Joint Object Segmentation and Behavior Classification in Image Sequences

Laura Gui[1], **Jean-Philippe Thiran**[1] and **Nikos Paragios**[2]

[1]Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2]Laboratoire MAS, Ecole Centrale de Paris, Chatenay-Malabry, France

Abstract

In this paper, we propose a general framework for fusing bottom-up segmentation with top-down object behavior classification over an image sequence. This approach is beneficial for both tasks, since it enables them to cooperate so that knowledge relevant to each can aid in the resolution of the other, thus enhancing the final result. In particular, classification offers dynamic probabilistic priors to guide segmentation, while segmentation supplies its results to classification, ensuring that they are consistent both with prior knowledge and with new image information. We demonstrate the effectiveness of our framework via a particular implementation for a hand gesture recognition application. The prior models are learned from training data using principal components analysis and they adapt dynamically to the content of new images. Our experimental results illustrate the robustness of our joint approach to segmentation and behavior classification in challenging conditions involving occlusions of the target object before a complex background.

1. Introduction

In the classical computer vision paradigm, the problems of image segmentation and object behavior [1] classification lie at different levels of abstraction. At a basic level, segmentation aims at extracting meaningful objects from the target image(s). A higher level image understanding task is to infer the behavior class of the object(s) extracted from each image, based on prior knowledge about behavior classes. For instance, one may want to classify object motion (e.g., car turn directions at an intersection), classify motion and deformation (e.g., hand gestures, body motions), or classify intensity changes in a brain activation map for diagnostic purposes. Generally, this inference is formulated in terms of a set of relevant attributes (e.g., color histogram, object position, orientation, shape, size, etc.), which have been extracted from the image sequence in a preceding phase. Thus, attribute extraction, which may or may not involve image segmentation, is conventionally performed separately from classification.

This paper pursues a joint solution to the problems of image segmentation and object behavior classification. Clearly, a precise segmentation of the target object would greatly facilitate behavior classification by making explicit any object attributes relevant to the classification task. Moreover, image segmentation could be dramatically improved by exploiting the knowledge which is available to the behavior classification task. This knowledge is typically represented in the form of probabilistic models of attribute values corresponding to behavior classes, and it can be used to guide the segmentation of the target object(s) in challenging conditions (e.g., images affected by noise or occlusions).

These considerations have motivated us to introduce a general framework for joint object segmentation and behavior classification in image sequences [11]. We formulated the segmentation in a variational setting, which enables the smooth integration of both prior knowledge (in the the form of behavior class models) and specific segmentation criteria for the target images. We further develop our framework in the present paper by introducing more complex prior models.

Variational methods offer a solid mathematical basis for the formulation and solution of many computer vision problems. In particular, the image segmentation problem has been formulated in terms of energy minimization, allowing the seamless blending of various criteria describing the desired solution, such as smoothness, region homogeneity, edge correspondence, etc. Starting with the original active contour (snakes) model [12], variational segmentation has been steadily advancing through the introduction of the Mumford-Shah model [16], the level set approach [17], geodesic active contours [3] and, more recently, versatile segmentation approaches such as [25, 18]. The segmentation of familiarly shaped objects in difficult cases has been made possible by the introduction of statistical shape priors into active contours [6] and also into level set active

[1]By "behavior" of an object in an image sequence, we refer to the temporal evolution of the object, as observed in the image sequence.

1-4244-1180-7/07/$25.00 ©2007 IEEE

contours [15, 5, 20] and the Mumford-Shah model [8, 9, 2]. Variational methods for contour evolution have also been adapted to object tracking (e.g., [12, 19, 8]). The coherence between frames has been exploited by approaches based on Kalman filtering [23], particle filtering [22], and autoregressive models [7].

Our framework fuses segmentation and behavior classification over image sequences. To our knowledge, this idea is novel in the context of variational image sequence analysis, and it capitalizes on existing developments in the use of shape priors. Segmentation has been combined with object recognition, yielding good results in the case of single, static images, both in variational [9] and non-variational [24, 14, 10, 13] settings. Our work makes a significant contribution in that we address *image sequences* and the temporal problem of *object behavior classification*. To tackle this problem, we introduce a variational framework that incorporates dynamic probabilistic priors automatically obtained via a machine learning approach. These priors are based on training data available for classification and they evolve dynamically as more information is accumulated from newly segmented images. Cooperation and shared access to all the available information (new images and priors) substantially improve both segmentation and behavior classification.

Note that in this paper we develop a *general framework* for the joint resolution of the two tasks—segmentation and behavior classification—which can have a wide range of applications by adapting its components and parameters according to the specific need. In particular, we illustrate the power of our approach in a gesture recognition application, where the combination of segmentation and classification dramatically increases the tolerance to occlusion and background complexity present in the input image sequence.

The remainder of the paper is organized as follows. Section 2 details the collaborating halves of our general framework, first behavior classification and then segmentation. In Section 3 we propose a particular implementation of our framework, which employs a specific image term and a dynamic prior component, for the purposes of gesture recognition. Experimental results are presented at the end of Section 3. Section 4 concludes the paper.

