Escaping local minima through hierarchical model selection: Automatic object discovery, segmentation, and tracking in video

Nebojsa Jojic, John Winn, and Larry Zitnick
Microsoft Research (web site: www.research.microsoft.com/∼jojic/mm.html)

Abstract

Recently, the generative modeling approach to video segmentation has been gaining popularity in the computer vision community. For example, the flexible sprites framework has been studied in, among other references, [11,13,14,24]. In general, detailed generative models are vulnerable to intractability of inference and local minima problems when approximations are made (see, e.g., [25]). Recent approaches to dealing with these problems focused on inference techniques for increasingly more expressive models. Simpler models, on the other hand, while less precise, are often not just faster, but less prone to local minima. In addition, while many different models may be based on similar hidden variables, some models may be more amenable to inference of some of the shared variables, while other models lead to efficient and accurate inference of other components of the hierarchical data description. In this paper, we empirically illustrate that forcing multiple models to share the posterior distribution leads to inference less prone to local minima. We define a set of key hidden variables that describe aspects of the data that we care about. The relationships among these key variables are defined through multiple conditional distribution models on the same pairs of variables, controlled by switch variables. The posterior distribution over the key hidden variables is shared, and inference of the switch variables serves as a mechanism for combinatorial model selection. The key observation here is that while the most expressive model often ends up a winner by the end of the iterative learning of model parameters, early iterations are dominated by simpler model components, and upon convergence, the free energy is lower than the ones reached by switching on all the most complex components from the beginning of the learning. We illustrate the performance of this approach on the unsupervised video segmentation task.

1. Introduction

Generative models, due to their hierarchical nature, and the ease of incorporating known structure of the data in their design, have provided popular tools for analysis of natural signals, e.g., images, video, audio, gene expression data, sequence data, etc. Through probabilistic inference, the generative modeling paradigm provides a way of teasing apart a variety of hidden causes of data variability. However, inference in most interesting models of natural phenomena is intractable, and so recent research in the area has focused on a variety of approximate inference and learning techniques, such as structured variational methods [5, 6], (loopy) belief propagation, sampling, etc. We argue in this paper that instead of focusing on inference in the most expressive model of the data, it is often possible to build a (redundant) structured model that is amenable to simple inference techniques. Such a model can be constructed out of the sets of *alternative* conditionals for each of the hidden and observed variables. Some of the alternative conditionals use simplified parameterizations and/or dependence structures in order to regularize learning, replace more complex but failing explanations, and, most importantly, avoid local minima to which approximate inference is particularly prone. On the other hand, the more expressive forms of the conditionals provide detailed explanations for the variability in the data. We show empirically that during variational inference in such models, for most random initializations, simple components are automatically chosen in early iterations, while the more expressive components prevail in later iterations, once the model parameters have sufficiently evolved away from inferior local minima.

Combinations of models or experts have been proposed before. For example, [1–3] propose mixing simple experts, usually in a supervised framework. Defining the probability distribution over data as a normalized product of individual distributions was proposed in [4]. In these approaches, the experts are independent and do not share the results of their internal inference. Variational Bayesian model selection, e.g., [7], provides a way to share the posterior distribution over models, but the models typically have identical parameterizations with only varying complexity. Furthermore, this paradigm allows for selection of a single model over all variables, while we want to be able to select for parts of the model, leading to combinatorial model selec-

tion. We note that the earlier work on structure learning, e.g., the structural EM in [8], has this property not in the set up, but in a way the search for the structure is conducted. However, the goal in structure learning is to properly evaluate the posterior distribution over the models, while we are interested in optimizing the likelihood of the data. Optimizing the likelihood of the data typically leads to selecting the most complex structure among the ones provided, but in our recipe, that structure is usually known a priori to be the best, and finding it is *not* the main goal. Instead, our goal is to use other models to steer the learning away from local minima of the likelihood *under the best model*. In addition, a hierarchical combination of models can be more expressive as was the case in other switching models, e.g., [10], which switch between the models to adapt better to various data samples, rather than to discover a single model for the entire dataset.

To illustrate how the hierarchical model selection *during* iterations of general EM helps avoid local minima, we focus on one of the unsolved computer vision problems, the unsupervised segmentation of video into moving and possibly occluding objects. In [11], it is argued argued that the solution to this problem would have to involve joint modeling of several most important causes of variability, such as appearance variability, object motion and segmentation variables, and a new framework for video segmentation, dubbed "flexible sprites", was derived starting from the deterministic setup of [12]. The flexible sprites are meant to be used in a fully unsupervised manner, without separating the training and testing procedures. The input to the algorithm is simply a video sequence, and the output are the appearance models of individual objects, and their particular positions and segmentation in various frames. However, each of the hidden variables in the model can be modeled in several ways. For example, appearance models of individual objects can be based on per-pixel mean and variance maps as in [11], or they could be made more expressive as in [15] or significantly simplified as in [20]. Among these techniques, the more complex ones can capture object structure in more detail, but the simplest ones can deal with large deviations from the basic object structure and lead to faster convergence. Adapting the model of the appearance from frame to frame has also been attempted, e.g., [9], but the robustness of online techniques is usually highly dependent on the amount of frame-to-frame change in motion and appearance, and precludes re-registering to the objects when they reappear in the scene. Sprite mask models and motion models reported in literature are equally diverse [11, 13–16, 18, 21, 24]

