SIFT and color feature fusion using localized maximum-margin learning for scene classification

Jianzhao Qin and Nelson H.C. Yung
Laboratory for Intelligent Transportation Systems Research
Department of Electrical & Electronic Engineering
The University of Hong Kong, Hong Kong
e-mail: jzhqin@eee.hku.hk, nyung@eee.hku.hk

***Abstract*—In this paper, we proposed a method that uses localized maximum-margin learning to fuse SIFT and color features during the bags of visual words modeling process for eventual scene classification. It offers a more flexible way in fusing these features through determining the similarity-metric locally by localized maximum-margin learning. The proposed method has been evaluated experimentally and the results indicate its effectiveness.**

Keywords-scene categorization; feature fusion; similarity-metric learning

I. INTRODUCTION

Scene categorization or classification is a task of labeling a given image to a specific scene category (e.g., coast, forest, highway, office, kitchen, etc.). Automatically categorizing an image to a scene category not only can help us organize our images, but also can help us retrieve a set of images relating to a specific scene from the internet or an image database [1-3]. Moreover, recognizing the scene category of a place is a rather important ability for an intelligent vehicle or robot to take correct actions under the scene [4, 5] (e.g. slow down when it is moving on a inner city street while speed up when it is moving on a highway.). Furthermore, scene categorization can also provide vital contextual information to many computer vision tasks, such as object recognition, image segmentation and video surveillance [6, 7].

In the early research work of scene categorization, global feature [2-4, 8] based methods have been proposed. This approach takes an image as a whole entity, and then describes the image by the distribution of the color [2, 3] and/or texture [2] and/or gradients [4, 8] over the entire image region. It achieved a certain success, especially in separating outdoor scenes from indoor scenes. However, it fails in classifying scenes that share similar global properties (e.g. bedroom vs. sitting room, or open country vs. coast). In recent years, local feature based methods [9-13] become more popular because of its robustness to occlusions, geometric deformation and illumination variations. This approach models a scene image by the co-occurrences of a number of visual components or the co-occurrences of a certain number of visual topics (intermediate representation). One of the most popular and successful models is called the *bags of visual words* [11-13], which has a number of variants [14, 15] [16, 17]. This model is also successful for visual object recognition and detection. In order to further enhance the performance of the visual recognition system, some methods has been proposed to combine different types of features [18-20]. In the methods of Varma etc. and Bosch etc. [18, 20], they first create several bags of words models corresponding to different types of features (e.g. bags of visual words model based on color feature, bags of visual words model based on SIFT feature and etc.). Then, these bags of visual words models are represented by different feature vectors which denote the existence or distribution of visual words respectively. Next, the multiple-kernel learning method is employed to learn a linear weighting of different kernels corresponding to different types of features. As it is, the fusion of features is performed after representing the whole image by the *bags of visual words*. In other words, the feature fusion is carried out in a global manner. One of the weaknesses of fusion globally is that the ambiguity of local patches of the image caused by single feature representation may not be compensated by introducing other types of features. This is because the information of other features is globally coded without any relation to any specific local image region. A further weakness of this approach is that the multiple-kernel learning step can only produce fixed weightings to different features. In practice, in order to differentiate a region from other regions, we may have to give more weight to SIFT feature while in other cases, we may have to give more weight to color feature instead. Horster & Lienhart [19] proposed three generative models to fuse features for image retrieval. Model A in their paper takes the fusion at the decision level, which is global in nature. Model B models the joint distribution of the visual words from SIFT color features, which is fusion at the visual word level. However, this fusion occurred after visual word assignment. If the visual word that represents the image region is wrongly assigned due to the single feature description, it may result in a wrong joint visual words distribution. Furthermore, their proposed model is generative model and assumes the joint distribution is multi-norminal. If the number of samples for distribution estimation is not large enough as the extracted feature usually has high dimension or the joint distribution is not multi-norminal, it may result in poor performance.

ISBN: 978-1-4244-8889-6

In this paper, we proposed a local feature fusion method using localized maximum margin learning. Given an unknown image region, the SIFT and color features are extracted from the image region. Based on the SIFT feature, we select a set of nearest neighbor visual words using Euclidean distance measurement by the SIFT feature. Subsequently, these candidate visual words are taken as classes. The SIFT and color features of the training image regions which form these visual words are taken as the features of the samples belonging to each classes. Next, a Support Vector Machine (SVM) classifier with a linear kernel is employed to learn linear weightings for each element of the feature which can result in a maximum margin separation of the samples belonging to each class. This trained classifier is then used as a similarity measurement to measure the similarity of the feature of the unknown image region to the candidate visual words. The classification result determines which visual word is used to represent this unknown region. Based on the maximum margin criteria, the classifier determines the weights for the SIFT and color feature based on different nearest neighbor candidates.

