ACTIVE BAYESIAN MIXTURE LEARNING FOR IMAGE MODELING AND SEGMENTATION USING LOW LEVEL FEATURES

Constantinos Constantinopoulos, Aristidis Likas

Department of Computer Science
University of Ioannina
GR 45110, Ioannina, Greece
e-mail: {ccostas, arly@cs.uoi.gr}

ABSTRACT

Gaussian mixture models (GMM) have been shown an effective tool for image representation and segmentation. However, several issues related to GMM training for image modeling have not been adequately resolved such as the specification of the number of mixture components and the increased complexity for images of typical size (e.g. 256x256). We present an approach for GMM-based image modeling employing an incremental variational algorithm for Bayesian mixture learning that automatically specifies the number of mixture components. Moreover, we integrate the method in an active learning framework which allows to gradually build the GMM using only a small fraction of the image pixels.

1. INTRODUCTION

Statistical image representation, ie. the modeling of an image based on the distribution of various features at pixel level, constitutes an important task in computer vision and image analysis. Several approaches have been proposed, with earlier work mainly focusing on the use of histograms, while later more sophisticated statistical modeling tools such as mixture models have been employed. Mixture models are now considered as a convenient tool for image modeling based on low level features such as color, texture etc [1]. More specifically, the assumption under the GMM framework is that an image is considered as a set of regions (segments) where each region is represented by a Gaussian distribution and the set of all regions in an image is represented by a GMM.

Without loss of generality, in this work we consider that features are computed at the pixel level and in our experiments we have considered the color feature computed in the (L, a, b) space, although it is straightforward to employ any sets of features (e.g. Gabor coefficients for texture modeling). To build the GMM for an image, a dataset X is constructed that contains one feature vector $x_n \in R^d$ for each image pixel n. Then an efficient training method is applied to this dataset to provide the final GMM for the image.

Let f be a mixture with J Gaussian components

$$f(x) = \sum_{j=1}^{J} \pi_j \, \mathcal{N}(x|\mu_j, T_j) \tag{1}$$

where $\pi = \{\pi_j\}$ are the mixing coefficients (priors), $\mu = \{\mu_j\}$ the means (centers) of the components, and $T = \{T_j\}$ the precision (inverse covariance) matrices. After training, it is possible to segment the image, ie. assign a group pixels to each GMM component by finding for the feature vector x of each pixel the component j with maximum posterior, (ie. with maximum $\pi_j \, \mathcal{N}(x|\mu_j, T_j)$).

An important issue is how to impose the requirement for spatial smoothness, ie. that in most cases neighboring pixels should be assigned to the same component (cluster). This can be achieved in two ways. The first is by imposing an MRF prior on the posteriors [2]. The MRF approach is more difficult to handle from the learning point of view, and requires all image pixels to be included in the training set. It provides as outcome the posterior probabilities for each pixel. In addition, no effective method has been proposed for automatically determining the number of GMM components in the MRF framework.

The second way to impose spatial smoothness is the *direct approach* [3], which considers the spatial location (x, y) of a pixel as an additional feature to be included in the feature vector describing this pixel, ie. $x_n = (L, a, b, x, y)$ for the n-th pixel with image coordinates (x, y) and color vector (L, a, b). Thus the spatial distance between two pixels contributes significantly to the total distance between the corresponding feature vectors. In this way, adjacent pixels tend to have similar cluster labels since their spatial distance is small, thus spatial smoothing is achieved. On the

This research was cofunded by the European Union in the framework of the Program "Heraklitos" of the "Operational Program for Education and Initial Vocational Training" of the third Community Support Framework of the Hellenic Ministry of Education, funded by 25 percent from national sources and 75 percent from the European Social Fund (ESF).

other hand, this approach forces large image areas (with approximately the same color) that can be considered as one segment to be splitted into smaller segments, because spatially distant pixels cause the distance of the corresponding feature vectors to be large, thus they cannot be modeled by a single Gaussian component. However, this fragmentation problem can be easily resolved through a simple post-processing stage that merges image regions corresponding to GMM components with similar color mean.