In the next section we setup a general framework for using various representations of model parts in concert during learning, switching between them in a way that optimizes the free energy of the hierarchical model. Then, in Section 3, we illustrate the framework on the task of video segmentation, using two different representations (conditional distribution forms) for appearance and two different representations for object mask. Finally, in Section 4 we describe various alternative components that we have included in a large hierarchical switching model capable of capturing object appearance and deformation changes that are considerably harder than the ones tackled by graphical models in the past.

2. Hierarchical switching of conditionals

Consider a model with three multidimensional hidden variables h_1, h_2, h_3 and one multidimensional observed variable x, Fig. 1(A). Suppose that we believe that the model

$$p_1(x, h_1, h_2, h_3) = p_1(h_1)p_1(h_2|h_1)p_1(h_3|h_2)p_1(x|h_2, h_3),$$

and particular parameterized forms of the conditionals $p_1(h_1; \theta_{1,1})$, $p_1(h_2|h_1, \theta_{2,1})$, $p_1(h_3|h_2, \theta_{3,1})$, $p_1(x|h_2, h_3, \theta_{x,1})$ are considered to capture well all or at least most of the data. The subscript 1 indicates that this is the first variant of the model we consider, as we will shortly introduce others. To estimate the parameters θ of the model from the data $x^1, x^2, ..., x^t, ...$, we can use the EM algorithm, or if that is intractable, we can use variational learning, or some other approximate technique. In any case, the exact maximization of the likelihood of the data is typically not guaranteed to be achieved within the computational resources allocated to the task, and some algorithms, esp. a single run of the EM or variational learning only guarantee local optimality regardless of the amount of computation performed. Also, the chances of getting stuck in a local minimum of the free energy is empirically known to increase with the complexity of the model.

However, a number of simpler models potentially less prone to local minima problems, but also less precise, may be easily expressible. In fact, each of the hidden variables can be generated in multiple ways using different conditional distributions. For instance, variable h_2 may only slightly depend on h_3, and so it can also be generated (although sometimes incorrectly) by another distribution, $p_2(h_2; \theta_{2,2})$. Alternatively, we can generate h_2 from a conditional distribution dependent on h_1 but of a simplified form $p_3(h_2|h_1, \theta_{2,3})$. Variable h_3 may be generated not from h_1, but from h_2 through $p_2(h_3|h_2, \theta_{3,2})$. Or, h_3 can be independently generated from $p_3(h_3; \theta_{3,3})$. Finally, we can describe generation of the data x in multiple ways, too, e.g., $p_2(x|h_2, \theta_{x,2})$, or $p_3(x|h_1, \theta_{x,3})$.

To select among the conditionals, we introduce switch variables s_1, s_2, s_3, s_x, so that, for example, conditional distribution over h_3 becomes $p(h_3|h_1, h_2, s_3) = p_1(h_3|h_1, \theta_{3,1})^{[s_3=1]} p_2(h_3|h_2, \theta_{3,2})^{[s_3=2]} p_3(h_2; \theta_{3,3})^{[s_3=3]}$, or simply $p(h_3|h_1, h_2, s_3) = p_{s_3}(h_3|h_1, h_2, \theta_{3,s_3})$, and the

joint distribution is

$$p(x, h_1, h_2, h_3, s_1, s_2, s_3) = p(s_1, s_2, s_3, s_x)p_{s_1}(h_1; \theta_{1,s_1})p_{s_2}(h_2|h_1, \theta_{2,s_2}) \times p_{s_3}(h_3|h_2, h_1, \theta_{3,s_3})p_{s_x}(x|h_1, h_2, h_3, \theta_{0,s_x}), \quad (2)$$

where the prior $p(s_1, s_2, s_3, s_x)$ puts zero mass to configurations where $s_1 \neq 1$, since in this example we have assumed only one form of the distribution on h_1. The nine possible configurations of the switch variables define nine possible models.

In general, a graphical model that switches among different forms of each conditional can be written as:

$$p(x = h_0, \{h_k, s_k\}) = p(\{s_k\}) \prod_k p_{s_k}(h_k|pa(h_k), \theta_{k,s_k}), \quad (3)$$

where we allow dual notation for the data $x = h_0$ when needed for simplification. For different values of $s_k \in \{1, ..., S_k\}$, the individual conditionals $p_{s_k}(h_k|pa(h_k), \theta_{k,s_k})$ may have different dependence structures (the dependence on different subsets of $pa(h)$), or completely different forms and parameterizations, even when the dependence structures are the same. Some of these models of h_k are simple, while others are complex and they are defined by parameters θ_{k,s_k}, which, in general, we aim to learn from the data.