II. IMAGE REGION FEATURE EXTRACTION

This section briefly introduces the SIFT feature and color feature extraction procedure.

In order to capture image information from different scales, the image is regularly divided into patches at different scales from the coarsest scale (i.e. the whole image) to consecutive finer scales. The image features (SIFT and color) are extracted from all these patches. In our work, the SIFT feature descriptor first proposed by Lowe [21] is employed. For the color feature, we use a similar extraction procedure. After transforming the image into Lab color space, the ROI is divided into 4x4 blocks. Then, the mean values of the L, a, b components are calculated. Next, these mean values for each block are concatenated together after weighted by a Gaussian function.

Meanwhile, the contextual information is also integrated to describe the ROI [17]. Such contextual information provides useful cue about the ROI. This potentially reduces the ambiguity when employing visual words to represent the local regions. We combine the image feature from the region at coarser scale (but with the same sampling point) and the image features from the neighbor regions at the same scale with the feature of ROI to describe the ROI. That is, let $\mathbf{P}_L \in \mathbb{R}^{m_L \times n_L}$ denotes the ROI, $\mathbf{P}_C \in \mathbb{R}^{m_C \times n_C}$ denotes the region having the same sampling point as the ROI but at a coarser scale level and $\mathbf{P}_N \in \mathbb{R}^{m_N \times n_N}$ denotes the neighbor regions of the ROI at the same scale level. For local visual word, the ROI is represented by $\mathbf{f} = f(\mathbf{P}_L)$ where f denotes the feature extraction function. For the contextual visual word, we represent the ROI as $\mathbf{f} = f(\mathbf{P}_L, \mathbf{P}_C, \mathbf{P}_N)$. We linearly combine these features. The feature of the ROI is then represented as:

$$\mathbf{f} = [f(\mathbf{P}_L), w_C \cdot f(\mathbf{P}_C), w_N \cdot f(\mathbf{P}_N)], \quad (1)$$

where w_C and w_N are the weighting parameters that control the significance of features from the coarser scale and the neighborhood regions. The weighting parameters for different contextual information are determined using cross-validation.

III. LOCALIZED MAXIMUM-MARGIN LEARNING FOR SIFT AND COLOR FEATURE FUSION

This section introduces the proposed localized maximum margin learning for SIFT and color feature combination. Figure 1 shows the procedure of combining color information with SIFT information using localized maximum margin learning to select the best visual words in the bag of visual words model forming procedure . Based on a given image region, the SIFT feature and color feature are extracted from this region. Using the SIFT feature vector, we calculate its similarity with the visual words in the codebook based on the Euclidean distance measurement. Constrained by the ratio to the shortest distance and maximum number, K -nearest neighbor candidate visual words are chosen. Then, the SIFT features which were clustered to form these visual words and their corresponding color feature are retrieved and concatenated to form a feature vector $\mathbf{f} = [\mathbf{f}_{SIFT}, \mathbf{f}_{color}]$. The features belonging to a visual word are taken as the features of a class. Next, the maximum margin learning is employed to learn weighting values of the elements in the feature vector. The 2-class maximum margin learning problem is equivalent to the 2-norm minimization problem showed in equation (2), where $\mathbf{w}$ is the weighting vector, ϕ denotes a linear or nonlinear transform to the feature vector, $\mathbf{x}_i$, L is the total number of training samples, y_i takes two values, i.e. 1, -1, corresponding to two classes respectively and P is the penalty parameter . It can be easily extended to multiple-class problem using the one against one strategy. Since the maximum margin learning is based on the features belonging to K -nearest neighbor candidate visual words, we call it localized maximum margin learning. The localized learning enables us finding a weighting vector $\mathbf{w}$ which maximizes the distances between the candidate visual words after introducing the color information. After that, the learned weighting value is used to measure the similarity of the feature of the unknown region to the candidate visual words for selecting the best representative visual words for the unknown region. Obviously, this weighting vector can change and adapt to different nearest neighbor structures for finding a suitable weighting values. For instance, if the SIFT features forming the candidate visual words are very similar, bigger weighting values will be put on the color feature in order to separate the samples belonging to different candidate visual words. Otherwise, if

the color features are similar, more weighting values will be put on SIFT feature.

