for estimating the hyperparameters of the stochastic deformable model. For notation conveniences, the joint distribution of \mathbf{Y} and \mathbf{d} is redefined as

$$\mathbf{P}(\mathbf{Y}, \mathbf{d} | \Theta) = \frac{1}{Z} \exp - E_\Theta(\mathbf{Y}, \mathbf{d}), \quad (15)$$

where, as already noticed, Z does not depend on Θ or on \mathbf{Y} .

2.4.1 Marginalized Maximum Likelihood estimation

The segmentation problem is formulated as the joint estimation of \mathbf{Y} and of the (unknown) set of hyper-parameters Θ [14, 29]. Since Θ is unknown, a standard criterion for estimating Θ is the Marginalized Maximum Likelihood (MML) criterion [29]:

$$\Theta^* = \arg \max_{\Theta} \int_{\mathbf{Y}} \mathbf{P}(\mathbf{Y}, \mathbf{d} | \Theta) d\mathbf{Y} = \arg \max_{\Theta} \int_{\mathbf{Y}} \mathbf{P}(\mathbf{d} | \mathbf{Y}, \Theta) \mathbf{P}(\mathbf{Y} | \Theta) d\mathbf{Y}. \quad (16)$$

The estimate of \mathbf{Y} is then computed in turn, using the MML estimate Θ^* and the already considered MAP criterion:

$$\mathbf{Y}^* = \arg \max_{\mathbf{Y}} \mathbf{P}(\mathbf{Y}, \mathbf{d} | \Theta^*). \quad (17)$$

Estimation of Θ Θ is estimated according to Eq. (16). We can derive a simple estimator for Θ if we assume $\mathbf{Y}^* = \mathbf{X}(\Theta^*)$ as an initial estimate for \mathbf{Y} . This simplifying assumption expresses the fact that $\delta = 0$ when Θ is estimated. The estimation of \mathbf{Y} will be revised once Θ^* is derived. Under this assumption, the gaussian distribution of \mathbf{Y} , may be approximated by a Dirac distribution:

$$\mathbf{P}(\mathbf{Y} | \Theta) = \frac{1}{Z_p} \exp -\frac{1}{2} (\mathbf{Y} - \mathbf{X}(\Theta))^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{X}(\Theta)) \approx \delta(\mathbf{Y} - \mathbf{X}(\Theta)). \quad (18)$$

It follows that

$$\begin{aligned} \Theta^* &\approx \arg \max_{\Theta} \int_{\mathbf{Y}} \mathbf{P}(\mathbf{d} | \mathbf{Y}, \Theta) \delta(\mathbf{Y} - \mathbf{X}(\Theta)) d\mathbf{Y} \\ &= \arg \max_{\Theta} \mathbf{P}(\mathbf{d} | \mathbf{Y} = \mathbf{X}(\Theta)) \\ &= \arg \min_{\Theta} E_d(\mathbf{X}(\Theta), \mathbf{d}). \end{aligned} \quad (19)$$

Thus the MML criterion reduces to the minimization of the data-related energy term $E_d(\mathbf{Y}, \mathbf{d})$, for $\mathbf{Y} = \mathbf{X}(\Theta)$.

Estimation of \mathbf{Y} The MAP estimate of \mathbf{Y} is easily derived, given the estimate Θ^* . The MAP criterion (see Eq. (17)) comes to the minimization of the global energy function

$$\mathbf{Y}^* = \arg \min_{\mathbf{Y}} E_{\Theta^*}(\mathbf{Y}, \mathbf{d}), \quad (20)$$

where $E_{\Theta}(\mathbf{Y}, \mathbf{d})$ is defined by Eq. (10).

To summarize, the segmentation of the structure of interest requires the estimation of the hyper-parameters according to Eq. (19) and the minimization of the global energy function according to Eq. (20). These (nonlinear) optimization steps are performed in different ways, depending on the clues used in the segmentation process. In practice, only one iteration of the optimization loop is performed, provided stable segmentation results.

3 Unsupervised learning of deformation modes

Given the above estimation scheme, the goal is to update the object (*i.e.* its original structure $\bar{\mathbf{x}}$ and deformation modes Φ) in order to converge towards a more and more reliable and compact representation for the segmentation task. In our approach, the modal amplitudes \mathbf{b} associated to the updated deformation modes are estimated *on-line* at each frame of the learning sequence.

3.1 Temporal evolution of the deformable template

Let us recall the general form of the deformable model at time t

$$\mathbf{Y}(t) = \mathbf{M}(k(t), \theta(t)) [\bar{\mathbf{x}}(t) + \Phi(t) \mathbf{b}(t)] + \mathbf{T}(t) + \delta(t). \quad (21)$$

$\bar{\mathbf{x}}(t)$ and $\Phi(t)$ respectively designates the mean shape and the matrix of the $m(t)$ unit eigenvectors corresponding to the $m(t)$ largest eigenvalues computed from the shapes up to time t . $\mathbf{b}(t)$ denotes the vector ($m(t) \times 1$) of modal amplitudes associated to the $m(t)$ most significant deformation modes at time t . Given the estimate of $\delta(t)$ and the hyperparameters $\Theta(t) = (\mathbf{M}(k(t), \theta(t)), \mathbf{T}(t), \mathbf{b}(t))$ of the model, deformations will be analyzed in a common reference coordinate system as described in Section 3.3.