Variational inference minimizes the free energy $F = \int q \log q - \int q \log p$, with respect to the posterior q, where the integral denotes both integration over continuous variables and summation over discrete variables. In addition, given a set of observations $\{x^t\}$, we can estimate the model parameters. For example, if the data points x^t are independently generated, then

$$F = \sum_t \int q^t \log q^t - \int q^t \log p(\{s_k^t\}) \prod_k p_{s_k^t}(h_k^t|pa(h_k^t), \theta_{k,s_k}), \quad (4)$$

and iterative minimization of F with respect to the q^t distributions and model parameters θ leads to approximate parameter learning. The q distribution over hidden variables $\{h_k\}_{k=1}^K, \{s_k\}_{k=0}^K$ can have any form as long as it is normalized, but choosing certain factorized forms leads to faster algorithms, and the ability to use switches to adapt the model structures differently to different data points is likely to be useful in many applications. We will focus on the case where

$$q^t(\{h_k^t\}_{k=1}^K, \{s_k^t\}_{k=0}^K) = q(\{h_k^t\}_{k=1}^K) q(\{s_k^t\}_{k=0}^K) \quad (5)$$

There is a combinatorial explosion of possible models achievable by independently selecting conditionals, and if the switch variables are assumed to be identical for all data points, then computing the posterior $q(\{s_k^t\}_{k=0}^K) = p(\{s_k^t\}_{k=0}^K|x^1, x^2, ...x^t)$ while learning the parameters would be equivalent to the structure learning task in [8]. Using different switch variable configurations for different data points would be a generalization of the previously proposed adaptation to different regimes, for example, [10]. The key property that we are after, however, is the sharing of the posterior among different models, which is considered an approximation in variational inference, e.g., [7], but actually helps avoid local minima of generalized EM, as shown in the next section. Note also that approximating the posterior distribution, e.g. as $q(\{s_k^t\}_{k=0}^K) = \prod_k q(s_k)$, leads to more efficient adaptation than in [7], regardless of whether the goal is structure posterior estimation or regime adaptation.

An interesting empirical observation, which we wish to emphasize in this paper, is the following. Suppose that a single model structure is by far the best, e.g. $s_k = 1$, for all k. This structure is either the true model, i.e. the data was indeed generated from it, or at least it is the best among the alternatives, so that variational optimization of (5) using the factorized posterior yields upon convergence in parameters θ the deterministic posterior[1] over models $q(\{s_k^t\}_{k=0}^K) = \prod_k [s_k = 1]$, where [] is an indicator function. Plugging this posterior back into (5), we get

$$F = T \times p(\{s_k\} = 1) + \sum_t \int q(\{h_k^t\}_{k=1}^K) \log q(\{h_k^t\}_{k=1}^K) - \int q(\{h_k^t\}_{k=0}^K) \log \prod_k p_1(h_k^t|pa(h_k^t), \theta_{k,1}), \quad (6)$$

where $p(\{s_k\} = 1)$ denotes the prior probability of the model where each $s_k = 1$, i.e., $p(s_0 = 1, s_1 = 1, ..., s_K = 1)$. Apart from the constant additive term $T \times p(\{s_k\} = 1)$[2], this is exactly the free energy of the single best model. However, direct iterative variational optimization of this free energy for the single best model may consistently result in higher free energy than indirect optimization of (5), even when optimizing the latter leads upon convergence to $q(\{s_k^t\}_{k=0}^K) = \prod_k [s_k = 1]$. In other words, the use of other models during the early iterations of variational optimization sets the algorithm in the direction towards better local minima. Optimization of (5), however, may not be significantly more expensive, as severe approximations of $q(\{h_k^t\}_{k=1}^K)$, with appropriate factorizations of $q(\{s_k^t\}_{k=0}^K)$, can lead to algorithms whose complexity increases only marginally.

3. A quantitative analysis of flexible sprite learning by conditional switching

In this section we give a practical example of the above claim. As indicated in the introduction, we focus on learning

[1] If we have an appropriate prior over the model structures $p(\{s_k\})$, then the posterior may not be able to reach a deterministic distribution, but will often get fairly close.

[2] T is the number of training samples here.

flexible sprites from video [11], but for simplicity limit ourselves to the case of just foreground and background layers. This simplification yields four (possibly multidimensional) hidden variables f, b, m, T for each frame x in a sequence. In this section we also assume that the frames are independently generated. The mask image $m = \{m_i\}$ is a set of binary variables m_i, one for each pixel i in the foreground image f. The variable m_i indicates if pixel i is opaque and thus contributes to the image x, or is (fully) transparent, in which case the appropriate pixel in the background image b makes a contribution. The foreground image f undergoes a transformation before contributing to the frame x, and for simplicity we focus (in this section only), on translations as the only allowable transformations. Then the conditional distribution over the data pixels can be written as

$$p_1(x|f,b,m,T) = \prod_i \mathcal{N}(x_i; f_{i-T}, \sigma^2)^{[m_{i-T}=1]} \mathcal{N}(x_i; b_i, \sigma^2)^{[m_{i-T}=0]} \quad (7)$$

The prior over all possible discrete shifts T is set to be uniform $p_1(T) = const$, and as indicated above the transformation only transforms the foreground and the mask, but leaves the background unaffected. The distribution over the static background image b is given by the same Gaussian sprite model as in [11], i.e., $p_1(b) = \prod_i \mathcal{N}(b_i; \mu_i^b, \phi_i^b)$. For the purposes of this section, the above defined models are the only versions of the appropriate conditionals. However, we allow two different ways of specifying the distribution over the foreground image f, and two different different ways of specifying the distribution over the mask m.