A notable advantage of using the direct approach for spatial smoothness is that it results in a GMM that takes as input the image location (x, y) and can assign feature labels (for example color) to every location (x, y) independent of whether this pixel has been used for training or not. This has several significant consequences:

- it is easy to obtain a model for a specific region of the image by considering only the mixture components that are active in this region. For example we can easily derive a mixture model for the color density in an arbitrary image region
- it is straightforward to marginalize the (x, y) coordinates and obtain a GMM for the distribution of the low level features
- it is easy to assign feature values (for example color) to image locations (x, y) not used for training, simply by computing the component j with highest posterior and assigning the mean color value $(\mu_{j,L}, \mu_{j,a}, \mu_{j,b})$ of this component as the 'representative' color for location (x, y) in the segmented image. This allows to exploit the redundancy in the pixel information and train the mixture model using only a representative subset of the pixels. The rest of the pixels could be used as a *test set* to evaluate "segmentation accuracy" by computing the difference between the original image and the "segmented" image.

Going one step further, we propose the use of an *active learning* methodology where training starts with a small number of feature vectors and more (appropriately selected) feature vectors are gradually added to the training set as learning proceeds. In this way it is possible to build a representative image model using only a small fraction of the pixels.

Despite the advantages obtained by using the direct approach for spatial smoothness, a major difficulty is introduced that relates to the specification of the number of mixture components. For example, it is possible that an image with four colors cannot be modeled using a GMM with four components, especially in the case where the image segments with the same color have large size or are disconnected. In this case the direct approach requires more components to model the image. Therefore *in the direct approach it is difficult to specify in advance the number of GMM components* and it is essential to use GMM training methods that incorporate a built-in mechanism for automatically assessing the number of components (as for example our method used in this work [4]) or the MML-based approach in [5]. However, to implement active learning it is preferable to use a learning algorithm that gradually increases the number of components as is the case with our method described next.

2. THE INCREMENTAL VARIATIONAL BAYESIAN METHOD

In this section we describe a Bayesian method for Gaussian mixture learning [4] that is deterministic, does not depend on the initialization, and resolves adequately the model selection problem, ie. the specification of the number of components. The method is an incremental one: it starts with one component and progressively adds components to the model. The procedure for component addition is based on a "splitting test" applied to each of the existing mixture components. According to this test, a component is replaced by two sub-components and then variational Bayesian learning is applied to the specific pair of components, while the rest components remain "fixed". Due to the introduction of priors on the parameters of the Gaussians, a competition takes place between the components. If the data distribution in the region of the tested component strongly suggests the existence of more than one clusters, then both sub-components will "survive" and the number of model components will be increased. Otherwise, the competition among the two components will cause one of them to be eliminated and the initial component will be recovered. This strategy of incremental component addition also facilitates the specification of the parameters of the priors, since it can be based on the parameters of the component to be splitted. In order to apply this idea, a modification of the typical Bayesian mixture model is required that is described in the graphical model of Figure 1. Note that a prior has been imposed only on the $J - s$ "fixed" mixing coefficients $\tilde{\pi}$. Let $X = \{x_n\}$ the set of training points containing the feature vectors of the image pixels.The hidden variables $Z = \{z_{jn}\}$ capture the missing information of which component has generated a given data point. More specifically, $z_{jn} = 1$ if component j is responsible for generating x_n, otherwise $z_{jn} = 0$. Therefore it holds that:

$$p(X|Z, \mu, T) = \prod_{n=1}^{N} \prod_{j=1}^{J} [\mathcal{N}(x_n|\mu_j, T_j)]^{z_{jn}} \qquad (2)$$

The distribution of Z is a product of multinomials

$$p(Z|\pi, \tilde{\pi}) = \prod_{n=1}^{N} \prod_{j=1}^{s} \pi_j^{z_{jn}} \prod_{j=s+1}^{J} \tilde{\pi}_j^{z_{jn}} \qquad (3)$$

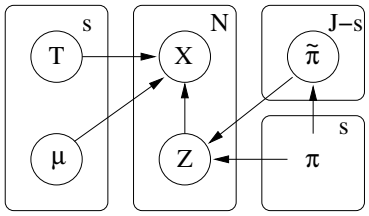

Fig. 1. The graphical model.

given the subset $\tilde{\pi} = \{\tilde{\pi}_j\}$ of "fixed" mixing coefficients and the subset $\pi = \{\pi_j\}$ of "free" mixing coefficients. For notational convenience and assuming J mixing components, we can always rearrange the indexes so that the first s components are the "free" ones.

