Low-level Image Segmentation Based Scene Classification

Emre Akbas and Narendra Ahuja
Beckman Institute, University of Illinois at Urbana-Champaign
{eakbas2,n-ahuja}@illinois.edu

Abstract

This paper is aimed at evaluating the semantic information content of multiscale, low-level image segmentation. As a method of doing this, we use selected features of segmentation for semantic classification of real images. To estimate the relative measure of the information content of our features, we compare the results of classifications we obtain using them with those obtained by others using the commonly used patch/grid based features. To classify an image using segmentation based features, we model the image in terms of a probability density function, a Gaussian mixture model (GMM) to be specific, of its region features. This GMM is fit to the image by adapting a universal GMM which is estimated so it fits all images. Adaptation is done using a maximum-aposteriori criterion. We use kernelized versions of Bhattacharyya distance to measure the similarity between two GMMs and support vector machines to perform classification. We outperform previously reported results on a publicly available scene classification dataset. These results suggest further experimentation in evaluating the promise of low level segmentation in image classification.

1. Introduction

This paper is primarily aimed at evaluating the semantic information content of low-level segmentation of images. Specifically, we wish to evaluate the power of features directly derived from the segmentation to model semantics of image classes, versus other features that are obtained by other operators. We do this by comparing their classification performances for datasets previously classified into different semantic classes. Further, automatic prediction of the semantic class of a given scene is an important problem in its own right. It could potentially be utilized in applications such as web-scale image search and retrieval, and it could also help in other vision problems, for example in disambiguating the context for object recognition/detection. It remains to be a challenging problem due to the large variability in the properties and spatial distribution of the objects that constitute a single scene class, lighting conditions, viewpoint and scale changes, etc. Thus, in addition to evaluating the information content of segmentation, this paper therefore also simultaneously presents an alternative approach to the semantic classification problem.

The most commonly used image representations for scene classification are in terms of properties of image patches. Rectangular or circular patches are sampled densely along a regular grid overlaid onto the image or at points detected by an interest point detector. These patches are described in various ways, including in terms of normalized intensity values, SIFT features, color and texture histograms, or filter responses [10][7] [4][13][3]. These descriptions then serve as the basis for class representations. They are either directly fed into discriminative algorithms [10] such as support vector machines (SVM), or first, generatively modeled by bag-of-words models [7][3], probabilistic graphical models [7], or latent topic models [13][3], followed by learning of some discriminative aspects of these generative models using a classifier.

As an alternative to the patch-, grid-, and filter-based representations above, in this paper we use features derived from a low-level segmentation of the image. We use the multiscale segmentation algorithm given in [1] which is designed to detect image regions regardless of their shape and size, spatial distribution, and contrast. The algorithm organizes all detected regions hierarchically into a tree data structure where the root node corresponds to the whole image. Nodes closer to the root correspond to larger regions, while their children nodes capture embedded details. Our representation consists of intrinsic properties of the image regions (capturing region geometry and its photometric appearance), as well as properties of their mutual embedding properties which are captured in the tree. Together the two sets of properties constitute our feature space whose capabilities we wish to evaluate via semantic scene classification.

Overview of Our Approach Our approach consists of the following major steps. We first obtain parametric models of the aforementioned two sets of proper-

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.884

ties of image segmentation, to represent them concisely. We achieve this by fitting a Gaussian mixture model (GMM) to the region properties observed within an image. This avoids the need for choosing a specific vocabulary per image and selecting the number of histogram bins or sizes associated with the widely used bag-of-words model or histogramming methods. Another advantage is that by utilizing kernel functions which measure the similarity between GMMs we can let the SVMs directly exploit the generative models in a discriminative setting.

However, the usual problem with such estimation of high-dimensional probability density functions (pdf), in this case a GMM, is the scarcity of the training samples. One approach to overcoming this problem is to first capture the gross distributional characteristics of the entire set of samples, e.g., via a universal GMM. This approach is popular in speech processing [8][12], and is also now becoming popular in computer vision [11][14]. We adapt the universal GMM to each image using a maximum-aposteriori (MAP) criterion. The goal of this adaptation is to maximize the posterior probability of the parameters of the image-generative model (GMM) given the universal GMM and the new data, *i.e.* regions of the image to be modeled. This is accomplished using an expectation-maximization (EM) procedure [8].

Once each image is modeled by a GMM, we use SVMs for classification. For SVMs to work on these GMMs, we need a kernel function which measures the similarity between two GMMs. We use kernelized versions of the Bhattacharyya distance [2] for this purpose.