3.2 Exercising the learning algorithm

The *on-line* training algorithm updates the number of deformation modes $m(t)$ over time. The different steps of the training algorithm exercised on a video sequence are the following:

- **Processing of the 1st image ($t = 1$):**
 - estimation of parameters $\mathbf{M}(k(1), \theta(1))$ and $\mathbf{T}(1)$
 - estimation of local deformation process $\delta(1)$
- **Processing of the 2nd image ($t = 2$):**
 - estimation of parameters $\mathbf{M}(k(2), \theta(2))$ and $\mathbf{T}(2)$
 - estimation du local deformation process $\delta(2)$
 - KL expansion of estimated shapes $\mathbf{Y}(t)$ ($t = 1, 2$) and computation of $\bar{\mathbf{x}}(2)$ et $\Phi(2)$
-
-
-
- **Processing of the i^{th} image ($t = i$):**

- estimation of parameters $\mathbf{M}(k(i), \theta(i))$ et $\mathbf{T}(i)$
- estimation of modal amplitudes $\mathbf{b}(i)$
- estimation of local deformation process $\delta(i)$
- KL expansion of estimated shapes $\mathbf{Y}(t)$ ($t = 1, \dots, i$) and updating of $\bar{\mathbf{x}}(i)$ and $\Phi(i)$.
- Prediction of hyperparameters at time $t + 1$.

In our approach, the processing of the first sequence is specific because no *a priori* knowledge about deformation modes is available. Global optimization techniques are thus necessary to estimate reliably $\mathbf{M}(k(1), \theta(1))$ and $\mathbf{T}(1)$ and finally $\mathbf{Y}(1)$ (*i.e.* $\delta(1)$ in practice). Due to the large size of the the space of configurations, the computation of the MAP estimate is generally computationally demanding but is justified insofar as the solution to the optimization problem is not constrained enough (no initial guess about the location and the deformations is known).

On the other hand, the hyperparameters and the local deformation process can be estimated efficiently and quickly on the subsequent frames by means of a deterministic optimization algorithm (Iterated Conditional Modes – ICM [4]) and a temporal prediction scheme: the hyperparameters are initialized at time $t + 1$ using final estimates obtained at time t . The updating of these hyperparameters is described in the next section.

Finally, the tracking procedure can be completed by a statistical detection of abrupt temporal changes, described in Section 3.4. When an abrupt change is detected at time t , a global optimization step is performed in order to obtain reliable estimates for $\mathbf{Y}(t)$. The computational cost is about 10 min of CPU time on SUN/SPARC10 workstation for one 256×256 image (see Fig. 6-8) when $\delta(t)$ is estimated using a *simulated annealing* algorithm. The other frames are analyzed using an ICM algorithm coupled with a tracking procedure between two successive abrupt changes. As a consequence, the average CPU time falls down to less than 1 min for one frame.

3.3 Identification of deformation modes

Given the estimate of $\delta(t)$ and the hyperparameters $\Theta(t)$ of the model at time t , deformations are analyzed in a common reference coordinate system by compensating the parameters from the group of similarity, at each frame. In this reference coordinate system the new model $\mathbf{y}(t)$ becomes

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{M}^{-1}(k(t), \theta(t)) [\mathbf{Y}(t) - \mathbf{T}(t)], \\ \mathbf{y}(t) &= \bar{\mathbf{x}}(t) + \Phi(t) \mathbf{b}(t) + \mathbf{M}^{-1}(k(t), \theta(t)) \delta(t). \end{aligned} \quad (22)$$

The template is easily updated at time $t + 1$ according to a Maximum Likelihood (ML) estimator

$$\bar{\mathbf{x}}(t + 1) = \frac{N(t)}{N(t + 1)} \bar{\mathbf{x}}(t) + \frac{1}{N(t + 1)} \mathbf{y}(t), \quad (23)$$

$$\bar{\mathbf{x}}(t+1) = \bar{\mathbf{x}}(t) + \frac{1}{N(t+1)} [\mathbf{\Phi}(t)\mathbf{b}(t) + \mathbf{M}^{-1}(k(t), \theta(t)) \boldsymbol{\delta}(t)] \quad (24)$$

where $N(t)$ is the current number of processed frames (corresponding to the number of shapes analyzed from the beginning of the sequence).

The deformation modes of the model are described by the unit eigenvectors of the covariance matrix $\mathbf{C}(t+1)$, defined at time $t+1$ according to a ML estimator

$$\mathbf{C}(t+1) = \frac{1}{N(t+1)} \sum_{i=1}^{t+1} (\mathbf{y}(i) - \bar{\mathbf{x}}(t+1))^T (\mathbf{y}(i) - \bar{\mathbf{x}}(t+1)). \quad (25)$$

An accurate description of the main variation modes is obtained by retaining in matrix $\mathbf{\Phi}(t+1)$ only the $m(t+1)$ eigenvectors associated to the $m(t+1)$ largest eigenvalues. The number of eigenvalues retained in this representation is adjusted through time in order to set the loss of the information to a constant (and small) value. Typically, between 95% and 99% of the total variability is preserved in the truncated representation. The number of significant modes $m(t)$ is then observed to increase over time at the beginning of the sequence, before becoming constant when $t \rightarrow \infty$ (for instance the final number of deformation modes in Fig. 3 is five).

Figure 3 to be placed here

From Eqs. (24) and (25), it is easily shown that $\bar{\mathbf{x}}(t+1)$ and $\mathbf{C}(t+1)$ converge to constant values when $t \rightarrow \infty$. In practice the infinite time corresponds to a sequence of more than many hundred frames where all representative deformations of the object class have occurred and thereby $\mathbf{\Phi}(t)$ and its associated eigenvalues are completely defined. Fig. 4 shows for instance the evolution of the eigenvalues, associated to the four first deformation modes, over a long image sequence composed of sixty frames when 98% of the total variability is preserved.