The typical Bayesian framework assumes conjugate Dirichlet priors over the entire set of mixing coefficients. However, in order to apply our idea, it is necessary to define the conditional joint distribution $p(\tilde{\pi}|\pi)$ of the "fixed" mixing coefficients given the "free". It is known that if the joint distribution of a set of variables is Dirichlet, then the marginal joint distribution of a subset of the variables is also Dirichlet. Using Bayes theorem the conditional joint distribution $p(\tilde{\pi}|\pi)$ can be derived, which is a non-standard Dirichlet with parameters α_j $(j = s+1, \ldots, J)$:

$$\begin{aligned} p(\tilde{\pi}|\pi) &= \left(1 - \sum_{j=1}^{s} \pi_j\right)^{-J+s} \frac{\Gamma(\sum_{j=s+1}^{J} \alpha_j)}{\prod_{j=s+1}^{J} \Gamma(\alpha_j)} \\ &\times \prod_{j=s+1}^{J} \left(\frac{\tilde{\pi}_j}{1 - \sum_{k=1}^{s} \pi_k}\right)^{\alpha_j - 1} \end{aligned} \quad (4)$$

and constitutes a conjugate prior of the "fixed" coefficients. Completing the specification of our Bayesian model we assume Gaussian and Wishart priors for μ and T respectively [6]:

$$p(\mu) = \prod_{j=1}^{s} \mathcal{N}(\mu_j|0, \beta\,\mathcal{I}) \quad (5)$$

$$p(T) = \prod_{j=1}^{s} \mathcal{W}(T_j|\nu, V). \quad (6)$$

Learning in the Bayesian framework can be achieved through maximization of the marginal likelihood of the data which is obtained by integrating out the hidden variables of the model. In our case, the marginal likelihood of X given π is obtained by integrating out $\theta = \{Z, \mu, T, \tilde{\pi}\}$ as follows

$$p(X|\pi) = \sum_{Z} \int p(X, \theta|\pi)\, \mathrm{d}\mu\, \mathrm{d}T\, \mathrm{d}\tilde{\pi}. \quad (7)$$

Following the Variational Bayes methodology, we maximize a lower bound $\mathcal{L}$ of the logarithmic marginal likelihood $\log p(X|\pi)$:

$$\mathcal{L}[q, \pi] = \sum_{Z} \int q(\theta) \log \frac{p(X, \theta|\pi)}{q(\theta)}\, \mathrm{d}\theta \quad (8)$$

where q is an arbitrary distribution that approximates the posterior distribution $p(\theta|X, \pi)$. The maximization of $\mathcal{L}$ is performed in an iterative way, where at each iteration two steps take place (in analogy to the EM approach): first maximization of the bound with respect to q, and subsequently maximization of the bound with respect to π. To implement this maximization with respect to q the mean-field approximation [6] has been adopted, which assumes that q is constrained to be a product of the form: $q(\theta) = q_Z(Z) q_\mu(\mu) q_T(T) q_{\tilde{\pi}}(\tilde{\pi})$. The resulting update equations for the parameters of the q distributions (E-step) and the parameters π (M-step) [4] are omitted due to space limitations.

Using the above idea the incremental algorithm for Bayesian mixture model learning proceeds as follows. Mixture components are sequentially added to the mixture model using the following component splitting procedure: one of the mixture components is selected and is appropriately splitted in two components. We treat the resulting two components as "free" and the rest as "fixed" according to the terminology introduced previously. Next we set the precision prior $p(T)$ based on the characteristics of the splitted component, and apply variational learning as described in the previous section. In case that the two components provide a much better fit to the data in their region, then both components are retained in the mixture model, otherwise the update equations will eliminate one of them. The splitting test is applied sequentially to all components and the method terminates when all mixture components have been unsuccessfully tested for splitting. In the case where a successful split is encountered, then the number of mixture components increases and a new round of split tests for all components is initialized.

