Simultaneous Segmentation and Object Behavior Classification of Image Sequences

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

[1] Signal Processing Institute (ITS),
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland

[2] MAS - Ecole Centrale de Paris,
Chatenay-Malabry, France

Technical Report TR_ITS_ 2007.04

Abstract

In this paper, we advance the state of the art in variational image segmentation through the fusion of bottom-up segmentation and top-down classification of object behavior over an image sequence. Such an approach is beneficial for both tasks and is carried out through a joint optimization, which enables the two tasks to cooperate, such 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 with prior knowledge. The prior models are learned from training data and they adapt dynamically, based on segmentations of earlier images in the sequence. We demonstrate the power of our approach in a hand gesture recognition application, where the combined use of segmentation and classification dramatically improves robustness in the presence of occlusion and background complexity.

1 Introduction

Image segmentation is one of the most basic yet most challenging problems of computer vision. Segmentation requires finding in an image semantically salient regions (or their bounding contours) associated with "objects". Behavior classification (or recognition) in image sequences is an important higher level task towards comprehensive visual perception. By the "behavior" of an object in an image sequence, we mean the temporal evolution of one or more of its attributes (such as position, orientation, shape, color, texture, etc.) apparent in the image sequence. Thus, classifying object behavior means associating with each of its temporal evolution instances one of several possible behavior class

labels. For example, we would like to classify object motion (e.g., tell whether a car in an intersection is turning right, left, going straight or performing a succession of these motions), classify motion and deformation (e.g., tell whether a person is running or walking), or classify successive intensity changes in a brain activation map for clinical purposes.

The conventional approach is to solve the segmentation and behavior classification problems separately and sequentially—i.e., segment the image sequence, extract the relevant features, and finally classify the time evolution of these features. However, behavior classification can be greatly facilitated if segmentation information is available. Reciprocally, image segmentation can greatly benefit from the consideration of additional information about the targeted object(s), such as shape, color, texture, etc. This kind of information is usually brought to bear on classification tasks in the form of *a priori* models of the classes to be distinguished, generally based on training instances of these classes. Therefore, benefits should accrue from a collaboration between image segmentation and behavior classification.

Our contribution in this paper is a *joint* solution of the two problems of image sequence segmentation and classification of object behavior, which enables the information related to each of them to enhance the results of both. To this end, we develop a new *variational framework* that smoothly integrates the two main sources of information: the target image sequence and the prior behavior models, which adapt dynamically as a function of the segmented images, through the classification strategy.

Variational methods underlie the mathematical formulation of numerous computer vision problems. The image segmentation problem has been formulated in terms of energy minimization, where one can seamlessly introduce various criteria describing the desired solution, such as smoothness, region homogeneity, edge correspondence, etc. Starting with the original active contour (snakes) model [12], continuing with the Mumford-Shah model [16], the introduction of the level set approach [17] and geodesic active contours [3], recent work has yielded versatile segmentation approaches such as [25, 18]. Statistical shape priors were introduced into active contours [6] and also into level set active contours [15, 5, 20] and the Mumford-Shah segmentation [9, 10, 2]. These techniques have made it possible to successfully segment a familiarly-shaped object in difficult cases. Variational methods for contour evolution have also been adopted for object tracking (e.g., [12, 19, 8, 9]). Coherence between frames has been exploited by approaches based on Kalman filtering [23], particle filtering [22], and autoregressive models [7].

Our new variational framework deals simultaneously with the issues of image sequence segmentation and object behavior classification, thus fusing the levels of image analysis and image understanding. By performing the two tasks cooperatively for a given image sequence, we enable them to benefit from all of the available information, which mutually increases their chances of success. On the one hand, segmentation is improved by guidance towards the target object via probabilistic priors, offered by classification. These priors are based on training data available for classification and they are able to evolve dynamically as more

information is accumulated from newly segmented images. On the other hand, classification is improved from the consideration of segmentation results, captured from new images, while also maintaining consistency with prior knowledge and with previous segmentations in the sequence. To our knowledge, the fusion of segmentation and behavior classification over image sequences is novel in the domain of variational image analysis, while it of course capitalizes on existing experience in the use of shape priors. The idea of combining segmentation and object recognition has previously yielded good results in the case of single, static images both in variational [10] and non-variational settings [24, 14, 11, 13]. 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. We illustrate the power of our proposed 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.

