Probabilistic Classification of Image Regions using an Observation-Constrained Generative Approach

Sanjiv Kumar[1], Alexander C. Loui[2], and Martial Hebert[1]

[1]The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Imaging Science and Technology Lab, Eastman Kodak Company, Rochester, NY, USA
Email: {skumar, hebert}@ri.cmu.edu, alexander.loui@kodak.com

Abstract— **In generic image understanding applications, one of the goals is to interpret the semantic context of the scene (e.g., beach, office etc.). In this paper, we propose a probabilistic region classification scheme for natural scene images as a priming step for the problem of context interpretation. In conventional generative methods, a generative model is learnt for each class using all the available training data belonging to that class. However, if a set of newly observed data has been generated because of the subset of the model support, using the full model to assign generative probabilities can produce serious artifacts in the probability assignments. This problem arises mainly when the different classes have multimodal distributions with considerable overlap in the feature space. We propose an approach to constrain the class generative probability of a set of newly observed data by exploiting the distribution of the new data itself and using linear weighted mixing. A KL-Divergence-based fast model selection procedure is also proposed for learning mixture models in a sparse feature space. The preliminary results on the natural scene images support the effectiveness of the proposed approach.**

***Keywords*—Image region classification, generative model, semantic interpretation, image segmentation**

.

I. Introduction

AUTOMATIC extraction of the semantic context of a scene is useful for image indexing and retrieval, robotic navigation, surveillance, robust object detection and recognition, and auto-albuming etc. Recent literature reveals increasing attention to this field [1]-[3]. However, most of the attempts have been made to extract the context of a scene at a high level in the abstraction hierarchy. For example, Torralba et al. [1] represent the context by using the power spectra at different spatial frequencies while Vailaya [2] uses the edge coherence histograms to differentiate between natural and urban scenes. Limited attention has been paid to the task of specific context generation from a scene (scene classification), e.g., if a scene is a *beach* or an *office*. The main hurdle in such context generation is that it requires not only the knowledge of the regions or objects in the image, but the semantic information contained in their spatial arrangement as well.

The motivation for our work comes from the following paradox of scene classification. In absence of any a-priori information, the scene classification task requires the knowledge of regions and objects contained in the image. On the contrary, it is increasingly being recognized in vision community that context information is necessary for reliable extraction of the image regions and objects [1], [3]. To solve this paradox, an iterative feedback scheme can be envisaged, which refines the scene context and image region hypotheses iteratively. Under this paradigm, an obvious choice to the region classification scheme is one that allows easy modification of the initial classification without requiring to classify the regions afresh. Probabilistic classification of image regions can provide great flexibility in future refinement using the Bayesian approach as the context information can be encoded as improved priors.

In this paper, we deal with the probabilistic classification of the image regions belonging to the scenes primarily containing natural objects, e.g., water, sky, sand, skin, etc. as a priming step for the problem of scene context generation. Several techniques have been proposed to classify various image regions in distinct categories, e.g., [4], [5]. However, these are primarily based on discriminative approaches leading to *hard* class assignments and hence are not suitable for the iterative refinement scheme mentioned above.

In the conventional informative schemes, a generative model for each class is learnt using all the available training data belonging to that class and newly observed data is assigned a probability based on the learnt model. However, it is possible that a set of newly observed data may have been generated because of the subset of the full generative model support, and using the full model to assign generative probabilities can produce serious artifacts in the probability assignments. For example, in the training set, various pixels associated with the class *tree* may have wide variations in the color and texture depending on the season, illumination conditions, and the scale. A generative model for class *tree* will try to capture all these variations in the same model. Hence, while assigning the generative probabilities to the newly observed data, the learnt generative model may assign high probability to not only the pixels belonging to the class *tree,* but also to those that may have some semblance to tree data (in some arbitrary combination of illumination and other physical conditions). Similar analogy holds for the data belonging to other classes also. This problem arises mainly when different classes have multimodal distributions that are close in the feature space.

Another problem with the generative models is that they tend to give more weight to the regions in the feature space that contain more data, instead of emphasizing on the discriminative boundary between the data belonging to different classes. This implies that the data near the boundary will be assigned almost similar probabilities irrespective of their class affiliations.

We propose to alleviate these problems by using the simple observation that the newly observed data is usually generated by a small support of the overall generative model. In our previous *tree* example, this means that the data belonging to the *tree* class in a new test image is usually generated during the same season as well as under relatively homogeneous illumination and other physical conditions. Thus, the distribution of the newly observed data can be used to constrain the overall generative model while producing the generative class density maps for that data. That is why the proposed technique has been named as the 'observation-constrained generative approach'.

II. Generative Approach

In the first stage of our approach, we use two well-known techniques, i.e., supervised learning using the labeled training data, and the unsupervised learning using the newly observed test data. The main contribution of our approach lies in the next stage where the outputs of these two techniques are merged using a Bayesian scheme. In the beginning of the following discussion, we briefly review the unsupervised and supervised learning applied in the present context.

