Describing Visual Scene through EigenMaps

Shizhi Chen, *Student Member, IEEE,* and YingLi Tian, *Senior Member, IEEE*

***Abstract*—We propose a novel approach to describe and recognize visual scene categories. Inspired by the success of Bag of Words approach, we represent a scene image using a collection of EigenMaps, which incorporate both appearance and spatial information for scene analysis. Each EigenMap captures the location likelihood of a visual word through the kernel density estimation method. By collecting EigenMaps of all visual words, our approach can effectively integrate both local features and their global correspondences. Experimental results demonstrate significant performance improvement as compared with the standard Bag of Words approach and the Latent Dirichelet Allocation model, which also utilizes a codebook of visual words, over each type of features including both region features and interest point features. The proposed method achieves the state of the art performance on both the UIUC Sport Scene database and the Natural Scene database.**

***Index Terms*— Bag of Words, EigenMap, Kernel Density Estimation, Scene Classification,**

I. Introduction

GIVEN an image of a complicated scene, can a computer recognize the scene category? The problem of scene classification is very challenging. As shown in Figure 1, even for the same scene category, there are significant variations of background, lighting, scale, rotation and viewpoint *etc*. Nevertheless, the ability to recognize visual scene reliably facilitates a large number of applications including object detection, image retrieval, and video surveillance *etc*.

One of the main challenges is to develop an effective visual scene description. How do we represent a visual scene, such that the scene representation is discriminative enough among different scene categories while robust enough to tolerate the large variations within the same category? Another challenge is that a scene representation should be general enough so that it can easily apply to different types of features.

There have been significant research efforts to develop feature representations in order to incorporate richer form of image understanding. Many researchers explore context information to improve recognition performance [5, 11, 19, 20, 30]. The most common method is to include the surrounding pixels of interest regions [9, 39], since the surrounding pixels or patches usually carry useful information for recognition. Other approaches include attribute learning [12, 13, 22, 23] and hierarchies [10, 17, 28].

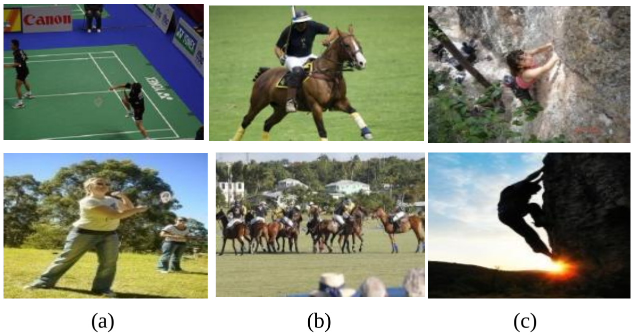


Figure 1: Sample scene images from the UIUC Sport Scene database [25]. (a) Badminton; (b) Polo; (c) Rock-climbing; Note that there are significant variations of background, lighting, scale, rotation and viewpoint for scene recognition.

Several approaches of scene classification have been developed recently [1, 7, 16, 26, 34]. Oliva and Torralba proposed a method using Spatial Envelope to represent the shape of a scene image [29]. They used a set of perceptual dimensions such as openness and naturalness *etc.* as the properties of the Spatial Envelope. However, the authors did not consider local features which are robust to partial occlusion and clustered backgrounds [8].

Recently, the bag of words representation has been successfully applied in computer vision applications, such as image retrieval and visual scene classification, owing to its simplicity and good performance [8, 33, 37, 38, 40]. The bag of words representation can also easily adapt to a wide range of feature types [8].

However, one major limitation using the bag of words model as an image representation is that it only models an image as a collection of local features without considering features' location information in the image. As proven by many researchers, knowing spatial relationship among different objects or object parts can be very important in visual scene classification [4, 24, 32].

In order to incorporate the spatial information into the bag of words representation, a few interesting models have been

Manuscript received September 20, 2011. This work was supported in part by NSF grant IIS-0957016, EFRI-1137172, ARO grant W911NF-09-1-0565, and NOAA CREST Grant NA11SEC4810004.

Shizhi Chen is with the Electrical Engineering Department, The City College, City University of New York, New York, NY, 10031, USA (e-mail: schen21@ccny.cuny.edu).

YingLi Tian is with the Electrical Engineering Department, The City College, City University of New York, New York, NY, 10031, USA (phone: 212-650-7046; fax: 212-650-8249; e-mail: ytian@ccny.cuny.edu). Prior to joining The City College in September 2008, she was with IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598, USA.

proposed. Savarese *et al.* [32] borrowed the idea of color correlograms [21] to develop visual word correlograms. The correlograms capture spatial correlation between all possible pairs of visual words by forming a co-occurrence matrix of visual words as a function of distance. However, the correlograms matrix requires expensive computation cost even after utilizing the integral image techniques [32]. Furthermore, it is not clear how to extend the correlograms feature representation to describe sparsely detected interest points.

Lazebnik *et al.* [24] proposed a Spatial Pyramid Matching model as visual scene descriptor by partitioning an image into successively sub-regions and by calculating the vocabulary histogram over each sub-region. Then, the model concatenates all histograms together with an appropriate weight. However, the Spatial Pyramid Matching model captures relatively weak spatial information by hard-assigning features to each region. It does not consider a coupling effect among the adjacent regions. As the number of pyramid level increases, the resulted feature dimension increases exponentially.