Note that in this paper we propose 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. The next section details the collaborating halves of our general framework, first behavior classification and then segmentation. A particular implementation of the framework is then proposed in Section 3, which employs a specific image term and 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 Variational Framework

Our goal is to *segment* an image sequence and *classify* it in terms of object behavior. As illustrated in Fig. 1, the key idea in our framework is to *interweave* the classification and segmentation processes while iterating through the given image sequence. This enables them to collaborate in exploiting the available prior knowledge and to improve each other by sharing partial results obtained throughout the image sequence. More concretely, for each image in the sequence, classification offers dynamic probabilistic attribute priors to guide segmentation. These priors, which are based on training, adapt in time according to knowledge gained from past segmentations. In turn, segmentation detects, and supplies to classification, object attributes that best explain the image evidence, consistently with the prior knowledge. These object attributes are used in the subsequent step of the classification, and so on, until the entire sequence is segmented and classified.

Note that 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 differen-

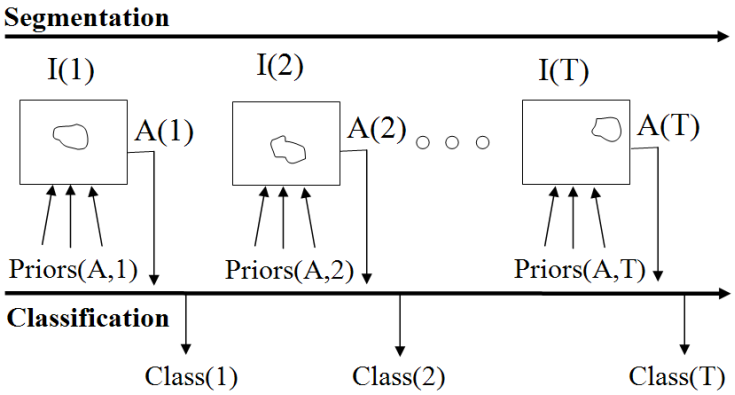


Figure 1: Our approach: cooperation of segmentation and classification along the image sequence.

tiable with respect to C). The palette of such attributes is 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 Classification and its Cooperation with Segmentation

The behavior classification task amounts to estimating, for each time instance of an image sequence, the behavior class of the object, based on its attributes. Supposing for the moment that the attribute values are known, we need only find the generating behavior classes. We solve this problem using the machine learning concept of generative models [1], in particular Hidden Markov Models (HMMs) [21], where the observations are attribute values and the hidden states are the unknown behavior classes. Once trained on typical attribute evolution sequences, an HMM classifies new attribute sequences by estimating the most likely state sequence generating them.

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):

$$P(A(t) \,|\, q_t = S_i) = P_i(A(t)), \; i = 1..M. \tag{1}$$

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 having estimated the ensemble λ of HMM parameters from training data, the HMM can be used to classify new attribute sequences. In order to assign a behavior class to each observation in a new sequence $A_{1..T} = \{A(1), A(2), \ldots, A(T)\}$, we estimate the state sequence $q_{1..T}^{\text{opt}} = \{q_1, q_2, \ldots, q_T\}^{\text{opt}}$ that best explains the observation sequence:

$$q_{1..T}^{\text{opt}} = \arg\max_{q_{1..T}} P(q_{1..T}|A_{1..T}, \lambda) = \arg\max_{q_{1..T}} P(q_{1..T}, A_{1..T}|\lambda), \tag{2}$$

using the Viterbi algorithm [21]. At each time step t and for each state S_i, the Viterbi algorithm calculates the quantity

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

representing the highest probability along a state sequence, at time t, which explains the first t observations and ends in state S_i. This quantity is computed by initializing the δs and then using the following recursion:

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

Finally, the optimal state sequence is retrieved by backtracking from these maximization results.

Thus, the Viterbi algorithm iterates through the attribute sequence, computing its best estimate for the probability of different generating classes, given the knowledge accumulated in the HMM. We can use these estimates to guide the segmentation process. The idea is to run this algorithm synchronously with the segmentation, using the attribute of the segmented object as the next observation, as soon as it becomes available. Then, we incorporate the algorithm's best momentary class estimations as attribute priors for the segmentation of the next image in the sequence.