2. Formulation of the General Framework

Our general framework for joint segmentation and behavior classification is based on the idea of cooperation between the two processes along the target image sequence (Fig. 1). In an interleaved fashion, classification is performed, providing probabilistic attribute priors to guide segmentation towards the most likely objects at the given time instance, followed by segmentation, which identifies these objects in the new image (consistently with prior knowledge) and provides their attributes to classification. The priors offered by classification are learned from training data and they are updated dynamically according to newly processed images.

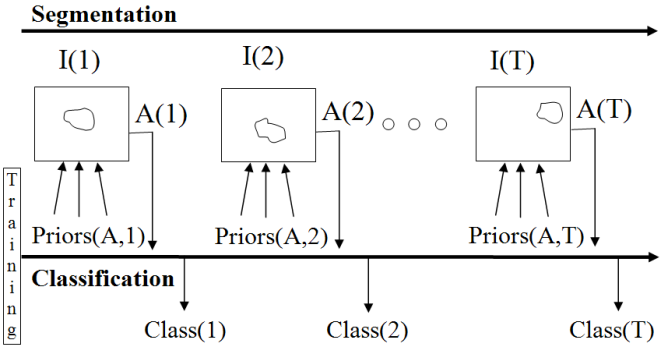


Figure 1. Our approach: Cooperation of segmentation and classification along the image sequence.

We use the generic term "attribute" to designate a visual property of the object of interest, which can be expressed as a functional $A(C, I)$ of the image I and of the object's segmenting contour C (A is assumed to be differentiable with respect to C). The palette of such attributes can be quite large, including all properties computable with boundary-based and/or region-based functionals, such as position, orientation, average intensity/color, or higher order statistics describing texture.

2.1. Behavior Classification and its Cooperation with Image Segmentation

Suppose that the segmentation problem is solved and we know the attribute values for the given image sequence. Then, behavior classification translates to finding the behavior classes which best account for the generation of these attributes in each image. We approach this problem in the machine learning framework of generative models [1]. In particular, we use Hidden Markov Models (HMMs) [21], where the hidden states (stochastic processes that generate observations) correspond to behavior classes and the observations are the attribute values. Once the HMM parameters are estimated from training samples, they can be used to classify new attribute sequences via the Viterbi algorithm. We can run this algorithm jointly with the segmentation of the image sequence in order to achieve the intended cooperation, as we show in the following.

We denote the states of the HMM (each corresponding to a behavior class) by $S = \{S_1, S_2, \ldots, S_M\}$, the state at time t by q_t, and the attribute value at time t by $A(t)$. The HMM parameters are

1. the initial state distribution $\pi = \{\pi_i\}$, with $\pi_i = P(q_1 = S_i), \quad i = 1..M$,

2. the state transition probability distribution $T = \{t_{ij}\}$, with $t_{ij} = P(q_{t+1} = S_j | q_t = S_i), \quad i, j = 1..M$, and

3. the state observation probability distributions (class

likelihoods):

$$P(A(t) \,|\, q_t = S_i) = P_i(A(t)),\ i = 1..M. \quad (1)$$

To support cooperation with the segmentation process, we require that these class likelihood functions $P_i(A(t))$ be differentiable with respect to $A(t)$.

Once the ensemble λ of HMM parameters have been estimated from training data, we can use the Viterbi algorithm [21] to classify new attribute sequences. For a new observation sequence $A_{1..T} = \{A(1), A(2), \ldots, A(T)\}$, the algorithm estimates the most likely state (behavior class) sequence $q^{\text{opt}}_{1..T} = \{q_1, q_2, \ldots, q_T\}^{\text{opt}}$ that generated it, as follows:

$$\begin{aligned} q^{\text{opt}}_{1..T} &= \arg\max_{q_{1..T}} P(q_{1..T} | A_{1..T}, \lambda) \\ &= \arg\max_{q_{1..T}} P(q_{1..T}, A_{1..T} | \lambda). \end{aligned} \quad (2)$$

This estimation translates to the evaluation, for each time step t and for each state S_i, of the quantity

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_{1..t-1}, q_t = S_i, A_{1..t} | \lambda). \quad (3)$$

It represents the highest probability along a state sequence, at time t, which explains the first t observations and ends in state S_i. After proper initialization, the following recursion is used to compute the δs:

$$\delta_t(i) = (\max_j \delta_{t-1}(j)\, t_{ji}) \cdot P_i(A(t) \,|\, \lambda). \quad (4)$$

These maximization results are stored and can be used at any time instance within the sequence to retrieve the (currently) optimal state sequence by backtracking.

