A Two Level Approach for Scene Recognition

Le Lu
Computer Science Department
Johns Hopkins University
Baltimore, MD 21218

Kentaro Toyama
Microsoft Research
One Microsoft Way
Redmond, WA 98052

Gregory D. Hager
Computer Science Department
Johns Hopkins University
Baltimore, MD 21218

Abstract

Classifying pictures into one of several semantic categories is a classical image understanding problem. In this paper, we present a stratified approach to both binary (outdoor-indoor) and multiple category of scene classification. We first learn mixture models for 20 basic classes of local image content based on color and texture information. Once trained, these models are applied to a test image, and produce 20 probability density response maps (PDRM) indicating the likelihood that each image region was produced by each class. We then extract some very simple features from those PDRMs, and use them to train a bagged LDA classifier for 10 scene categories. For this process, no explicit region segmentation or spatial context model are computed.

To test this classification system, we created a labeled database of 1500 photos taken under very different environment and lighting conditions, using different cameras, and from 43 persons over 5 years. The classification rate of outdoor-indoor classification is 93.8%*, and the classification rate for 10 scene categories is* 90.1%*. As a byproduct, local image patches can be contextually labeled into the 20 basic material classes by using Loopy Belief Propagation [33] as an anisotropic filter on PDRMs, producing an image-level segmentation if desired.*

1 Introduction

Classifying pictures into semantic types of scenes [24, 26, 22] is a classical image understanding problem which requires the effective interaction of high level semantic information and low level image observations. Our goal is to build a very practical prototype for scene classification of typical consumer photos, along the lines of the Kodak system [22]. Thus, we are interested in systems that are accurate, efficient, and which can work with a wide range of photos and photographic quality.

Given the extremely large within-category variations in typical photographs, it is usually simpler and thus easier to break the problem of scene classification into a two-step process. In this paper, we first train local, image patch based color-texture Gaussian Mixture models (GMM) to detect each of 20 materials in a local image patch. These models are used to scan an image and generate 20 local responses for each pixel. Each response map, called a Probability Density Response Map (PDRM), can be taken as a real-valued image indicating the relative likelihood of each material at each image location. We then compute moments from the response maps and form a feature vector for each photo. By employing the random subspace method [12, 28] and bootstrapping [31], we obtain a set of LDA scene classifiers over these feature vectors. These classification results are combined into the final decision through bagging [2]. After learning the local and global models, a typical 1200×800 image can be classified in less than 1 second with our unoptimized Matlab implementation. Therefore there is a potential to develop a real-time scene classifier upon our approach. A complete diagram of our approach is shown in Figure 1.

There are several related efforts in this area. Luo et al. [19, 22] propose a bottom-up approach to first find and label well-segmented image regions, such as water, beach, sky, and then to learn the spatial contextual model among regions. A Bayesian network codes these relational dependencies. By comparison, we do not perform an explicit spatial segmentation, and we use relatively simple (LDA-based) classification methods. Perona et al. [8, 30] present a constellation model of clustered feature components for object recognition. Their method works well for detecting single objects, but strongly depends on the performance and reliability of the interest detector [13]. In the case of scene classification, we need to model more than one class of material, where classes are non-structural and do not have significant features (such as foliage, rock and et al.) [13]. This motivates our use of a GMM on the feature space. In order to maintain good stability, we estimate the GMM in a linear subspace computed by LDA. These density models are quite flexible and can be used to model a wide variety of image patterns with a good compromise between discrimination and smoothness.

Kumar et al. [14, 15] propose the use of Markov random field (MRF)-based spatial contextual models to detect man-made buildings in a natural landscape. They build a multi-scale color and textual descriptor to capture the local dependence among building and non-building image blocks and use MRF to model the prior of block labels. In our work, we have found that simple local labeling suffices to generate good classification results; indeed regularization using loopy belief propagation method [33] yields no significant improvement in performance. Thus, we claim that there is no need to segment image regions explicitly for scene classification as other authors have done [22, 19, 15].

Linear discriminant analysis (LDA) is an optimization method to compute linear combinations of features that have more power to separate different classes. For texture modeling, Zhu et al [35] pursue features to find the marginal distributions which are also the linear combinations of the basic filter banks, but they use a much more complex method (Monte Carlo Markov Chain) to stochastically search the space of linear coefficients. In our case, the goal is not to build a generative model for photos belonging to different scenes, but simply to discriminate among them. We show a simple method such as LDA, if designed properly, can be very effective and efficient to build a useful classifier for complex scenes.

We organize the rest of the paper as follows. In section 2, we present the local image-level processing used to create PDRMs. In section 3, we describe how PDRMs are processed to perform scene classification. Experimental results and analysis on the performance of patch based material detector and image based scene classification on a database of 1500 personal photos taken by 43 users using traditional or digital cameras over the last 5 years are given in section 4. Finally we summarize the paper and discuss the future work in section 5.

2 Local Image-Level Processing

The role of image-level processing is to roughly classify local image content at each location in the image. The general approach is to compute feature vectors of both color and texture, and then develop classifiers for these features. In our current implementation, we have chosen to perform supervised feature classification. Although arguably less practical than corresponding unsupervised methods, supervised classification permits us to control the structure of the representations built at this level, and thereby to better understand the relationship between low-level representations and overall system performance.