Other researchers introduce a hidden layer, such as object or theme between visual words and a visual scene [14, 25, 31]. These intermediate layers are learned either in an unsupervised or a weakly supervised manner. Each object or theme induces a probability density of visual words, while each scene category learns the distribution of objects or themes. These topic discovery models are usually more flexible and easier to integrate with multiple types of features, including non-visual features such as text information. Li *et al*. [25] integrate both visual words with tag information to automatically classify scene images. One drawback is that the additional intermediate layer usually makes the model more complicated.

In this paper, we propose a new type of feature representation, EigenMap, to describe a visual scene. Different from the previous work [4, 24], which divide a visual scene image to several regions, then form the Bag of Words on each region, we model the location likelihood of each visual word on the whole scene image through the kernel density estimation [2]. The location density map of each visual word is further projected into a very small dimensional space using the Principal Component Analysis (PCA). We define the location density map of a visual word in the principal component space as the EigenMap.

Describing a visual scene with a collection of EigenMaps, the approach not only incorporates the spatial information of visual words in a scene image, but also effectively integrates local appearance features and their global correspondences together. Our experiments demonstrate promising results of visual scene classification on both the UIUC Sport Scene database and the Natural Scene database. As compared with the Bag of Words (BOW) model and the Latent Dirichlet Allocation (LDA) model, which utilizes a codebook of visual words, the proposed EigenMap method significantly improves the performance over both region features and interest point features.

The paper is organized as the following. Section II describes the proposed method in details including feature extraction and the EigenMap Generation for the extracted features. Section III presents the databases and our experimental results. We then conclude our proposed method in section IV.

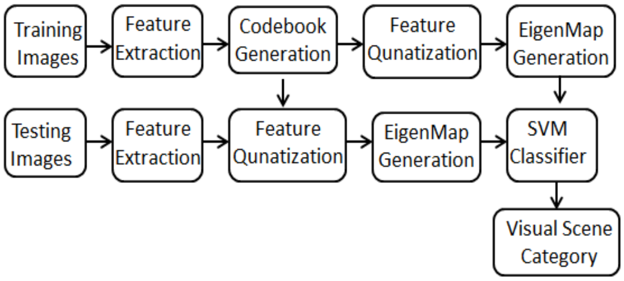


Figure 2: Flowchart of the proposed approach using EigenMap representation in visual scene classification.

II. METHOD

A. Overview

The flowchart in Figure 2 illustrates our overall approach in visual scene classification using EigenMap representation.

We first extract features from images. A codebook is generated from the training image features by using an unsupervised clustering algorithm such as the K-Mean method. The center feature vectors in the codebook are called visual words. Then each feature in both training and testing images is vector-quantized to one of the visual words in the codebook.

We then construct location map for each visual word in a scene image using the kernel density estimation method [2]. The EigenMap of a visual word is then generated by projecting the location map to the principal component space. The concatenation of every visual word's EigenMap in the scene image forms an input feature vector of a SVM (Support Vector Machine) classifier [6]. Finally we can classify an unknown scene image to different scene categories. The proposed EigenMap representation of a scene image not only incorporates spatial information in the appearance features, but also effectively integrates both local features and their global interactions.

B. Feature Extraction

In order to verify the effectiveness of the proposed model, we extract five types of different features, which include both region features and interest point features. In other words, for each type of features, we evaluate the performance improvement of the EigenMap model. In our experiments, we extract three types of region features and two types of interest point features.

B.1. Region Features: Texture, Shape, and Color

Three types of region features are extracted in our experiments: texture, shape, and color. Before generating any region features, we first perform segmentation on images using the algorithm proposed by Felzenszwalb and Huttenlocher [15]. As shown in Figure 3, connected pixels with same color are used to represent one segmented region. At each segmented region, the above three types of region features are extracted.

Texture features are generated by passing the original image with S filter bank [35]. S filter bank is rotationally invariant with 13 isotropic. There are 13 responses for each image. The means and standard deviations of each response are calculated for individual segmented region in the image. In other words, each segmented region has 13 means and standard deviations of the filter responses. These means and standard deviations are combined together as texture features of one segmented region.

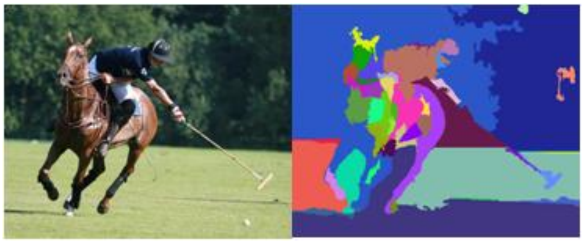


Figure 3: Segmentation example of a Polo scene. Connected pixels with same color belong to the same segment.

A simple type of shape features is extracted in this experiment following the approach in [25]. The size of each segmented region is found by calculating the maximum length of a segmented region in x and y directions. Then the shape feature of each segmented region is formed by combining the size and the number of pixels in each segmented region.

Color features are formed by calculating color histograms of each segmented region over the RGB color space. Each color space is divided into 10 bins. Therefore, the color feature vector of each segmented region has 1000 dimensions.