The first model for the foreground image f is again based on the set of position-specific Gaussians as in, for example, [11, 13, 14],

$$p_1(f) = \prod_i \mathcal{N}(f_i; \mu_i^f, \phi_i^f),$$

but the second one assumes position-independent mixture model,

$$p_2(f) = \prod_i \sum_{c=1}^{C} \lambda^c \mathcal{N}(f_i; \mu^{f,c}, \phi^{f,c}).$$

The latter is akin to histograms often used as simplified tools for sorting through images by using global color similarity and ignoring the actual structure of the object. This model can represent much more variability in object appearance at the cost of over-generalizing and thus providing little appearance separation from the background.

The first model for the mask is also position-specific, as in [11, 13, 14],

$$p_1(m) = \prod_i p_1(m_i),$$

where $p_1(m_i = 1)$ is the probability that the i-the pixel in the foreground is opaque (i.e., it occludes the background pixel it happens to be in front). The second model for the mask ties the masking probabilities for all pixels to the coordinates of the object center η, and the elliptical shape described by the projection matrix R

$$p_2(m) = \prod_i e^{-(\epsilon + \|R(i-\eta)\|^2)[m_i=1]} (1 - e^{-(\epsilon+\|R(i-\eta)\|^2)})^{[m_i=0]},$$

where ϵ is a fixed small constant that ensures that no pixel in the foreground will be fixed as opaque. R is the orthogonal, but not orthonormal, matrix that controls the orientation and an aspect of an ellipse centered at η, and the amplitude of its rows determine how quickly the probability of opaqueness drops from the object's center η along the ellipse's axes. It is clear how this model can serve to regularize the learning – most objects have a localized spatial content which can roughly be approximated with a probabilistic elliptical shape. While this model has little chance of fully explaining complex objects, it is less likely to lock onto local minima in which several disconnected parts of a scene are associated with the foreground.

Finally, hidden switch variables $s_f \in \{1, 2\}$ and $s_m \in \{1, 2\}$, choose the particular forms of the alternative conditionals for each frame. This leads to a model of the form (3), which we trained on the data illustrated in Fig. 1, approximating the true posterior conditioned on *all* frames $\{x^t\}$ by a fully factored form $q^t = q(s_f^t)q(s_m^t)q(T^t)\prod_i q(b_i^t)q(f_i^t)q(m_i^t)$, with Gaussian distributions for $q(f_i)$ and $q(b_i)$, and discrete distributions for $q(m_i)$ and $q(T)$. We minimized the free energy (5) for several sets of constraints on $q(s_f^t)$, $q(s_m^t)$, always using the same update schedule for the unconstrained parts of q. These sets of the constraints are as follows:

- Ellipse/histogram – $q(s_m^t) = [s_m^t = 2]$, $q(s_f^t) = [s_f^t = 2]$, for all t
- Mask/histogram – $q(s_m^t) = [s_m^t = 1]$, $q(s_f^t) = [s_f^t = 2]$, for all t
- Ellipse/Sprite – $q(s_m^t) = [s_m^t = 2]$, $q(s_f^t) = [s_f^t = 1]$, for all t
- Mask/Sprite – $q(s_f^t) = [s_f^t = 1]$, $q(s_m^t) = [s_f^t = 1]$ for all t
- All equal – $q(s_f^t) = 0.5$, $q(s_m^t) = 0.5$ for all t
- Adaptive – no constraints on discrete distributions $q(s_f^t)$, $q(s_m^t)$

We generated 15 random uninformative initalizations for model parameters and q distributions. The means of the posterior distributions, and the means of the background and foreground prior are set to the mean intensity of the dataset with added random noise of small intensity, and the variances are set to unity. Color histogram (mixture) model is set to the histogram of the whole dataset. Mask ellipse

center η is chosen randomly in the image, and the shape matrix R is set to be wide, and so on. In the case of the adaptive model, which is free to vary the structure during learning so as to best optimize (5) in each step, we initialize the structure posterior to uniform $q(s_m^t) = q(s_f^t) = 0.5$. For each of generated random initializations, and for each of the above sets of constraints on the structure posterior, we run the same variational learning algorithm which iterates updates on model parameters and q functions in the same order (except that in the non-adaptive cases, $q(s_m^t)$ and $q(s_f^t)$ are not updated).