In this step, we compute 20 data driven probabilistic density models to describe the color-texture properties of image patches of 20 predefined materials[1]. These 20 categories are: building, blue sky, bush, other (mostly trained with human clothes), cloudy sky, dirt, mammal, pavement, pebble, rock, sand, skin, tree, water, shining sky, grass, snow, carpet, wall and furniture.

To prepare the training data, we manually crop image regions for each material in our database, and randomly draw dozens of 25 by 25 pixel patches from each rectangle. Altogether, we have 2000 image patches for each material. Some examples of the cropped images and sampled image patches are shown in Figure 2. For simplicity, we do not precisely follow the material boundaries in the photos while cropping. Some outlier features are thus included in the training patches. Fortunately these outliers are smoothed nicely by learning continuous mixture density models.

Multi-scale image representation and automatic scale selection problem has been a topic of intense discussion over the last decade [17, 20, 13, 6, 14]. In general, the approach of most authors has been to first normalize images with respect to the estimated scale of local image regions before learning. However it is not a trivial problem to reliably recover the local image scales for a collection of 1500 family photos. We instead choose to train the GMM using the raw image patches extracted directly from the original pictures. For the labeled image patches with closer and coarser views, their complex color-texture distributions can will be approximated by a multi-modal Gaussian mixture model during clustering.

2.1 Color-Texture Descriptor for Image Patches

Our first problem is to extract a good color-texture descriptor which effectively allows us to distinguish the appearance of different materials. In the domain of color, experimental evaluation of several color models has not indicated significant performance differences among color representations. As a result, we simply represent the color of an image patch as the mean color in RGB space.

There are also several methods to extract texture feature vectors for image patches. Here we consider two: filter banks, and the Haralick texture descriptor. Filter banks have been widely used for 2 and 3 dimensional texture recognition. [16, 5, 27]. We apply the **Leung-Malik (LM) filter bank** [16] which consists of 48 isotropic and anisotropic filters with 6 directions, 3 scales and 2 phases. Thus, each patch is represented by a 48 component feature vector.

The Haralick texture descriptor [10] is designed for image classification and has been adopted in the area of image retrieval [1]. Haralick texture measurements are derived from the Gray Level Co-occurrence Matrix (GLCM). GLCM is also called the **Grey Tone Spatial Dependency Matrix** which is a tabulation of how often different combinations of pixel brightness values (grey levels) occur in an image region. GLCM texture considers the relation between two pixels at a time, called the reference and the neighbor pixel. Their spatial relation can be decided by two fac-

[1] The vocabulary of materials to be detected is designed by considering their popularity in the usual family photos. This definition is, of course, not unique or optimized.

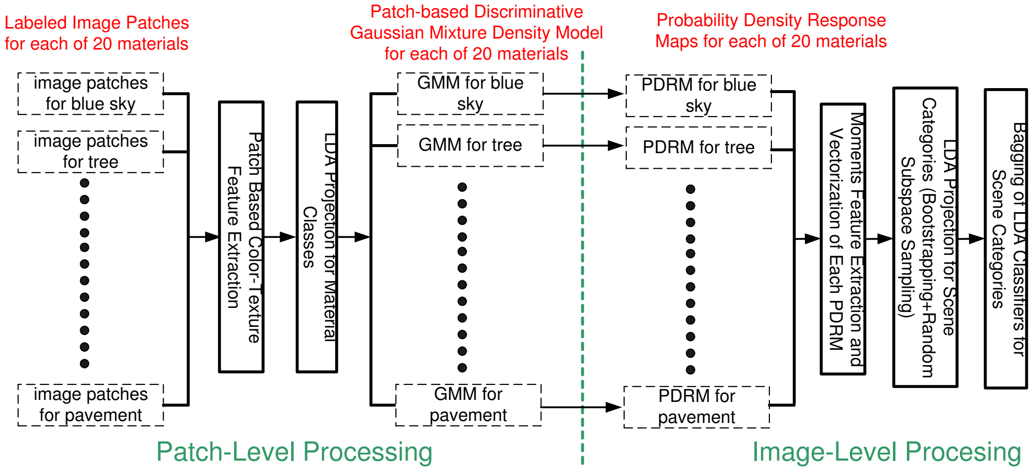

Labeled Image Patches for each of 20 materials
image patches for blue sky
image patches for tree
image patches for pavement
Patch Based Color-Texture Feature Extraction
LDA Projection for Material Classes
Patch-based Discriminative Gaussian Mixture Density Model for each of 20 materials
GMM for blue sky
GMM for tree
GMM for pavement
Probability Density Response Maps for each of 20 materials
PDRM for blue sky
PDRM for tree
PDRM for pavement
Moments Feature Extraction and Vectorization of Each PDRM
LDA Projection for Scene Categories (Bootstrapping+Random Subspace Sampling)
Bagging of LDA Classifiers for Scene Categories
Patch-Level Processing
Image-Level Procesing

Figure 1: *The diagram of our two level approach for scene recognition. The dashed line boxes are the input data or output learned models; the solid line boxes represent the functions of our algorithm.*