We implement the cooperation between behavior classification and segmentation by using the probability estimates of the Viterbi algorithm at each step to guide the segmentation of each image. To this end, we run the algorithm and segmentation in an interleaved manner along the image sequence, using as observations the attributes of newly segmented images as soon as they become available. Suppose that we have completed step $t-1$ of both the segmentation and the Viterbi algorithm, so that attributes $A_{1..t-1}$ and $\delta_{t-1}(j),\ j = 1..M$ are available. To guide the segmentation of $I(t)$, we use the maximum amount of a priori knowledge offered by classification:

1. the predictions of each class i for the next attribute $A(t)$; i.e., the likelihood functions $P_i(A(t) \,|\, \lambda),\ i = 1..M$ (1), and

2. our relative confidence in the prediction of each class i, given by the Viterbi algorithm, i.e., the maximum probability of reaching state S_i at time step t, after having observed attributes $A_{1..t-1}$:

$$\begin{aligned} w_{t+1}(j) &= \max_{i=1..N} \delta_t(i) t_{ij} \\ &= \max_{q_1, q_2, \ldots, q_t} P(q_{1..t}, q_{t+1} = S_j, A_{1..t} | \lambda). \end{aligned} \quad (5)$$

More specifically, we use the product of these two quantities as prior information about the target object offered by each behavior class i. According to (4), this product is actually

$$\delta_t(A(t), i) = w_t(i)\, P_i(A(t) \,|\, \lambda),\ \ i = 1..M; \quad (6)$$

i.e., δ_t as a function of the unknown attribute $A(t)$. Next, we explain how to introduce these class contributions into the segmentation framework.

2.2. Image Segmentation and its Cooperation with Behavior Classification

To guide the segmentation process of an image, probabilistic attribute priors $\delta_t(A(t), i)$ are associated with each behavior class i. Our philosophy for dealing with these multiple priors is to create a competition between them, resulting in a final segmented object belonging to the class which best accounts for its generation, given the image evidence.

We formulate segmentation in a variational framework and use a labeling mechanism motivated by [9] to create competition between the multiple priors. Supposing that we have run our joint segmentation / behavior classification framework on the first $t-1$ frames of an image sequence, we employ the following energy functional in order to segment $I(t)$:

$$E(C, \mathcal{L}, I(t)) = E_{\text{data}}(C, I(t)) + \alpha E_{\text{prior}}(C, \mathcal{L}, I(t)), \quad (7)$$

where C is the segmenting contour, $\mathcal{L} = (L_1, \ldots L_M)$ is the set of labels (defined below) and α is a positive weighing constant. Energy $E_{\text{data}}(C, I(t))$ can be any boundary-based or region-based segmentation energy that best suits the application at hand (e.g., the energy proposed in [4]). The energy due to the priors is

$$\begin{aligned} E_{\text{prior}}(C, \mathcal{L}, I(t)) = &- \sum_{i=1}^{M} \log\left(\delta_t(A(C, I(t)), i)\right) L_i^2 \\ &+ \beta \left(1 - \sum_{i=1}^{M} L_i^2\right)^2, \end{aligned} \quad (8)$$

where β is a positive constant and the δ function is defined in (6). For each prior i, we associate a label L_i, a scalar variable that varies continuously between 0 and 1 during energy minimization and converges either to 1 (for the winning prior) or to 0 (for the other priors). The winner among

attribute priors is the one whose probability has been maximized through segmentation. Each of the prior terms carries a label factor equal to L_i^2 that controls its contribution to segmentation according to its relative probability with respect to the other priors. Competition is enforced by the soft constraint that the values of these factors should sum to 1, which is introduced by the term $(1 - \sum_{i=1}^{M} L_i^2)^2$ in energy (8).

We minimize (7) simultaneously with respect to the segmenting contour C and the labels $\mathcal{L}$ using the calculus of variations and gradient descent. The contour C is driven by image forces (intensity, gradients, etc.) due to $E_{\text{data}}(C)$ and by the M attribute priors due to $E_{\text{prior}}(C, \mathcal{L})$:

$$\frac{\partial C}{\partial \tau} = -\frac{\partial E_{\text{data}}(C, I(t))}{\partial C} - \alpha \frac{\partial E_{\text{prior}}(C, \mathcal{L}, I(t))}{\partial C}. \quad (9)$$

Here $\partial E_{\text{data}}(C, I(t))/\partial C$ can be derived through the calculus of variations for the particular chosen form of $E_{\text{data}}(C, I(t))$. The second term can be written as:

$$\frac{\partial E_{\text{prior}}(C, \mathcal{L}, I(t))}{\partial C} = -\sum_{i=1}^{M} \left(\frac{L_i^2}{\delta_t(A(C, I(t)), i)} \frac{\partial \delta_t(A(C, I(t)), i)}{\partial A} \frac{\partial A(C, I(t))}{\partial C} \right), \text{ with}$$
$$\frac{\partial \delta_t(A(C, I(t)), i)}{\partial A} = w_t(i) \frac{\partial P_i(A(C, I(t)) \mid \lambda)}{\partial A}. \quad (10)$$

The derivatives $\partial P_i/\partial A$ and $\partial A(C, I(t))/\partial C$ are computed according to the particular likelihood function and attribute employed.

The evolution equation for the label L_i is

$$\frac{\partial L_i}{\partial \tau} = \sum_{i=1}^{M} \delta_t(A(C, I(t)), i)\, L_i - \beta\, L_i \left(1 - \sum_{i=1}^{M} L_i^2 \right). \quad (11)$$

The effect of these equations is that the label L_i corresponding to the maximum $\delta_t(A(C, I(t)), i)$ will be driven towards 1—i.e., the maximum δ_t will be extremized—while the other labels will be driven to 0.