Let X_ω be the set of training data associated with class ω, where $X_\omega = \left\{x_k^\omega : x_k^\omega \in \Re^d\right\}_{k=1}^{n_\omega}$ and C be the classes of interest. The set containing all the C classes is referred to as the *recognition vocabulary*. The dimension of the feature space, d is the number of bands in the multiband input images. Each band represents a color or a texture feature associated with the image pixels. To capture this inherent multimodality of the data, we assume a mixture of Gaussian generative model for the data. Mixture models can approximate any continuous density to arbitrary accuracy provided the model has sufficiently large number of components and the parameters of the model are chosen correctly [6]. The class conditional density function of a data point $x_\omega^k \in X_\omega$ is given by,

$$p(x_k^\omega \,|\, \omega) = \sum_{m=1}^{M} p(x_k^\omega \,|\, m,\omega)\, P(m \,|\, \omega) \qquad (1)$$

where M is the number of components in the mixture model. In this model, each data point x_k^ω belonging to class ω is generated by first choosing a Gaussian component with probability $P(m \,|\, \omega)$ and then generating the data point with probability $p(x_k^\omega \,|\, m, \omega)$, which is a Gaussian given by,

$$p(x_k^\omega \,|\, m,\omega) = \frac{1}{\left|2\pi C_m^\omega\right|^{1/2}} exp\left\{-\frac{1}{2}(x_k^\omega - \mu_m^\omega)^T C_m^{\omega^{-1}} (x_k^\omega - \mu_m^\omega)\right\} \qquad (2)$$

where μ_m^ω is the mean and C_m^ω is the covariance matrix of the component m, belonging to class ω.

The parameters θ_ω of the generative model in (1), i.e., μ_m^ω, C_m^ω and $P(m \,|\, \omega)$ for each component m are learnt using the standard Maximum Likelihood formulation using the Expectation Maximization (EM) optimization technique [7]. Thus,

$$\hat{\theta}_\omega = \underset{\theta_\omega}{arg\,max} \prod_{k=1}^{n_\omega} \sum_{m=1}^{M} p(x_k^\omega \,|\, m,\omega)\, P(m \,|\, \omega) \qquad (3)$$

In the above formulation, the data points in the set X_ω have been assumed to be conditionally independent given θ_ω.

For a given test image I, the newly observed dataset is defined as $X_I = \left\{x_t^I : x_t^I \in \Re^d\right\}_{t=1}^{N}$ where N is the number of pxiels in I. The probability of association of each x_t^I with a given class in the recognition vocabulary (given by (1)) yields a Class Density Map over the image I. Now, the aim is to constrain the probabilities contained in these maps by enforcing the statistics of the newly observed data obtained from the test image. To do this, at first, the newly observed data is *soft* clustered in different groups in an unsupervised manner. For soft clustering, we again use the mixture of the Gaussian model. According to this, the probability of a newly observed data point, x_t^I is given as,

$$p(x_t^I) = \sum_{j=1}^{K} p(x_t^I \,|\, j)\, P(j) \qquad (4)$$

where $j = 1 \ldots K$ are the clusters and $p(x_t^I \,|\, j) \sim N(\mu_j, C_j)$ similar to (2). However, instead of traditional clustering (where each image pixel is assigned to a particular cluster), we utilize the probability of association of each pixel with a cluster j, i.e., $P(j \,|\, x_t^I)$. This leads to *soft* or probabilistic clustering and the maps representing these probabilities for each cluster are called Cluster Probability Maps. These maps enable the refinement of the Class Density Maps by constraining the overall generative model. To estimate the parameters of the unsupervised learning model mentioned in (4), similar EM formulation is used as in (3) except that m and x_k^ω are replaced by j and x_t^I respectively along with the suitable cardinalities of their corresponding sets.

Given the test image *I*, the first step towards constraining the Class Density Maps using soft clustering involves finding the class that has highest probability of being represented by a Cluster Probability Map. This is done by maximizing the conditional probability of class ω_c, $(c = 1 \ldots C)$ given a cluster *j*, i.e. $P(\omega_c|j)$. To compute the class posterior, first the marginal density of the cluster *j* given class ω_c is considered from the joint density of the newly observed data *x* and the cluster as,

$$p(j|\omega_c) = \int_{S_x} p(j, x|\omega_c)\, dx$$

or,

$$p(j|\omega_c) = \int_{S_x} p(j|x, \omega_c)\, p(x|\omega_c)\, dx \qquad (5)$$

It should be noted that *j* is defined only over the support of newly observed data S_x, which has been explicitly mentioned in the above integrals. In the first term of the integrand of (5), we further make a fair assumption that clustering is conditionally independent of the class given the data, i.e., $p(j|x,\omega_c) = p(j|x)$. Thus,

$$p(j|\omega_c) = \int_{S_x} p(j|x)\, p(x|\omega_c)\, dx$$

Since the image data is discrete, the integral over S_x is approximated by the finite sum and the cluster conditional is given by,

$$p(j|\omega_c) = \sum_{x \in X_I} P(j|x)\, p(x|\omega_c) \qquad (6)$$

Using (6), the class posterior can be computed easily from the Bayes rule,

$$P(\omega_c|j) = \frac{p(j|\omega_c)P(\omega_c)}{\sum_{c=1}^{C} p(j|\omega_c)P(\omega_c)} \qquad (7)$$

Given the class posterior for a cluster, the class $\hat{\omega}$ that maximizes this posterior is chosen as the consistent class for that cluster. This is equivalent to a MAP selection of the class given a cluster under the assumption of zero-one loss function. This procedure yields the most consistent class for each cluster. An implicit assumption has been made that each cluster belongs to one of the classes in the recognition vocabulary. Let K_ω be the set of the cluster maps that are consistent with the class ω. In other words, K_ω contains all those cluster maps that yield class ω as the MAP estimate of their corresponding class posterior given by (7). The Constrained Class Density Map is obtained simply by linear weighting of each pixel density of the Class Density Map by the corresponding pixel probability of the consistent Cluster Probability Maps, i.e.,

$$p_{cons}(x_t^I|\omega) = \sum_{j \in K_\omega} p_{orig}(x_t^I|\omega)P(j|x_t^I) \qquad (8)$$

where, $p_{orig}(.|.)$ is the original and $p_{cons}(.|.)$ is the constrained class conditional density. An intuitive explanation of (8) is that the constrained map is a linearly mixed, weighted density map where weights follow Gaussian distribution in the feature space because each cluster map approximately corresponds to a Gaussian distribution. It can be noted, that the final constrained density map for a given class tends to enhance those areas in the original density map that are supported by the statistics of the regions in the test image that show strong association with the given class.