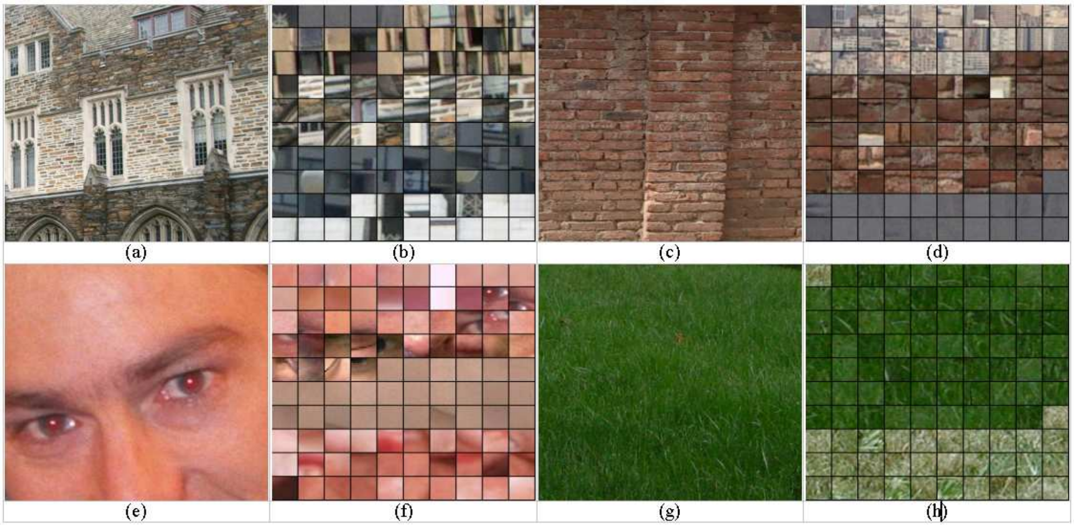


Figure 2: *(a, c, e, g) Examples of cropped subimages of building, building under closer view, human skin, and grass respectively. (b, d, f, h) Examples of image patches of these materials including local patches sampled from the above subimages. Each local image patch is 25 by 25 pixels.*

tors, the orientation and offset. Given any image patch, we search all the pixel pairs satisfying a certain spatial relation and record their second order gray level distributions with a 2 dimensional histogram indexed by their brightness values[2]. Haralick also designed 14 different texture features [10] based on the GLCM. We selected 5 texture features including dissimilarity, Angular Second Moment (ASM), mean, standard deviation (STD) and correction. Definitions for these can be found in Appendix A.

There is no general argument that the filter bank features or Haralick feature is a better texture descriptor. We evaluate their texture discrimination performances experimentally in section 4 and find Haralick features generally perform better.

2.2 Discriminative Mixture Density Models for 20 Materials

The color and texture features for 2000 image patches form, in principle, an empirical model for each material. However, classifying new patches against the raw features would require the solution to a high-dimensional nearest-neighbor problem, and the result would be sensitive to noise and outliers. Instead, we compute a continuous membership function using a Gaussian mixture model.

Although we have 2000 training samples, our feature vectors have 40 dimensions, so the training set is still too sparse to learn a good mixture model without dimensional reduction. Because one of our purposes is to maximize the

[2] The reference and neighbor pixel intensities normally need to be quantized into 16 or less levels instead of 256 which results in not too sparse GLCM.

discrimination among different materials, Linear Discriminant Analysis (LDA) [31] was chosen to project the data into a subspace where each class is well separated. The LDA computation is reviewed in appendix B.

When each class has a Gaussian density with a common covariance matrix, LDA is the optimal transform to separate data from different classes. Unfortunately the material color-texture distributions all have multiple modes because the training image patches are sampled from a large variety of photos. Therefore we have two options: employ LDA to discriminate among 20 material classes; or use LDA to separate all the modes of materials. Although the latter seems closer to the model for which LDA was designed, we found its material classification rate is worse because the optimal separation among the multiple modes within the same material class is irrelevant. Therefore we choose the former.

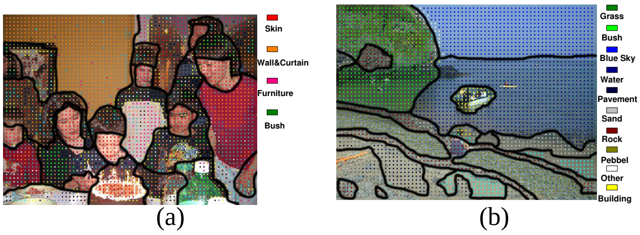
The LDA computation provides a projection of the original feature space into a lower-dimensional feature space $\mathcal{Z}$. We assume that the color-texture features of each material class is described by a finite mixture distribution on $\mathcal{Z}$ of the form

$$P(z|c) = \sum_{k=1}^{g^c} \pi_k^c \mathcal{G}(z; \mu_k^c, \Sigma_k^c), \quad c = 1, 2, ..., 20 \qquad (1)$$

where the π_k^c are the mixing proportions ($\sum_{k=1}^{g_c} \pi_k^c = 1$) and $\mathcal{G}(z; \mu_k^c, \Sigma_k^c)$ is a multivariate Gaussian function depending on a parameter vector θ_k^c. The number of mixtures g_c and the model parameters $\{\pi_k^c, \theta_k^c\}$ for each material class c are initialized by spectral clustering [21] and learned in an iterative Expectation-Maximization manner [31, 7] where g_c ranged from 4 to 8 depending on the material class. As a summary, discriminative Gaussian mixture models are obtained by applying LDA across the material classes and learning the GMM within each material class, respectively.