From a probabilistic perspective, the minimization of our proposed energy using competing priors can be interpreted as the maximization of the probability $\delta_t(A(C, I(t)), i)$ with respect to both the attribute $A(C, I(t))$ and the class i, subject to image-based constraints imposed through the energy $E_{\text{data}}(C, I(t))$. Then the segmentation of image $I(t)$ can be regarded as the joint estimation of the attribute value $A^*(t)$ and the class i^* as:

$$(A^*(t), i^*) = \arg \max_{A(C, I(t)), i} \delta_t(A(C, I(t)), i), \quad \text{subject to image constraints via } E_{\text{data}}(C, I(t)). \quad (12)$$

Thus, segmentation works concurrently towards the same goal as classification—maximizing the joint probability of the class and the observation at time t, while remaining consistent with previous observations, according to prior knowledge (through the HMM), and incorporating new information from image $I(t)$.

From the segmentation of $I(t)$ we obtain $A(t)$, so that we can estimate $\delta_t(i)$ and $w_{t+1}(i)$ (through the Viterbi recursion), then segment $I(t+1)$, and repeat the cycle to the end of the image sequence. We obtain the classification of the image sequence as the most probable state sequence given the observations, by backtracking from the results of the Viterbi algorithm.

3. A Specific Implementation of our Framework for Hand Gesture Recognition

In the following, we demonstrate the potential of our general framework through a particular implementation for a hand gesture recognition application. We begin by describing the problem that we wish to address, then we detail the specific models that we use within our framework and, finally, we present our results.

3.1. Application

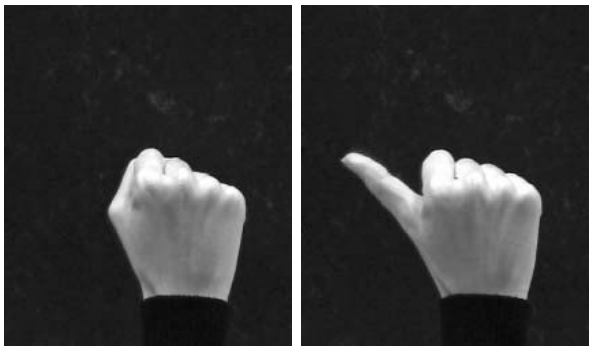
In our application, we identify four gesture classes consisting of a right hand going through four finger configurations: fist (Class 0), thumb extended (Class 1), thumb and index finger extended (Class 2) and thumb, index, and middle finger extended (Class 3). An example image of each gesture class is shown in Fig. 2.

Our gesture image sequences depict finger-counting from 1 to 3 (starting from the fist position) and from 3 to 1 (ending with the fist position), which is, in terms of gesture class successions, 0,1,2,3 and 3,2,1,0. Our aim is to perform joint segmentation and classification of image sequences containing such successions; i.e., for each image, extract the segmenting contour of the hand and determine the gesture class to which it belongs. To achieve this, we instantiate our general framework with particular segmentation and probability models, we use training data to estimate the HMM parameters and then we test the resulting implementation on new gesture image sequences.

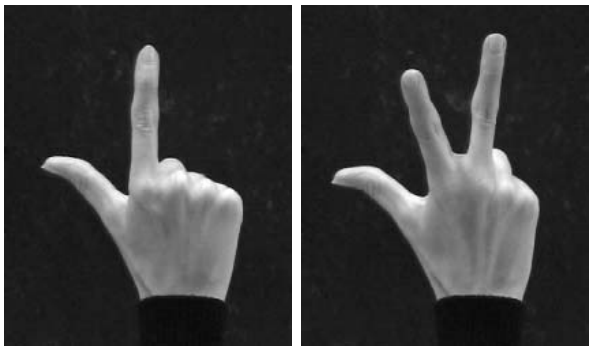
3.2. PCA-Based Prior Modeling

The object attribute that we use for this application is the contour segmenting the hand $A(C, I) = C$. We represent the contour using the level set approach [17], via the level set function (LSF) $\phi : \Omega \to \mathbb{R}$, chosen to be the signed distance function to the contour, so that $C \equiv \{(x, y) : \phi(x, y) = 0\}$.

The likelihood model $P_i(C \mid \lambda)$ for each class i relies on a shape distance function, motivated by [2], between the segmenting contour C and a prior contour corresponding to that class. While in our previous work [11], the likelihood model was a local Gaussian model whose parameters were



(a) Class 0 (b) Class 1



(c) Class 2 (d) Class 3

Figure 2. Samples from the four gesture classes that we use in our application.

fixed from training, here we use a prior contour for each class, which is computed from the training data via principal components analysis (PCA), that evolves dynamically during segmentation so as best to match the image information. We improve the distance function proposed in [2] by making it symmetric, so that the resulting likelihood models are suitable for classification.

