Context-based Segmentation of Image Sequences

Jacob Goldberger   Hayit Greenspan

Abstract— We describe an algorithm for context-based segmentation of visual data. New frames in an image sequence (video) are segmented based on the prior segmentation of earlier frames in the sequence. The segmentation is performed by adapting a probabilistic model learned on previous frames, according to the content of the new frame. We utilize the maximum a-posteriori version of the EM algorithm to segment the new image. The Gaussian mixture distribution that is used to model the current frame, is transformed into a conjugate-prior distribution for the parametric model describing the segmentation of the new frame. This semi-supervised method improves the segmentation quality and consistency and enables a propagation of segments along the segmented images. The performance of the proposed approach is illustrated on both simulated and real image data.

Index Terms— image-sequence analysis, video segmentation, model adaptation, conjugate prior, MAP, context-based segmentation

I. INTRODUCTION

Automatic segmentation of images is one of the central challenges in computer vision and also plays an important role in human vision perception. In general, image segmentation is a key step towards image understanding, and serves in the variety of applications including object recognition, image coding and image indexing. Analysis of image-sequences, or video data, is an extension of the single image frame analysis to an analysis of a set of images that have both spatial and temporal coherency and causality.

In most of the existing works in the field, the first step towards segmenting an image is extracting a feature vector (e.g. color, texture, position) from each pixel. The segmentation task can then be treated as a clustering task, with the clustering performed on the extracted feature vectors. A variety of clustering approaches can be found in the literature. One possible clustering approach is based on mixture models in which one assumes the data was sampled from multiple models such that each model corresponds to one of the segments in the image. The assignment of points to segments can then be easily performed by computing the posterior probability of a pixel affiliation to a cluster. Learning the mixture model can be done by utilizing the well known EM algorithm [4] [12]. An alternative approach is to use spectral clustering methods based on finding eigen-vectors of affinity matrices [21] [23]. These methods are non-iterative and could be also efficiently computed [9], and thus avoid the necessity of a good initial condition.

While significant progress has been made in feature extraction and grouping algorithms, human vision perception is (still) much better than state-of-the-art computer vision image

J. Goldberger is with the Bar-Ilan University, Ramat-Gan 52900, Israel. E-mail: goldbej@eng.biu.ac.il

H. Greenspan is with the Tel-Aviv University, Tel Aviv 69978, Israel. E-mail: hayit@eng.tau.ac.il

segmentation algorithms. One of many possible explanations is that the human perception is also based on prior familiarity with objects and shapes that appear in previously seen similar images. In other words, human vision perception is based on high level prior knowledge about the semantic meaning of the objects that compose the image. In contrast, most computer based segmentation methods are based on totally unsupervised, per image or per frame procedures. The current work incorporates the above intuition and suggests a methodology for image segmentation, based on cues taken from the segmentation of related images. The most natural example of similar images occurs in image-sequence (video) segmentation where there is a continuity in the time axis within a single video shot. The technical contribution of this study is a method that transforms the MoG-based representation of a frame into a conjugate prior distribution for the next frame.

The paper proceeds as follows. In section 2 we review recent works in image and video segmentation that utilize context in the segmentation process. In section 3 we review the probabilistic image segmentation method based on mixture of Gaussians. Section 4 describes the transformation of a mixture of Gaussians (MoG) into a conjugate prior for the maximum a-posteriori (MAP) version of the EM algorithm as applied to learning a MoG. In section 5 we apply the MAP-EM algorithm to video segmentation and present experimental results on image sequences.

II. RELATED WORKS

Recently a number of authors suggested algorithms for segmenting a novel image based on the available segmentation of other similar images. In [2] hand-labeled image fragments are used from a given class (horses) to guide the segmentation task. A similar approach was proposed by [1] in which the normalized-cut algorithm was applied to a graph which is obtained by matching parts of a novel image to parts of pre-segmented images. A related approach is to improve the segmentation quality of similar images by sharing segmentation information among the images. In [11] the earth mover distance was used for merging and splitting a mixture model of one image in the context of a mixture model of a similar image. In [8] an iterative method was proposed, that assigns a semantic meaning to segmented images based on keywords supplied with the images. Most of the current approaches to supervised image segmentation are based on an un-supervised clustering procedure that is performed separately for each image. The information extracted from similar images is used in a post-processing step that refines the clustering and relates the clusters with high-level objects.