3 Global Image Processing

Once we obtain 20 Gaussian mixture models $\{\pi_k^i, P(z; \theta_k^i), i = 1, 2, ..., 20\}$ for 20 material classes, we can evaluate the membership density values of image patches for each material class. For any given photo, we scan local image patches, extract their color-texture feature vector, normalize each of its components from 0 to 1 [1], project it to the lower dimensional subspace $\mathcal{Z}$ computed by LDA, and finally compute the density value given by equation (1) for all 20 material classes. The result is 20 real-valued grid maps[3] representing membership support for each of the 20 classes. An example is shown in Figure 3. Two examples of the local patch labeling for indoor and outdoor photos are shown in Figure 4.



(a) (b)

Figure 4: *(a) The local patch material labeling results of an indoor photo. (b) The local patch material labeling results of an outdoor photo. Loopy belief propagation is used for enhancement. The colored dots represent the material label and the boundaries are manually overlayed for illustration purpose only.*

Our next goal is to classify the photos into one of ten categories: cityscape, landscape, mountain, beach, snow, other outdoors, portrait, party, still life and other indoor. In order to classify photos, we must still reduce the dimension of the PDRMs to a manageable size. To do this, we compute the zeroth, first, and second order moments of each PDRM. Intuitively, the zeroth moment describes the prevalence of a given material class in an image; the first moment describes where it occurs, and the second moment its spatial ”spread”. The moment features from the 20 PDRMs are combined in a global feature vector Y.

Using the scene category labels of the training photos, we now compute the LDA transform that attempts to separate the training feature vectors of different categories. For the indoor-outdoor recognition, the LDA projected subspace has only one dimension. As a typical pattern classification problem, we can find the optimal decision boundary from the training data and apply it to the other testing data. Finding decision boundaries for 10 scene category recognition is more complex. In practice, it is very difficult to train a GMM classifier because of the data is too sparse over the 10 categories. As a result, we have used both the nearest neighbor and Kmeans [31] classifiers for this decision problem.

We have found that the standard method for creating an LDA classifier works well for indoor-outdoor scene classification, but the classification results for 10 scene categories is not good enough to constitute a practical prototype. To improve the classification rate, we have implemented variations on random subspace generation [12, 28] and bootstrapping [31] to create multiple LDA classifiers. These classifiers are combined using bagging [2]. Recall that LDA is a two step process that first computes the singular value decomposition (SVD) [9] of the within-class scatter matrix $\mathbf{S}_W$, then, after normalization, computes SVD on the between-class scatter matrix $\mathbf{S}_B^{'}$. After the first step, $\mathbf{S}_W$ is divided into the principal subspace $\mathbf{S}_P$ of the nonzero eigenvalues Λ_P and their associated eigenvectors $\mathbf{U}_P$, and the null subspace $\mathbf{S}_N$ with the zero eigenvalues Λ_N and corresponding eigenvectors $\mathbf{U}_N$. In the traditional LDA transform, only $\mathbf{S}_P$ is used for the whitening of $\mathbf{S}_W$ and nor-

[3] The size of the map depends on the original photo size and the patches' spatial sampling intervals.

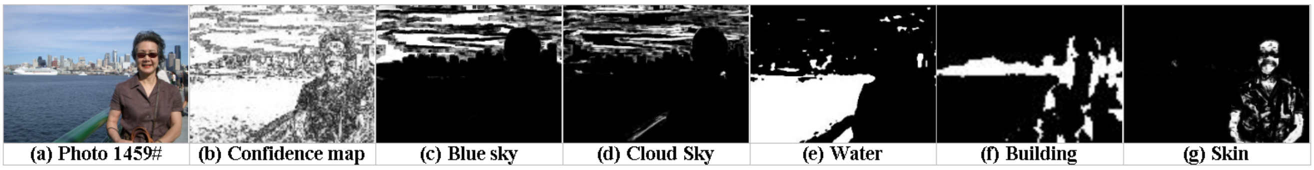

(a) Photo 1459# (b) Confidence map (c) Blue sky (d) Cloud Sky (e) Water (f) Building (g) Skin

Figure 3: *(a) Photo $1459\#$. (b) Its confidence map. (c, d, e, f, g) Its support maps of blue sky, cloud sky, water, building and skin. Only the material classes with the significant membership support are shown.*

malization of $\mathbf{S}_B$ while $\mathbf{S}_N$ is discarded (see equation 10 in Appendix B). Chen et al. [4] have found that the null subspace $\mathbf{S}_N$ satisfying $\mathbf{U}_P^T\mathbf{S}_W\mathbf{U}_P = 0$ also contains important discriminatory information. Here we make use of this observation by uniformly sampling an eigenvector matrix $\mathbf{U}_r$ from $\{\mathbf{U}_P \cup \mathbf{U}_N\}$ and use it in place of $\mathbf{U}$ in the initial LDA projection step. Several projections (including the original LDA projection matrix) are thus created.

In the second step of LDA, the subset $\mathbf{V}_P$ of the full eigenvector matrix $\mathbf{V}$ with the largest eigenvalues, normally replaces $\mathbf{V}$ in equation (10). It is also possible that there is useful discriminative information in the subspace $\{\mathbf{V} - \mathbf{V}_P\}$. Therefore we employ a similar sampling strategy as [28] in the context of PCA by first sampling a small subset of eigenvectors $\mathbf{V}_r$ of $\{\mathbf{V} - \mathbf{V}_P\}$, then replacing $\mathbf{V}$ with the joint subspace $\{\mathbf{V}_P \cup \mathbf{V}_r\}$ in equation 10.