The purpose of PCA is to reduce redundant information and summarize the main variations of a training set. Given a training set of discrete LSFs $\{\boldsymbol{\phi}_1, \dots \boldsymbol{\phi}_n\}$, which have been discretized on a rectangular grid, its principal directions of variation are captured by the eigenvectors $\{\mathbf{e}_1, \dots \mathbf{e}_n\}$ of the covariance matrix $\boldsymbol{\Sigma} = \frac{1}{n-1}\mathbf{M}\,\mathbf{M}^T$, where the column vectors of the matrix $\mathbf{M}$ are the n training LSFs. The singular value decomposition of the covariance matrix $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ is computed. An approximate representation of the training set is then obtained in the reduced space of the $p < n$ eigenvectors $\{\mathbf{e}_1, \dots \mathbf{e}_p\}$, which are the columns of $\mathbf{U}$ corresponding to the p largest singular values in the diagonal singular matrix $\mathbf{S}$. This enables us to approximate a new level set function $\hat{\boldsymbol{\phi}}$ using the p-dimensional vector of eigencoefficients $\mathbf{c}$, as:

$$\hat{\boldsymbol{\phi}} = \overline{\boldsymbol{\phi}} + \mathbf{E}\,\mathbf{c}, \tag{13}$$

where $\overline{\boldsymbol{\phi}} = (1/n)\sum_{i=1}^{n} \boldsymbol{\phi}_i$ is the mean of the training level set functions and $\mathbf{E}$ is a matrix whose columns are the eigenvectors $\{\mathbf{e}_1, \dots \mathbf{e}_p\}$.

Our shape distance function between the current segmenting contour ϕ and a continuously interpolated version of the PCA-represented level set function $\hat{\boldsymbol{\phi}}$ of a prior contour is given by:

$$d(\phi, \mathbf{c}, \boldsymbol{\tau}) = \iint_{\Omega} \left(\hat{\phi}^2 |\nabla\phi|\delta(\phi) + \phi^2 |\nabla\hat{\phi}|\delta(\hat{\phi}) \right) dx\, dy. \tag{14}$$

Here, δ is the Dirac function and $\hat{\phi}(\mathbf{c}, h_{\boldsymbol{\tau}})$ is the interpolated level set function of the prior contour, which depends on the eigencoefficient vector $\mathbf{c}$, according to (13), and on the parameters $\boldsymbol{\tau} = \{s, \theta, T_x, T_y\}$ of a similarity transformation

$$h_{\boldsymbol{\tau}}\left([x \;\; y]^T\right) = s \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix}. \tag{15}$$

This transformation aligns the prior contour with contour ϕ by scaling the former by s, rotating it by θ, and translating it by T_x, T_y. Since $\iint_{\Omega} |\nabla\phi|\delta(\phi)\, dx\, dy$ represents the length of the zero level set of ϕ, we can readily observe that the first term of (14) approximates the minimal Euclidian distance to the prior contour, integrated along the segmenting contour. This is an approximation because the level set function $\hat{\phi}$ resulting from PCA is not the exact distance function, but just a reasonable approximation of it. The second term of (14), which exchanges the roles of ϕ and $\hat{\phi}$ relative to the first term, makes the distance function symmetric and thus suitable for use in classification.

In our application, we use one PCA-based prior contour $\hat{\phi}_i$ for each class i, learned from training samples corresponding to that class and adapting to each new image in terms of its coefficients $\mathbf{c}^i(t)$ and $\boldsymbol{\tau}^i(t)$, as will be shown in the next section. Based on the shape distance function (14), we define the likelihood of the segmenting contour represented by ϕ, for time t (image $I(t)$) and class i as

$$P_i(\phi(t)) = e^{-d(\phi(t), \mathbf{c}^i(t), \boldsymbol{\tau}^i(t))}, \tag{16}$$

where $\mathbf{c}^i(t)$ are the PCA coefficients corresponding to class i and $\boldsymbol{\tau}^i(t)$ are the transformation parameters aligning the prior contour $\hat{\phi}_i$ of class i with $\phi(t)$.

3.3. Segmentation Process

As the data term in the segmentation energy (7), we use the piecewise constant Mumford-Shah model ([4]) to guide the evolution of the segmenting contour ϕ and of the prior contours $\hat{\phi}_i$ for each class i, in terms of their PCA coefficients $\mathbf{c}^i$ and similarity alignment parameters $\boldsymbol{\tau}^i$, as fol-

lows:

$$E_{\text{data}}(\phi,\, \mathbf{c}^{i=1..M}, \boldsymbol{\tau}^{i=1..M}) = \iint_\Omega (I-\mu_{\phi+})^2 H(\phi) + (I-\mu_{\phi-})^2 H(-\phi)\, dx\, dy + \sum_{i=1}^{M} \iint_\Omega (I-\mu_{\hat{\phi}_i+})^2 H(\hat{\phi}_i) + (I-\mu_{\hat{\phi}_i-})^2 (H(-\hat{\phi}_i))\, dx\, dy + \nu \iint_\Omega |\nabla H(\phi)|\, dx\, dy. \tag{17}$$

Here H is the Heaviside function and $\mu_{\phi+}$, $\mu_{\hat{\phi}_i+}$ and $\mu_{\phi-}$, $\mu_{\hat{\phi}_i-}$ are the mean image intensities over the positive, respectively negative, regions of the LSFs ϕ and $\hat{\phi}_i$, where $\hat{\phi}_i$ is the continuously interpolated LSF of the prior contour $\hat{\phi}_i(\mathbf{c}^i, h_{\boldsymbol{\tau}^i}(x,y)) = \overline{\boldsymbol{\phi}}_i(h_{\boldsymbol{\tau}^i}(x,y)) + \mathbf{E}_i(h_{\boldsymbol{\tau}^i}(x,y))\, \mathbf{c}^i$. The mean LSF $\overline{\boldsymbol{\phi}}_i$ and eigenvectors $\mathbf{E}_i$ have been obtained from the training set of class i. The last term of (17) imposes smoothness of the segmenting contour.