$$\min_{\mathbf{w},\eta} P\sum_{i=1}^{L}\eta_i+\frac{1}{2}\|\mathbf{w}\|_2 \qquad (2)$$

$$s.t.\ \ y_i\mathbf{w}^T[\phi(\mathbf{x}_i^T)\ 1]^T+\eta_i\geq 1,\ \ \eta_i\geq 0, i=1,\ldots,L$$

The steps for the selection of the best representative visual words by combining the color feature using the localized maximum margin learning are as follows:

Step1: Given a list of visual words, $\mathbf{V}=\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$, and the SIFT feature of an image patch , $\mathbf{f}_{SIFT}$ (this feature is the combination of the local image feature with the contextual feature and transformed by the PCA transformation matrix), calculate the Euclidean distances between the feature and the visual words, $\{d_1,d_2,\ldots,d_n\}$.

Step2: Choose the minimum distance, $d_m=\min\{d_1,d_2,\ldots,d_n\}$. Then calculate the ratios of the distance to the minimum distance, $r_i=\frac{d_i}{d_m}, i=1,2,\ldots n$.

Ste 3: Select the candidate visual words whose distance to $\mathbf{f}_{SIFT}$ satisfy the ratio $r_j\leq T_r, j=1,2,\ldots,K$ ($T_r=1.15$ in this paper) and its corresponding visual words, $v_j, j=1,2,\ldots,K$ (in order to reduce the computational burden, the maximum number of the preliminary selected visual words can be up-bounded by number N_b.

Step4: Retrieve the SIFT features $\{\mathbf{f}^{SIFT}_{1j},\mathbf{f}^{SIFT}_{2j},\ldots,\mathbf{f}^{SIFT}_{3j}\}, j=1,2,\ldots,K$ and corresponding color features $\{\mathbf{f}^{color}_{1j},\mathbf{f}^{color}_{2j},\ldots,\mathbf{f}^{color}_{3j}\}, j=1,2,\ldots,K$ from the training set that are clustered to form the visual words, $\mathbf{v}_j, j=1,2,\ldots,K$. Normalize these features by their norm, i.e., $\mathbf{f}^{SIFT}_{ij}=\frac{\mathbf{f}^{SIFT}_{ij}}{\|\mathbf{f}^{SIFT}_{ij}\|_2}, \mathbf{f}^{color}_{ij}=\frac{\mathbf{f}^{color}_{ij}}{\|\mathbf{f}^{color}_{ij}\|_2}$. Then, concatenate them, $\mathbf{f}_{ij}=[\mathbf{f}^{SIFT}_{ij};\mathbf{f}^{color}_{ij}]$.

Step5: Take the retrieved features $\{\mathbf{f}_{1j},\mathbf{f}_{2j},\ldots,\mathbf{f}_{3j}\}, j=1,2,\ldots,K$ as the training set of K classes. Then, train a classifier based on equation (2) for these K classes (in our experiment, we used Radial Basis Function as kernel).

Step 6: Classify the given patch feature $\mathbf{f}$ by the trained classifier to a class c. Then, the image patch is represented by the visual word corresponding to class c.

IV. EXPERIMENTAL RESULTS

This section reports the experimental results of the proposed method. The performance of the proposed scene classification method is tested on an outdoor scene dataset which has been widely used in previous research [14, 22-24]. This dataset consists of 2688 color images from 8 categories (SCENE-8): coast (360 samples), 328 forest (328 samples), mountain (274 samples), open country (410 samples), highway (260 samples), inside city (308 samples), tall buildings (356 samples), and streets (292 samples). The average size of each image is 256×256.

In the experiment, we perform a 10-fold cross-validation to achieve more accurate performance estimation. And, except perform the experiment that combines all the scale levels, we also perform the experiments at scale 1, 2, 3, 4 respectively in order to investigate how the scale level influence the performance of the proposed method. Table Ⅰ shows the 10-fold cross-validation results at scale level 1 to scale level 4 and the comparison results that obtained only using the SIFT feature. These results show that the average accuracy results are improved by 4.53%, 5.04%, 1.02% and 0.87% at scale 1, 2, 3 and 4 respectively by combining the color feature. The results reveal that the proposed method is more effective at coarser scales. At finer scales, since the image is divided into smaller regions and each image consists of larger number of image patches, even some of these patches are correctly represented by the suitable visual words using the proposed method, the influence of this correction to the final recognition rate becomes smaller.