Finally we also perform bootstrapping [31] by sampling subjects of the training set and creating LDA classifiers for these subsets. By the above three random sampling processes, we learn a large set of LDA subspaces and classifiers which we combine using the majority voting (bagging) methods [2]. In Section 4, we show the bagged recognition rates of 20 classifiers from bootstrapping replicates and 20 from random subspace sampling.

4 Experiments

Our photo collection currently consists of 540 indoor and 860 outdoor customer photos. We randomly select half of them as the training data and use other photos as the testing data. We have also intentionally minimized redundancy when collecting photos, i.e., only one photo is selected when there are several similar pictures.

We first address the problem of the image patch based color-texture feature description and classification. Comparison of the recognition rates of 1200 testing image patches for each material class for different color-texture descriptors, different numbers of training patches and different classifiers is provided in Figure 6 (a,b). In particular, we have also benchmarked the LDA+GMM model against a brute-force nearest neighbor classifier. Let x_j and z_j represent an image patch feature vector before and after the LDA projection, respectively. The nearest neighbor classifier computes the class label of a testing patch j as the label of that training patch l such that $\|x_j - x_l\| = \min_i\{\|x_j - x_i\|\}$ where i ranges over the training image patches of all material classes. The GMM classifier simply chooses the maximal class density, i.e. the class c^* such that $P(z_j|c^*) = \max_{c=1,2,...,20}\{P(z_j|c)\}$.

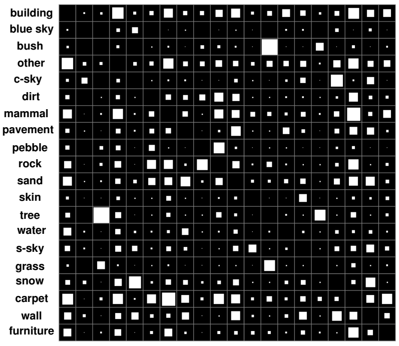


Figure 5: *The pairwise confusion matrix of 20 material classes. The indexing order of the confusion matrix is shown on the left of the matrix. The indexing order is symmetrical.*

Comparing the plots shown in Figure 6, the classifier based on the Maximum Likelihood of GMM density functions outperforms the Nearest Neighbor classifier, thus validating the use of the LDA+GMM method. We also compared the recognition rates of 4 different feature combinations and found that the Haralick texture descriptor combined with the mean color of the image patch yields the best results. Finally, in Figure 6 (b), we see that the LDA+GMM method improves the recognition rate significantly when increasing the training image patch from 500, becoming stable after 2000 patches.

Figure 5 shows the confusion rate using the GMM classifiers learned from 2000 training image patches per class. The size of the white rectangle in each grid is proportional to the pairwise recognition error ratio. The largest and smallest confusion rates are 23.6% and 0.24%, respectively. From Figure 5, we see that pebble, rock and sand classes are well separated which shows that our patch-level learning process achieves a good balance of Haralick texture and color cues by finding differences of the material classes with the similar color. There is significant confusion among grass, bush and tree due to their similar color and texture distribution. For some material classes, such as furniture, carpet, and other, the overall confusion rates are also high.

For global classification, we have found that first order

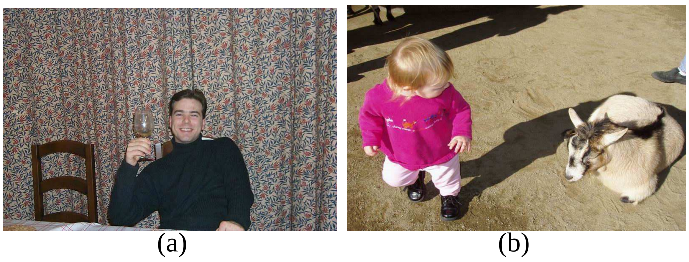

(a) (b)

Figure 7: *(a) An misclassified indoor photo. (b) An misclassified outdoor photo.*

moment features of PRDMs are useful in outdoor scenes, but reduce the recognition rate for indoor scenes. This makes sense since in most outdoor scenes spatial contextual constraints, for instance the sky above grass, are useful cues. This naturally suggests a hierarchical classification scheme (first determine indoor/outdoor followed by categorization), however we have not yet pursued this approach. Thus, we confine ourselves to zeroth order moments for the remainder of this paper.

Our global image moment features after LDA projection are very easy to visualize in the indoor/outdoor case as they become points in a 1-dimensional LDA subspace (6 (c)). In this case, the 1-D indoor-outdoor decision boundary is simply determined by fitting a scaled exponential function to each of the indoor or outdoor histogram distributions and calculating the point of intersection.