III. Model Selection

In the proposed generative approach, we need to know the number of the components in the Gaussian mixture models in (1) as well as in (4), which amounts to the problem of model selection. The maximum likelihood approach is not appropriate for this task, as it would always prefer more components. Several techniques have been proposed under the topic of model selection [8]-[11]. Full Bayesian model selection techniques provide a more principled method of model selection and generally use a parametric or hierarchical form to approximate the prior distribution over the parameters [8]. The BIC or MDL approach, as proposed by Rissanen [11] can be shown to be asymptotically consistent version of the full Bayesian model selection techniques. In the MDL technique, the description length (*DL*) is given as,

$$DL = -\log p(X/\theta) + (l/2) \log n \qquad (9)$$

where *X* is the data, θ is the parameter vector containing all the model parameters, *l* is the number of parameters and *n* is the dataset size. The first term in the right hand side of (9) is the negative log likelihood (i.e., code length of likelihood) and the second term is code length of the parameters, which acts as a penalty term as the number of parameters increases.

However, there are two main limitations while implementing the full Bayesian or MDL criterion. First, they generally need iterative schemes to compute the model evidence or the likelihood, which is prone to getting stuck in local extrema and second, they are fairly slow and become almost impractical for the online soft clustering applications. In the present work, we propose a Kullback-Leibler Divergence (KLD) based method to estimate the number of components in the mixture model based on certain assumptions. The KL Divergence (KLD) is defined as ,

$$KL(\widetilde{p} \| p) = \int \widetilde{p}(x) log \frac{\widetilde{p}(x)}{p(x)} dx \qquad (10)$$

where $\widetilde{p}(x)$ is the true density of data x and $p(x)$ is the model density.

To apply the KLD in the component selection procedure, the model density is given by the mixture of Gaussians, but the true density is not known. We assume the data histogram (empirical distribution) to be the representative of the true density. This assumption is reasonable in the case of natural scene images, as they are mostly composed of smooth and homogeneous regions. The true density is further approximated by incrementally fitting Gaussians on the modes of the data histogram. For this purpose, we use second order normal approximation of the empirical distribution [12]. Parameters of the Gaussian are obtained by matching the curvature of the Gaussian with that of the empirical density at the mode. This is equivalent to using the inverse of the information matrix at the mode of the empirical density as the covariance matrix of the Gaussian. The expected information matrix is given as,

$$J(x) = E\left\{\left[\frac{\partial}{\partial x} log\, p(x)\right]\left[\frac{\partial}{\partial x} log\, p(x)\right]^T\right\}$$

where E is the expectation with respect to x. $J(x)$ is further approximated by its value at the mode $\hat{x}$ of $p(x)$, and the stochastic or observed information matrix is given as:

$$I(x) = \left[\frac{\partial}{\partial x} log\, p(x)\right]\left[\frac{\partial}{\partial x} log\, p(x)\right]^T\Bigg|_{\hat{x}} \qquad (11)$$

By using $I(x)$, the covariance matrix C_g of the approximated Gaussian is given as,

$$C_g = I^{1}(x) \qquad (12)$$

Because we are using the discrete empirical distribution as $p(x)$, computing $I(x)$ in (11) is equivalent to computing the negative of the hessian of the best fit curve at the mode of the distribution. The neighborhood of $log\ p(x)$ at the mode was assumed to be locally quadratic for computational purposes. Let $\Omega(\hat{x}) = \left\{x_\rho : x_\rho \in \Re^d\right\}_{\rho=1}^{n_\Omega}$ be the set of all data points in the neighborhood of the mode $\hat{x}$. In the local neighborhood of $\hat{x}$, the log-density can be expressed as:

$$log\, p(x_\rho) = \frac{1}{2} x_\rho{}^T H x_\rho + B^T x_\rho + c \qquad (13)$$

where H is the hessian matrix, and B and c are other constants. The hessian can be computed using Singular Value Decomposition (SVD) on the local neighborhood of the mode of the histogram. Let us denote the combined vector for all the neighbors of $\hat{x}$ as $Y = [log\ p(x_1), \ldots, log\ p(x_{n_\Omega})]^T$, the data matrix as D, each component of which is a monomial of order two or less, and ϑ as the vector that contains all the components of H, B, and c. The quadratic fitting amounts to sloving the linear set of equations given by $Y = D\vartheta$, which is solved using SVD as $\vartheta = (V\Sigma^{1}U^T)Y$, where $D = U\Sigma V^T$.

In the proposed method, the first estimate of the model density is obtained by fitting the first Gaussian at the mode of the histogram and the KLD between the empirical and the model density is computed. The integral in (10) is approximated by the finite sum over the discrete bin data. Next, a new density is computed by fitting one more Gaussian on the next highest mode of the histogram and mixing the two Gaussians. The mixing parameters have been assumed to be uniform. The modes are selected to be the maxima in their local neighborhood of a prespecified size. As per the above formulation, as the number of Gaussians is increased, KLD decreases and starts increasing when the number of Gaussians grows beyond that supported by the data distribution. In addition, it should be noted that if the KLD remains stable as the components are increased, it is an indicator that one or more of the previous components has already explained the new modes. This technique works fairly well in low dimensional, sparse feature space (as is the case with the images containing natural regions) as finding the modes in the local neighborhood of a histogram is relatively easy.