The results are summarized in Fig. 1(B), where we show the average over 15 runs of the log likelihood bound (negative free energy $-F$) after each iteration and for each of the model structures above. One iteration consists of one pass through updating q distributions, and then updating all model parameters once, corresponding to one variational E step and one M step of a generalized EM algorithm. It should come as no surprise that for this data the mask/sprite model outperforms other fixed model combinations. However, the adaptive model which allows updating of the model structure in each iteration, as described in the previous section, reaches an even higher average bound (and perceptually always near-perfect segmentation of the foreground object). This is true even though in all fifteen runs, the estimated $q(s_m^t)$ and $q(s_f^t)$ distributions end up being the same as for the mask/sprite model (choosing the first variant of each conditional in *each* frame). This is due to the help the adaptive model has from the simplified models, more suitable to inference with only slightly evolved parameters in the early iterations of learning.

During the first iteration of variational learning, the adaptive model puts most of the probability mass into the elliptical shape model, i.e. $q(s_m^t) = 2$, as well as into the histogram model of appearance $q(s_f^t) = 2$, for many frames in the sequence. After the second or third iteration, the elliptical shape model is typically turned off, and almost all the mass goes to the fully expressive mask model. The fully expressive sprite model replaces the histogram for most data points by the fifth iteration, and by the end of learning no single frame is explained by the simpler conditionals. Given the data that fits, and the overwhelmingly stronger modeling power of the mask/sprite model, without any punishment for model complexity, it is not surprising that this model prevails at the end of joint structure and parameter optimization. However, it is an important observation that an opportunity to use simpler descriptions of the data in early iterations of learning, while parameters are still imperfect, leads to escaping local minima more efficiently than random restarts.

Intelligent initializations for complex models, annealing or other types of direct intervenion on the outputs of each EM iteration, have all been used to nudge EM out of local minima. We believe that the use of switching conditionals is a cleaner, more automatic, and powerful way of escaping local minima, as this framework guarantees convergence and local optimality of the result, since each iteration reduces the free energy. Furthermore, the switching models interact in a much more sophisticated and adaptive way during variational learning than is true in most heuristic implementations. Finally, the modularity of variational algorithms allows easy combination of existing variational inference engines by passing q distributions and expectations under q distributions between modules [17].

To further illustrate this qualitatively, in addition to the above small scale quantitative study, we combine in the next section several ways of representing a video sequence and run the variational hierarchical switching model learning on it, achieving a reasonable unsupervised extraction of a video object undergoing considerably harder shape and appearance changes against a much more confusing background clutter than the previous cited graphical models could cope with.

4. Flexible sprite learning using hierarchical switching among many conditionals

In this section, we describe a hierarchical generative model which uses the same basic hidden variables as the flexible sprite model [11] limited to two moving sprites (foreground/backround), but describes relationships among them through many different conditional distributions, automatically selected so as to maximally minimize the free energy in each step of learning. These different modeling strategies lead to automatic mining of various image cues, such as consistency in motion, the extent of color variability within a sprite in a single frame, global color consistency of the sprite, shape consistency across frames, and shape contiguity within a frame.

As before, the basic variables include foreground sprite f, background sprite b, discrete mask m, and transformation T. The transformation model is enriched to include scale in addition to shift, so instead of notation $i - T$, we will use $T(i)$ to denote the change in coordinate due to transformation. The global appearance histogram model in the previous section is replaced by a local (per-frame appearance model),

$$p_2(f^t) = \prod_i \sum_{c=1}^{C} \lambda^{c,t} \mathcal{N}(f_i^t; \mu^{f,c,t}, \phi^{f,c,t}), \qquad (11)$$

to allow palette change as in the PIM model of [19]. Furthermore, the background model b also has the same variants as the foreground model, and has its own transformation variable T_b. In addition to per-pixel mask prior and the elliptical model of the previous section, the mask distribution for each frame is also expressed by

$$p_3(m^{t+1}|m^t, T^t, T^{t+1}) = \prod_i \epsilon^{\left[m^{t+1} \neq m^t_{T^{t+1}(T^t)^{-1}(i)}\right]} (1-\epsilon)^{\left[m^{t+1} = m^t_{T^{t+1}(T^t)^{-1}(i)}\right]},$$

which creates an expectation that the inferred masks in neighboring frames should be similar (T^{-1} denotes the inverse transformation).