We begin the description of our approach by first presenting in Section 2 the low-level segmentation based image representation we use, and the classification algorithm. Then in Section 3 we present experimental results. We conclude the paper with a discussion of the results in Section 4.

2. Models and Algorithms

2.1 Image Representation

We represent an image by the list of its regions obtained by a low-level multiscale segmentation algorithm [1]. Each region is described by the following 20 features: (1) area, (2) mean intensity, (3) standard deviation of the intensity, (4) (perimeter)^2 / area, (5) outer-ring area, *i.e.* the area of the region except its children, (6) orientation, (7) eccentricity, (8) solidity, *i.e.* the proportion of the pixels in the convex hull that are also in the region, (9-12) the first four central moments, (13) mean contrast of the boundary, (14) standard deviation of the boundary contrast, (15-16) x-y coordinates of the center of mass expressed in the image coordinate system, (17) perimeter, (18) extent, *i.e.* the ratio of pixels in the region to pixels in the total bounding box, (19) major axis length, (20) minor axis length. As mentioned earlier, these features capture different aspects of image segmentation, including intrinsic geometric (1,4,6-12, 15-20) and photometric properties (2,3) of the regions, their relative properties (13, 14), and a topological property (5). More diversity in the selection of these features is possible, and will be a part of our future work which will be guided by the results obtained in this paper.

Images as adapted GMMs We want to model an image by a GMM of its region properties. However, as mentioned above, robustly fitting a GMM to a small number of regions (some images have only 30-40 regions) is a problem. To overcome this, we employ the MAP adaptation (or Bayesian adaptation). A previously trained *universal GMM* is used as a prior mixture and adapted to the image that we want to model. Then, the image is represented by this adapted GMM. For this purpose, we train a universal GMM using all the regions of all training images by using EM. We denote this universal GMM by its parameters $\Theta^u = \{c_i^u, \mu_i^u, \Sigma_i^u\}_{i=1}^{N^u}$, where c_i^u is the mixture coefficient, μ_i^u is the mean, and Σ_i^u is the covariance matrix of the i^{th} Gaussian. N is the number of Gaussian components in the mixture.

MAP adaptation Once the universal GMM is trained, we adapt it to the regions of each image that we want to model. Let this image be I and let it have R regions: $I = \{r_i | i = 1, 2, \ldots, R\}$. Recall that r_i is a 20-dimensional vector. For completeness, we provide the equations for the MAP adaption here (see [8] or [11] for details). The adaptation is achieved by an EM procedure.

In the E-step, occupancy probabilities, *i.e.* the probability that an observed region r is generated by the i^{th} Gaussian component is estimated:

$$w_i(r) = \frac{c_i p_i(r|\Theta)}{\sum_{j=1}^{N} c_j p_j(r|\Theta)} \tag{1}$$

where

$$p_i(r|\Theta) = \frac{exp\{-\frac{1}{2}(r-\mu_i)'\Sigma_i^{-1}(r-\mu_i)\}}{(2\pi)^{(D/2)}\sqrt{|\Sigma_i|}} \tag{2}$$

In the M-step, the parameters of the GMM are re-estimated (adapted):

$$\hat{c}_i = \frac{\sum_{j=1}^{R} w_i(r_j) + \tau}{R + N \cdot \tau} \tag{3}$$

$$\hat{\mu}_i = \frac{\sum_{j=1}^{R} w_i(r_j) r_j + \tau \mu_i^u}{\sum_{j=1}^{R} w_i(r_j) + \tau} \tag{4}$$

$$\hat{\Sigma}_i = \frac{\sum_{j=1}^{R} w_i(r_j) r_j r_j' + \tau(\Sigma_i^u + \mu_i^u \mu_i'^u)}{\sum_{j=1}^{R} w_i(r_j) + \tau} - \hat{\mu}_i \hat{\mu}_i' \tag{5}$$

version	# train per class	[10]	[4],[3]	Our best	Our average $(N = 1)$	Our average $(N = 75)$
gray	100	83.70%	-	$\mathbf{86.61}\%$ $(N = 1)$	$85.32 \pm 0.71\%$	$84.87 \pm 0.92\%$
gray	50%	-	84.39% [4]	$\mathbf{87.88}\%$ $(N = 1)$	$86.81 \pm 0.82\%$	$85.95 \pm 0.68\%$
color	50%	-	86.65% [4], 87.80% [3]	$\mathbf{88.31}\%$ $(N = 1)$	$87.19 \pm 0.97\%$	$86.46 \pm 0.82\%$

Table 1. Comparison of classification performances of our method and others.