IV. Feature Extraction

The type of features to be extracted from an image depends on the nature of the scene classification task. In the present work, we deal with the scene images primarily containing natural regions. Although not sufficient, low-level features such as color and texture contain good representation power for the region classification of natural scenes.

A. Color Features

Color is an important component of the natural scene classes. However, the color-based features suffer from the problem of color constancy. For the natural scenes, we argue that given enough variations in the training data set, we can capture the class color distribution in varying illumination conditions. In addition, in the outdoor natural scenes, the problem of artificial illuminants is not as restrictive as the changes in brightness.

To extract the color features, we need to represent the color in a suitable space. In the present work, we found that the results with three different spaces i.e., rectangular HSV, g-RGB, and L*uv* were fairly similar. We chose generalized RGB (g-RGB) space, as it normalizes the effects of brightness variations effectively. The g-RGB space is given by two coordinates, ($r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$) where (R, G, B) are the coordinates in the RGB space. It is clear that one degree of freedom is lost in this conversion because the third coordinate of this space is simply (1 - *r* - *g*). This loss of one degree of freedom is true for almost all the color spaces, which either normalize or dispense with the luminance information.

B. Texture Features

In contrast to the color, texture is a not a point property. Instead, it is defined over a spatial neighborhood. Texture can provide good discrimination between natural classes that resemble in color domain e.g., water and sky. However, in our case, texture should be dealt with more care, as any single class does not have a unique texture. Within a semantically coherent region, there might be areas of high or low textures, with different scales and directional uniformity. In the classification of such regions, a very strong texture measure can sometimes undo the good work done by the color features.

Several techniques have been reported in the literature to compute the texture in a pixel neighborhood. The famous ones include Multiresolution Simultaneous Autoregressive (MSAR) model [13], Gabor Wavelets [14] and the Second Moment Eigenstructure (SME) [15, 16]. In the present work, we have used a weaker measure of texture yielded by the Second Moment Eigenstructure (SME), which can capture the essential neighborhood characteristics of a pixel. The second moment matrix at each image pixel *(i, j)* is given by:

$$M(i,j)=\begin{bmatrix}\sum_{c}\sum_{W}I_i^{c\,2} & \sum_{c}\sum_{W}I_i^cI_j^c \\ \sum_{c}\sum_{W}I_i^cI_j^c & \sum_{c}\sum_{W}I_j^{c\,2}\end{bmatrix} \quad (14)$$

where I_k^c is the gradient of the image in spatial direction *k* over the color channel *c* for *k = i, j* and *c = R, G, B*. *W* is the window over which these gradients are summed. We use Gaussian weighting in window *W* around the pixel of interest to give more weight to the pixels near it. The texture obtained from (14) is a *colored texture* because the above matrix captures the texture in color space instead of usual intensity space. The second moment matrix can be shown to be a modification of bilinear, symmetric positive definite metric defined over the 2D image manifold embedded in a 5D space of (*R, G, B, i, j*) [16]. The eigenstructure of the second moment matrix represents the textural properties. Two measures have been defined in [15] using the eigenvalues of the matrix, (a) anisotropy = $1 - \lambda_2/\lambda_1$, and (b) normalized strength = $2\sqrt{(\lambda_1 + \lambda_2)}$, where λ_1 and λ_2 are the two eigenvalues of matrix *M(i, j)* and $\lambda_1 > \lambda_2$. In the present work, we have used the combination of anisotropy and the strength called *texture strength (S)* as the texture measure. It is given as the product of anisotropy and normalized strength.

V. Results and Discussion

We have divided this section in two subsections. The first subsection discusses the simulation results of the proposed KLD-based model selection scheme on the synthetic data, and the second subsection contains the results of the proposed observation-constrained generative approach applied to real natural scenes.

A. Model Selection Results on the Synthetic Data

To verify the effectiveness of the proposed KLD based model selection scheme and compare the results with the MDL based approach, we applied these techniques on a simulated dataset containing 100,000 samples drawn from a mixture of three Gaussians in two-dimensional space. The mixing parameters used in the mixture model were 0.7, 0.2, and 0.1. The KL Divergence based model selection scheme was applied to the data histogram and the size of neighborhood was chosen to be 3 × 3. Fig. 1 shows the change in KLD as the number of components is increased in the Gaussian mixture. The KLD falls sharply when the components are increased from 1 to 3 and then starts increasing. Thus, this method correctly finds the number of components in the mixture data.

To compare the results of KLD with MDL approach, MDL was computed for the given dataset using (9). The maximum likelihood estimates of the parameters were obtained using EM. Because EM is sensitive to the initialization, we performed the MDL computation several times. Fig. 2 (a) and (b) show two typical plots of the change in description length as the number of components is increased. In (a), the MDL metric incorrectly favors two components, possibly, because the EM algorithm was stuck in a local maximum.