B.2. *Interest Point Features: Uniform Grids and Harris Corners*

In addition to the region features above, two types of interest point features are evaluated in our experiments: the uniform grids and the Harris corners. In our evaluations, the uniform grid method is used to sample interest points every 10 pixels in x and y directions. The number of interest points generated for a typical image (resolution of 300 by 500) is around 1500.

Unlike the uniform grid method, the Harris corners utilize gradient information to detect more stable interest points in an image [18]. The average number of the Harris corners in one image is approximately 100 in our experiments, which is significantly less than the number of the uniform grid interest points.

The Scale Invariant Feature Transform (SIFT) descriptor [27] is used to describe all interest points regardless of their detection methods. A square patch window with each interest point at its center is extracted. The patch window size is 24 by 24 pixels. 4 by 4 center points are uniformly sampled from the patch window. For each center point, an 8-Bin orientation histograms of gradients within the patch window is constructed. The gradient magnitudes are further weighted by a Gaussian function with the mean corresponding to the center point. Then all histograms of the 16 center points are concatenated together to form an interest point descriptor, which has 128 dimensions.

C. *Codebook Formation and Feature Quantization*

After extracting feature vectors from the training images, the K-Mean clustering algorithm is used to group the feature vectors together based on the Euclidean distance. As a result, a set of center feature vectors are called visual words. The resulting visual words form the codebook vocabulary [8]. The codebook sizes of the texture, shape and color features are 120, 100 and 30 respectively. Both the uniform grid and the Harris corner features have the codebook size of 150.

The features in each image are then vector-quantized to one of visual words in the codebook. The vector quantization process of a feature is to find a visual word in the codebook with the smallest Euclidean distance. Then the feature is represented by the closest visual word in the codebook.

D. *EigenMap Generation*

In order to effectively incorporate spatial information into these visual words and describe their global correspondence within a scene image, we generate an EigenMap for each visual word in the scene image. The flowchart of EigenMap generation for each visual word is shown in Figure 4.

Given an input image and a codebook of visual words generated from the K-Mean clustering algorithm as described in the last section, we first locate a visual word *Vi* in the input image and mark the corresponding positions at the visual word *Vi*'s location map. The location map has fixed size of 50 by 50 pixels. Then we use kernel density estimation [2] to model the location likelihood of the visual word *Vi* in its location map, as illustrated in the examples shown in Figure 5. The kernel we used is the normal distribution with the standard deviation of 2.

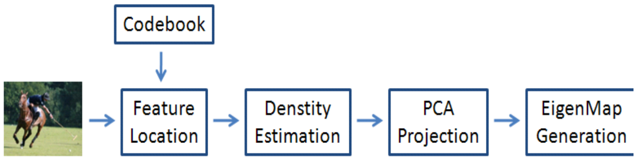
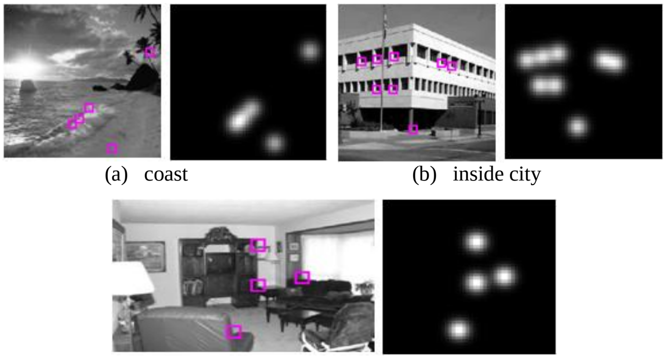


Figure 4: The flowchart of EigenMap Generation for each visual word from the codebook.



(a) coast

(b) inside city

(c) living room

Figure 5: A visual word *Vi*'s locations in an input image and its location likelihood on the corresponding location map using the kernel density estimation analysis.

The next step is to project the constructed location map to a lower dimensional space using the principal component analysis, as shown in Equation (1):

$$\boldsymbol{s} = \boldsymbol{\phi}^T * (\boldsymbol{m} - \mu)\,, \tag{1}$$

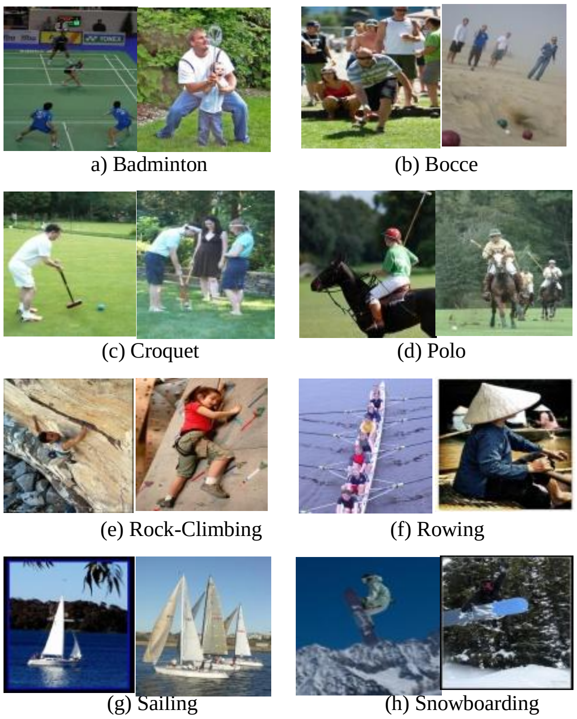
where $\boldsymbol{s}$ is the location map projected in the eigen-space, $\boldsymbol{m}$ is the location map, and μ is the average value of the location map $\boldsymbol{m}$. Each column vector of $\boldsymbol{\phi}$ is an eigenvector of the location maps' covariance matrix obtained from the training images in the order of descending eigenvalues of the covariance matrix. Finally, the EigenMap $\boldsymbol{\eta}$ is constructed as the concatenation of $\boldsymbol{s}$ and μ, as shown in Equation (2):