Now, suppose 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. In order to segment $I(t)$, we use the maximum available *a priori* knowledge:

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)

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 $A_{1..t-1}$:

$$w_t(i) = \max_{j=1..M} \delta_{t-1}(j) t_{ji} = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_{1..t-1}, q_t = S_i, A_{1..t-1}|\lambda). \tag{5}$$

As prior information offered by each behavior class i, we shall use the product of these two quantities, which according to (4) is actually

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

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 Segmentation and its Cooperation with Classification

We take a variational approach to segmentation that incorporates the dynamic probabilistic priors offered by classification. For an image $I(t)$, these priors consist of the delta functions of the object attribute corresponding to each class i; i.e., $\delta_t(A(t), i)$. We introduce these class contributions into the segmentation model by means of a competition mechanism, since we are searching for a single "winning" class that best accounts for the generation of the next observation.

To create a "competition" during segmentation among the priors associated with different classes, we employ a labeling mechanism similar to the one in [10]. For each of the priors i we use one label L_i, a scalar variable that varies continuously between 0 and 1 during energy minimization and converges either to 0 or 1. The value of the ensemble of labels $\mathcal{L} = (L_1, \dots L_M)$ after convergence designates a "winner" among the attribute priors, corresponding to the probability which has been maximized through segmentation. Each of the prior terms carries a label factor equal to L_i^2 and the competition between them is enforced by the constraint that the values of these factors should sum to 1. This constraint is introduced by adding the term $(1 - \sum_{i=1}^{M} L_i^2)^2$ to the segmentation energy.

Once having run our joint segmentation/classification framework on the first $t-1$ frames of an image sequence, we segment $I(t)$ by minimizing with respect to the contour C and the labels $\mathcal{L}$ the following energy functional:

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

where α is a positive weighing constant. Here $E_{\text{data}}(C, I(t))$ can be any boundary-based or region-based segmentation energy, suitable to the application at hand (e.g., the energy proposed in [4]). The energy due to the priors is

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

where β is a positive constant and the δ function is defined in (6).

The minimization of (7) simultaneously with respect to the segmenting contour C and the label vector $\mathcal{L}$ is performed 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 $E_{\text{prior}}(C, \mathcal{L})$. At the same time, the labels evolve according to the competition between the priors, so as to maximize the probability of the most likely prior, given image evidence.

The evolution equation for the contour C is:

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

Here $\partial E_{\text{data}}(C, I(t))/\partial C$ represents the contribution of the image-based term to segmentation and can be derived through calculus of variations for the particular

form of $E_{\text{data}}(C, I(t))$ that we choose. The second term can be written as:

$$\begin{aligned}\frac{\partial E_{\text{prior}}(C, \mathcal{L}, I(t))}{\partial C} = -\sum_{i=1}^{M} \Bigg(& \frac{1}{\delta_t(A(C, I(t)), i)} \\ & \cdot \frac{\partial \delta_t(A(C, I(t)), i)}{\partial A} \cdot \frac{\partial A(C, I(t))}{\partial C} \cdot L_i^2 \Bigg), \text{ with} \\ \frac{\partial \delta_t(A(C, I(t)), i)}{\partial A} = & \; w_t(i) \, \frac{\partial P_i(A(C, I(t)) \,|\, \lambda)}{\partial A}.\end{aligned} \tag{10}$$

Both 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 labeling function L_i is

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

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 the perspective of the cooperation between segmentation and classification, the minimization of our proposed energy, where the priors from different classes are in competition with each other, amounts to the maximization of the probability $\delta_t(A(t), i)$ with respect to both the attribute $A(t)$ and class i, subject to image-based constraints. 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:

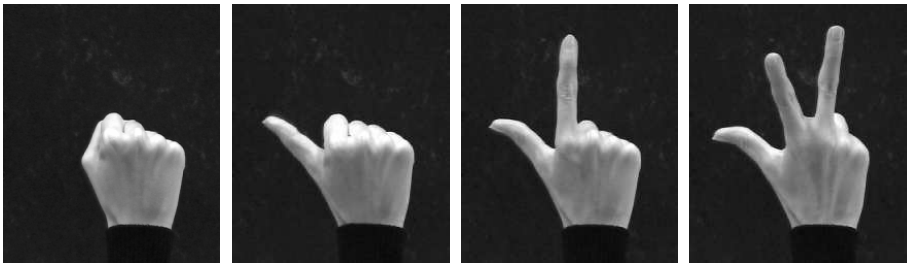
$$\begin{aligned}&(A^*(t), i^*) = \arg \max_{A(t), i} \delta_t(A(t), i), \\ &\text{subject to image constraints } (A(t), I(t)).\end{aligned} \tag{12}$$

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 $I(t)$.