The prior term of the energy is given by:

$$E_{\text{prior}}(\phi, \mathcal{L}) = -\sum_{i=1}^{M} \log\left(\delta_t(\phi, i)\right) L_i^2 + \beta \left(1 - \sum_{i=1}^{M} L_i^2\right)^2, \tag{18}$$

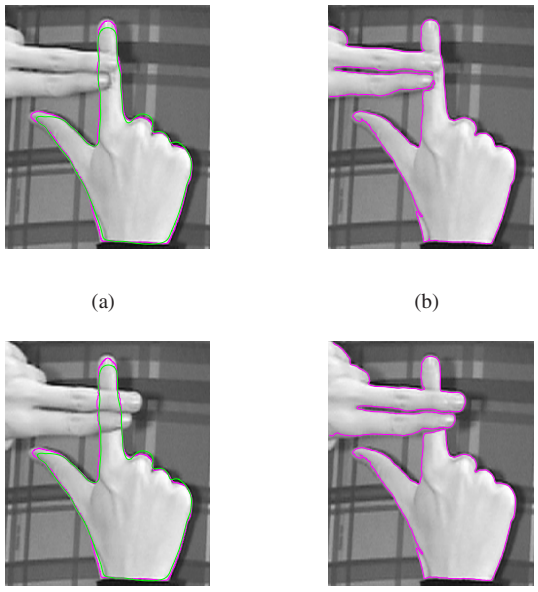
where $\delta_t(\phi, i) = w_t(i)\, P_i(\phi)$. Substituting likelihoods $P_i(\phi)$ with (16), the prior energy becomes

$$E_{\text{prior}}(\phi, \mathcal{L}, \mathbf{c}^{i=1..M}, \boldsymbol{\tau}^{i=1..M}) = \sum_{i=1}^{M} (d(\phi(t), \mathbf{c}^i(t), \boldsymbol{\tau}^i(t)) - \log w_t(i))\, L_i^2 + \beta \left(1 - \sum_{i=1}^{M} L_i^2\right)^2. \tag{19}$$

The total energy (7), combining (17) and (19), is minimized via the calculus of variations and gradient descent, yielding evolution equations for the contour ϕ, the labels $\mathcal{L}$, the PCA coefficients $\mathbf{c}^{i=1..M}$ and the alignment parameters $\boldsymbol{\tau}^{i=1..M}$.

3.4. Training the Model

In the training phase, we estimate the parameters of the HMM (see, e.g., [21]) using labeled sequences of LSFs obtained as segmentations of the mentioned training gesture sequences (0,1,2,3 and 3,2,1,0). The state observation probability distributions are given by the likelihoods $P_i(\phi)$, $i = 1..M$ from (16). Parameter estimation for these likelihoods amounts to PCA of the respective training LSFs, yielding the corresponding mean LSF $\overline{\boldsymbol{\phi}}_i$ and eigenvectors $\mathbf{E}_i$ for each class i. For this application, we used the first 5 eigenvectors corresponding to the largest eigenvalues, which account for 94.8%, 97.6%, 96.5%, and 95.5% of the variance of the training sets for classes 0, 1, 2, and 3, respectively.



(a) (b) (c) (d)