$$\boldsymbol{\eta} = [\boldsymbol{s};\ \mu]\,. \tag{2}$$

Typical dimension of EigenMap $\boldsymbol{\eta}$ is below 8, which results a very compact representation of a scene image, as compared with previous work [24, 32]. After each visual word's EigenMap $\boldsymbol{\eta}$ of a scene image is constructed, we then concatenate the EigenMaps of all visual words together to represent the scene image. This concatenated feature vector is also an input to the SVM classifier.

E. *Classifier*

We employ the SVM with the RBF kernel as our multi-class classifier [6]. The SVM is to find a set of hyper-planes which separates each pair classes of data with the maximum margin. That is to assign a scene category to an unknown image based on the collections of visual words' EigenMaps.


a) Badminton (b) Bocce

(c) Croquet (d) Polo

(e) Rock-Climbing (f) Rowing

(g) Sailing (h) Snowboarding

Figure 6: Sample images of the 8 scene categories from the UIUC Sport Scene Database [25].

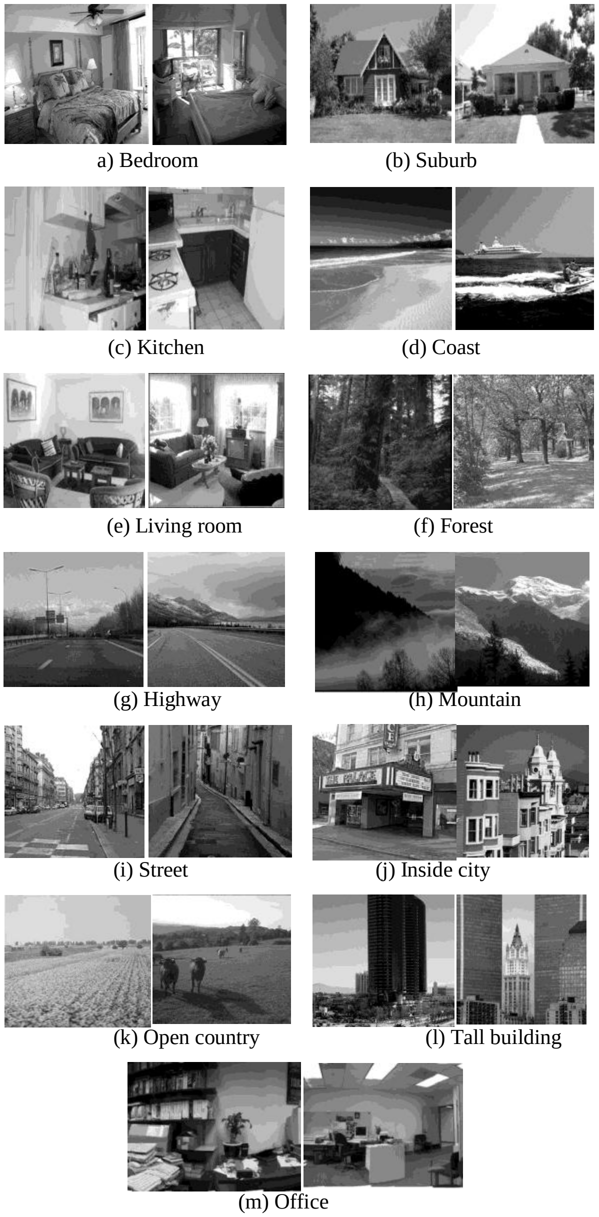
III. Experiments

A. *Databases*

Experiments are performed over two databases: the UIUC Sport Scene database [25] and the Natural Scene database [14].

A.1. *UIUC Sport Scene Database*

The UIUC Sport Scene database is a very challenging visual scene database with significant intra-class variations in the background, scale and lighting etc. As shown in Figure 6, the badminton scene can happen on the badminton court or at the backyard of a house. The lighting and scale can also be very different. The database consists of 8 categories of sport scenes with 500 images in each category.


a) Bedroom (b) Suburb

(c) Kitchen (d) Coast

(e) Living room (f) Forest

(g) Highway (h) Mountain

(i) Street (j) Inside city

(k) Open country (l) Tall building

(m) Office

Figure 7: sample images of the Natural Scene database, which has 13 categories [14].

A.2. *Natural Scene Database*

Natural Scene database consists of 13 categories, with 210 scene images in each category, as shown in Figure 7. Most of them are gray images. Therefore, we cannot evaluate the color feature on this dataset.