An extension of single frame analysis to multiple frames, as in the spatio-temporal segmentation and tracking of image-sequences, is a well-known challenging research problem. The

many algorithms proposed in the literature may be divided into two schools-of-thought: (1) approaches that track regions from frame to frame, and (2) approaches that consider the whole 3D volume of pixels and attempt a segmentation of pixel volumes in that block. An updated survey of spatio-temporal grouping techniques can be found in [19]. The current work focuses on frame-by-frame analysis. There are many works that use frame-by-frame segmentation and tracking (e.g.,[7], [5], [14], [22]). Many approaches use optical flow methods [13] to estimate motion vectors at the pixel level, and then cluster pixels into regions of coherent motion to obtain segmentation results. In [22], moving images are decomposed into sets of overlapping layers using block-based affine motion analysis and the K-means clustering algorithm. Each layer corresponds to the motion, shape and intensity of a moving region. Due to the complexity of object motion in general videos, pure motion-based algorithms cannot be used to automatically segment and track regions through image sequences. The drawbacks include the fact that optical flow does not cope well with large motion and that regions of coherent motion may contain multiple objects and require further segmentation for object extraction.

In works that incorporate spatial segmentation along with motion segmentation, it is commonly the case that the spatio-temporal segmentation task is decomposed into two separate tasks of spatial segmentation (based on in-plane features such as color and texture) within each frame in the sequence or within a selected frame of the sequence, followed by a motion segmentation phase. In [5], color and edge features are used to segment a frame into regions. Optical flow is utilized to project and track color regions through the video sequence. Optical flow is computed for each pair of frames. Given color regions and the generated optical flow, a linear regression algorithm is used to estimate the affine motion for each region. In [6] a six-parameter, two-dimensional affine transformation is assumed for each region in the frame and is estimated by finding the best match in the next frame. Multiple objects with the same motion are separated by spatial segmentation.

Recent works include statistical models for representing video content. Each frame is represented in feature space (color, texture etc.) via models such as the MoG model. Tracking across frames is then achieved by extended models, such as HMMs [3] or by model adaptation from frame to frame [16]. A related method, based on online learning of MoG models, is real-time segmentation and tracking of moving regions using background subtraction [15] [18].

In the current work, a video segmentation scheme is proposed that is based on pairing between adjacent frames in the video sequence, such that one frame serves as the context for the following frame. The segmentation of the first frame (frame1) is used as a (conjugate) prior distribution for the probabilistic representation of the next frame (frame2) in the sequence. The key point is that video is continuous in time. This ensures that the current frame is a perfect prior for the next frame. Unless there is a cut, the changes are small and can be modeled by adapting the blobs of the current frame. Unlike 3D batch-based schemes [19], the adaptation method is a frame by frame method so it can be used online.

Fig. 1. Input image (left). Image modeling via a mixture of Gaussians (center). Image segmentation using the learned model (right).

A natural choice for probabilistic image (or image-frame) representation is a mixture of Gaussians model (MoG). The MoG parameters of frame1 are used as hyper-parameters for the MoG learned from frame2. In other words, the segmentation algorithm is based on adaptation of an approved model, that can contain high level knowledge about objects and shapes in a given frame, to the next frame. In addition to improving the segmentation, this method also yields a local-based correspondence between the objects that appear in both frames.

III. Image Modeling with MoG

We consider an image-frame as a set of coherent regions. Each convex homogeneous region in the image plane is represented by a Gaussian distribution, and the set of all the regions in the image is represented by a Gaussian mixture model. The image, therefore is viewed as an instance of a mixture of Gaussians model. We focus here on the color feature. In order to include spatial information, the (x, y) position of the pixel is appended to the feature vector. It should be noted that the representation model is a general one, and can incorporate any desired feature space (such as texture, shape, etc) or combination thereof. Color features are extracted by representing each pixel with a three-dimensional color descriptor in a selected color space. In this work we choose to work in the $L * a * b$ color space which was shown to be approximately perceptually uniform; thus distances in this space are meaningful. Including the position enables a local based representation. Thus each pixel is represented by a five-dimensional feature vector. Pixels are grouped into homogeneous regions, by grouping the feature vectors in the selected five-dimensional feature space. The Expectation-Maximization (EM) algorithm can be used to determine the maximum likelihood parameters of a mixture of k Gaussians. The MoG model induces a segmentation in a natural manner by assigning each pixel to the Gaussian component whose posterior probability is maximal. Figure 1 shows an example of learning a MoG model for a given input image. In this visualization the Gaussian mixture is shown as a set of ellipsoids. Each ellipsoid represents the support, mean color and spatial layout, of a particular Gaussian in the image plane. Using the learned model (center) each pixel of the original image is affiliated with the most probable Gaussian, providing for a probabilistic image segmentation (right). A detailed description of MoG modeling for image segmentation can be found in [12].