We show the recognition results of our method in Figure 6 (d), compared with the direct low-level color or texture based scene recognition methods[4] without LDA learning as the baselines. Our indoor-outdoor recognition rate is 93.8%, which is comparable or slightly better than the Kodak's recently published classification system [22], although our approach is tested on a 40% larger photo database. It is interesting that the bagging algorithm does not significantly improve the recognition performance of for indoor-outdoor classification. The likely explanation is that the individual indoor-outdoor LDA classifiers have nearly achieved the best possible recognition rate. Figure 7 shows 2 examples of misclassified photos. The first photo consists of a person sitting indoors, but in front of a curtain of tree leaves. In the second, the playground is incorrectly classified as ”carpet” not ”dirt”. The appearance of people and animals are irrelevant for indoor-outdoor classification — their associated moment features are assigned with near zero weights.

As shown in Figure 6 (e), feature points of some scene categories are well separated from others and thus easy to be recognized in a certain LDA subspace, while some categories are not. Fortunately, Figure 6 (f) demonstrates that the individual LDA classifiers capture the complimentary discriminative information in different random subspaces. Finally, it results that the combined (nearest neighbor and Kmeans) classifiers both show improved performances of $6 - 10\%$ on average. As a comparison, Boutell et al. [3] achieve less than 80% classification accuracy for 923 images in 5 categories. In their work, model-based graph matching techniques are used to learn the explicit scene configuration consisting of semantic image regions.

5 Conclusions & Discussion

This paper makes three contributions. First, we propose an efficient, yet effective, approach for scene recognition for both indoor-outdoor and multiple photo categories. In practice, this approach can handle the photos' spatial complexity both in the local patch-level and the global image-level successfully. All the training and testing processes are based upon a challenging photo database. Second, we describe a combination of LDA and Gaussian mixture models that achieves a good balance of discrimination and smoothness. Finally, we study the use of moment features of PDRMs as an effective image-level representation for scene classification, and the bagging [2] method to combine the individual scene classifiers obtained by the random subspace algorithm [12]. The bagging method has shown success in our experiments, especially for 10 category scene recognition.

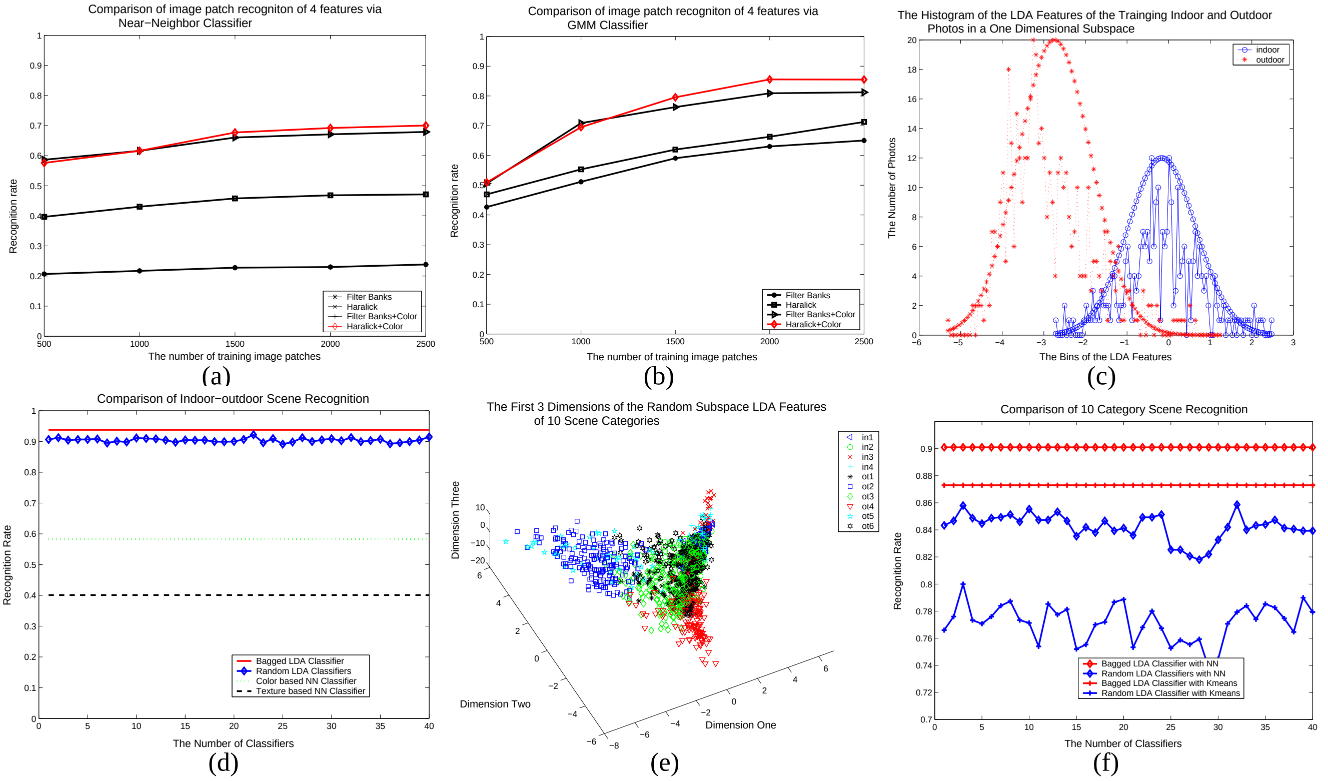
Although we have used supervised methods to create the local image patch classifiers, a practical system would like learn at least some of these classifiers using unsupervised methods. However we believe that the supervised material detectors provide the best scene recognition performance, and as such provide a “benchmark” against which unsupervised methods can be evaluated. In future work, we intend to investigate unsupervised clustering methods for low-level image patch classification. In particular, we plan to apply our unsupervised, iterative LDA-GMM algorithm [18]. We also plan to investigate a hybrid approach where classified images are used as labeled data to compute an initial LDA projection, which is then subsequently refined with new, unlabeled images using iterative LDA-GMM. Finally, because LDA is only optimal when each class has a Gaussian density with a common covariance matrix, the non-parametric discriminant analysis (proposed in [34]) will be tested as a means to generalize our approach to a more comprehensive image database which may contain thousands of various kinds of photos.