We also use an alternative MRF mask prior $p_4(m|f)$ which favors short segmentation boundaries aligned with intensity gradients in the image. In addition to extending the number of conditionals, we also introduce new hidden variables for each frame, which we adopt from the over-segmentation model of [23]. This model was found to be amenable to variational inference of dense optical flow, but insufficient for grouping the segments into coherent objects. The added variables include the pixel displacement (flow) field $\{d_i\}$, the image *oversegmentation* $\{g_i\}$, $g_i \in \{1, ..., K\}$ into a large number of segments that have coherent motion and color within a frame, and the matching variables $\{h_k\}_{k=1}^K$, where h_k is the index of the segment in the next frame that corresponds to the segment k. The oversegmentation is treated as an alternative model p_2 to generation of frame pixels in (7), where each pixel observation x_i is treated here as a combination of color c_i and position $r_i = i$, and generation of a pixel is assumed to follow

$$p_2(x_i = \{c_i, r_i\}|g_i) = \mathcal{N}(c_i; \xi_{g_i}, \Sigma_{g_i})\mathcal{N}(r_i; \zeta_{g_i}, \Delta_{g_i}),$$

where g_i denotes the segment to which pixel i belongs, and the ξ, and ζ are the color and spatial means of the segments, similar to [21] which used many fewer segments (in our experiments we use several hundred segments per image). Then the alternative conditional to (7) is

$$p_2(x|g) = \prod_i p_2(x_i = \{c_i, r_i\}|g_i). \quad (12)$$

The segmentation g_i is similar to the foreground-background segmentation m_i, except that m_i is binary, and the number of possible segments K is large. We would expect that grouping of segments leads to inference of mask m and object extraction, and so the oversegmentation g provides the fifth way of generating the mask. Each segment k has associated probability of its pixels being opaque $p_5(m_i|i \text{ belongs to } k)$ and these can be used as an alternative way of generating masks:

$$p_5(m|g, T) = \prod_i p_5(m_i|g_{T^{-1}(i)}) \quad (13)$$

In addition, by looking at the segment mapping one or more frames into the future or into the past can provide additional conditional distributions like the ones above. (In our experiment, we use the entire segment track).

The segments' color means are generated by either a broad Gaussian distribution, or a Gaussian process $p(\xi_{h_k}^{t+1}|\xi_k^t)$ based on mapping variables h_k.

One of the conditionals on segmentation is simply flat $p_1(g_i) = const$, while the other copies the segmentation from the previous frame using the displacement field d_i

$$p_2(g_i^{t+1}|d_i, g^t) = \rho^{\left[g_i^{t+1} = h_{g_{i-d_i^t}^t}\right]}(1-\rho)^{\left[g_i^{t+1} \neq h_{g_{i-d_i^t}^t}\right]}$$

Parameter ρ, like the parameter ϵ in the mask case, controls the strength of the influence of the previous frame's segmentation on the current segmentation.

The displacement field is defined by several conditional distributions in the hierarchical model. The first of them is the analogue of equation (7),

$$p_1(d|d^f, d^b, m, T) = \prod_i \mathcal{N}(d_i; d^f_{T(i)}, \sigma^2)^{[m_{T(i)}=1]}\mathcal{N}(d_i; d^b_i, \sigma^2)^{[m_{T(i)}=0]} \quad (15)$$

which introduces hidden foreground and background displacement fields d^f, d^b, generated by the same set of multiple conditional forms as their appearance equivalents, just replacing 3-D color with 2-D flow vectors.

In addition, the displacement field variables are connected into an MRF $p_2(d)$, and also into a model $p_3(d|g, h)$ which enforces that each pixel's flow is equal to one of the segment displacements [23],

$$p_3(d|g, h) = \prod_i \left(1 - \prod_{k \in \varepsilon_i} \left(1 - [d_i = \zeta_k - \zeta_{h_k}]\right)\right), \quad (16)$$

where ε_i denotes the segment assignments of the neighboring pixels, and includes g_i.

As before, switch variables, one per the above described hidden variables with alternative conditionals, complete the hierarchical model. Because of the presence of normalized MRF-s as conditionals, some parameters of the model cannot be estimated in the M step of generalized EM so as to guarantee reduction in free energy. These parameters, namely the ones controlling the MRFs, can apparently be varied without significant impact on the learning of the other more important parameters, so we set the MRF parameters by hand and do not update them. Using a fully factorized variational posterior, we iterated free energy minimization until convergence on the video whose few frames are shown in Fig. 2. This resulted in a fully unsupervised segmentation of the person in video, without any manual input into the algorithm, as illustrated in Fig. 2 and in the videos available in the supplemental material. The video ***finalcolor.avi*** illustrates both the input video and the final inferred segmentation mask m: the color in the background regions (where $m = 0$) is suppressed, while the foreground layer (where $m = 1$) has its original color. The video ***segOutSmall.avi***, illustrates some intermediate variables. Boundaries in the fine segmentation map described by the variables g are shown by white lines, and each segment is colored in the mean segment color ξ. The segmentation tends to be consistent for at least several frames due to local constraints in the model (equations 9-11), but not consistent enough to lead directly to foreground-background segmentation on its own.[3]

[3]The match variables h describe motion tracks consisting of matched segments. About 30% of these tracks are longer than 20 frames.