IV. GENERATING A CONJUGATE PRIOR FROM A GIVEN MOG

In this section we derive an EM algorithm for finding the maximum a posteriori (MAP) parameters in the case of a mixture of Gaussian density. The (conjugate) prior parameter set is extracted from a given similar MoG model and from a single additional scalar parameter that quantizes the amount of similarity that exists between the given MoG and the model we intend to learn from the observed data. To motivate the EM updating expressions, we start with the simpler case of MAP of a single Gaussian density where there is a closed-form solution. Let $N(\mu, \Sigma)$ be a d-dimensional Gaussian density. Using a Bayesian approach, we apply the following conjugate prior. The precision (inverse of the covariance matrix) is Wishart with m degrees of freedom:

$$\Sigma^{-1} \sim \text{Wishart}(\Sigma_0^{-1}, m) \qquad (1)$$

i.e.

$$f(\Sigma|\Sigma_0, m) \propto |\Sigma|^{-\frac{m-d-1}{2}} \exp(-\frac{1}{2}Tr(\Sigma_0\Sigma^{-1}))$$

and the mean (conditioned on the precision) is normal with a scaling factor β:

$$\mu \sim N(\mu_0, \frac{1}{\beta}\Sigma) \qquad (2)$$

where μ_0, Σ_0, m and β are the hyper-parameters of the prior density. Given n independent samples $x_1, ..., x_n$ from $N(\mu, \Sigma)$, the MAP estimation of the normal density parameters is:

$$\hat{\mu} = \frac{\sum_{t=1}^{n} x_t + \beta\mu_0}{n + \beta} \qquad (3)$$

$$\hat{\Sigma} = \frac{\sum_{t=1}^{n}(x_t - \hat{\mu})^2 + \beta(\hat{\mu} - \mu_0)^2 + \Sigma_0}{n + m - d}$$

(where the square operator a^2 stands for $aa^\top$). In this derivation the hyper-parameters of the prior distribution are given as part of the set-up of the problem. Alternatively, assume that instead of hyper-parameters, we are given another normal distribution $N(\mu', \Sigma')$ which we know that is "similar" to the normal distribution we want to learn from the samples. We can extract hyper-parameters from the given normal distribution in the following way:

$$\mu_0 = \mu' \quad , \quad \Sigma_0 = \beta\Sigma' \quad , \quad m = \beta + d \qquad (4)$$

such that β is a parameter that controls the influence of the prior on the learned model. It can be set according to our prior knowledge about the amount of similarity that exists between the two models. Substituting definition (4) in equation-set (3), we obtain the following MAP estimation:

$$\hat{\mu} = \frac{\sum_{t=1}^{n} x_t + \beta\mu'}{n + \beta} \qquad (5)$$

$$\hat{\Sigma} = \frac{\sum_{t=1}^{n}(x_t - \hat{\mu})^2 + \beta((\hat{\mu} - \mu')^2 + \Sigma')}{n + \beta}$$

These expressions can be viewed as the ML estimation of a Gaussian distribution computed from the observations $x_1, ..., x_n$ and from another set composed of β virtual observations sampled from $N(\mu', \Sigma')$. The empirical average and variance of the virtual sample set coincide with μ' and Σ' respectively.

We shall now generalize this Bayesian approach to the case of a Gaussian mixture model. A d-dimensional distribution composed of a mixture of k Gaussians has the following form:

$$f(x|\theta) = \sum_{i=1}^{k} \alpha_i f_i(x|\mu_i, \Sigma_i) \qquad (6)$$

such that f_i is the normal distribution $N(\mu_i, \Sigma_i)$ and $\theta = \{\alpha_i, \mu_i, \Sigma_i, i = 1, .., k\}$.