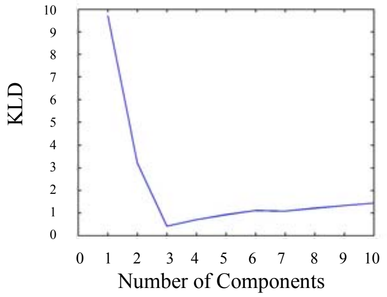
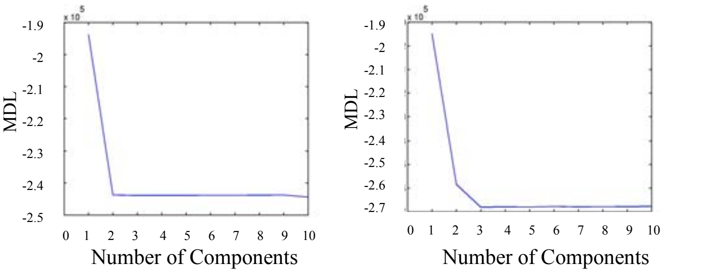


Fig. 1. KL-Divergence based model selection on the synthetic data. KLD correctly favors three components.



(a) (b)

Fig. 2. Two typical runs of the MDL based model selection technique on the synthetic data. (a) MDL incorrectly favors two components. (b) MDL correctly favors three components

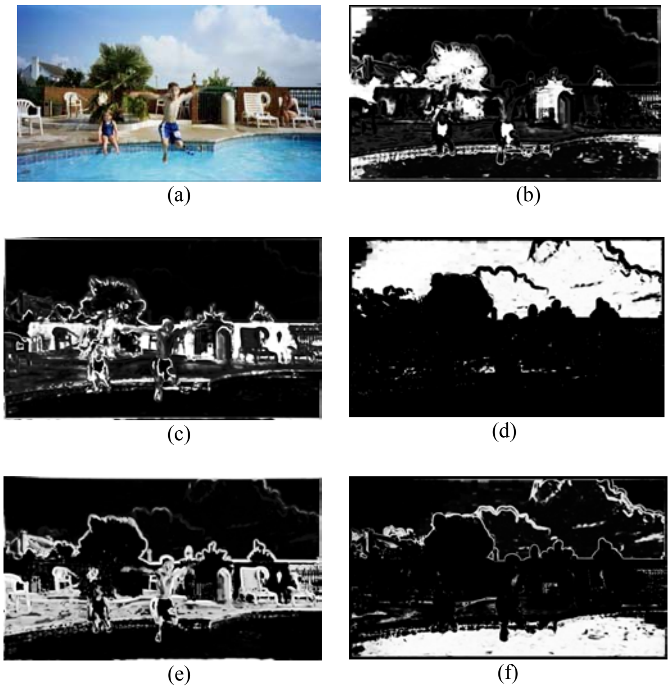


Fig. 3. Cluster Probability Maps corresponding to different clusters (*j*) obtained from the soft clustering of the input image. (a) Input color image. (b) *j* = 1 (c) *j* = 2 (d) *j* = 3 (e) *j* = 4 (f) *j* = 5

In another typical run when EM yielded the correct parameters caused by convergence to the global maximum, the MDL metric correctly favors three components. Thus, it can be seen that even for a very low dimensional space, MDL is not robust when EM is not initialized in proximity of the global maximum. In addition, the average time taken by a typical MDL run was about 935 s (in Matlab, 733 MHz) while that by KLD was only 0.21 s. Thus, we can get substantial savings in terms of speed by using KLD. This is especially important for the unsupervised clustering in the test image, which has to be done online unlike the supervised learning.

B. Results on Real Images

The proposed observation-constrained generative approach was tested with several images primarily containing natural regions. The class recognition vocabulary contained five classes: *sky, water, skin, sand/soil,* and *grass/tree*. A total of 130 pixelwise labeled images, each of size 499 × 874 pixels, were used as a training set for the supervised learning.

An input color image is shown in Fig. 3 (a). It can be noted that the scene is fairly cluttered and contains regions with varying illumination intensity, e.g., tree regions. Sky pixels in the image show bimodal distribution due to the presence of white textured cloud in a blue sky. At first, the soft clustering is performed on the image and the Cluster Probability Maps are generated. For this purpose, the number of clusters in the image is estimated using the KLD based technique. Fig. 4 shows the change in KL Divergence when the number of clusters (K) is increased. As K is increased from 1 to 2, sharp decrease in KLD can be noticed. After K = 5, the KLD almost stabilizes indicating the non-significance of further increments in K. Hence, we have chosen five as the number of clusters in the original image. Intuitively, one can observe five broad categories in the input image i.e., *sky, water, red wall, floor/skin,* and *tree/grass.*

Once the number of clusters has been determined, unsupervised learning of the mixture model generates the Cluster Probability Maps i.e., $P(j \left| x_t^I \right.)$. These maps for five clusters in the original image are given in Fig. 3 (b)-(f). In the Cluster Probability Maps, a brighter pixel indicates a higher probability of association of that pixel with the given cluster. It is clear from Fig. 3 that different clusters have captured the similarities in the image pixels. Increasing order of j, the maps broadly represent *grass, red wall, sky, floor,* and *water*. The semantic classes, which represent relatively small regions, have been merged with other clusters, e.g., *skin* regions were merged with red wall or floor.

There are some intuitively obvious incorrect associations in the cluster maps, e.g., for *j* = 1, parts of sky, water, and swimming costume have been clustered along with the *grass/tree* class. However, it should be emphasized that at this stage the algorithm has no notion of a semantic class. The results of probabilistic clustering are purely based on the *similarity* within the newly observed data.