Appendices

A The GLCM

Let us denote GLCM a $N \times N$ matrix $P_{i,j}$ where N is the quantized level of pixel intensity and $i, j = 0, 1, ..., N-1$. The diagonal elements $(i = j)$ all represent pixel pairs with

[4] We divide each image as a 9 by 9 grid, and extract the mean color or the DOG (Derivative of Gaussian) filtered texture features within each grid. Each photo is then formulated as a feature vector by combining cues in all grids. A nearest neighbor classifier is later employed for recognition based on the feature vectors' distances of the training and testing photos.



(a) (b) (c) (d) (e) (f)

Figure 6: *(a) Comparison of the image patch based recognition of 4 kinds of features (filter banks feature, Haralick texture feature and their joint features with color) via Nearest-Neighbor Classifier. (b) Comparison of the image patch based recognition of 4 kinds of features via GMM Classifier. (c)The 1D feature histogram distributions of indoor-outdoor photos after LDA projection. (d) The comparison of indoor-outdoor recognition rates of 4 methods. (e) The first 3D feature point distributions of 10 category photos after LDA projection. (f) The comparison of 10 categories recognition rates of 4 methods.*

no grey level difference; while the off-diagonal cells ($i \neq j$) represent pixel pairs with dissimilarity $|i-j|$ increasing linearly away from the diagonal. Therefore we have $dissimilarity = \sum_{i,j=1}^{N-1}(P(i,j) \times |i-j|)$. Furthermore $ASM = \sum_{i,j=1}^{N-1} P(i,j)^2$ measures the uniformity of the distribution of GLCM. $\mu_i = \sum_{i,j=1}^{N-1}(P(i,j) \times i|)$ and $\mu_j = \sum_{i,j=1}^{N-1}(P(i,j) \times j|)$ are the means of the reference pixels or neighbor pixels. Similarly, $\sigma_i = \sqrt{\sum_{i,j=1}^{N-1}(P(i,j) \times (i-\mu_i)^2)}$ and $\sigma_j = \sqrt{\sum_{i,j=1}^{N-1}(P(i,j) \times (j-\mu_j)^2)}$ are the respective standard deviations, and $correlation = \sum_{i,j=1}^{N-1}(P(i,j) \times (i-\mu_i)(j-\mu_j))/(\sigma_i \times \sigma_j)$. If the above means and standard deviations are calculated from symmetrical GLCM, $\mu = \mu_i = \mu_j$ and $\sigma = \sigma_i = \sigma_j$. Finally the output of 5 Haralick features are $\{dissimilarity, ASM, \mu, \sigma, correlation\}$ for each GLCM [5].

B LDA

The following objective function

$$J(\omega) = \frac{\omega^T \mathbf{S}_B \omega}{\omega^T \mathbf{S}_W \omega} \tag{2}$$

is maximized by solving a generalized eigenvector equation

$$\mathbf{S}_B \omega = \lambda \mathbf{S}_W \omega \tag{3}$$

where

$$\mathbf{S}_W = \frac{1}{M} \sum_{i=1}^{C} \sum_{j=1}^{M} z_{ij}(X_j - m_i)(X_j - m_i)^T \tag{4}$$

$$\mathbf{S}_B = \sum_{i=1}^{C} \frac{M_i}{M}(m_i - m)(m_i - m)^T \tag{5}$$

Denote that $\mathbf{S}_B$ and $\mathbf{S}_W$ are respectively named the between-class or within-class scatter matrix, x_j is a feature vector, m_i is the mean of class i and m is the global mean of the data X, $i = 1...C$ is a class number (C is the total number of classes) and the binary membership function

$$z_{ij} = \begin{cases} 1, & if \;\; x_j \in class\ i \\ 0, & otherwise \end{cases} \tag{6}$$

The LDA algorithm firstly perform the singular value de-

[5] Note that we choose the pair of reference and neighbor pixels according to 4 directions (45 degree each) and 1 or 2 pixel offsets. Therefore we have 8 GLCMs for any image patch which results in a 40 component feature vector.

composition (SVD) of $\mathbf{S}_W$

$$\mathbf{S}_W = \mathbf{U}\Lambda\mathbf{U}^T \tag{7}$$

then transform $\mathbf{S}_B$ into

$$\mathbf{S}_B^{'} = \Lambda^{-\frac{1}{2}}\mathbf{U}^T\mathbf{S}_B\mathbf{U}\Lambda^{-\frac{1}{2}} \tag{8}$$

and compute the eigenvectors of

$$\mathbf{S}_B^{'}\mathbf{V} = \mathbf{V}\hat{\Lambda} \tag{9}$$

where $\hat{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{S}_B^{'}$. The optimal feature vectors **Z** are therefore

$$\mathbf{Z} = \mathbf{A}^T\mathbf{X} \tag{10}$$

through the projected transform $\mathbf{A}^T = \mathbf{V}^T\Lambda^{-\frac{1}{2}}\mathbf{U}^T$. For dimension reduction, only the subset of eigenvectors **V** and **U** with large eigenvalues are used in the transform. The dimension of the LDA projected subspace is at most $C-1$.