Utilizing the Bayesian methodology, we can consider the MoG parameters as random variables whose prior distribution represent our prior knowledge on their values. We use conjugate priors on the parameter set θ (more precisely we use a prior model that is conjugated to the complete version of the MoG where the latent random variable is also observed). The following priors are used. The mixing coefficients are jointly Dirichlet:

$$(\alpha_1, ..., \alpha_k) \sim \text{Dirichlet}(\alpha_{01}, ..., \alpha_{0k}) \qquad (7)$$

The precisions (inverse of the covariance matrix) are Wishart with m_i degrees of freedom and the means (conditioned on the precisions) are normal with scaling factors β_i. More explicitly, the prior distribution $f(\theta)$ is:

$$f(\theta) = f(\alpha|\alpha_0) \prod_{i=1}^{k} f(\Sigma_i|\Sigma_{0i}, m_i) f(\mu_i|\mu_{0i}, \Sigma_i, \beta_i) \qquad (8)$$

such that:

$$f(\alpha|\alpha_0) \propto \prod_{i=1}^{k} \alpha_i^{(\alpha_{0i}-1)} \qquad (9)$$

$$f(\Sigma_i|\Sigma_{0i}, m_i) \propto |\Sigma_i|^{-\frac{m_i-d-1}{2}} \exp(-\frac{1}{2}Tr(\Sigma_{0i}\Sigma_i^{-1}))$$

$$f(\mu_i|\mu_{0i}, \Sigma_i, \beta_i) \propto$$

$$|\frac{1}{\beta_i}\Sigma_i|^{-\frac{1}{2}} \exp(-\frac{\beta_i}{2}(\mu_i - \mu_{0i})^\top \Sigma_i^{-1}(\mu_i - \mu_{0i}))$$

where $\alpha_{0i}, \mu_{0i}, \Sigma_{0i}, m_i$ and β_i are the hyper-parameters of the Bayesian model, i.e. they are the (known) parameters of the prior parameters density. Given n independent samples $x_1, ..., x_n$ from the MoG $f(x|\theta)$, we can utilize the EM algorithm to obtain the maximum a posteriori parameter estimation.

$$\theta_{map} = \arg\max_{\theta} f(\alpha) \prod_{i=1}^{k} f(\mu_i, \Sigma_i) \times \prod_{t=1}^{n} f(x_t|\theta) \qquad (10)$$

The expectation step of the MAP-EM algorithm is (same as the maximum likelihood version of the EM):

$$w_{it} = p(i|x_t) = \frac{\alpha_i f_i(x_t|\mu_i, \Sigma_i)}{\sum_{j=1}^{k} \alpha_j f_j(x_t|\mu_j, \Sigma_j)} \qquad (11)$$

We shall use the abbreviation:

$$n_i = \sum_{t=1}^{n} w_{it} \qquad i = 1, ..., k \qquad (12)$$

such that n_i is the expected number of data points sampled from the i-th Gaussian component f_i. The Maximization step is:

$$\hat{\alpha}_i = \frac{\alpha_{0i} - 1 + n_i}{\sum_j (\alpha_{0j} - 1) + n} \tag{13}$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^{n} w_{it} x_t + \beta_i \mu_{0i}}{n_i + \beta_i}$$

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^{n} w_{it} (x_t - \hat{\mu}_i)^2 + \beta_i (\hat{\mu}_i - \mu_{0i})^2 + \Sigma_{0i}}{n_i + m_i - d}$$

(where the square operator a^2 stands for $aa^\top$).