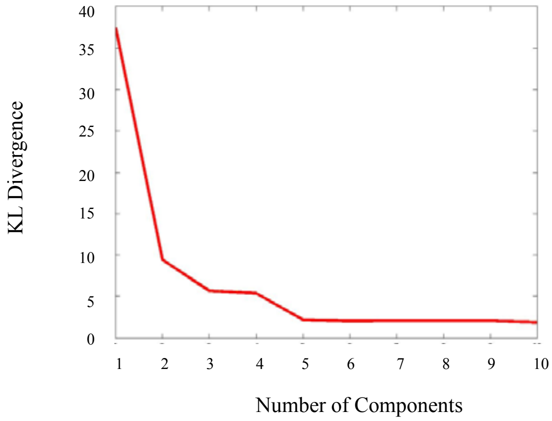


Fig. 4. KLD based component selection for the image from Fig. 3 (a)

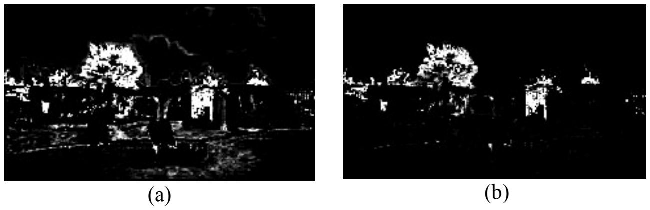


Fig. 5. For the input image shown in Fig. 3 (a.), (a) Class Density Map corresponding to the class *tree/grass.* (b) Corresponding Constrained Class Density Map.

The main aim of the soft clustering was not to obtain perfect clustering, but to obtain a probabilistic estimate of cohesiveness in various pixels, which could later be used for constraining the results of the class generative models. These errors in the clustering do not have limiting influence on the final Constrained Class Density Maps. This will be shown in the later part of this section.

As the next part of the process, generative mixture models were learnt for each class in the recognition vocabulary. The feature sets corresponding to all the pixels belonging to each class (in 130 training images) were used. The number of components in each class conditional mixture model was learnt using the KLD-based technique explained before. In the present results, we focus on the class *grass/tree* to explain all the steps involved in the observation-constrained generative approach. For this class, the KLD-based model selection favored five components in the mixture model. A Class Density Map displaying $p(x_k^{\omega} | \omega)$ for every pixel in the input image (Fig. 3 (a)) corresponding to class *grass/tree* is given in Fig. 5 (a). A brighter pixel indicates a higher density. It can be seen that the generative model has correctly predicted high density for the *grass/tree* regions. However, there are several non-grass areas with significant probability of being grass, e.g., parts of water, wet floor etc. These areas share the features close to those of *grass/tree* in some arbitrary illumination or physical conditions like scale, season etc. They have the potential of misleading the Bayesian refinement of class maps while using improved priors obtained from the scene context cues. Fortunately, the statistics of the observed *grass/tree* pixels in the given test image has enough information to remove these outliers by constraining the overall generative model.

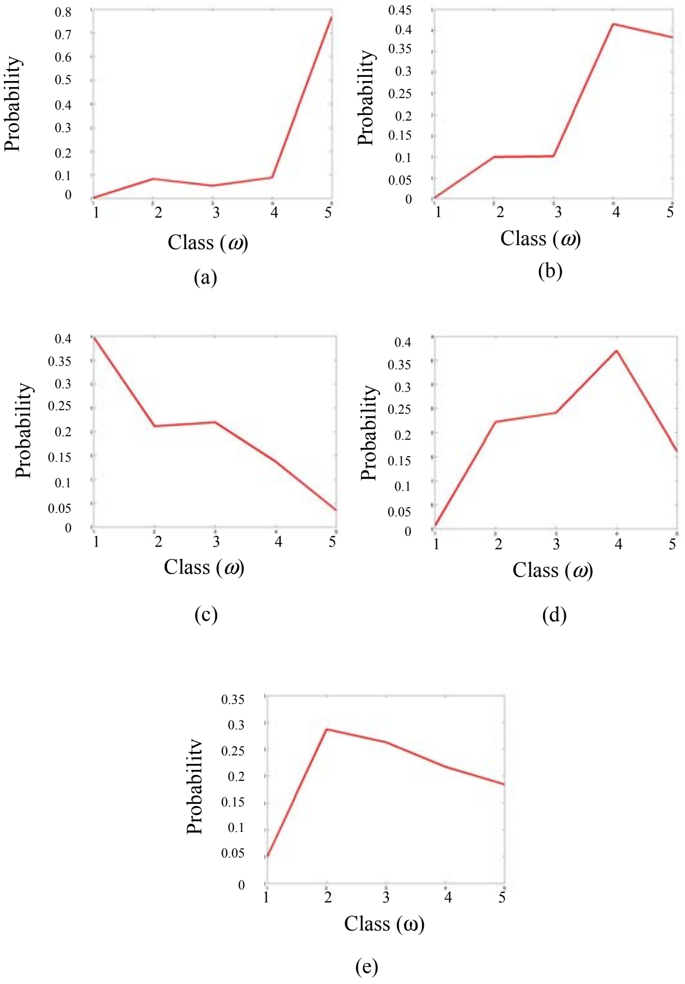


Fig. 6. Posterior distribution $P(\omega | j)$ over classes *(ω)* given the different Cluster Probability Maps *j* from Fig. 3. (a) *j* = 1, (b) *j* = 2, (c) *j* = 3, (d) *j* = 4, (e) *j* = 5