Figure 4 to be placed here

Besides, we take advantage of the temporal coherence of the object movements to predict the value $\hat{\mathbf{b}}(t+1)$ of the deformation modes at time $t+1$ from the estimate at time t and from the updated representation. $\mathbf{b}(t)$ is the projection of the shape $\mathbf{y}(t)$ onto the eigenmodes basis $\mathbf{\Phi}(t)$:

$$\hat{\mathbf{b}}(t+1) = \mathbf{\Phi}^T(t+1) (\mathbf{y}(t) - \bar{\mathbf{x}}(t+1)). \quad (26)$$

i.e.

$$\hat{\mathbf{b}}(t+1) = \mathbf{\Phi}^T(t+1) (\bar{\mathbf{x}}(t) + \mathbf{\Phi}(t)\mathbf{b}(t) + \mathbf{M}^{-1}(k(t), \theta(t)) \boldsymbol{\delta}(t) - \bar{\mathbf{x}}(t+1)). \quad (27)$$

When $t \rightarrow \infty$, we observe that $\|\Phi^T(t+1) - \Phi^T(t)\| \simeq 0$ and $\|\bar{\mathbf{x}}(t+1) - \bar{\mathbf{x}}(t)\| \simeq 0$, so $\hat{\mathbf{b}}(t+1)$ no longer depends on the template $\bar{\mathbf{x}}(t)$:

$$\hat{\mathbf{b}}(t+1) = \mathbf{b}(t) + \Phi^T(t) \mathbf{M}^{-1}(k(t), \theta(t)) \boldsymbol{\delta}(t). \quad (28)$$

$\hat{\mathbf{b}}(t+1)$ is used as a good initial estimate for the deformation modes $\mathbf{b}(t+1)$ in the next frame. By using the temporal coherence of the movement of the deformable structure, fast local optimization techniques can be used to obtain reliable MAP estimates. The experimental results show in this case that the optimal solution provided by global optimization techniques (stochastic algorithm) is indeed close to the initial estimate given by $\hat{\mathbf{b}}(t+1)$ in Eq. (27). For the parameters from the similarity group, we adopt the predictions at time $t+1$

$$\begin{aligned} \widehat{\mathbf{M}}(k(t+1), \theta(t+1)) &= \mathbf{M}(k(t), \theta(t)) \\ \widehat{\mathbf{T}}(t+1) &= \mathbf{T}(t) \end{aligned}$$

This method is able to provide an accurate and very compact representation of deformations. In Fig. 3, one can see that a very low number of deformation modes is required to integrate 98% of the total variability on a typical test sequence. However, this representation can be less compact than the method based on the alignment of extracted shapes according to a *Generalized Procrustes Analysis* [12, 3]. This remark will be completed in Section 3.5.

3.4 Detection of abrupt changes

When the tracked object undergoes slight deformations, we observe that $\|\boldsymbol{\delta}(t)\|$ remains nearly constant and is confirmed to be a pertinent measure to detect abrupt temporal changes. We wish to detect significant changes in the mean of $\|\boldsymbol{\delta}(t)\|$ since we observe that $\|\boldsymbol{\delta}(t)\|$ suddenly tends to increase if the shape variation between two successive frames is noticeable. In that case, we propose to re-estimate the local deformation process $\boldsymbol{\delta}(t)$ using a stochastic algorithm (*simulated annealing*) to avoid local maxima of the MAP criterion. In return, $\boldsymbol{\delta}(t)$ is estimated by means of an ICM algorithm [4] between two successive jumps instants.

We resort to a Cumulative sum test, Hinkley's test [21], to detect significant jumps of the variable $\|\boldsymbol{\delta}(t)\|$ among all meaningless small variations. This test was originally designed to the analysis of a gaussian white noise sequence in signal processing. In our case, it examines the sequence of observed quantities $\|\boldsymbol{\delta}(t)\|$ and so provide the significant jump instant. In practice, the computational load of a such test is low and requires no precise adjustment of involved thresholds. Since, we are interested only in the increase of $\|\boldsymbol{\delta}(t)\|$, only one test is performed to look for upwards jumps. We compute at each

instant the quantities

$$S_k = \sum_{p=0}^k \left(\|\delta(t)\| - m_0 - \frac{\nu_{min}}{2} \right) \quad (k \geq 0) \quad (29)$$

$$M_k = \min_{0 \leq l \leq k} S_l \quad (30)$$

in which m_0 is the on-line estimated mean of $\|\delta(t)\|$ and ν_{min} the expected minimal change magnitude. We accept the hypothesis that a change has occurred if

$$S_k - M_k > \gamma \quad (31)$$

where γ is the threshold of the test. The estimate of jump instant is the last minimum time before detection. In practice, this test is set off from the second frame.

3.5 Discussion

The advantage of the described method is that performance to identify statistical deformation modes thanks to bayesian estimators in comparison with an increase of the computational load. Note that the algorithm complexity essentially depends on the number of key points describing the shape outline. Some additional comments about our training algorithm are introduced in this section:

- It is clear than the training cannot reliably be performed in presence of occlusions of the object in the image sequence. The missing data would be sources of noise that will be incorporate as relevant signals in the training set [6]. The number of training examples must be very large in order to reduce this sensitivity [3].
- Our method is insensitive to the initial location of the model in the first image and is naturally robust to noise. In practice, we observe that the learning/tracking procedure is robust to partial occlusions if they occur once most of deformation modes have already correctly been identified.
- In [12, 10] and [3], a normalization step of training examples is necessary before the determination of statistical modes: the parameters from the group of similarity (rotation, translation and scale) enable to match all examples by minimizing a weighted mean square error between sets of feature points. In our approach, we have adopted bayesian estimators to perform this matching. In figure 5a, we see that the removing of the mean square error between two synthetic shapes induces a mismatching of feature points. Therefore, \bar{x} and Φ are not correctly determined given the ground truth (Fig. 5b). In real world examples, the temporal coherence of the motion enables to remove the ambiguity emphasized in figure 5. In addition, our technique gets round the difficulty of mismatching: the deformations are analyzed in a common referential and the matching

is implicit since all shapes are parametric deformed versions of the initial template. However, the KL expansion yields to eigenvalues upper to those obtained by minimization of a mean square error [10], increasing the complexity of the shape representation.