The segmentation of $I(t)$ yields $A(t)$, enabling the Viterbi algorithm to estimate $\delta_t(i)$ and $w_{t+1}(i)$, so that we can continue by segmenting $I(t+1)$ and repeat the cycle to the end of the image sequence. Finally, 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 Implementing our Framework for Hand Gesture Recognition

We now demonstrate the strength of our framework of Section 2 in a hand gesture recognition application. We begin by describing the problem that we



(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.

wish to address. Then, we detail two particular implementations of our general framework, including the specific models that we use. Finally, we present the results obtained with these implementations.

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.

Given an image sequence of such gestures, our goal is to perform joint segmentation and classification; i.e., for each image, extract the segmenting contour of the hand and determine the gesture class to which it belongs. Note that our gesture image sequence depicts finger-counting from 1 to 3 and back to 1, ending with the initial fist position; i.e., the following succession of gesture classes: 0,1,2,3,2,1,0. Our strategy is first to train a 4-class HMM with such sequences and then to incorporate the HMM into our framework in order to segment and classify new sequences of this sort.

3.2 Solutions using the proposed framework

For this application, the object attribute employed within our framework is the contour segmenting the hand $A(C, I) = C$. Using the level set approach [17], we represent the contour by the level set function (LSF) $\phi : \Omega \to \mathbb{R}$, chosen to be the signed distance function to the contour, so that $C \equiv \{(x, y) : \phi(x, y) = 0\}$.

As a *data term* in the segmentation energy (7), guiding the evolution of the

LSF ϕ, we use the piecewise constant Mumford-Shah model as in [4]:

$$\begin{aligned} E_{\text{data}}(\phi) &= E_{\text{MS}}(\phi)\,, \\ E_{\text{MS}}(\phi) &= \iint_{\Omega} (I-\mu_+)^2 H(\phi) dxdy + \iint_{\Omega} (I-\mu_-)^2 (1-H(\phi)) dxdy \\ &\quad + \nu \iint_{\Omega} |\nabla H(\phi)| dxdy, \end{aligned} \tag{13}$$

where H is the Heaviside function and μ_+, μ_- are the mean intensities corresponding to the positive, respectively negative regions of the level set function ϕ. The *prior term* of the energy is given by:

$$E_{\text{prior}}(\phi, \mathcal{L}) = -\sum_{i=1}^{M} \log\left(\delta_t(\phi, i)\right) L_i^2 + \beta \left(1 - \sum_{i=1}^{M} L_i^2\right)^2, \tag{14}$$

where $\delta_t(\phi, i) = w_t(i)\, P_i(\phi)$.

For the prior class models $P_i(\phi)$, we have investigated both a Gaussian likelihood model and a PCA-based likelihood model. The two models and the corresponding results that we have obtained are detailed in the following.

Implementation using Gaussian likelihood model For the first implementation, we use a local Gaussian model of the level set function as a probability model for each class i:

$$p_i^{(x,y)}(\phi) = \frac{1}{\sqrt{2\pi}\sigma_i((x,y))} e^{-\dfrac{(\phi(x,y)-\phi_i(x,y))^2}{2\sigma_i^2(x,y)}}, \tag{15}$$

where $(x,y) \in \Omega$ is an image location, ϕ_i is the average level set function of class i and the variance $\sigma_i(x,y)$ models the local variability of the level set at location (x,y). Assuming that densities are independent across pixels, the likelihood function offered by class i for a level set function ϕ is given by the product of these densities over the image domain:

$$P_i(\phi) = \prod_{(x,y)\in\Omega} p_i^{(x,y)}(\phi). \tag{16}$$

Substituting likelihoods $P_i(\phi)$ in the prior energy (14) and augmenting by similarity transformations $h_{\boldsymbol{\tau}^i}$ (including translation, rotation, and scale) that align each prior i with contour ϕ, the prior energy becomes:

$$\begin{aligned} E_{\text{prior}}(\phi, \mathcal{L}, \boldsymbol{\tau}^{i=1..M}) = \sum_{i=1}^{M} \Bigg(&- \log w_t(i) + \iint_{\Omega} \Bigg(\log \sigma_i(h_{\boldsymbol{\tau}^i}(x,y)) \\ &+ \frac{(\phi(x,y) - \phi_i(h_{\boldsymbol{\tau}^i}(x,y)))^2}{2\sigma_i^2(h_{\boldsymbol{\tau}^i}(x,y))} \Bigg) dxdy \Bigg) L_i^2 + \beta \left(1 - \sum_{i=1}^{M} L_i^2\right)^2. \end{aligned} \tag{17}$$

Here $\boldsymbol{\tau} = \{s, \theta, T_x, T_y\}$ are the parameters of a similarity transformation

$$h_{\boldsymbol{\tau}}\left([x \;\; y]^T\right) = s \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix}. \tag{18}$$

Each transformation $h_{\boldsymbol{\tau}^i}$ aligns prior i with contour ϕ by scaling the former by s^i, rotating it by θ^i, and translating it by T_x^i, T_y^i.