As can be seen in the figure and the videos, the foreground object exhibits significant variability in illumination, pose (both rigid and articulated), position, scale, and nonuniform deformation. The sunny conditions caused the shirt of the subject to change color from saturated white to dark blue in the video, and the overall color distribution on the body is complex, including skin and hair hues. The background scene is even more complex and includes several objects in motion, and equally complex changes in illumination. Finally, the camera is in motion, making both layers non-static. In fact, in this case, even the most highly parameterized combination of the conditionals described above may not be capable of discerning the foreground from the background in all frames. However, the hiearachical model switching allows for automatic individual model component adaptations to different frames, while polling of simpler models helps avoid local minima in early iterations of the variational EM (iterative reduction of free energy with respect to posterior and model parameters). The algorithm automatically uses multiple descriptions of the scene which differently express general video characteristics, such as smoothness of motion, slow change of color and loose consistency of object appearance across frames, or slow evolution of object shape.

The resulting segmentation goes beyond what graphical models reported in literature can accomplish in unsupervised settings. (Note that the object was not segmented out in the first frame as in some tracking setups, nor was any model parameter initialized in an informative way). In this case, the hierarchical switching not only helped in avoiding local minima, but also provided additional expressive power, as some of the frames in the video are much better expressed by the added alternatives to the basic flexible sprite description.

5. Conclusions

We provide an approach for inference in detailed graphical models, which does not improve on techniques for approximating posterior distributions, but rather proposes inclusion of switchable additional conditionals in various parts of the basic model. The ability to switch some of these models on during early iterations of variational learning then helps to avoid local minima even when simple variational approximations are used. Furthermore, as in some previous specialized cases of switching models, some alternative conditionals may be necessary to improve the overall modeling power. We find that this approach dramatically improves the performance on a very hard unsupervised video segmentation task, while essentially using similar individual conditional models to the one previously studied in isolation in the literature. It may be possible to extend the switching model of the last section not only to improve segmentation, but also so that it can be used for recognition of recurring objects in long sequences or multiple sequences.

References

[1] S. Nowlan and G. Hinton, “Evaluation of adaptive mixtures of competing experts,” NIPS 3: 774-780, 1990.

[2] M. Jordan and R. Jacombs, “Adaptive mixtures of local experts ,” Neural Computation 3: 79-87, 1991.

[3] R. Jacobs and M. Jordan and S. Nowlan and G. Hinton, “Hierarchical mixtures of experts and the EM algorithm,”A.I. Memo No. 1440, MIT, 1993.

[4] G. Hinton, “Product of experts,”ICANN 99.

[5] R. Neal and G. Hinton, “A view of the EM algorithm that justifies incremental, sparse and other variants,” in *Learning in Graphical Models*, M. I. Jordan, Ed. Kluwer Academic Publishers, Norwell MA., 1998.

[6] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” in *Learning in Graphical Models*, M. I. Jordan, Ed. Kluwer Academic Publishers, Norwell MA., 1998.

[7] N. Ueda and Z. Ghahramani, “Bayesian model search for mixture models based on optimizing variational bounds,” Neural Networks 15: 1223-1241, 2002.

[8] N. Friedman, “Bayesian structural EM algorithm,” UAI 1998.

[9] A. Jepson, D. Fleet, and T. El-Maraghi, “Robust online appearance models for visual tracking,” CVPR 2001.

[10] Z. Ghahramani and G. Hinton,“Variational learning for switching state-space models,” Neural Computation 12(4): 963-996, 1998.

[11] N. Jojic and B. Frey, “Learning flexible sprites in video layers,” IEEE Conf. on Computer Vision and Pattern Recognition (CVPR ’01)

[12] E. Adelson and P. Anandan, “Ordinal characteristics of transparency,” AAAI-90 Workshop on Qualitative Vision, 1990.

[13] C. Williams and M. Titsias, “Greedy learning of mutliple objects in images using robust statistical factorial learning,” Neural Computation, 16(5): 1039-1062, 2004.

[14] J. Winn and A. Blake, “Generative affine localization and tracking,” NIPS 2004.

[15] B. Frey, N. Jojic, and A. Kannan, “ Learning appearance and transparency manifolds of occluded objects in layers,” CVPR 2003.

[16] A. Kannan, N. Jojic and B. Frey, “Generative model for layers of deformation and appearance,” AISTATS 2005.

[17] N. Jojic and B. Frey, “A generative model for 2.5D vision” www.research.microsoft.com/∼jojic/gmkl.ps, 2002.

[18] B. Frey and N. Jojic, “Advances in probabilistic inference in strucutred models,” IEEE PAMI, January 2005.

[19] N. Jojic, Y. Caspi, and M. Reyes-Gomez, “Probabilistic index maps for modeling natural signals,” UAI 2004.

[20] J. Movellan, J. Susskind, J. Hershey, “Large-Scale Convolutional HMMs for Real-Time Video Tracking,” CVPR 2004.

[21] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-Time Tracking of the Human Body,” IEEE PAMI 19(7): 780-785, 1997.

[22] J. Y. A Wang and E. H. Adelson, “Representing moving images with layers,” IEEE Transactions on Image Processing 3(5): 625–638, 1994.

[23] L. Zitnick, N. Jojic, and S. B. Kang, “Consistent segmentation for optical flow estimation,” ICCV, 2005.

[24] P. Pawan Kumar, P.H.S. Torr, and A. Zisserman,“Learning layered motion segmentation of video,” ICCV 2005.