Figure 5 to be placed here

- No manual intervention is required which is helpful to reduce the tedious task of manual selection of feature points. The considerable amounts of expert intervention should be compared to the computational load necessary to automatically collect training shapes from an image sequence as described in this paper.
- Finally, the approach we propose is unfortunately limited since an initial template that approximates the target object outline must be provided by the expert. In addition, the number of feature points should be large enough to handle complex shapes with many degrees of freedoms (see Section 4).

4 Experimental results

In our experiments, we have considered the segmentation of deformable structures [18, 5, 10, 6, 22] corresponding to moving hands against static backgrounds (Figs. 6–8). This case study has already been considered in the literature [18, 5, 10, 6, 22, 1, 8] and is used here to demonstrate the validity of our training/tracking approach. In contrast to [1], the hand was considered here as a 2D structure undergoing 2D articulated motions. Henceforth, visual interpretation of hand gestures can help in achieving the ease and naturalness desired for man-machine interfaces [8, 36, 1].

The image sequences presented in Figs. 6-8 are composed of about 100 256×256 frames and the observations correspond to thresholded temporal gradients maps (see Section 2.3). The initial template of the hand is a 30-point model specified by the expert. We recall that the initial configuration of the template is defined at random on the first frame for the two experimental sequences. In our experiments, we have adopted a fast suboptimal exponential decreasing temperature schedule that controls the convergence of the *simulated annealing* algorithm. This temperature schedule yielded satisfactory final segmentations in practice.

Hand against an uniform background Figures 6a and 6b present respectively an intermediate step and the final results of the MAP segmentation for the first frame corresponding to the first experimental sequence. Figure 6a shows the estimation of the global transformations from the similarity group.

Global deformation modes are not yet available on this early stage of the training process. Figure 6b depicts the final segmentation including the local deformation process which exclusively contributes to the solution in this case. A time $t = 2$, a first training step is performed in order to derive the first eigenmodes $\Phi(2)$. This information will be exploited at time $t = 3$. The evolution of the number of eigenmodes and eigenvalues corresponding to this sequence is respectively showed in Figs.3 and 4. Figures 7a and 7d show the similar segmentation results for the third frame in which the deformation modes have been updated for the first time with the estimation obtained from the second frame. As can be seen, the local deformation process (Fig. 7d) is nearly unutilized on this frame because the configuration of the shape is close to the previous one. Hence the deformation modes captured on the second frame provide an excellent training for the third frame. Complementary results at time $t = 7$ and $t = 14$ are presented respectively in Figs. 7b-c) and F7e-f). We observe that the contribution of the local deformation process is low in presence of deformations already stored in the training set.

Figure 6 to be placed here

Figure 7 to be placed here

Hand against a textured background Figure 8 presents the same intermediate and final results for a second test sequence showing hand moving against a textured background. In this case, we have considered a cubic B-spline shape with 30 control points for the deformable model. Figures 8a-c show the results of the estimation of hyperparameters $\Theta(t)$ at time $t = 103$, $t = 105$ and $t = 109$. Figure 8d-f respectively correspond to the estimation of $\delta(t)$ at the same moment.

Figure 8 to be placed here

Here, the model has been built from a training set of 50 configurations estimated at each frame of the sequence. Figure 9 shows the 3 first deformation modes captured by the trained model. The template deformations are obtained by varying the model amplitudes b_k ($k = 1, 2, 3$) within two standard deviations $\sqrt{\lambda_k}$ where λ_k is the eigenvalue associated to the k^{th} variation mode. Figure 10 shows the

variation of the three first eigenmodes computed from 11 shapes selected manually.

Figure 9 to be placed here

Figure 10 to be placed here

In our experiments, the number of abrupt changes detected by the Cumulative sum test essentially depends on the complexity of the deformable motion. Although only a few abrupt changes were observed in practice on the two image sequences, the number of significant changes may be increased when the variation between two successive shapes is noticeable. In addition, the two tested sequences were too short to experimentally check the convergence of modes. In the future, more efforts should be done to improve and complete these preliminary results.

5 Conclusion

In this paper, we have presented a general framework for the modeling and unsupervised training of deformation modes of nonrigid objects. The technique relies on the definition of a prototype shape on which two deformation processes are applied. The deformations are described using statistical models and the optimal bayesian estimate of these deformations is computed using stochastic and deterministic optimization techniques.

The proposed modeling and algorithmic framework is comprehensive and suited to the representation of a large class of deformable objects. It may be adapted to segmentation problem based on other image attributes (luminance, color, texture, depth, etc.): the use of this technique for the learning of the nonrigid motion of a beating heart [27], is planned. The use of the learning procedure also yields promising future prospects as far as the characterization and the interpretation of the dynamic behavior of complex objects is concerned.

References

- [1] T. Ahmad, C.J. Taylor, A. Lanitis and T.F. Cootes, “Tracking and recognizing hand gestures, using statistical shape models”, *Image Vision Computing*, Vol. 15, 1997, pp. 345–352.
- [2] Y. Amit, U. Grenander, and M. Piccioni, “Structural image restoration through deformable templates”, *J. American Statist. Assoc.*, vol. 86, 1991, pp. 376–387.
- [3] A. Baumberg and D. Hogg, “Learning flexible models from image sequences”, *Proc. European Conf. Computer Vision*, Stockholm, 1994, pp. 299–308.