Implementation using PCA-based likelihood model In the second implementation of our framework, the probability model for each class is based on a shape distance function between the segmenting contour and the prior contour corresponding to that class [2]. The prior contour for each class is represented using principal components analysis (PCA) and it evolves dynamically during the segmentation of each image so as to best match new image information. We improve the distance function proposed in [2] by making it symmetric, so that the resulting probability 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 $\{\boldsymbol{\phi}_1, \ldots \boldsymbol{\phi}_n\}$, which have been discretized on a rectangular grid, its principal directions of variation are captured by the eigenvectors $\{\mathbf{e}_1, \ldots \mathbf{e}_n\}$ 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 can then be obtained in the reduced space of the $p < n$ eigenvectors $\{\mathbf{e}_1, \ldots \mathbf{e}_p\}$, 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:

$$\hat{\boldsymbol{\phi}} = \overline{\boldsymbol{\phi}} + \mathbf{W}\,\mathbf{c}, \tag{19}$$

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

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 the prior contour is given by:

$$d(\phi, \mathbf{c}, \boldsymbol{\tau}) = \iint_{\Omega} \left(\hat{\phi}^2 |\nabla\phi|\delta(\phi) + \phi^2 |\nabla\hat{\phi}|\delta(\hat{\phi})\right) dx\, dy. \tag{20}$$

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 (19), and on the parameters $\boldsymbol{\tau} = \{s, \theta, T_x, T_y\}$ of the similarity transformation 18 which 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 (20) 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 (20), 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.

Based on the shape distance function (20), we define the likelihood of the segmenting contour represented by ϕ, for time t (image $I(t)$) and class i as:

$$P_i(\phi(t)) = e^{-d(\phi(t),\mathbf{c}^i(t),\boldsymbol{\tau}^i(t))}, \tag{21}$$

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 $\phi(t)$. Both are obtained by dynamic evolution of the prior contour $\hat{\phi}_i$ in image $I(t)$, according to the piecewise constant Mumford-Shah model. Thus, the data term in energy (7) becomes:

$$\begin{aligned} E_{\text{data}}(\phi, \mathbf{c}^{i=1..M}, \boldsymbol{\tau}^{i=1..M}) = \; & E_{\text{MS}}(\phi) + \sum_{i=1}^{M} E_{\text{MS}}(\hat{\phi}_i) \\ = \; & \iint_\Omega (I-\mu_{\phi+})^2 H(\phi) + (I-\mu_{\phi-})^2 H(-\phi)\, dx\, dy \\ & + \sum_{i=1}^{M} \iint_\Omega (I-\mu_{\hat{\phi}_i+})^2 H(\hat{\phi}_i) + (I-\mu_{\hat{\phi}_i-})^2 (H(-\hat{\phi}_i))\, dx\, dy \\ & + \nu \iint_\Omega |\nabla H(\phi)|\, dx\, dy. \end{aligned} \tag{22}$$

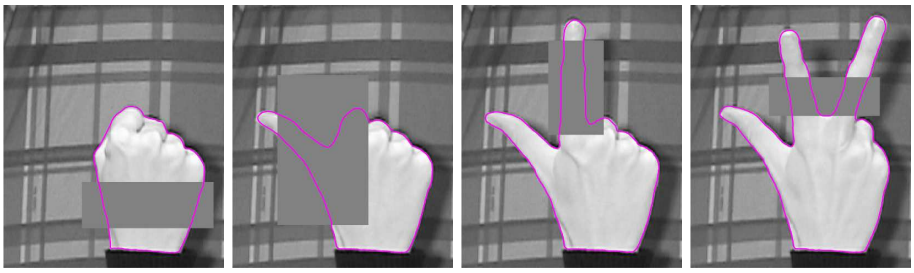
Here $\hat{\phi}_i$ is a function of $\mathbf{c}^i$ according to (19), evaluated at $h_{\boldsymbol{\tau}^i}(x,y)$, H is the Heaviside function and $\mu_{\phi+}$, $\mu_{\hat{\phi}_i+}$ and $\mu_{\phi-}$, $\mu_{\hat{\phi}_i-}$ are the mean intensities corresponding to the positive, respectively negative regions of the LSFs ϕ and $\hat{\phi}_i$. The last term imposes contour smoothness and is only needed for the main segmenting contour ϕ, as the PCA prior contours $\hat{\phi}_i$ are naturally smooth.