Assume we are given another MoG distribution which is known to be similar to the model we want to learn. Denote the parameter set of the given MoG model by $\theta' = \{\alpha'_i, \mu'_i, \Sigma'_i, i = 1, ..., k\}$. We can extract hyper-parameters from the given MoG distribution in the following form:

$$\beta_i = \beta \alpha'_i \quad , \quad m_i = \beta_i + d \tag{14}$$

$$\alpha_{0i} = \beta \alpha'_i + 1 \quad , \quad \mu_{0i} = \mu'_i \quad , \quad \Sigma_{0i} = \beta_i \Sigma'_i$$

Substituting definition (14) in equation-set (13), we obtain the following EM re-estimation equations:

$$\hat{\alpha}_i = \frac{n_i + \beta_i}{n + \beta} \tag{15}$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^{n} w_{it} x_t + \beta_i \mu'_i}{n_i + \beta_i}$$

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^{n} w_{it} (x_t - \hat{\mu}_i)^2 + \beta_i ((\hat{\mu}_i - \mu'_i)^2 + \Sigma'_i)}{n_i + \beta_i}$$

Using the following notation:

$$\bar{x}_i = \frac{\sum_t w_{it} x_t}{n_i} \quad , \quad \bar{x}_i^2 = \frac{\sum_t w_{it} x_t x_t^\top}{n_i} \quad , \quad c_i = \frac{n_i}{n_i + \beta_i}$$

the EM updating formulas can be written in the following simplified way:

$$\hat{\mu}_i = c_i \bar{x}_i + (1 - c_i) \mu'_i \tag{16}$$

$$\hat{\Sigma}_i = c_i \bar{x}_i^2 + (1 - c_i)(\mu'_i \mu'^\top_i + \Sigma'_i) - \hat{\mu}_i \hat{\mu}_i^\top$$

Note that we use only one extra parameter to build a prior distribution from the given MoG. This parameter, denoted by β, can obtain any value between zero and infinity. The value of the parameter β is based on our belief on the similarity between the models. The parameter β can be viewed as an adaptation coefficient which controls the balance between the known reference model and the new model we want to learn from the data. The first extreme case $\beta = \infty$ corresponds to enforcing the given model on the observations without any adaptation, i.e. $\theta \leftarrow \theta'$. The second extreme case, $\beta = 0$, corresponds to maximum-likelihood based learning of the new model without any reference to the prior information. The main advantage of the method presented in this paper is the ability to choose any learning scheme that exists between these two extreme cases. During the adaptation procedure we can eliminate a Gaussian component if the expected number of pixels in the new image that support this component (equation (12)) is less than a specified threshold.

The MAP-EM algorithm presented in this section monotonically increases the function $f(x|\theta) f(\theta|\theta', \beta)$ which is:

$$\prod_{t=1}^{n} f(x_t|\theta) \times \text{Dirichlet}(\alpha_1, ..., \alpha_k; \beta_1 + 1, ..., \beta_k + 1) \times$$

$$\prod_{i=1}^{k} (\text{Wishart}(\Sigma_i^{-1}; (\beta_i \Sigma'_i)^{-1}, \beta_i + d) \times N(\mu_i; \mu'_i, \frac{1}{\beta_i} \Sigma_i))$$

such that $\beta_i = \beta \alpha'_i$.

Similarly to the case of a normal distribution, we can interpret the MAP-EM as an algorithm that finds the maximum-likelihood parameter-set of an unknown MoG, based on the observations $x_1, ..., x_n$ and another set of β virtual observations, such that $\beta \alpha'_i$ of these virtual samples are sampled from the i-th component of the given MoG. The method presented in this section is related to a MoG adaptation method used for speaker verification tasks [10] [20]. Reynolds et al. [20] proposed an EM-based method to adapt a universal model to a speaker model given a short utterance. In addition to the different application domains, the derivation of the prior model from a given MoG is also different. In [20], the same number of virtual samples ($\beta = \beta_i = 16$ samples) was taken from each Gaussian component without taking into account the weights α'_i associated with the Gaussian components. In vision applications, MAP parameter estimation for MoG learning is currently used mainly as a penalized estimation. The conjugate prior expression is viewed as a penalty term that is added to the log-likelihood term as a regularizer (e.g. to penalize a non-identity or a non-isotropic covariance [17]).