Figure 2 shows an 'Open country' image which is wrongly classified as 'Coast' using SIFT feature only but correctly classified using the proposed method. The right part of figure 2 gives the patch samples that form the visual word selected to represent the red regions in the given image without using the proposed method and using the proposed method respectively. Some regions of the glass land of the given image are wrongly represented by the visual word which represents the region of sea water or sand of 'Coast' due to the similarity in the SIFT feature without incorporating the color. However, after using the proposed method to combine color locally, these glass land regions are correctly coded by the visual word which represents the glass land.

Table Ⅱ shows the result of the proposed method after combining 5-scale visual words, the result that uses multiple-kernel learning to combine SIFT and color bags of visual words models and the results of several previous representative scene classification method on this dataset. We can observe that the performance of the proposed localized maximum margin learning based method is superior to the combination method which combines the features in a global manner by 1% in average and improved the performance by 1.27% compared with using SIFT feature only. The results also show the superiority of the proposed method over the previous representative scene classification methods (The experimental setting (the composition of the

training set and test set) is the same for the comparison of these methods. And the parameters setting of other methods is the same as proposed in their papers [8, 16, 17, 23]).

TABLE I. ACCURACY RATES (MEAN (STANDARD DEVIATION)%) OF THE PROPOSED METHOD COMPARING WITH USING SIFT FEATURE ONLY AT SCALE 1, 2, 3 AND 4 RESPECTIVELY.

	Scale 1	Scale 2	Scale 3	Scale 4
SIFT only	66.63(4.54)	74.21(3.65)	82.68(3.00)	88.15(3.03)
Proposed method	**71.16(3.25)**	**79.25(3.79)**	**83.70(3.48)**	**89.02(3.48)**

TABLE II. COMPARISON WITH OTHER REPRESENTATIVE ALGORITHMS (IN AVERAGE ACCURACY RATE %) . (1) PROPOSED CONTEXTUAL VISUAL WORDS WITH FUSED SIFT AND COLOR FEATURES; (2) CONTEXTUAL VISUAL WORDS AND MKL BASED METHOD WITH FUSED SIFT AND COLOR FEATURES; (3) CONTEXTUAL VISUAL WORDS BASED METHOD WITH SIFT FEATURE ONLY; (4) SPATIAL PYRAMID MATCHING WITH SIFT FEATURE ONLY; (5) PROBABILITY LATENT SEMANTIC ANALYSIS (PLSA, COLOR-SIFT); (6) GIST FEATURE INCLUDING COLOR INFORMATION

(1)	(2)	(3)	(4)	(5)	(6)
91.57 (2.71)	90.57 (2.44)	90.30 (2.54)	88.19 (3.46)	84.78 (1.93)	80.48 (3.94)

V. CONCLUSION

In this paper, we have proposed a localized maximum margin learning method to fuse SIFT feature and color feature locally. After selecting K-nearest neighbor visual words using Euclidean distance measurement, the SIFT features and color features of the samples which forms the K-nearest neighbor visual words are retrieved to trained a classifier using maximum margin learning. Then, the classification result of the feature of given image region is used to select the best representative visual words. Comparing with the global feature combination method, the virtue of the proposed method is that it is capable of determining different combination strategy according to different local feature property. The experimental results show the effectiveness of the proposed method.

ACKNOWLEDGMENT

The authors would like to thank Dr. Kwan Wing Keung from the computer center of the University of Hong Kong for providing high performance computing support.

REFERENCES

[1] J. Z. Wang, L. Jia, and G. Wiederhold, "SIMPLIcity: semantics-sensitive integrated matching for picture libraries," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 947-963, 2001

[2] E. Chang, G. Kingshy, G. Sychay, and W. Gang, "CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines," IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, pp. 26-38, 2003.

[3] A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang, "Content-based hierarchical classification of vacation images," in IEEE International Conference on Multimedia Computing and Systems, 1999, pp. 518-523.

[4] C. Siagian and L. Itti, "Gist: A Mobile Robotics Application of Context-Based Vision in Outdoor Environment," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 1063-1069.

[5] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle Detection and Terrain Classification for Autonomous Off-Road Navigation," Autonomous Robots, vol. 18, pp. 81-102, 2005.

[6] A. Torralba, "Contextual priming for object detection," International Journal of Computer Vision, vol. 53, pp. 169-191, Jul 2003.