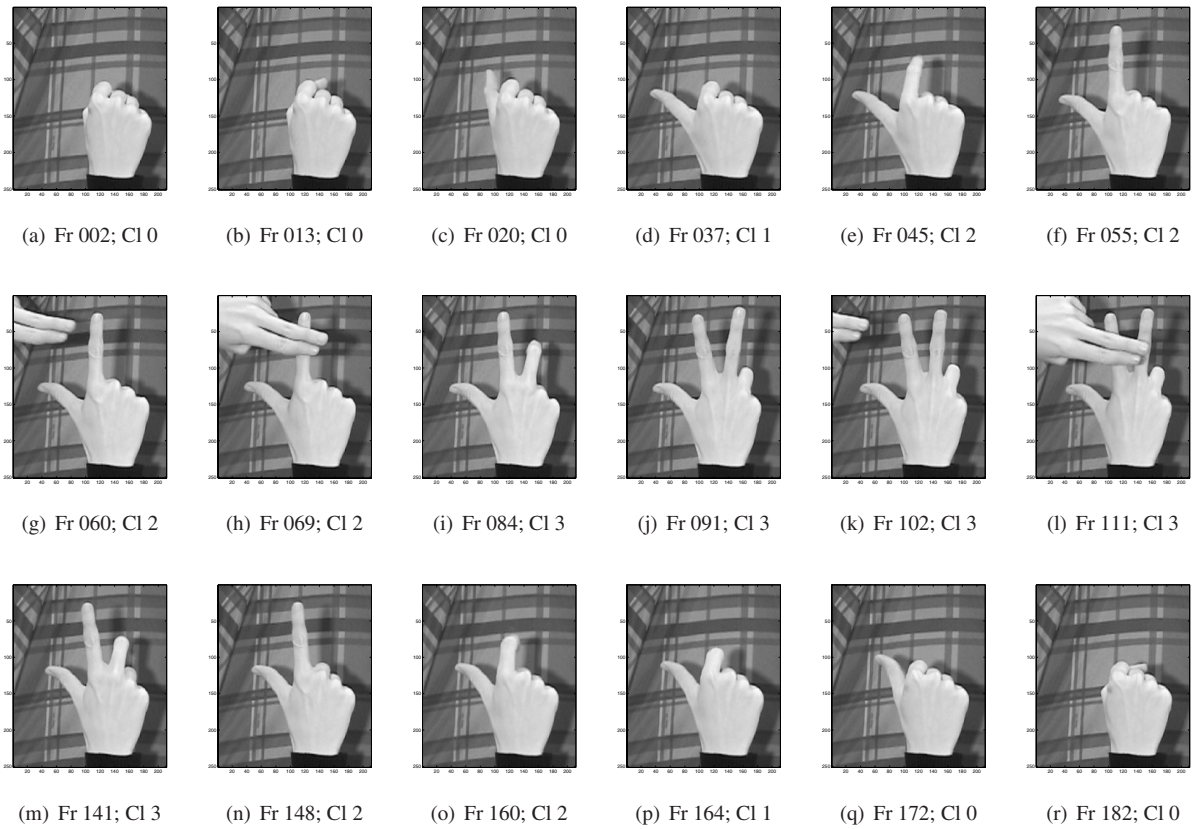
Figure 3. (Left) Segmentation (purple contour) with the proposed framework of an image in the presence of occlusion and background complexity. The green contour shows the best-fitting PCA prior model. (Right) Conventional segmentation of the image is confused by the occluding left hand.

3.5. Results

In the testing phase, we ran our implementation for joint behavior classification and segmentation on image sequences of a hand performing the succession of gestures 0,1,2,3,2,1,0 in front of a complex background, degraded by significant occlusions (Fig. 4). The frame number and resulting classification of each frame are indicated in the figure.

By virtue of the prior information supplied by the classification, segmentation with the PCA prior model is able to cope with occlusions as can be seen in Fig. 3(a). By contrast, Fig. 3(b) shows that conventional segmentation fails.

Figure 5 shows the classification results for all the frames of the sequence presented in Fig. 4, which correctly follow our understanding of the sequence in terms of the executed gestures. Moreover, the frame classification obtained by backtracking from the Viterbi algorithm corresponds to the partial classification results obtained throughout the sequence, which have been used to guide segmentation. This concordance can be seen in Fig. 5, which exhibits, as functions of time (frame), (a) the final classification, (b) the delta functions of each class, and (c) the prior confidence of each class (the w function) used as input to the segmentation. The w values have been scaled with respect to their maxi-



(a) Fr 002; Cl 0 (b) Fr 013; Cl 0 (c) Fr 020; Cl 0 (d) Fr 037; Cl 1 (e) Fr 045; Cl 2 (f) Fr 055; Cl 2

(g) Fr 060; Cl 2 (h) Fr 069; Cl 2 (i) Fr 084; Cl 3 (j) Fr 091; Cl 3 (k) Fr 102; Cl 3 (l) Fr 111; Cl 3

(m) Fr 141; Cl 3 (n) Fr 148; Cl 2 (o) Fr 160; Cl 2 (p) Fr 164; Cl 1 (q) Fr 172; Cl 0 (r) Fr 182; Cl 0

Figure 4. Frames sampled from a test image sequence of the right hand performing the 0,1,2,3,2,1,0 gesture in front of a complex background. Note the left hand enters the scene around Frame 60 and again around frame 102, significantly occluding the right hand around Frames 69 and and 111. The frame number and resulting classification of each frame are indicated.

mum value for every frame.

4. Conclusion

We have introduced and developed a novel and general framework that enables the joint segmentation of image sequences and classification of object behavior in these sequences. Cooperation between the segmentation and classification processes facilitates a mutual exchange of information, which is beneficial to their joint success. In particular, we employed a classification strategy based on generative models that provided dynamic probabilistic attribute priors to guide image segmentation. These priors enabled the segmentation process to work towards the same goal as classification, by outlining the object that best accounted both for the image data and for the prior knowledge encapsulated in the generative model.

We illustrated the effectiveness of our general framework in a hand gesture analysis application, where we successfully segmented and classified image sequences of a gesturing hand before a complex background in the presence of occlusions.