- [4] J. Besag, “On the statistical analysis of dirty pictures”, *J. Royal Statist. Soc. B*, Vol. 48, 1986, pp. 259–302.
- [5] A. Blake, R. Curwen, and A. Zisserman, “A framework for spatiotemporal control in the tracking of visual contours” *Int. J. Computer Vision*, Vol. 11, 1993, pp.127–145.
- [6] A. Blake, M. Isard and D. Reynard, “Learning to track the visual motion of contours”, *Artificial Intelligence*, Vol. 78, 1995, pp. 179–212.
- [7] D. Reynard, A. Wildenberg, A. Blake and J. Marchant, “Learning dynamics of complex motions from image sequences”,. *Proc. European Conf. on Computer Vision*, Cambridge, UK, April 1996, pp. 357–368.
- [8] M.A. Black and A.D. Jepson, “Eigentracking: robust matching and tracking of articulated objects using a view-based representation”, *Proc. European Conf. Computer Vision*, Cambridge, UK, April 1996, pp. 329–342.
- [9] P. Bouthemy and A. Benveniste, “Modeling of atmospheric disturbances in meteorological pictures”, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 6, 1984, pp. 587–600.
- [10] T.F. Cootes, C.J. Taylor, D.H. Cooper and J. Graham, “Active shape models – their training and application”, *CVGIP: Image Understanding*, Vol. 61, 1994, pp. 38–59.
- [11] T.F. Cootes and C.J. Taylor, “Combining point distribution models with shape models based on finite element analysis”, *Image and Vision Computing*, Vol. 13, 1995, pp. 403–409.
- [12] 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.
- [13] T.F. Cootes and C.J. Taylor, “Active shape models : smart snakes” *Proc. British, Machine Vision Conf.*, Leeds, UK, September 1992, pp. 266–275,.
- [14] G. Demoment, “Image reconstruction and restoration: overview of common estimation structures and problems” *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 37, 1989, pp. 2024–2036.
- [15] G.W. Donohoe, D.R. Hush, and N. Ahmed, “Change detection for target detection and classification in video sequences”, *Proc. Int. Conf. Acoust., Speech, Signal Processing*, New-York, 1988, pp. 1084–1087.
- [16] N. Friedland and D. Adam, “Automatic ventricular cavity boundary detection from sequential ultrasound images using simulated annealing”, *IEEE Trans. Medical Imaging*, Vol. 8, 1989, pp. 344–353.

- [17] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions and the bayesian restoration of images”, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 6, 1984, pp. 721–741.
- [18] U. Grenander, Y. Chow, and D.M. Keenan, *Hands. A Pattern Theoretic Study of Biological Shapes*, Springer Verlag, Berlin, Heidelberg, New-York, 1991.
- [19] U. Grenander and D.M. Keenan, “Towards automated image understanding”, *J. Applied Statistics*, Vol. 16, 1989, pp. 207–221.
- [20] F. Heitz and P. Bouthemy, “Multimodal estimation of discontinuous optical flow using Markov random fields”, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 15, 1993, pp. 1217–1232.
- [21] D.V. Hinkley, “Inference about the change-point from cumulative sum-tests”, *Biometrika*, Vol. 58, 1971, pp. 509–523.
- [22] M. Isard and A. Blake, “Contour tracking by stochastic propagation of conditional density”, *Proc. European Conference on Computer Vision*, Cambridge, UK, April 1996, pp. 343–356.
- [23] M. Kass, A. Witkin, and D. Terzopolous, “Snakes: Active contour models”, *Proc. Int. Conf. Computer Vision*, London, UK, 1987, pp. 259–268.
- [24] C. Kervrann and F. Heitz, “A hierarchical statistical framework for the segmentation of deformable objects in image sequences” *Proc. Conf. Comp. Vision Pattern Rec.*, Seattle, 1994, pp. 724–728.
- [25] C. Kervrann and F. Heitz, “Learning structure and deformation modes of nonrigid objects in long image sequences”, In *Proc. Int. Workshop on Automatic Face and Gesture Recognition*, Zurich, 1995, pp. 104–109.
- [26] C. Kervrann and F. Heitz, “A Markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics”, *IEEE Trans. Image Processing*, Vol. 4, 1995, pp. 856–862.
- [27] C. Kervrann and F. Heitz, “Statistical model-based segmentation of deformable motion”, *Proc. Int. Conf. Image Processing*, Lausanne, 1996, Vol. I, pp. 937–940.
- [28] C. Kervrann and F. Heitz, “A hierarchical markov modeling approach for the segmentation and tracking of deformable shapes” *Graphical Models Image Processing*, Vol. 60, 1998 (to appear).
- [29] A. Mohammad Djafari, “On the estimation of hyperparameters in bayesian approach of solving inverse problems”, *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Minneapolis, 1993, pp. 495–498.

- [30] J. Martin, A. Pentland and R. Kikinis, “Characterization of neuropathological shape deformations”, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 20, 1998, pp. 97–112.
- [31] T. McInerney, D. Terzopoulos, “Deformable models in medical image analysis: a survey”, *Medical Image Analysis*, Vol. 2, 1996, pp. 91–108.
- [32] H. Murase and S.K. Nayar, “Visual learning and recognition of 3D objects from appearance” *Int. J. Computer Vision*, Vol. 14, 1995, pp. 5–24.
- [33] C. Nastar and N. Ayache, “Fast segmentation, tracking and analysis of deformable objects”, *Proc. Int. Conf. Computer Vision*, Berlin, 1993, pp. 275–279.
- [34] J.M. Odobez and P. Bouthemy, “Detection of multiple objects using multiscale Markov random fields, with camera compensation”, *Proc. Int. Conf. Image Processing*, Austin, 1994, pp. 257–261.
- [35] E. Oja. – Subspace methods of pattern recognition, Research Studies Press, Hertfordshire, 1983.
- [36] V.I. Pavlovic, R. Sharma and T.S. Huang, “Visual interpretation of hand gestures for human-computer interaction: a review”, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 19, 1997, pp. 677–695.
- [37] A. Pentland and S. Sclaroff, “Closed-form solutions for physically based shape modeling and recognition”, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 13, 1991, pp. 715–729.
- [38] P. Perez and F. Heitz, “Restriction of a Markov Random Field on a graph and multiresolution statistical image modeling”, *IEEE Trans. Information Theory*, Vol. 42, 1996, pp. 180–190.
- [39] L.H. Staib and J.S. Duncan. – “Boundary finding with parametrically deformable models”, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 14, 1992, pp. 1061–1075.
- [40] D. Terzopoulos and R. Szelisky, “Tracking with Kalman snakes” in: A. Blake and A. Yuille, Ed., *Active Vision*, MIT-Press, Chapter 1, pp. 3–20, 1992.
- [41] M. Turk and A. Pentland, “Eigenfaces for recognition”, *J. of Cognitive neuroscience*, Vol. 3, 1991, pp. 71–86.
- [42] A. Yuille, P.W. Hallinan, and D.S. Cohen, “Feature extraction from faces using deformable templates”, *Proc. Int. J. Computer Vision*, Vol. 8, 1992, pp. 99–111.