Here the parameter τ is called *the relevance factor* and it regulates the balance between the universal GMM and the new data, *i.e.* the image that we want to model. In our experiments, we choose the value of τ using cross-validation.

2.2 Classification

We use SVMs for classification. As each image is represented by a GMM (adapted from the universal GMM), we need a similarity function between two given GMMs. For the special case of $N = 1$, we use the Bhattacharyya kernel [14] which has the following closed form solution for two Gaussians p and q:

$$K(p,q) = |\Sigma|^{1/2}|\Sigma_p|^{-1/4}|\Sigma_q|^{-1/4} \\ \exp\Big(- \tfrac{1}{4}\mu_p'\Sigma_p^{-1}\mu_p - \tfrac{1}{4}\mu_q'\Sigma_q^{-1}\mu_q + \tfrac{1}{2}\mu'\Sigma\mu\Big) \tag{6}$$

where $\mu = \frac{1}{2}(\Sigma_p^{-1}\mu_p + \Sigma_q^{-1})^{-1}\mu_q)$ and $\Sigma = (\frac{1}{2}\Sigma_p^{-1} + \frac{1}{2}\Sigma_q^{-1})^{-1}$.

For the case of $N > 1$, there is no exact solution. For this, various approximations have been proposed [14] in the literature. We use the following approximation from [5] which uses the fact that there is one-to-one correspondence between the Gaussians of the universal model and the adapted model. For two GMMs p and q:

$$K(p,q) \approx \textstyle\sum_{i=1}^{N} \Big\{ \Big[\frac{1}{2}\Big(\frac{\Sigma_i^p + \Sigma_i^u}{2}\Big)^{-\frac{1}{2}}(\mu_i^p - \mu_i^u)\Big]^T \\ \Big[\frac{1}{2}\Big(\frac{\Sigma_i^q + \Sigma_i^u}{2}\Big)^{-\frac{1}{2}}(\mu_i^q - \mu_i^u)\Big]\Big\} + \\ \textstyle\sum_{i=1}^{N} tr\Big[\Big(\frac{\Sigma_i^p + \Sigma_i^u}{2}\Big)^{\frac{1}{2}}(\Sigma_i^p)^{-\frac{1}{2}}\Big(\frac{\Sigma_i^q + \Sigma_i^u}{2}\Big)^{\frac{1}{2}}(\Sigma_i^q)^{-\frac{1}{2}}\Big] \tag{7}$$

Note that the second term on the right hand side of the equation can be written as an inner product of two vectors – since the covariance matrices are diagonal – which makes its implementation easy and its evaluation efficient.

3. Experiments

For experimental validation, we used the publicly available dataset of Oliva and Torralba[1] [10]. This dataset contains 2688 color images organized in 8 classes: coast, forest, mountain, open country, highway, inside city, tall building, and street. Each class has a different number of images, ranging between 292 to 410.

We compare our results with two previous methods: one that uses gist features and SVM [10] and another that uses SIFT features and a hybrid generative/discriminative method [4][3]. In [10], the dataset is split into training and testing subsets by randomly choosing 100 images for training from each class, and using the rest for testing. Although the images are in color, the authors use the grayscale versions since gist features cannot utilize color information. In [4][3], the authors split the dataset into two by randomly choosing half of the images per class for training, and they use the rest for testing. Results on both grayscale and color versions of the datasets are reported. In all experiments of [10][4][3], the classification accuracy is reported as the mean of the diagonal of the confusion matrix. However, the authors do not mention whether these reported numbers are the best results they get over different random splits of the dataset, or they are the average results over a number of trials. We present our best results as well as the average results we get over 10 random splits.

For each random split of the dataset, we ran our GMM system for four different choices of the number of components N: 1, 50, 75, and 100. The relevance factor, τ was fixed to 50, 5, 2, and 1, respectively. These values were found in our preliminary experiments by 5-fold cross-validation on the training sets.

When there is only a single component in our GMM ($N = 1$), *i.e.* the model is a single multidimensional Gaussian, we used a full-covariance matrix. When $N > 1$, we used diagonal-covariance matrices. There are two reasons for using diagonal-covariance matrices, both motivated by computational considerations: 1) It is easier to avoid singularities (in the covariance matrices) during the GMM training and/or adaptation, 2) It is much more efficient in terms of time and memory.