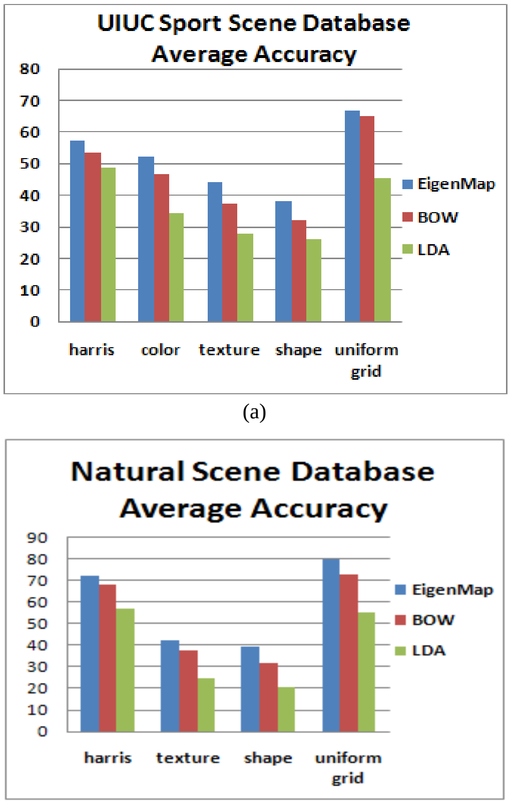
B. Experimental Setups

We divide the dataset of each category into five subsets. Then the images of one subset are used as testing set, while the images from the remaining four subsets are used as training set. The process is repeated five times with each of the five subsets used as the testing data once. All experimental results reported in the paper are the average accuracy of the five repeated testing.

C. Experimental Results

C.1. Compare to the Bag of Words Model and the LDA model

For each feature type, *i.e*., texture, shape, color, the Harris corner with the SIFT descriptor and the uniform grid interest point with the SIFT descriptor, we compare the proposed EigenMap approach with the Bag of Words model (BOW) [8] and the Topic Discovery model [25]. More specifically, we employ the Latent Dirichlet Allocation (LDA) [3] similar to the approach proposed by Li *et al.* [25]. They are all running under the same experimental setup.



(a)

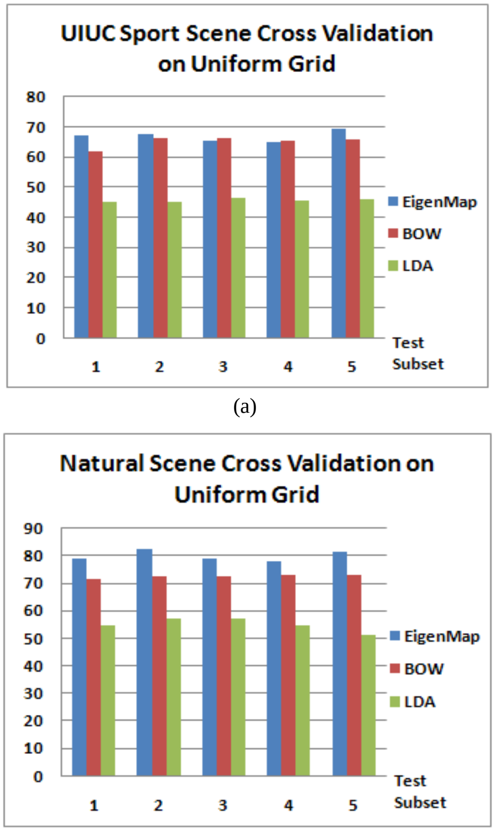
(b)

Figure 8: Comparing to the Bag of Words (BOW) model [8] and the Latent Dirichlet Allocation (LDA) model [25] on (a) the UIUC Sport Scene database; and (b) the Natural Scene database.

The detailed comparisons over different feature types are shown in Figure 8(a) and Figure 8(b) for the UIUC Sport Scene database and the Natural Scene database respectively. On the UIUC Sport Scene database, the EigenMap model outperforms the standard Bag of Words model by the average of 4.8% over all different feature types. It also outperforms the LDA model by the average of 15.3%.

We observe larger performance improvement over the other two models on the Natural Scene database. The proposed approach improves classification accuracy by the average of 6% and 19% as compared with the BOW and the LDA models respectively.

The consistent performance improvement over every feature type verifies the effectiveness of the proposed model in the visual scene classification. The spatial correspondences among local features, which the EigenMap model captures, contribute to the performance improvements. Figure 9 shows the detailed cross validation results as compared with the BOW and the LDA model.