The prior term of the energy, obtained by substituting likelihoods $P_i(\phi)$ in (14) with (21), is:

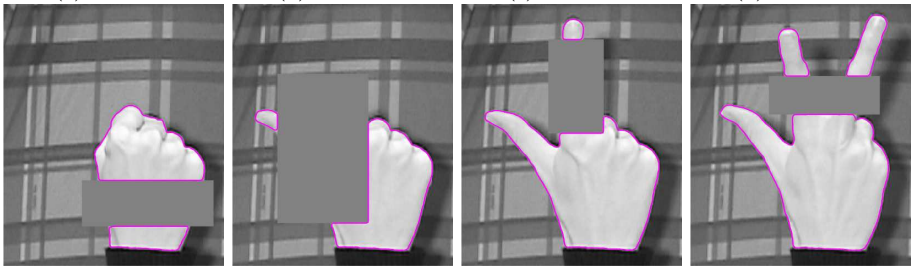
$$\begin{aligned} E_{\text{prior}}(\phi, \mathcal{L}, \mathbf{c}^{i=1..M}, \boldsymbol{\tau}^{i=1..M}) = & \sum_{i=1}^{M} (-\log w_t(i) + d(\phi(t), \mathbf{c}^i(t), \boldsymbol{\tau}^i(t))\, L_i^2) \\ & + \beta \left(1 - \sum_{i=1}^{M} L_i^2\right)^2. \end{aligned} \tag{23}$$

3.3 Training the model

In the training phase, we estimate the parameters of the HMM (see, e.g., [21]) using a labeled sequence of level sets corresponding to a manual segmentation of the mentioned gesture sequence (0,1,2,3,2,1,0). For the Gaussian likelihoods, we use the method described in [20] to obtain smooth estimates of the mean ϕ_i and



(a) Frame 2 (b) Frame 26 (c) Frame 51 (d) Frame 80



(e) Frame 2 (f) Frame 26 (g) Frame 51 (h) Frame 80

Figure 3: (a)–(d) Segmentation with the proposed framework of an image sequence in the presence of occlusion and background complexity. (e)–(h) Conventional segmentation of the same image sequence.

variance σ_i for each gesture class i. Parameter estimation for the PCA-based likelihoods amounts to PCA of the training level sets, yielding the corresponding mean level set $\overline{\phi}_i$ and eigenvectors $\mathbf{W}_i$ for each class i. For this application, we chose to use 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.

3.4 Results

In the testing phase, we ran classification and segmentation jointly on new image sequences of a hand performing the same succession of gestures in front of a complex background, this time degraded by occlusions. Our framework brings considerable improvements to the segmentation/classification task, even in the case of employing the unsophisticated Gaussian likelihood model. By virtue of the prior information supplied by the classification, segmentation is able to cope with severe occlusions, as can be seen in Fig. 3(a)–(d). Figure 3(e)–(h) shows that the results obtained on the same sequence with conventional segmentation are clearly inferior, since the desired shape of the object cannot be recovered because of the occlusions.

Figure 4 shows the classification results, which correctly follow the test gesture sequence and our understanding of the sequence in terms of the executed