[7] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in Adv. in Neural Information Processing Systems 17 (NIPS), Lawrence K. Saul and Yair Weiss and Leon Bottou, 2005, pp. 1401-1408.

[8] C. Siagian and L. Itti, "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 300-312, 2007:

[9] J. Luo and A. Savakis, "Indoor vs outdoor classification of consumer photographs using low-level and semantic features," in 2001 International Conference on Image Processing, 2001, pp. 745-748 vol.2.

[10] J. Vogel and B. Schiele, "A Semantic Typicality Measure for Natural Scene Categorization," in 2004 DAGM 2004, pp. 195-203.

[11] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in Tenth IEEE International Conference on Computer Vision, 2005, pp. 1816-1823 Vol. 2

[12] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, pp. 1475-1490, 2004

[13] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in Ninth IEEE International Conference on Computer Vision, 2003, pp. 1470-1477 vol.2

[14] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 524-531 vol. 2

[15] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in Tenth IEEE International Conference on Computer Vision, 2005, pp. 883-890 Vol. 1

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, 2006, pp. 2169-2178

[17] J. Qin and N. H. C. Yung, "Scene categorization via contextual visual words," Pattern Recognition, vol. 43, pp. 1874-1888, 2010

[18] M. Varma and D. Ray, "Learning The Discriminative Power-Invariance Trade-Off," in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1-8

[19] E. Horster and R. Lienhart, "Fusing Local Image Descriptors for Large-Scale Image Retrieval," in Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1-8

[20] A. Bosch, A. Zisserman, and X. Munoz, "Image Classification using ROIs and Multiple Kernel Learning," IJCV 2008, 2008

[21] D. G. Lowe, "Object recognition from local scale-invariant features," in The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, pp. 1150-1157 vol.2

[22] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol. 42, pp. 145-175, 2001

[23] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification Via pLSA," in ECCV 2006, 2006, pp. 517-530

[24] A. Bosch, A. Zisserman, and X. Muoz, "Scene Classification Using a Hybrid Generative/Discriminative Approach," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, pp. 712-727, 2008

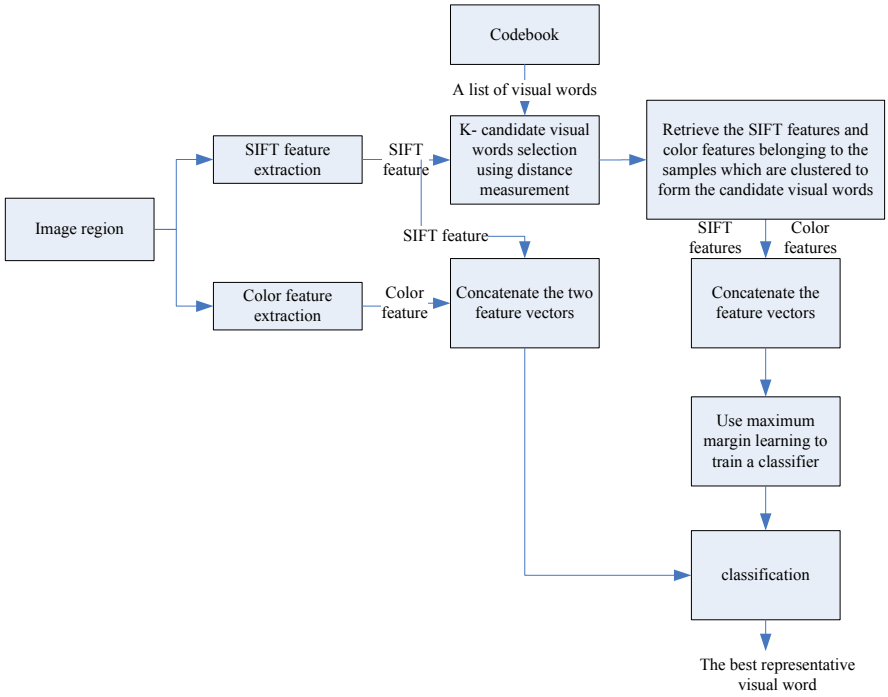

Figure 1. Fusion of SIFT and color features using localized maximum margin learning to select the best visual word for local region representation

Figure 2. The left 'Open country' image is wrongly classified as 'Coast' without using the proposed method to utilize color information but correctly classified as 'Open country' after using the proposed method to combine color information.