The classification performances of the previous methods and our method are given in Table 1. We outperform the previously reported results in all cases. Interestingly, in almost all experiments, the single full-covariance Gaussian model ($N = 1$) outperformed the GMM with multiple components. Very rarely was the GMM better than the single full-covariance model. We believe that this situation might be due to three reasons: 1) Since GMM ($N > 1$) uses only diagonal-covariance

[1] http://people.csail.mit.edu/torralba/code/spatialenvelope/

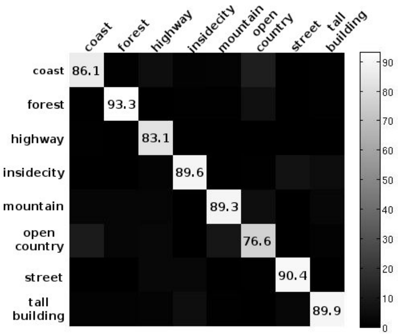

Figure 1. Class confusion matrix of our classifier for 8 classes.

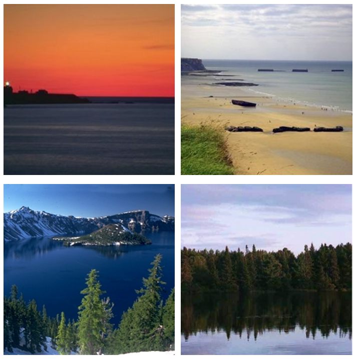

Figure 2. Misclassified image examples. Images in the upper row belongs to the "coast" class but they were labeled as "opencountry". And, images in the lower row belongs to the "opencountry" class but they were labeled as "coast".

matrices, it only takes into account the variances of the features, and this will tend to increase the required number of components, to compensate for the covariances of features. On the other hand, all the covariance information is captured in the single Gaussian case. 2) There is no exact similarity measure between two GMMs. The approximation we used might be degrading the classification performance. 3) Although we have not tested it statistically, the region properties of a given image might be truly distributed as a single multidimensional Gaussian. Other researches also reported that single, full-covariance Gaussian models give consistently good results, and are sometimes better than a GMM [6][9].

Figure 1 shows a typical confusion matrix for the "color, 50%" version of the dataset. The most confused two classes are "open country" and ”coast", which is also the case in previous work [10][4][3]. We give a couple of misclassified images from these classes in Figure 2 as an example of how challenging the dataset is.

4. Conclusions and Discussions

Experimental results suggest that the use of multi-scale, low-level image representation is promising for scene classification. In this preliminary investigation, we have made use of only a limited diversity of features available in the segmentation as is clear from Sec. 2.1 (e.g., we use only one topological feature). In future work, we plan to evaluate the information content of region based features more methodically, using greater diversity of segmentation features, for scene classification and other similar semantic tasks, and on other datasets. We also plan to use data-adaptive recognition and classification methods so the results of the evaluation are more representative of the power of the features instead of being a consequence of differences in the recognition methods used.

References

[1] E. Akbas and N. Ahuja. From Ramp Discontinuities to Segmentation Tree. In *Asian Conference on Computer Vision (ACCV'09)*, Xi'an, China, 2009. Springer.

[2] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99 – 109, 1943.

[3] A. Bosch, X. Muñoz, and A. Zisserman. Scene classification using a hybrid generative/discriminative approach. *IEEE TPAMI*, 30(4):712–27, 2008.

[4] A. Bosch, A. Zisserman, and X. Muñoz. Scene Classification Via pLSA. In *ECCV 2006*, LNCS, pages 517–530, 2006.

[5] Y. Chang Huai, L. Kong Aik, and L. Haizhou. A GMM supervector Kernel with the Bhattacharyya distance for SVM based speaker recognition. In *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, 2009.

[6] J. D. H. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving ”bag-of-keypoints" image categorisation: Generative Models and PDF-Kernels. In *Technical Report*, University of Southampton, 2005.

[7] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *CVPR'2005*, volume 2, pages 524–531. IEEE, 2005.

[8] J.-L. Gauvain and L. Chin-Hui. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[9] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[10] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV*, 42(3):145–175, May 2001.

[11] F. Perronnin. Universal and adapted vocabularies for generic visual categorization. *IEEE TPAMI*, 30(7):1243–56, 2008.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models,. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[13] J. Vogel and B. Schiele. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. *IJCV*, 72(2):133–157, April 2007.

[14] L. Yan and F. Perronnin. A similarity measure between unordered vector sets with application to image categorization. In *IEEE CVPR'2008*, pages 1–8.