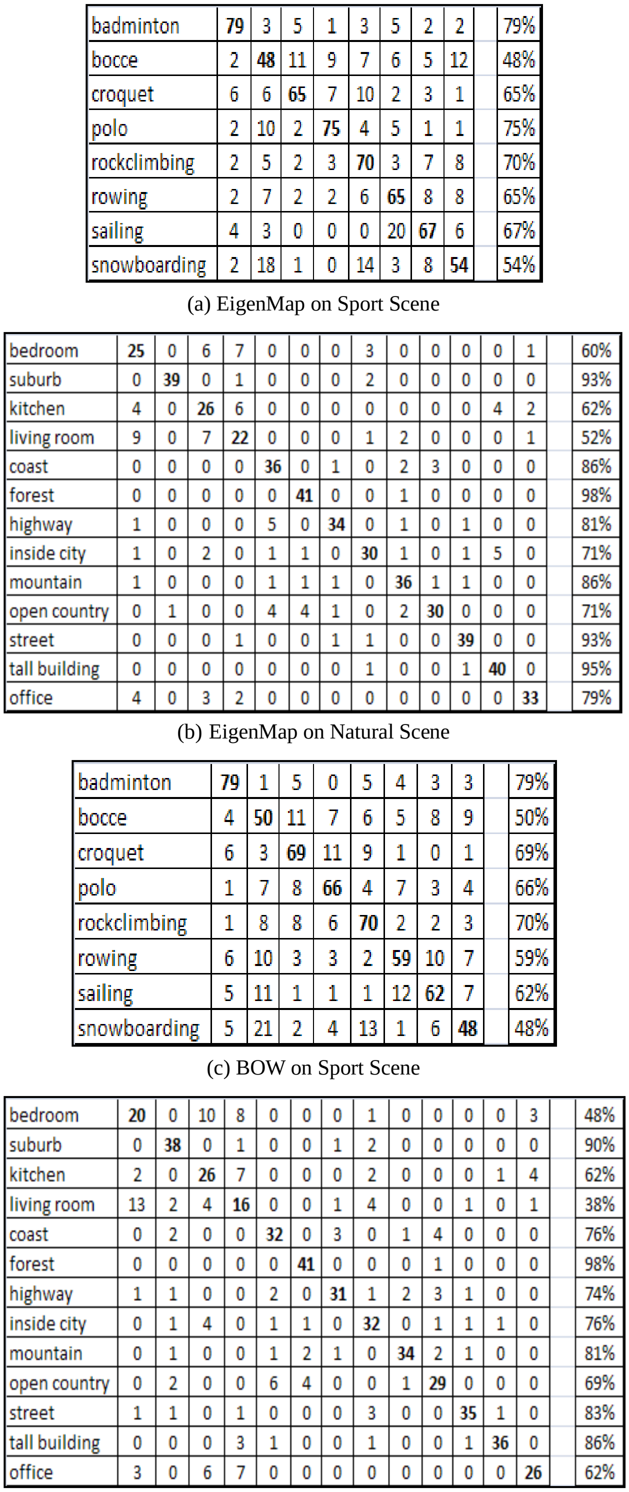
(a)

(b)

Figure 9: Comparing the BOW and the LDA models using the five-fold cross validation results of the uniform grid interest point features on (a) the UIUC Sport Scene database; and (b) the Natural Scene database.

The sample confusion matrices of the uniform grid interest point feature are also shown in Figure 10 for both the UIUC Sport Scene database and the Natural Scene database. The true positive rate for each category is shown in the last column next to the corresponding confusion matrix. As we can see from the confusion matrices of the EigenMap and the BOW, the

EigenMap approach achieves higher performance on most of the scene categories.



badminton	**79**	3	5	1	3	5	2	2		79%
bocce	2	**48**	11	9	7	6	5	12		48%
croquet	6	6	**65**	7	10	2	3	1		65%
polo	2	10	2	**75**	4	5	1	1		75%
rockclimbing	2	5	2	3	**70**	3	7	8		70%
rowing	2	7	2	2	6	**65**	8	8		65%
sailing	4	3	0	0	0	20	**67**	6		67%
snowboarding	2	18	1	0	14	3	8	**54**		54%

(a) EigenMap on Sport Scene

bedroom	**25**	0	6	7	0	0	0	3	0	0	0	0	1		60%
suburb	0	**39**	0	1	0	0	0	2	0	0	0	0	0		93%
kitchen	4	0	**26**	6	0	0	0	0	0	0	0	4	2		62%
living room	9	0	7	**22**	0	0	0	1	2	0	0	0	1		52%
coast	0	0	0	0	**36**	0	1	0	2	3	0	0	0		86%
forest	0	0	0	0	0	**41**	0	0	1	0	0	0	0		98%
highway	1	0	0	0	5	0	**34**	0	1	0	1	0	0		81%
inside city	1	0	2	0	1	1	0	**30**	1	0	1	5	0		71%
mountain	1	0	0	0	1	1	1	0	**36**	1	1	0	0		86%
open country	0	1	0	0	4	4	1	0	2	**30**	0	0	0		71%
street	0	0	0	1	0	0	1	1	0	0	**39**	0	0		93%
tall building	0	0	0	0	0	0	0	1	0	0	1	**40**	0		95%
office	4	0	3	2	0	0	0	0	0	0	0	0	**33**		79%

(b) EigenMap on Natural Scene

badminton	**79**	1	5	0	5	4	3	3		79%
bocce	4	**50**	11	7	6	5	8	9		50%
croquet	6	3	**69**	11	9	1	0	1		69%
polo	1	7	8	**66**	4	7	3	4		66%
rockclimbing	1	8	8	6	**70**	2	2	3		70%
rowing	6	10	3	3	2	**59**	10	7		59%
sailing	5	11	1	1	1	12	**62**	7		62%
snowboarding	5	21	2	4	13	1	6	**48**		48%