To illustrate the details of splitting, assume that some component $\hat{\jmath}$ has to be splitted, with density $\mathcal{N}(x|\mu_{\hat{\jmath}}, T_{\hat{\jmath}})$. The idea is that in order to form the new mixture, we remove component $\hat{\jmath}$ and insert two new components with densities $\mathcal{N}(x|\mu_{\hat{\jmath}1}, T_{\hat{\jmath}1})$ and $\mathcal{N}(x|\mu_{\hat{\jmath}2}, T_{\hat{\jmath}2})$ respectively. We have selected to place the centers of the two components along the dimension of the principal axis of the covariance $T_{\hat{\jmath}}^{-1}$ and at opposite directions with respect to the center $\mu_{\hat{\jmath}}$. The mixing coefficients of the two components are set equal $\pi_{\hat{\jmath}1} = \pi_{\hat{\jmath}2} = \pi_{\hat{\jmath}}/2$, and their parameters are set according to: $\mu_{\hat{\jmath}1} = \mu_{\hat{\jmath}} + \sqrt{\lambda}\, u$, $\mu_{\hat{\jmath}2} = \mu_{\hat{\jmath}} - \sqrt{\lambda}\, u$, $T_{\hat{\jmath}1} = T_{\hat{\jmath}}$ and $T_{\hat{\jmath}2} = T_{\hat{\jmath}}$, where λ is the maximum eigenvalue of $T_{\hat{\jmath}}^{-1}$ and u the corresponding eigenvector. An important issue in the proposed method is the specification of the scale parameter V of the prior $\mathcal{W}(\nu, V)$ over the precision matrices, based on the splitted component. We set $\nu = d$ (which is the min-

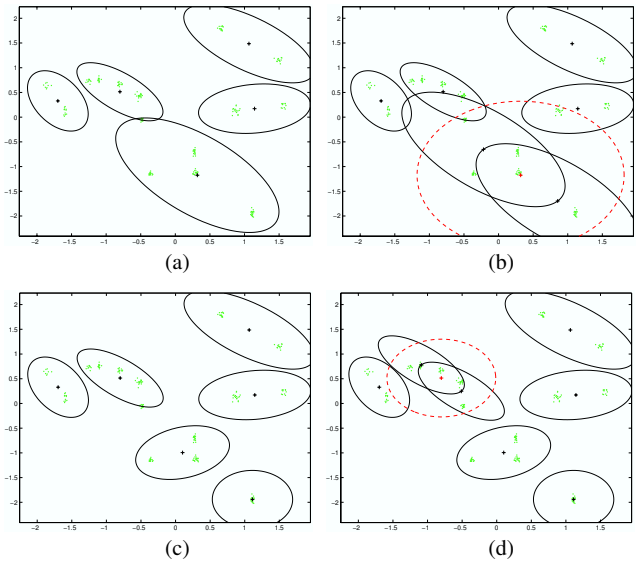

Fig. 2. Four steps of the incremental training procedure. The expected covariance w.r.t. the Wishart prior is depicted with a dashed line. (a) An intermediate solution with 5 components. (b) One component is splitted in two. (c) The mixture after variational learning. (d) Another component is selected and splitted.

imum allowed value) and $V = \nu\lambda\mathcal{I}$, where λ is the highest eigenvalue of T_j^{-1}. The value of β was set to 10^{-10}. An example of component splitting is illustrated in Figure 2.

3. ACTIVE GMM LEARNING

While the above procedure can be applied to the whole image dataset to provide the image GMM, we propose the use of an *active learning* approach. More specifically, we start with a small set of feature vectors and build an initial mixture model using the method described previously. Next, additional feature vectors are selected using an appropriate selection criterion and added to the training set. Then the learning algorithm is applied to the augmented training set starting from the mixture model obtained in the previous iteration. This procedure is repeated several times until a stopping criterion is satisfied.

More specifically, given an image with N pixels, let $\Omega = \{x_n |\ n = 1, \ldots, N\}$ be the set of all feature vectors. In each active learning iteration, there are two disjoint sets X and X_c ($X \cup X_c = \Omega$). X contains the set of feature vectors that we use to learn the mixture, and X_c is the rest dataset (pool of samples) that we use to augment X in the next iteration. In order to add points, we compute the 'reconstruction error' for the points in X_c. The reconstruction error for a feature vector (L, a, b, x, y) is computed by determining the mixture component j with the highest posterior and computing the distance between the actual color vector (L, a, b) and the mean color vector $(\mu_{j,L}\mu_{j,a}, \mu_{j,b})$ of component j. Next we augment X with the subset of X_c that contains the feature vectors with the largest reconstruction error. Of course these feature vectors are removed from X_c.

To terminate the active learning, we compute (as described above) at the end of each iteration the total reconstruction error for all the points in Ω. If this error has been increased, then we reject the current mixture and adopt the mixture of the previous iteration. If the error has not improved for four consecutive iterations, we terminate the GMM learning procedure. This active learning method significantly accelerates learning since usually a small fraction of the samples (e.g. 2%) is used for building the GMM model of the image.