List of footnotes

- C. Kervrann was with IRISA/INRIA Rennes, FRANCE. He is now with INRA, Biométrie et Intelligence Artificielle, Domaine de Vilvert, 78352 Jouy-en-Josas, FRANCE (e-mail: ck@jouy.inra.fr).

List of figure captions

- Figure 1: *Description of deformations.*
- Figure 2: *Observation maps: a) original image; b) thresholded temporal gradients.*
- Figure 3: *Number of deformation modes over time in the unsupervised training procedure (moving hand against an uniform background, see Fig. 6-7)*
- Figure 4: *Evolution over time of eigenvalues associated to the four first deformation modes (moving hand against an uniform background, see Fig. 6-7).*
- Figure 5: *Matching of two synthetic shapes: a) matching by removing of the mean square error between the two sets of points (unique rotation); b) matching of the two shapes corresponding to the ground truth.*
- Figure 6: *Model-based segmentation of a moving hand against an uniform background at time $t = 1$; a) Estimation of $\mathbf{M}(k(1), \theta(1))$ and $\mathbf{T}(1)$; b) Estimation of $\delta(1)$.*
- Figure 7: *Model-based segmentation of a moving hand against an uniform background.*
- Figure 8: *Model-based segmentation of a moving hand against a textured background.*
- Figure 9: *Deformations of a hand captured automatically (see text).*
- Figure 10: *Deformations of a hand captured manually (see text).*

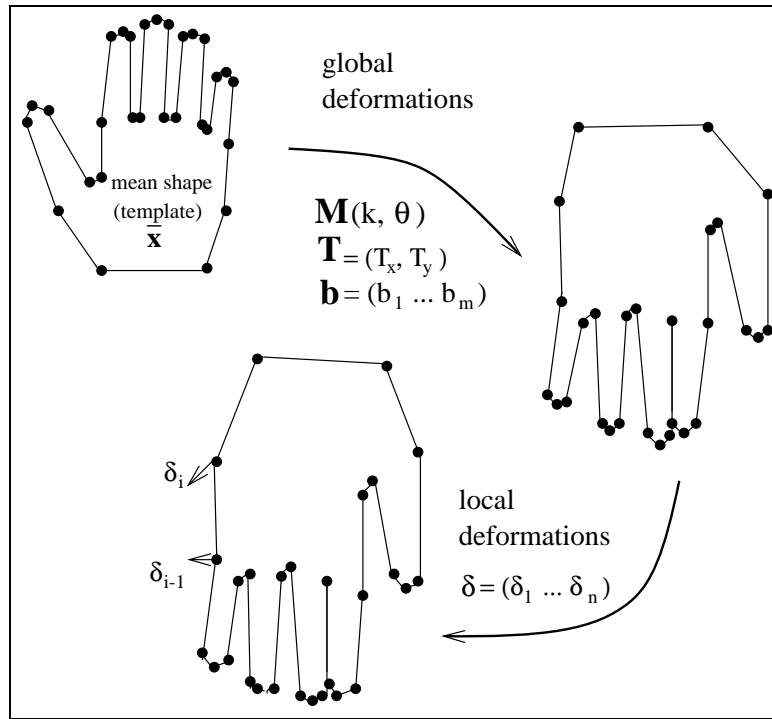


Figure 1: *Description of deformations.*



Figure 2: *Observation maps: a) original image; b) thresholded temporal gradients.*

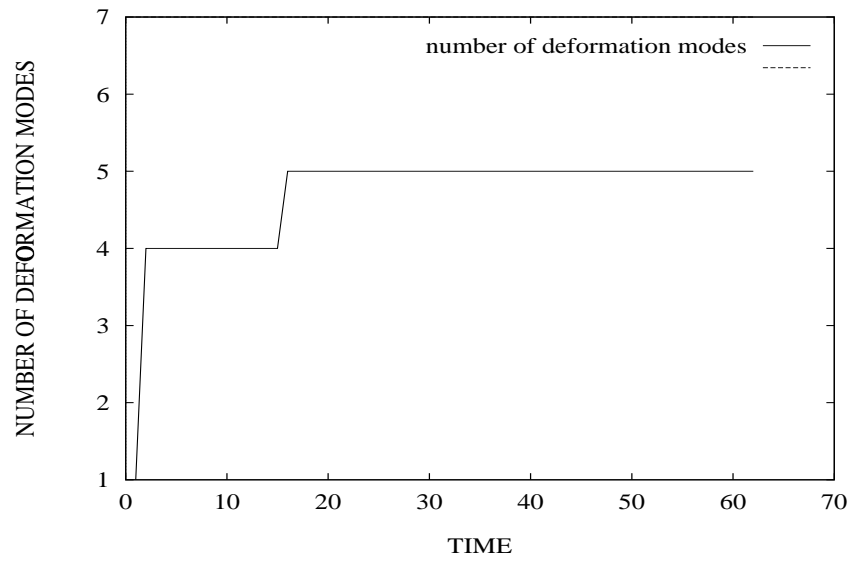


Figure 3: *Number of deformation modes over time in the unsupervised training procedure (moving hand against an uniform background, see Fig. 6-7).*