To obtain the Constrained Class Density Maps for class *grass/tree*, first those Cluster Probability Maps are found that are consistent with class *grass/tree*. For this purpose, posterior distributions over the classes given each Cluster Probability Map are obtained using (7). Fig. 6 displays the posterior distributions for Cluster Maps *j* = 1…5 displayed in Fig. 3 (b)–(f). Each plot is a discrete valued graph where horizontal axis displays the five different classes i.e., 1. *sky,* 2. *water,* 3. *skin,* 4. *sand/soil,* and 5. *grass/tree.* For each cluster map, the maximally consistent class is obtained using the MAP estimate of the posterior. It is clear from the plots that the cluster map for *j* = 1 has much higher probability of being from the class *grass/tree* ($\omega = 5$) than the other four classes. Also, this is the only cluster that is consistent (in the sense of MAP) with the class *grass/tree*. This correspondence between the chosen class and the Cluster Probability Map is supported intuitively from Fig. 3 (b) and Fig. 5 (a). All other cluster maps have natural semantics favoring other classes in the recognition vocabulary. Thus, the set of consistent cluster maps, K_{ω} for the class *grass/tree* has cardinality one.

The Cluster Probability Map corresponding to *j* = 1 (Fig. 3 (b)) was used to constrain the original Class Density Map for *grass/tree* (Fig. 5 (a)) using (8). The Constrained Class Density Map containing $p_{cons}(x_t^I | \omega)$ for each pixel in the input image corresponding to the class *grass/tree* is given in Fig. 5 (b). The densities of false grass regions, e.g., parts of wet floor, water etc. have been significantly reduced. In case, more than one cluster would have been consistent with the class *grass/tree*, the constrained map could be easily computed using all the consistent cluster maps in (8). Similar constrained maps were also obtained for the remaining classes in the recognition vocabulary.

To display the results of the probabilistic classification of the image regions in Fig. 3 (a) using the observation-constrained generative approach, we have used the MAP paradigm. Each image pixel is classified as belonging to class ω_i, which has highest posterior $P(\omega_i | x_t^I)$. Posteriors are computed using the constrained class conditional densities given in (8) and assuming equal priors for all the classes. The resulting classification is given as binary images in Fig. 7 (b) –

(f) where white pixels represent the pixels belonging to the given class. The overall good dicriminative classification results are indicative of good Constrained Class Density Maps. *Sky, water,* and *grass/tree* regions have been classified almost perfectly except on the edges, because of the problem of *edginess* in the texture measure. Use of improved combinations of the eigenvalues of the second moment matrix to avoid this problem has been left as a future work. In the training data, the features of the sky and water regions are distributed close to each other and classification just based on the training data would be quite misleading. Reasonably good classification results for these regions were obtained using our approach (Fig. 7 (b) and (c)), because the statistics of these two regions in the test image could reinforce their respective class associations, generating good Constrained Density Maps.

It can be noted that the above MAP based classification scheme assigns each pixel in the test image to one of the classes in the recognition vocabulary. But some of the image pixels may not belong to any of the classes in the vocabulary, e.g., in the given test image, pixels pertaing to the red wall and associated steel gate, chairs etc. In such cases, the MAP scheme assigns those pixels to the best possible classes in the vocabulary. Other minor artifacts in the classified regions are due to the above reason. In future, we plan to expand the recognition vocabulary to contain more classes.

In Fig. 7 (e), the concrete floor has been classfied as 'sand' because there is no perceptual difference between the two. This is a good example, which advocates the use of the scene context to discriminate between regions that are almost impossible to disambiguate using purely low level features. Because there was no exclusive class for the 'red wall' in the vocabulary, it has been classfied in the semantically nearest class 'skin'. Similarly, skin and sand regions have been partially misclassified caused by extreme overlap in low-level features.

However, it should be noted that MAP-based classification was used purely for evaluation purposes. The main results of the proposed generative approach are the Constrained Class Density Maps for each class that are to be further used in the iterative scene context generation scheme.

The proposed generative method was applied to another image given in Fig. 8 (a). The Class Density Map corresponding to the class *sky* has been shown in Fig. 8 (b). High density has been assigned to most of the regions of the sky, although the map shows some regions of water as potential candidates of being sky. The Constrained Class Density Map in Fig. 8 (c) shows improvement over the original Class Density Map, as the false regions have shrunk significantly. However, there are still some non-sky regions that remain candidates for class *sky*. A careful investigation of these regions reveals that these are watery regions that are either splashes or reflections in the intermixed region. Based on the color and texture features, used in the present work, it is almost impossible to differentiate these pixels from those belonging to the bright cloudy sky. Thus, if the statistics of the newly observed image also supports the hypothesis obtained from the Class Density Maps, there is little one can do except either enhancing the feature set, or using some kind of high-level contextual cues.

VI. Conclusions

We have proposed and successfully demonstrated the use of an observation-constrained generative approach for the probabilistic classification of image regions. A probabilistic approach towards clustering and classification leads to a useful technique, which is capable of refining the classification results obtained using the generative models. The proposed scheme is robust to the errors in the clustering. A KL Divergence-based fast component selection procedure has been proposed for natural scene images. In the future, we intend to work towards evolving efficient schemes to generate distribution over scene hypothesis using the Constrained Class Density Maps. The proposed probabilistic classification forms the first step of a promising feedback framework to iteratively refine the scene context as well as the region hypothesis.

Acknowledgment

The first author would like to thank Amit Singhal and Zhao Hui Sun from Eastman Kodak Company, Rochester, and Daniel Huber and Goksel Dedeoglu from Carnegie Mellon University, Pittsburgh for their useful suggestions.