4. EXPERIMENTAL RESULTS

To illustrate the performance of the proposed method, we have conducted experiments using both artificially generated and natural images. For each image the following steps were taken. First the dataset containing the feature vectors for all pixels was constructed and next the feature vectors were preprocessed so that each feature distribution has zero mean and unit standard deviation. The resulting dataset Ω was then used for building the mixture model using the active learning methodology. Using the resulting mixture model, the 'segmented' image is produced by assigning to each pixel (x, y) the mean color value of the GMM component with highest posterior. In all experiments, to initiate active learning, 500 uniformly distributed pixels are selected and the corresponding feature vectors are added to the training set X. At each active learning iteration 100 points were selected and added to X.

Figure 3 illustrates the segmentation result for two artificial images. The depicted mixture components and the feature vectors have been projected on the spatial coordinates (x, y). To demonstrate performance on natural images, we provide segmentation results for two images from the Berkeley Segmentation Data Set (BSDS) [7]. Figure 4 illustrates the segmentation of image 253036 (top row) and 118035 (bottom row). The first image was segmented using a mixture with 13 components, while the second with 8. It is clear that segmentation results are quite satisfactory In addition, we compared the proposed active data selection method for augmenting X with a method where X is augmented through uniform random selection of feature vectors. The total reconstruction (ie. segmentation) error and the number of components at each active learning iteration are illustrated in Table 1. It is clear that the proposed data selection criterion leads to solutions with much better

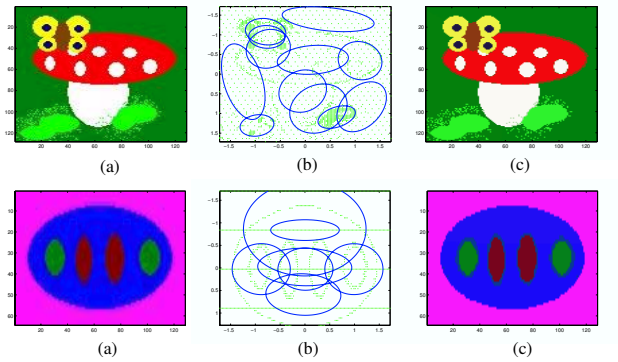

Fig. 3. Artificial images. (a) Original image. (b) The selected training points and components of the mixture model. (c) Image segmentation using the mixture.

Fig. 4. Segmentation of natural images from BSDS. (left) Original image. (right) Segmented image.

reconstruction error, while the random selection approach seems to terminate prematurely.

5. CONCLUSIONS

We have presented an efficient approach for image modeling using GMMs. The approach is fully automatic, makes no assumptions regarding the required number of GMM components and provides solutions that are spatially smooth. Through an active learning methodology it is possible to obtain representative GMMs using only a small fraction of the pixels. Future work will focus on testing the performance of the method in the case where several other features are also included in the feature vector (such as texture-related features). Also we plan to integrate in our approach a technique for the automatic selection of the most salient features.

Table 1. Comparison of the proposed active sampling method compared to random sampling. For each active learning iteration we report the reconstruction error, and in parentheses the number of GMM components.

BSDS image 253036		*BSDS image 118035*	
active	*random*	*active*	*random*
1516.2 (4)	1516.2 (4)	1143.7 (4)	1143.7 (4)
1163.2 (6)	1300.6 (5)	1017.5 (5)	1047.7 (4)
1099.0 (7)	1276.8 (5)	840.3 (6)	
903.6 (9)		827.4 (8)	
800.5 (11)			
744.5 (13)			

6. REFERENCES

[1] Goldberger J. Greenspan, H. and L. Ridel, "A continuous probabilistic framework for image matching," *J. Computer Vision and Image Understanding*, vol. 84, pp. 384–406, 2001.

[2] S. Z. Li, *Markov Random Field Molelling in Computer Vision*, Springer Verlag, 2001.

[3] Belongie S. Greenspan H. Carson, C. and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026–1038, 2002.

[4] C. Constantinopoulos and A. Likas, "Bayesian gaussian mixture learning based on variational component splitting," *Technical Report no 24–9/2005, Dept. of Computer Science, Univ. of Ioannina*, 2005.

[5] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[6] A. Corduneanu and C. M. Bishop, "Variational Bayesian model selection for mixture distributions," in *Artificial Intelligence and Statistics 2001*. 2001, pp. 27–34, Morgan Kaufmann.

[7] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, 2001, vol. 2, pp. 416–423.