Future directions of our work include the development of more complex attribute priors that would enable us to deal with more challenging application scenarios involving under-constrained problems, such as the understanding of 3D object behavior from 2D monocular images. We will also investigate the use of alternative optimization methods that are less sensitive to local minima.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 2

[2] X. Bresson, P. Vandergheynst, and J.-P. Thiran. A variational model for object segmentation using boundary information and shape prior driven by the Mumford-Shah functional. *International Journal of Computer Vision*, 28(2):145 – 162, July 2006. 2, 4, 5

[3] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. In *Proc. IEEE Intl. Conf. on Comp. Vis.*, pages 694–699, Boston, USA, 1995. 1

[4] T. Chan and L. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001. 3, 5

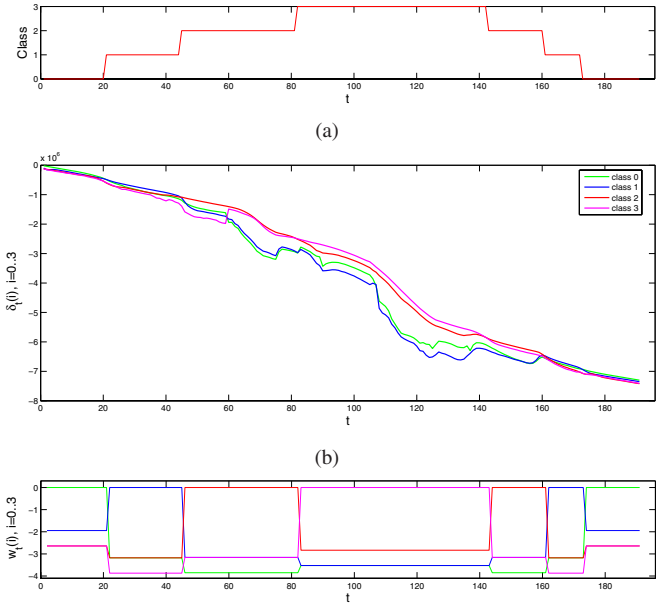


Figure 5. Classification results plotted per frame. (a) Final classification. (b) Delta functions of each class. (c) Prior confidence of each class used as input to the segmentation.

[5] Y. Chen, H. Tagare, S. Thiruvenkadam, F. Huang, D. Wilson, K. Gopinath, R. Briggs, and E. Geiser. Using prior shapes in geometric active contours in a variational framework. *International Journal of Computer Vision*, 50(3):315–328, 2002. 2

[6] T. Cootes, C. Beeston, G. Edwards, and C. Taylor. Unified framework for atlas matching using active appearance models. *Intl Conf. Inf. Proc. in Med. Imaging*, pages 322–333, 1999. 1

[7] D. Cremers and G. Funka-Lea. Dynamical statistical shape priors for level set based tracking. In S. LNCS, editor, *3rd. Workshop on Variational, Geometric and Level Set Methods in Computer Vision*, volume 3752, pages 210–221, 2005. 2

[8] D. Cremers, S. Osher, and S. Soatto. Kernel density estimation and intrinsic alignment for knowledge-driven segmentation: Teaching level sets to walk. *Pattern Recognition*, 3175:36–44, 2004. 2

[9] D. Cremers, N. Sochen, and C. Schnör. Multiphase dynamic labeling for variational recognition-driven image segmentation. In *European Conf. on Computer Vision*, volume 3024, pages 74–86, 2004. 2, 3

[10] V. Ferrari, T. Tuytelaars, and L. V. Gool. Simultaneous object recognition and segmentation by image exploration. In *ECCV*, 2004. 2

[11] L. Gui, J.-P. Thiran, and N. Paragios. A variational framework for the simultaneous segmentation and object behavior classification of image sequences. In *Proc. Scale Space and Variational Methods in Computer Vision*, 2007. In press. 1, 4

[12] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331, 1987. 1, 2

[13] I. Kokkinos and P. Maragos. An Expectation Maximization approach to the synergy between image segmentation and object categorization. In *ICCV*, pages 617–624, 2005. 2

[14] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshop on SLCV*, 2004. 2

[15] M. Leventon, W. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 316–323, June 2000. 2

[16] D. Mumford and J.Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications in Pure and Applied Mathematics*, 42:577–685, 1989. 1

[17] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on the Hamilton-Jacobi formulation. *Journal of Computational Physics*, 79:12–49, 1988. 1, 4

[18] N. Paragios and R. Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision*, 46(3):223–247, 2002. 1

[19] N. Paragios and R. Deriche. Geodesic active regions and level set methods for motion estimation and tracking. *Computer Vision and Image Understanding*, 97:259–282, 2005. 2

[20] N. Paragios and M. Rousson. Shape priors for level set representations. In *European Conference in Computer Vision*, volume 2, pages 78–92, 2002. 2

[21] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 1989. 2, 3, 6

[22] Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi. Particle filtering for geometric active contours with application to tracking moving and deforming objects. In *Proc. CVPR*, volume 2, pages 2–9, 2005. 2

[23] D. Terzopoulos and R. Szeliski. Tracking with Kalman snakes. *Active vision*, pages 3–20, 1993. 2

[24] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image parsing: Segmentation, detection, and recognition. In *ICCV*, pages 18–25, 2003. 2

[25] L. Vese and T. Chan. A multiphase level set framework for image segmentation using the Mumford and Shah model. *International Journal of Computer Vision*, 50(3):271–293, 2002. 1