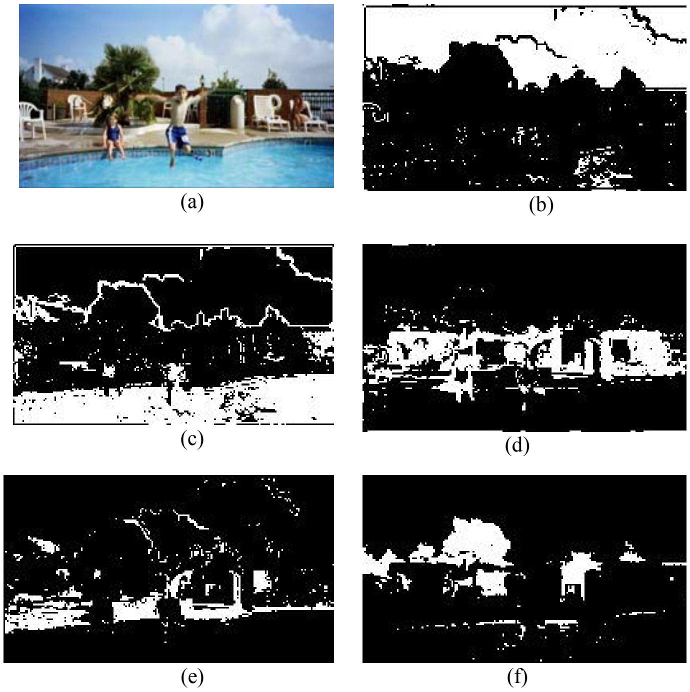
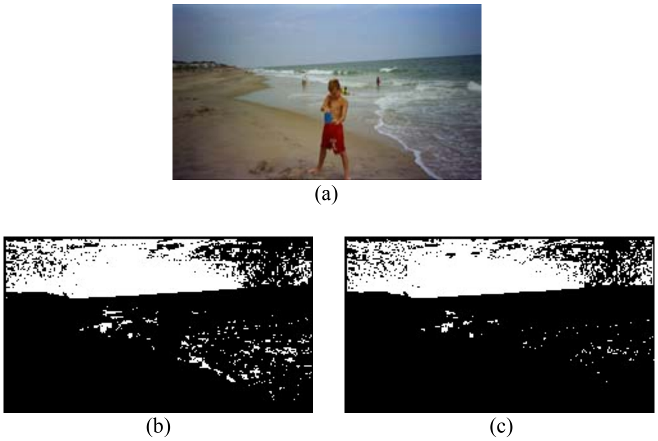


Fig. 7. Discriminative classification of image regions. The above maps are binary maps where a white pixel represents the presence of the corresponding class. (a) Input color image. (b)-(f) show class maps for *sky, water, skin, sand, and grass/tree respectively*



(a)

(b) (c)

Fig. 8. (a) Input color image. (b) Class Density Map corresponding to class *sky*. (c) Corresponding Constrained Class Density Map.

REFERENCES

[1] A. Torralba and P. Sinha, "Statistical Context Priming for Object Detection", *Proc International Conference on Computer Vision*, *ICCV'01,* 2001.

[2] A. Vailaya, "Semantic Classification in Image Databases", *PhD thesis*, Michigan State University, 2000.

[3] T. M. Strat, "Natural Object Recognition," *Springer-Verlag*, 1992.

[4] N. W. Campbell, W. P. J. Mackeown, B. T. Thomas, and T. Troscianko, "The Automatic Classification of Outdoor Images," *International Conference on Engineering Applications of Neural Networks*, pp 339-342, June 1996.

[5] M. Storring, H. Andersen, E. Granum, "Skin colour detection under changing lighting condition", H. Araujo and J. Dias (ed.), *7th Symposium on Intelligent Robotics Systems*, pp. 187-195, 1999.

[6] C. M. Bishop, "Neural Networks for Pattern Recognition", Oxford University Press, 1995.

[7] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm", *J. Royal Statistical Soc.*, Ser. B, vol. 39, no. 1, pp. 1 – 38, 1977.

[8] D. Heckerman and D. Chickering, "A Comparison of Scientific and Engineering Criteria for Bayesian Model Selection," *Technical Report MSR-TR-96-12,* Microsoft Research, June, 1996.

[9] M. H. Hansen, and B. Yu, "Model Selection and the Principle of Minimum Description Length". *JASA*, Vol. 96, No. 454, pp. 746-774, 1998.

[10] C. Farley and A. E. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", *Technical Report,* No. 329, Department of Statistics, University of Washington, 1998.

[11] J. Rissanen, "Modeling by Shortest Data Description", *Automatica*, vol. 14, pp. 465 – 471, 1978.

[12] M. A. Tanner, "Tools for Statistical Inference", 3rd Ed., *Springer-Verlag Inc.*, 1996.

[13] J. Mao and Jain A. K., "Texture Classification and Segmentation using Multiresolution Simultaneous Autoregressive Models", *Pattern Recognition*, vol. 25, no. 2, pp. 173-188, 1992.

[14] B. S. Manjunath and W. Y. Ma, "Texture Features for Browsing and Retrieval of Image Data", *IEEE Trans Pattern Analysis Machine Intelligence*, vol. 18, No. 8, pp. 837-842, 1996.

[15] C. Carson, S. Belongie, H. Grennspan, and J. Malik, "Region-Based Image Querying", *CVPR'97, Workshop on Content-Based Access of Image and Video Libraries*, 1997.

[16] N. Sochen, R. Kimmel and R. Malladi, "A General Framework for Low Level Vision", *IEEE Trans on Image Processing*, Vol. 7, No. 3, pp. 310-318, 1998.