V. EXPERIMENTAL RESULTS

The MAP estimation, developed in the previous section, can be utilized in context-based video segmentation. The case of video sequences is inherently adequate for the proposed adaptation method since video sequences are highly continuous along the time axis. Hence, in all frames along the sequence, that are not detected as a cut, each video frame can be used to construct a prior model for the next frame. Technically, the adaptation is performed by estimating a maximum a posteriori MoG model such that the MoG model of the reference image is used as a prior knowledge. The MAP-EM algorithm (Equation 15) is used for the parameter-set learning. The adaptation parameter β can be related in the case of video data to the amount of continuity present in a video (i.e. for higher frame rate we could expect to use a larger β). To evaluate the video segmentation and tracking performance we have compared 3 MoG learning methods. In the first case a MoG model was independently generated for each frame. The second option is using the previous frame to initialize the MoG learning for the current frame. The third option is the MAP-EM framework developed in this study, where the previous frame is used as a prior model. The ratio between β and the number of pixels in the current frame was manually optimized and set to 0.2 (however there is low sensitivity to the exact value of β) . The prior model was also used as an initial value for the MAP-EM algorithm. We have found empirically that when the prior model is used for initialization, it is better

Fig. 2. Video sequence, example 1: (a) selected frames (1,5,9,13) from original sequence; (b) ML model separately learned per frame; (c) ML model learned per frame, previous frame is used for initialization ($\beta = 0$) ; (d) MAP-EM model based on adaptation from previous frame ($\beta = 0.2$); (e) segmentation maps based on ML model learned separately for each frame; (f) segmentation maps using the previous frame for initialization (ML); (g) segmentation maps using the MAP-EM estimation.

to multiply all variance matrices by a constant greater than one (we used 4) to avoid converging to a local maximum point. We have also found empirically that there is no need to set prior values for the Gaussian weights a_i (i.e we can use a non-informative prior). Using a prior model enforces constraints on the colors and the relative location of objects within the images. Adding a constraint on the objects size (via the mixture coefficients) is too restrictive and does not correspond to the variability along the image sequence. The MAP-EM algorithm (written in C and run on a Pentium 4) takes less than a second per frame. Note that the two context-based learning methods are similar in terms of computational efficiency. A selected set of frames from a given input video

sequence is shown in Figure 2a. Results of frame-by-frame (independent) MoG model generation are presented in Figure 2b. Modeling results of using the previous frame to initialize the MoG learning for the current frame are shown in Figure 2c. Results of the MAP-EM estimation are shown in Figure 2d. The corresponding segmentation maps are presented Figure 2e, 2f and 2g, respectively. Several things may be noted from the presented results: In the modeling results, it is evident that there is not much consistency in the blob structure of Figure 2b. Increased consistency across the blob representations can be seen as we add context to the modeling framework. This is seen in Figure 2c and even more evident in Figure 2d. The MAP-EM learning method imposes strong consistency in the

Fig. 3. Video sequence, example 2: (a) selected frames (1,11,21,31) from original sequence; (b) ML model learned per frame, previous frame is used for initialization ($\beta = 0$); (c) MAP-EM model based on adaptation from previous frame ($\beta = 0.2$); (d) segmentation maps using the previous frame for initialization (ML); (e) segmentation maps using the MAP-EM estimation.

blob structure, especially in the non-moving segments. The MAP-EM method can also improve the segmentation quality. An example can be seen in the third frame of the presented sequence. The dark shirt (blue colored) and dark trousers (black colored) in the third frame of Figure 2f are merged together while in Figure 2g we note that with the MAP-EM based segmentation the two regions remain separated, as desired.

In order to track high-level objects across image sequences or video, a linkage is necessary between the individual frame models (or corresponding segmentation maps). The presented results exemplify that in the case of MAP-EM learning, strong correspondence of blobs across frames is evident as well as strong correspondence of segmented regions across frames in the sequence. This implies the ability to track regions or objects in the video stream. Figure 3 presents similar segmentation experiments performed on another image sequence. It can be seen that the segmentation quality is much better when the MAP-EM is used for segmentation of the current frame. Note that, similar to the previous example, if each frame is separately segmented, then the segmentation can be improved but there will be no linkage between models corresponding to consecutive frames.