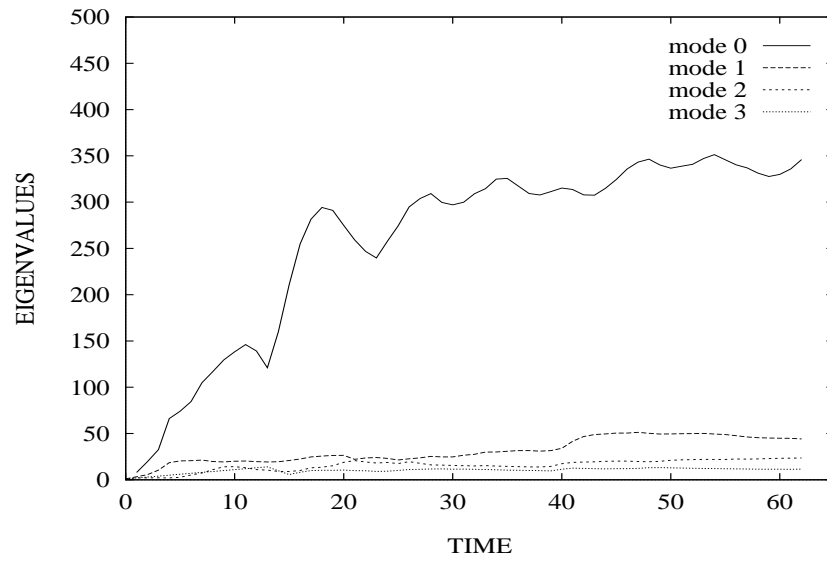


Figure 4: *Evolution over time of eigenvalues associated to the four first deformation modes (moving hand against an uniform background, see Fig. 6-7).*

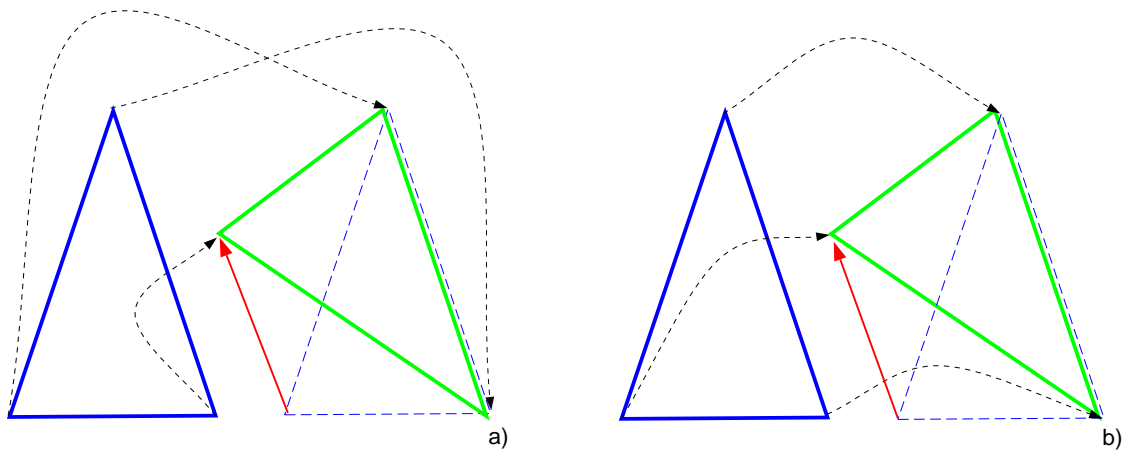


Figure 5: Matching of two synthetic shapes: a) matching by removing of the mean square error between the two sets of points (unique rotation); b) matching of the two shapes corresponding to the ground truth.

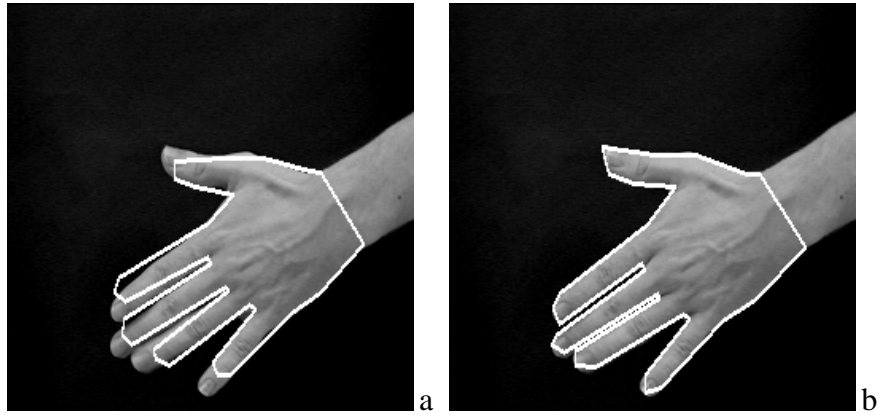


Figure 6: *Model-based segmentation of a moving hand against a uniform background at time $t = 1$;
a) Estimation of $\mathbf{M}(k(1), \theta(1))$ and $\mathbf{T}(1)$; b) Estimation of $\delta(1)$.*



Estimation of $M(k(t), \theta(t))$, $\mathbf{T}(t)$ and $\mathbf{b}(t)$ at $t = 3$ (Fig. 7a), $t = 7$ (Fig. 7b) and $t = 14$ (Fig. 7c).



Estimation of $\delta(t)$ at $t = 3$ (Fig. 7d), $t = 7$ (Fig. 7e) and $t = 14$ (Fig. 7f).

Figure 7: Model-based segmentation of a moving hand against an uniform background.



Estimation of $M(k(t), \theta(t))$, $\mathbf{T}(t)$ and $\mathbf{b}(t)$ at $t = 103$ (Fig. 8a), $t = 105$ (Fig. 8b) and $t = 109$ (Fig. 8c).



Estimation of $\delta(t)$ at $t = 103$ (Fig. 8d), $t = 105$ (Fig. 8e) and $t = 109$ (Fig. 8f).