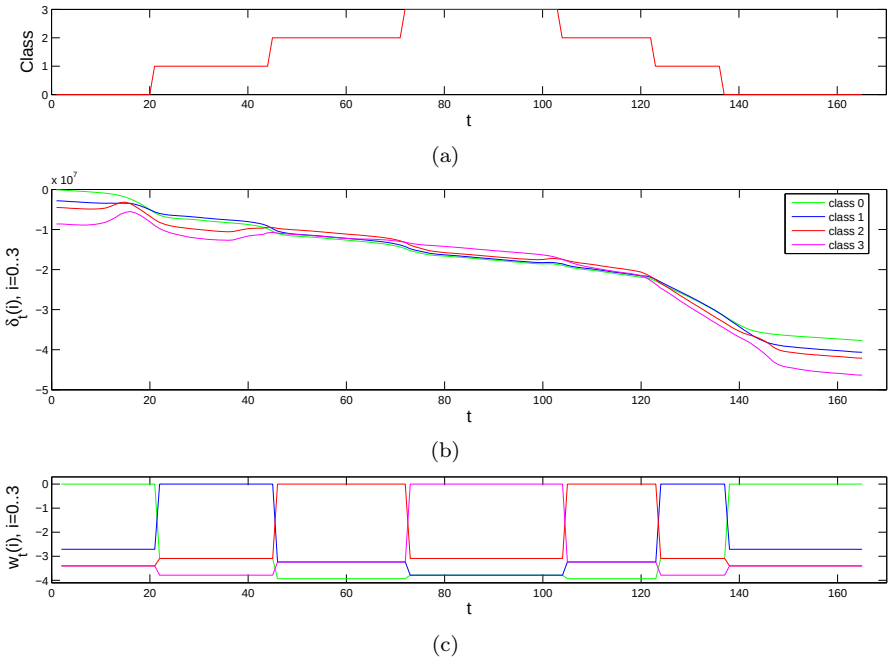
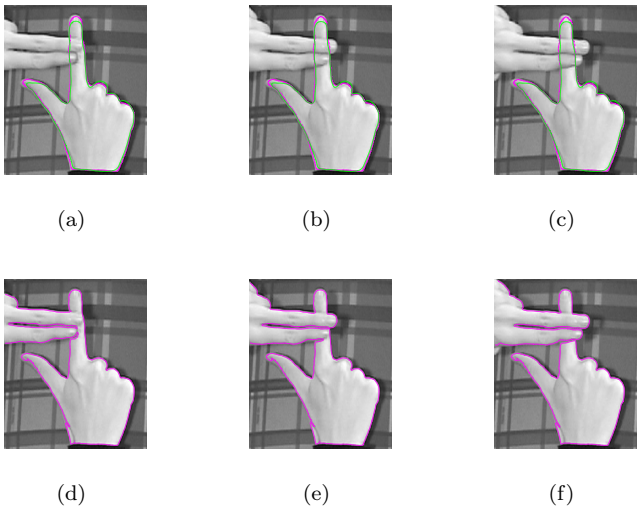


Figure 4: Classification results plotted per frame. (a) Final classification. (b) Delta functions of each class. (c) Prior confidence of each class used as input to the segmentation.

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. 4, 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 maximum value for every frame.

A limitation of the Gaussian likelihood model is the fact that the mean and variance of the prior corresponding to each class are fixed throughout the image sequence, thus the model doesn't adapt to varying shapes of the same class. This makes it difficult to obtain accurate segmentations for images where the winning class prior doesn't offer a close match to the image, even after the similarity transformation. We obtain an improvement with respect to this limitation by using the PCA-based likelihood models, because the prior contours adapt dynamically to the content of new images. We can thus perform the segmentation/classification of sequences with naturally occurring occlusions, such as the occlusion of the gesturing hand by the other hand, as can be seen in Fig. 5(a). By contrast, Fig. 5(b) shows that conventional segmentation fails.



(a) (b) (c)

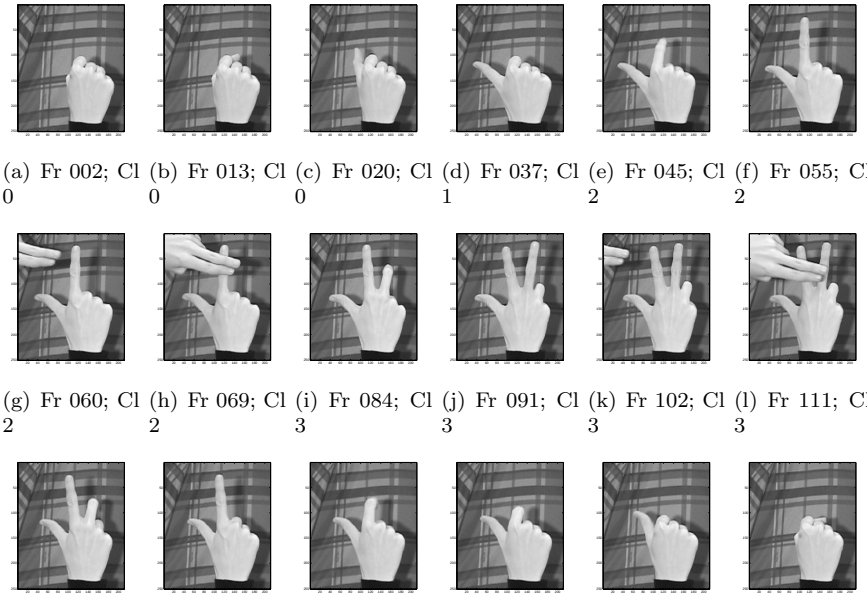
(d) (e) (f)

Figure 5: (Upper row) 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. (Lower row) Conventional segmentation of the image is confused by the occluding left hand.