We next wish to quantitatively compare between the two context-based model adaptation schemes, namely using a given (previous frame) MoG as an *initialization* for the maximum-likelihood iterative (EM) estimation of a new model versus using the MAP-EM method, in which the given MoG is used for *initialization and prior* for a MAP learning of a similar MoG. It is difficult to quantify the relative contribution of the two schemes, directly from the video data itself. Hence we performed the following simulation example: In each trial a 2D

mixture of 5 Gaussians $f(x)$ is randomly chosen. The mean of each Gaussian is sampled from $N(0, I)$ and the variance is set to be the identity matrix. A parameter β is used to transform f into a conjugate prior for a new MoG $g(x)$, i.e. the parameters of g are drawn from the conjugate prior model created from f. We don't observe g explicitly. We are only given $n = 200$ samples from g and the task is to estimate the parameters of g where f is used either as prior (and for initialization) or just for initialization. To measure the similarity between the learned MoG $\hat{g}$ and the true model g we use the relative entropy between the two MoGs. In both learning methods, we can assume that the correspondence between the Gaussian components of the true and the estimated model is correct. Hence we use (a symmetrical version of) the conditional relative entropy (conditioned on the latent random variable) which is the following expression:

$$\text{dist}(g, \hat{g}) = \text{dist}(\sum_{i=1}^{k} \alpha_i g_i, \sum_{i=1}^{k} \hat{\alpha}_i \hat{g}_i) = \sum_{i=1}^{k} \alpha_i D_{KL}(g_i || \hat{g}_i)$$

where $D_{KL}(g_i || \hat{g}_i)$ is the Kullback-Leibler divergence between the two Gaussian densities. The average results over 1000 trials for several values of β are shown Figure 4. The simulations clearly demonstrate that using the context-based model adaptation via the MAP-EM method achieves increased performance than using initialization-only adaptation within the ML framework.
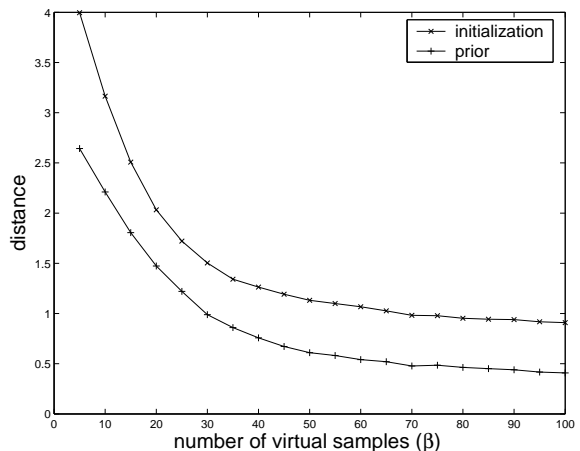

Fig. 4. The distance between the true MoG and estimation of the MoG learned from samples and a given similar MoG model, as a function of β. The graph presents a comparison between using the similar model just as an initialization for the EM algorithm or also as a prior for a MAP-EM.

VI. DISCUSSION

In this study we described an algorithm for segmenting a novel image based on available segmentation of a similar image. The algorithm was applied to video segmentation. Segmentation using apriori knowledge, or context, provided better segmentation results than unsupervised segmentation schemes. In addition, higher-level object segmentation is enabled, as well as tracking in time. In a video segmentation task, considering the previous video frame as a prior knowledge for the current frame is much more statistically well-founded

and can be more reliable for tracking an object along a video sequence. The differences detected between the MAP-EM method and initialization-based adaptation within an iterative parameter estimation process, may have been caused by the fact that the injected side-information in the initialization may lose its effect after several iterations of the learning process. In contrast the MAP-EM re-injects the prior knowledge in every iteration of the learning process.

The method proposed here can be used in several additional contexts, including segmenting a given image based on a set of similar pre-segmented images, or segmenting a given image based on an expert-based segmentation (and labeling) of a sample representative image. Interesting applications may be in the medical domain. For example, segmenting an MRI brain image based on pre-segmented brain atlas. The main challenge in general image segmentation (rather than image-sequence) is the invariance issue, be it in the general contrast of the respective images or in the positioning of the objects of interest within the image, i.e. the alignment issue. For example, if the high-level object (e.g. animal) is in the top corner of the image instead of the center, the segmentation will not necessarily benefit from the conjugate prior. An additional issue to consider is the case of scale variability within the image set. In general, the supervised approach that benefits from context within a pre-segmented and pre-labeled image set, is sensitive to invariance issues within the set. Such issues are much less prominent in image-sequences. Frames within a short video sequence are expected to be similar due to the continuity of the video signal in the time axis. Hence, a segmentation of one video frame can be easily propagated along the entire sequence.