Figure 8: *Model-based segmentation of a moving hand against a textured background.*

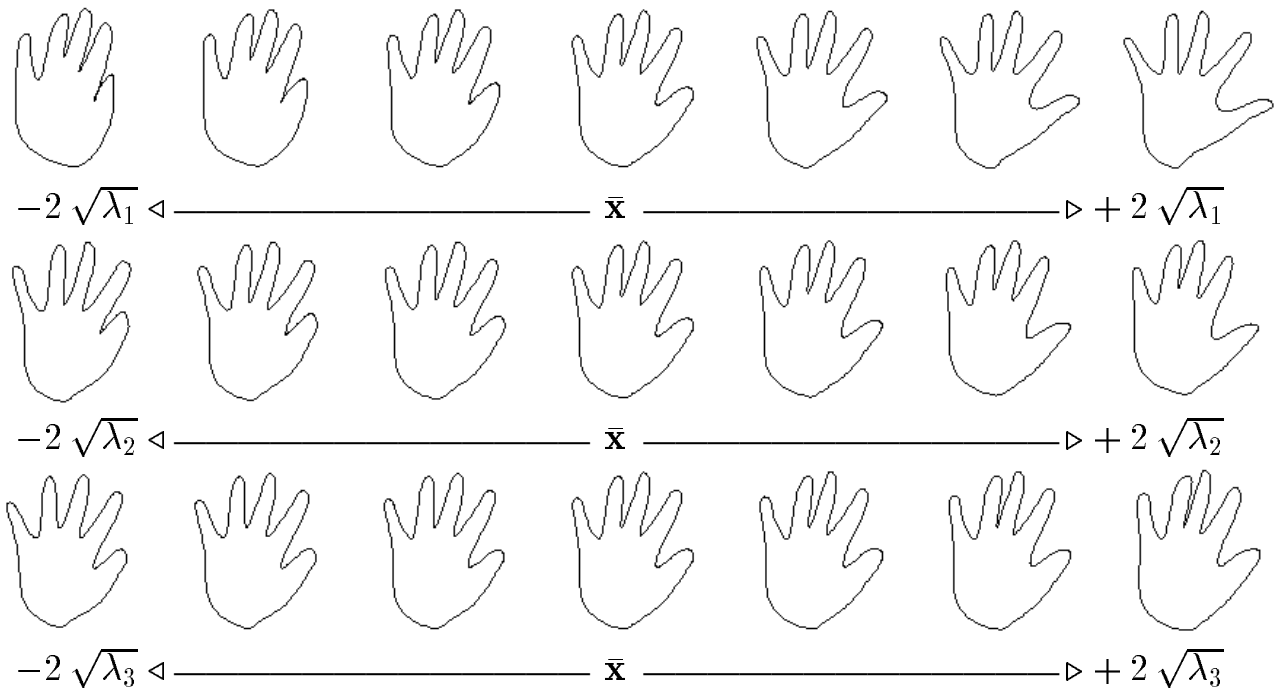


Figure 9: *Deformations of a hand captured automatically (see text).*

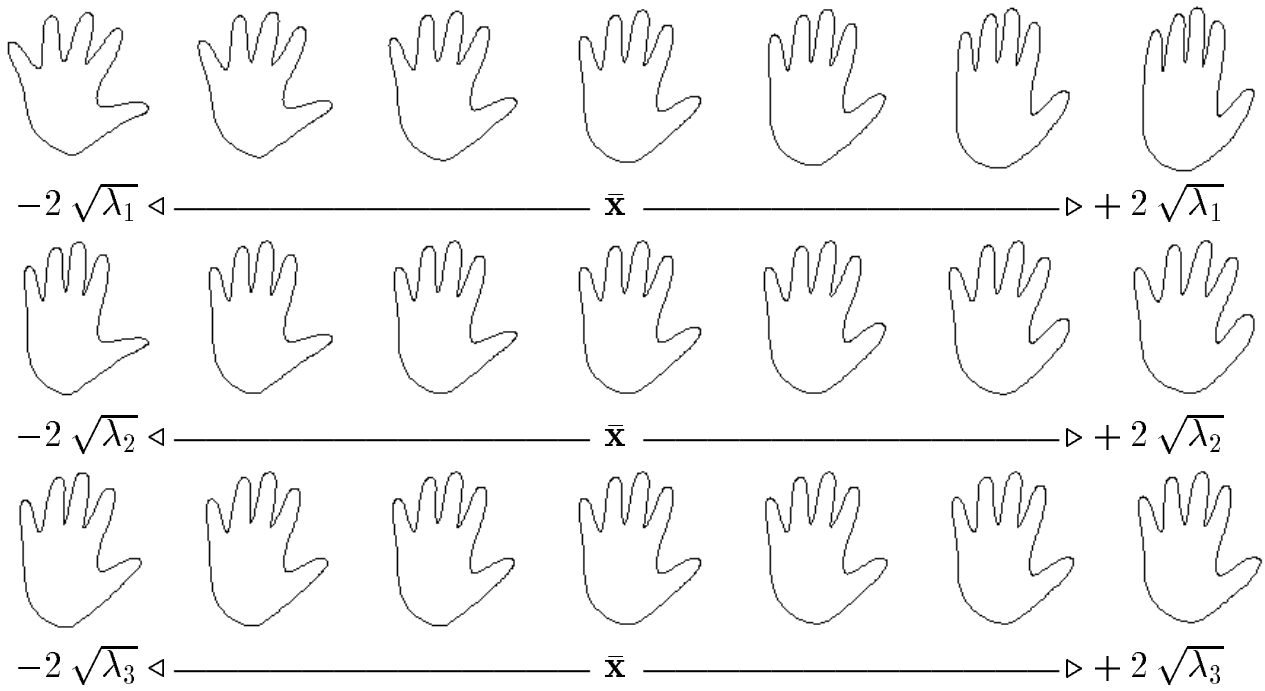


Figure 10: *Deformations of a hand captured manually (see text).*