(c) BOW on Sport Scene

bedroom	**20**	0	10	8	0	0	0	1	0	0	0	0	3		48%
suburb	0	**38**	0	1	0	0	1	2	0	0	0	0	0		90%
kitchen	2	0	**26**	7	0	0	0	2	0	0	0	1	4		62%
living room	13	2	4	**16**	0	0	1	4	0	0	1	0	1		38%
coast	0	2	0	0	**32**	0	3	0	1	4	0	0	0		76%
forest	0	0	0	0	0	**41**	0	0	0	1	0	0	0		98%
highway	1	1	0	0	2	0	**31**	1	2	3	1	0	0		74%
inside city	0	1	4	0	1	1	0	**32**	0	1	1	1	0		76%
mountain	0	1	0	0	1	2	1	0	**34**	2	1	0	0		81%
open country	0	2	0	0	6	4	0	0	1	**29**	0	0	0		69%
street	1	1	0	1	0	0	0	3	0	0	**35**	1	0		83%
tall building	0	0	0	3	1	0	0	1	0	0	1	**36**	0		86%
office	3	0	6	7	0	0	0	0	0	0	0	0	**26**		62%

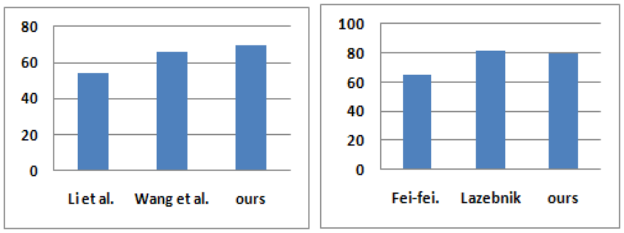
(d) BOW on Natural Scene

Figure 10: Sample confusion matrices of the EigenMap and the BOW models over both the UIUC Sport Scene and the Nature Scene database; Note that the last column of each confusion matrix indicates the true positive rate for each category. All the four confusion matrices are generated using the uniform grid interest point with the SIFT descriptor, where the rows are the ground truth while the columns are the classified categories.

In the UIUC Sport Scene dataset, the most confusion occurs between the "Rowing" and the "Sailing" categories since both sport scenes are very similar in the background, which contains water in the scene images. In the Natural Scene dataset, the most confusion occurs between the bedroom and the living room scene images.

C.2. Compare to the State-of-the-art Performance

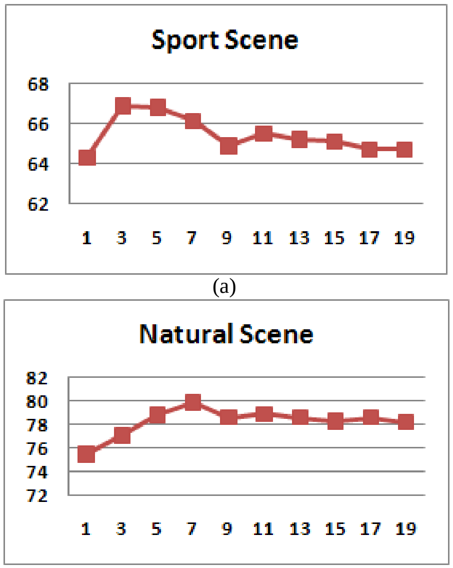
Figure 11(a) and 11(b) show the detailed comparison with the state-of-the-art performance on the UIUC Sport scene dataset [25, 36] and the Natural Scene dataset [14, 24] respectively. The results are directly cited from their papers. From Figure 11, the proposed EigenMap model achieves a state-of-the-art performance on both the UIUC Sport Scene database and the Natural Scene database.



(a) (b)

Figure 11: Compare with the state of the art reported by Fei-Fei and Perona [14], Lazebnik *et al*. [24], Li *et al*. [25], and Wang *et al*. [36] on both the UIUC Sport Scene database and the Natural Scene database.

C.3. Select Number of Principal Components for EigenMap



(a)

(b)

Figure 12: The effect of number of eigenvectors used in the PCA projection on the classification performance over (a) the UIUC Sport Scene database; (b) the Natural Scene database; Note that we used the uniform grid interest point with the SIFT descriptor for both databases.

We also evaluate the effect of number of eigenvectors used in the construction of the EigenMap on the classification performance. As we can see from Figure 12, the number of eigenvectors used in the PCA projection achieves the best performance when it is around 5. As the number of eigenvectors continues increasing, the performance degrades slightly. That suggests that we only need a very small dimensional space to represent each visual word's EigenMap.

IV. Conclusion

We have proposed a novel EigenMap representation of a scene image, which can not only incorporates the spatial information with the appearance features, but also integrates both local features and their global correspondences effectively. The EigenMap model has been evaluated on two public databases for scene image classification and outperforms both the standard Bag of Words model and the LDA model. The proposed model also achieves a state-of-the-art performance on both datasets with small feature dimension.

.

References

[1] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts", PAMI, 2002.

[2] S. Bernard, "Density Estimation for Statistics and Data Analysis", 1st Edition, Chapman and Hall, 1986.