Figure 6 shows classification results of our framework, using the PCA-based likelihood model, for a sequence containing natural occlusions of the gesturing hand by the left hand. The detected classes correspond to our perception of the gesture sequence. Figure 7 shows the classification results for all the frames of the sequence illustrated in Fig. 6, which reflect with accuracy our understanding of the executed gestures. Similarly with the results of the Gaussian model implementation, 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, as shown in Fig. 7.

4 Conclusion

We have introduced and developed a novel variational framework that enables the segmentation of image sequences simultaneously with the 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 allowed 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



(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 6: 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.

framework via two particular implementations for a hand gesture analysis application, employing Gaussian and PCA-based likelihood models, respectively. Our framework has allowed us to successfully segment and classify image sequences of a gesturing hand before a complex background, in the presence of occlusions.

References

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

[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.

[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.

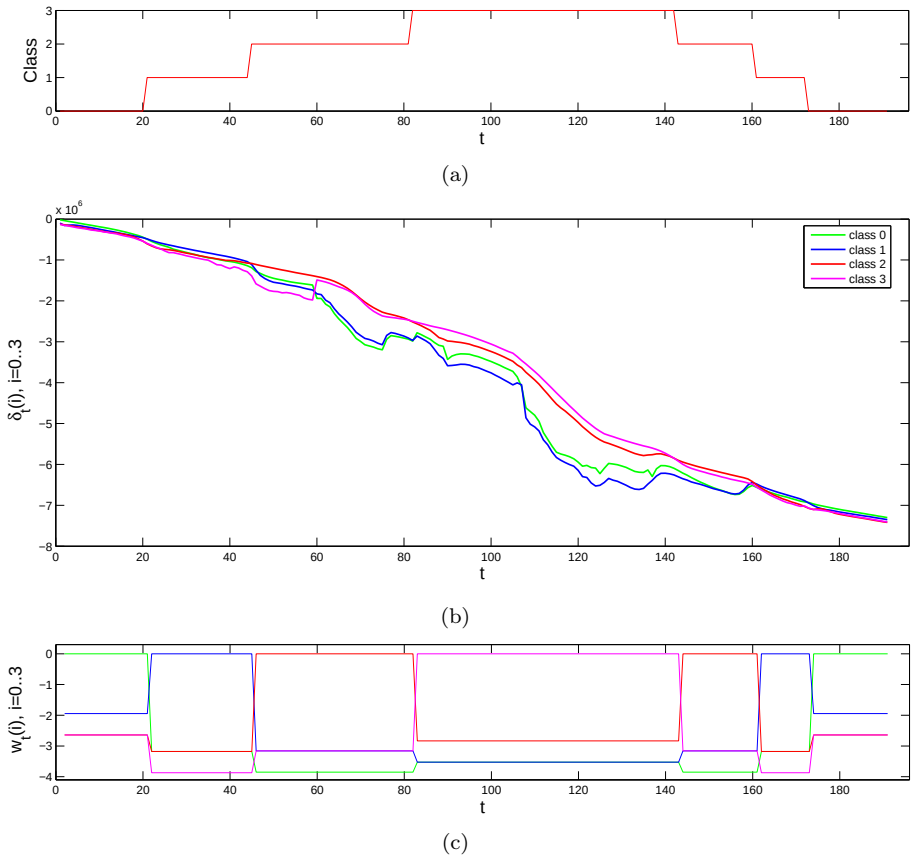


Figure 7: Classification results plotted per frame. (a) Final classification. (b) Delta functions of each class. (c) Prior confidence of each class used as input to the segmentation.

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

[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.

[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.

[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.

[8] D. Cremers, T. Kohlberger, and C. Schnör. Shape statistics in kernel space for variational image segmentation. *Pattern Recognition*, 36:1929–1943, 2003.

[9] 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.

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

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

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

[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.

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

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

[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.

[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.

[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.

[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.

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

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

[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.

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

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

[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.