References

[1] S. Aksoy and R. Haralick, Feature Normalization and Likelihood-Based Similarity Measures for Image Retrieval, *Pattern Recognition Letters*, 22(5):563-582, 2001.

[2] L. Breiman, Bagging Predictors, *Machine Learning*, 24(2):123-140, 1996.

[3] M. Boutell, J. Luo and C. Brown, Learning spatial configuration models using modified Dirichlet priors, *Workshop on Statistical Relational Learning*, 2004.

[4] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, A new LDA-based face recognition system which can solve the small sample size problem, *Journal of Pattern Recognition*, 33(10):1713-1726, 2000.

[5] O. Cula and K. Dana, Compact representation of bidirectional texture functions, *CVPR* I:1041-1047, 2001.

[6] J. De Bonet, P. Viola, A non-parametric multi-scale statistical model for natural images, *NIPS*, 1997.

[7] A. Dempster, N. Laird and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Series B, 39:1-38, 1977.

[8] R. Fergus, P. Perona and A. Zisserman, Object class recognition by unsupervised scale-invariant learning, *CVPR*, 2003.

[9] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1996.

[10] R. Haralick, K. Shanmugam, and I. Dinstein, Texture features for image classification. *IEEE Trans. on System, Man and Cybernatic*, 1973.

[11] X. He, R. Zemel and M. Carreira-Perpiñán, Multiscale Conditional Random Fields for Image Labeling, *CVPR*, 2004.

[12] T. K. Ho, The random subspace method for constructing decision forests, *IEEE Trans. on PAMI*, 20(8):832-844, 1998.

[13] T. Kadir and M. Brady, Scale, saliency and image description, *IJCV*, 45(2):83105, 2001.

[14] S. Kumar and M. Hebert, Man-made structure detection in natural images using a causal multiscale random field, *CVPR*, 1:119-126, 2003.

[15] S. Kumar, A. C. Loui and M. Hebert, An Observation-Constrained Generative Approach for Probabilistic Classification of Image Regions, *Image and Vision Computing*, 21:87-97, 2003.

[16] T. Leung and J. Malik, Representing and recognizing the visual appearance of materials using three-dimensional textons, *IJCV*, 2001.

[17] T. Lindeberg, Principles for automatic scale selection, *Handbook on Computer Vision and Applications*, 2:239–274, Academic Press, Boston, 1999.

[18] Authors, Authors' paper, to appear *NIPS*, 2004.

[19] J. Luo, A. Singhal, S. Etz, and R. Gray, A computational approach to determination of main subject regions in photographic images, *Image Vision Computing*, 22(3):227-241, 2004.

[20] J. Malik, S. Belongie, T. Leung and J. Shi Contour and texture analysis for image segmentation, *Int Journal of Computer Vision*, 43(1):7-27, 2001.

[21] A. Ng, M. Jordan and Y. Weiss, On Spectral Clustering: Analysis and an algorithm, *NIPS*, 2001.

[22] N. Serrano, A. Savakis and J. Luo, Improved scene classification using efficient low-level features and semantic cues, *Pattern Recognition* 37(9):1773-1784, 2004.

[23] A. Singhal, J. Luo and W. Zhu, Probabilistic Spatial Context Models for Scene Content Understanding, *CVPR*, 2003.

[24] M. Szummer and R. W. Picard, Indoor-outdoor image classification, *IEEE Int. Workshop Content-Based Access Image Video Databases*, 1998.

[25] A. Torralba, Contextual priming for object detection, *Int Journal of Computer Vision*, 53(2):169-191, 2003.

[26] A. Vailaya, M. Figueiredo, A. Jain and H.-J. Zhang, Image classification for content-based indexing, *IEEE Trans. Image Processing*, 10(1):117-130, 2001.

[27] M. Varma and A. Zisserman, Classifying images of materials: achieving viewpoint and illumination independence, *ECCV*, 2002.

[28] X. Wang and X. Tang, Random sampling LDA for face recognition, *CVPR*, 2004.

[29] X. Wang and X. Tang, Dual-space linear discriminant analysis for face recognition, *CVPR*, 2004.

[30] M. Weber, M. Welling and P. Perona, unsupervised learning of models for recognition, *ECCV*, 2000.

[31] A. Webb, *Statistical Pattern Recognition*, John Wiley & Sons, 2002.

[32] Y. Wu, Q. Tian, T. Huang, Discriminant-EM algorithm with application to image retrieval, *CVPR*, I:222-227, 2000.

[33] J. Yedidia, W. T. Freeman and Y. Weiss, Understanding belief propagation and its generalizations, *IJCAI*, 2001.

[34] M. Zhu and T. Hastie, Feature extraction for non-parametric discriminant analysis, *JCGS*, 12(1):101-120, 2003.

[35] S.C. Zhu, Y.N. Wu, and D. Mumford, Minimax entropy principle and its applications to texture modeling, *Neural Computation*, 9:1627-1660, 1997.