[3] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, 3:993-1022, 2003.

[4] Y. Cao, C. Wang, Z. Li, L. Zhang, L. Zhang, "Spatial Bag of Features", CVPR, 2010.

[5] P. Carbonetto, N. Freitas, and K. Barnard, "A statistical model for general contextual object recognition", ECCV, 2004.

[6] C. Chang and C. Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] O. Chum, A. Zisserman, "An exemplar model for learning object classes", CVPR, 2007.

[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, "Visual Categorization with Bag of Keypoints", ECCV, 2004.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", CVPR, 2005.

[10] J. Deng, A. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us", ECCV, 2010.

[11] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, "An empirical study of context in object detection", CVPR, 2009.

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes", CVPR, 2009.

[13] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization", CVPR, 2010.

[14] L. Fei-Fei and P. Perona, "A Bayesian hierarchy model for learning natural scene categories", CVPR, 2005.

[15] P. Felzenszwalb and D. Huttenlocher, "Efficient Graph-Based Image Segmentation", IJCV, 2004.

[16] P. Gehler, S. Nowozin, "On feature combination for multiclass object classification", ICCV, 2009.

[17] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization", CVPR, 2008.

[18] C. Harris and M. Stephens, "A Combined Corner and Edge Detector", *Proc. Alvey Vision Conf.*, Univ. Manchester, pp. 147-151, 1988.

[19] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things", ECCV, 2008.

[20] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective", IJCV, 2008.

[21] J. Huang, S. Kumar, M. Mitra, W. Zhu, R. Zabih, "Image Indexing Using Color Correlograms", CVPR, 1997.

[22] N. Kumar, A. Berg, P. Belhumeur, and S. K. Nayar, "Attribute and simile classifiesrs for face verification", ICCV, 2009.

[23] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer", CVPR, 2009.

[24] S. Lazebnik, C. Schmid, J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", CVPR, 2006.

[25] L. Li, R. Socher, and L. Fei-Fei, "Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework", CVPR, 2009.

[26] L. Li, H. Su, E. Xing and L. Fei-Fei, "Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification", NIPS, 2010.

[27] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", IJCV, 2004.

[28] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In CVPR, 2007.

[29] A. Oliva and A. Torralba, "Modeling the shape of the scene: A Holistic Representation of the Spatial Envelope", IJCV, 2001.

[30] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context", ICCV, 2007.

[31] N. Rasiwasia and N. Vasconcelos, "Scene Classification with Low-dimensional Semantic Spaces and Weak Supervision", CVPR, 2008.

[32] S. Savarese, J. Winn, A. Criminisi, "Discriminative Object Class Models of Appearance and Shape by Correlatons", CVPR, 2006.

[33] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos", ICCV, 2003.

[34] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient Object Category Recognition Using Classemes", ECCV, 2010.

[35] M. Varma and A. Zisserman, "Classifying Images of Materials: Achieving Viewpoint and Illumination Independence", ECCV, 2002.

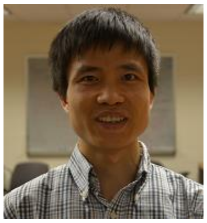
[36] C. Wang, D. Blei and L. Fei-Fei, "Simultaneous Image Classification and Annotation", CVPR, 2009.

[37] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan, "Categorizing nine visual classes using local appearance descriptors", IWLAVS, 2004.

[38] J. Winn, A. Criminisi and T. Minka, "Object Categorization by Learned Universal Visual Dictionary", ICCV, 2005.

[39] L. Wolf and S. Bileschi, "A critical view of context", IJCV, 2006.

[40] Y. Zhang, Z. Jia, T. Chen, "Image Retrieval with Geometry-Preserving Visual Phrases", CVPR, 2011.



Shizhi Chen (S'11) is a Phd student in the Department of Electrical Engineering at the City College of New York. His research interests include facial expression recognition, scene understanding, machine learning and related applications. He received the BS degree of Electrical Engineering from SUNY Binghamton, New York in 2004, and the MS degree of Electrical Engineering and Computer Science from UC Berkeley, California in 2006. From 2006 to 2009, he worked as an engineer in several companies including Altera, Supertex Inc., and US Patent and Trademark Office. He is a member of Eta Kappa Nu (electrical engineering honor society), and a member of Tau Beta Pi (engineering honor society). He also received numerous scholarships and fellowships, including Beat the Odds scholarship, Achievement Rewards for College Scientists (ARCS) Fellowship, and NOAA CREST Fellowship.

YingLi Tian (M'99–SM'01) received her BS and MS from TianJin University, China in 1987 and 1990 and her PhD from the Chinese University of Hong Kong, Hong Kong, in 1996. After holding a faculty position at National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, she joined Carnegie Mellon University in 1998, where she was a postdoctoral fellow at the Robotics Institute. Then she worked as a research staff member in IBM T. J. Watson Research Center from 2001 to 2008. She is one of the inventors of the IBM Smart Surveillance Solutions.

She is currently an associate professor in Department of Electrical Engineering at the City College of New York. Her current research focuses on a wide range of computer vision problems from motion detection and analysis, to human identification, facial expression analysis, and video surveillance. She is a senior member of IEEE.