In this work we used only color features. Texture characteristics of regions, as well as shape and other features, can be extracted as additional features. Such features can be associated with the components of the prior model and provide additional knowledge in the segmentation process.

REFERENCES

[1] S. Agarwal and S. Belongie. Segmentation by example. Technical report, UCSD CSE Dept. Tech Report, 2002.

[2] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *Proc. of the European Conf. Comput. Vision*, pages 109–122, 2002.

[3] C. Bregler. Learning and recognizing human dynamcs in video sequences. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, June 1997.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

[5] S-F Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):602–615, 1998.

[6] Y. Deng and B. S. Manjunath. Netra-v: Toward an object-based video representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):616–627, 1998.

[7] B. Duc, P. Schroeter, and J. Bigun. Spatio-temporal robust motion estimation and segmentation. In *6th Int. Conf. Comput. Anal. Images and Patterns*, pages 238–245, 1995.

[8] P. Duygulu, K. Barnard, N. de Freitas, and D.A. Forsyth. Object recognition as machine translation; learning a lexicon for a fixed image vocabulary. In *Proc. of the European Conf. Comput. Vision*, 2002.

[9] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystron method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

[10] J.L. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech Audio Processing*, pages 291–298, 1994.

[11] H. Greenspan, G. Dvir, and Y. Rubner. Context dependent image segmentation and image matching via emd flow. *Journal of Computer Vision and Image Understanding*, 93(1):86–109, 2004.

[12] H. Greenspan, J. Goldberger, and L. Ridel. A continuous probabilistic framework for image matching. *Journal of Computer Vision and Image Understanding*, 84:384–406, 2001.

[13] B. Horn and B. Schunck. Determining optical flow. *Artificial Intell.*, 17:185–203, 1981.

[14] G. Iyengar and A. B. Lippman. Videobook: An experiment in characterization of video. In *Proc. IEEE Int. Conf. Image Processing*, volume 3, pages 855–858, 1996.

[15] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Proc. of the European Workshop on Advanced Video Based Surveillance Systems*, 2001.

[16] S. Khan and M. Shah. Object based segmentation of video using color, motion and spatial information. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, volume II, pages 746–751, 2001.

[17] S. McKenna and H. Nait-Charif. Learning spatial context from tracking using penalized likelihoods. In *Int. Conference on Pattern Recognition (ICPR)*, pages 138–141, 2004.

[18] S. McKenna, Y. Raja, and S. Gong. Object tracking using adaptive colour mixture models. In *Asian Conference on Computer Vision (ACCV)*, pages 615–622, 1998.

[19] R. Megret and D. DeMenthon. A survey of spatio-temporal grouping techniques. Technical report, UMIACS-2002-83 Technical report, 2002.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.

[21] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[22] J. Y. Wang and E. H. Adelson. Spatio-temporal segmentation of video data. In *SPIE*, volume 2182, pages 120–131, 1994.

[23] Y. Weiss. Segmentation using eigenvector: a unifying view. In *Proc. of the Int. Conference on Computer Vision*, 1999.

PLACE PHOTO HERE

Jacob Goldberger received the B.Sc. degree in mathematics in 1985 from Bar-Ilan-University, Israel and the M.Sc. degree in mathematics and the Ph.D. degree in electrical engineering in 1989 and 1998, respectively both from Tel-Aviv University, Israel. He did a Postdoc at the Weizmann institute and at the University of Toronto. In 2004 he joined the School of Engineering, Bar-Ilan University where he is currently a faculty member. His research interests include machine learning, information theory, computer vision and speech recognition.

PLACE PHOTO HERE

Hayit Greenspan received the B.Sc. and M.Sc. degrees from the Electrical Engineering Department of the Technion, Israel, in 1986 and 1989, respectively, and the Ph.D. degree from the Electrical Engineering Dept. at CALTECH - California Institute of Technology, in 1994. Following the Ph.D. she was a Postdoc at the Computer Science Division at U.C. Berkeley. In 1997 she joined the Biomedical Engineering Dept. of the Faculty of Engineering, Tel-Aviv University, Israel, where she is currently a faculty member. Her research interests include medical image processing and analysis, content-based image and video search and retrieval, statistical image modeling and segmentation.