[25] C. Sminchisescu and B. Triggs, “Building roadmaps of minima transitions in visual models,” Intl. Journal of Computer Vision, vol.61, No.1,2005.

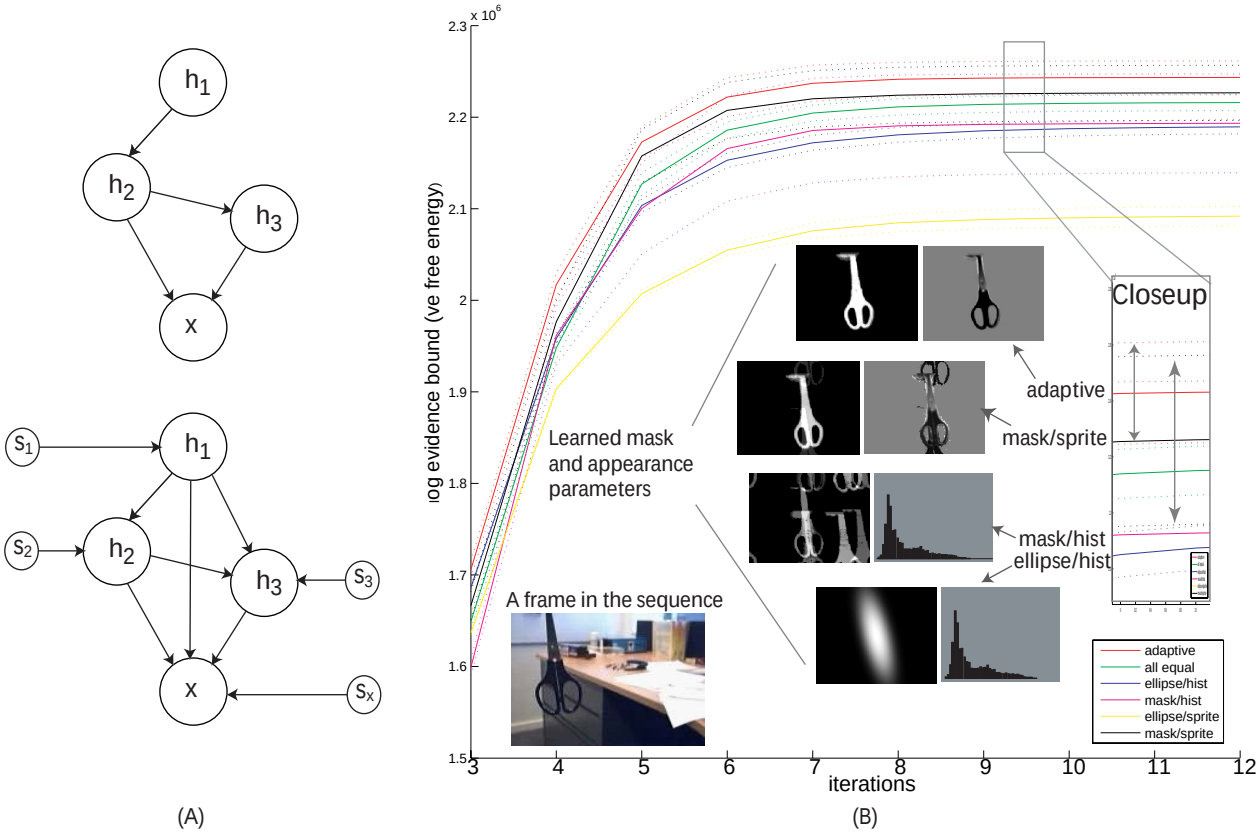

Figure 1. (print or view in color) (A) The graphical model example from Section 2 extended with multiple conditionals and switching variables so that setting $s_1 = s_2 = s_3 = 1$ reduces it back to the original model, while the additional possible models obtained combinatorially for three settings of s_2 and three settings for s_3 can help reduce local minima prolems in variational parameter learning. (B) A comparison of different settings of switch variables in Section 3 on the performance of unsupervised extraction of the scissors in the video whose one frame is shown above. The scissors move translationally from left to right and back. The Y coordinate in the plots denotes the negative of the free energy (log likelihood bound), and the X coordinate is the iteration number (from 3 to 12). A closeup of the curves after 10th iteration is also shown with an indicated two deviations spans for the top two models - adaptive and mask/sprite. The two deviation marks for all curves are drawn in dashed lines. Learned parameters after typical runs are shown for four models for qualitative comparison of the results. To make task more difficult, the models worked with gray-level frames. The metal scissors reflect light in variable directions creating difficult appearance variability.

Figure 2. (print or view in color) A few frames from a 180-frame long video sequence with a foreground object undergoing severe deformations in appearance and shape due to nonrigid motion and illumination effects on a sunny day. The background itself contains complex motion and illumination changes. The middle column shows the mode of the posterior distribution over pixel masks and the right column shows the resulting segmentation. (See also the supplemental videos at www.research.microsoft.com/∼jojic/mm.html)