LEARNING DISCRIMINATIVE VISUAL DICTIONARY FOR NATURAL SCENE CATEGORIZATION

Ying Huang, Wenmin Wang, Ronggang Wang*

School of Electronic and Computer Engineering
Shenzhen Graduate School, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, China 518055
ecehuangying@sz.pku.edu.cn, {wangwm, rgwang}@ece.pku.edu.cn

ABSTRACT

Many successful systems for scene recognition transform low-level descriptors into complex representations. This process consists of the two steps: 1) feature coding, which performs a pointwise transformation of the descriptors into a representation adapted to the task, and 2) image pooling, which summarizes the coded features. Even though these two steps have been paid so much attention, but there are still some problems in combining scene semantic with local features. The goal of this paper is threefold: to address the problem by modifying the traditional bag-of-features (BoF) framework; to show how to achieve the best performance by learning a semi-supervised discriminative dictionary; and to provide theoretical and empirical insight into the remarkable performance. By teasing apart components shared by modern scene categorization pipeline, our approach aims to facilitate the design of better scene recognition architectures.

***Index Terms*—** Dictionary learning, Scene categorization, Incremental learning, Bag-of-features (BoF), Discriminative dictionary

1. INTRODUCTION

Scene categorization is an important problem for computer vision and multimedia analysis. Content based image/video retrieval could benefit from the semantic knowledge [1, 2]. Bag-of-features (BoF) model [3] has been extremely practical for addressing this problem, which is inspired by the Bag-of-Words (BoW) model used in text categorization. However in text domain, the word is explicitly expressed, there is no given vocabulary for the visual categorization problem. And it has to be learned from a random sampling set of local features like SIFT [4]. The pipeline of BoF contains four main phases: 1) extracting local visual features, 2) learning a representative dictionary to code the local features, 3) image representation generating and 4) classifier training and testing. The phases 2) and 3) are named as Coding and Pooling in modern researches respectively [5, 6, 7]. The accuracy of BoF can be improved by extracting robust features [8, 9], learning a more representative dictionary to code the local features or giving a more generalized image representation. [10] set a baseline for evaluating the recognition approaches employing this pipeline. It proposed a "spatial pyramid matching (SPM)" representation for the image of an orderless BoF. The method partitions an image into $2^l \times 2^l$ segments in different scales $l = 0, 1, 2$, then computes the BoF histogram of these $21(1 + 4 + 16)$ segments, and finally concatenates all the histograms to form a vector representation of the image. In the case where only the scale $l = 0$ is considered, SPM reduces to BoF. A remarkable success is accomplished by involving the pyramid kernel [11]. [5, 6] developed an extension of the SPM method by generalizing vector quantization to sparse coding [12]. And [13] presented a variation of BoF method in combination with soft assignment BoF histogram based on the theory of kernel density estimation.

These methods in BoF pipeline perform well in general. However, there are still some weaknesses in current approaches. Firstly, there are massive scene categories in real world. When a new category is added to the system, the dictionary should be re-learning again on whole data, which will carry heavy computation cost [14]. Secondly, the visual features clustering and scene classification are disconnected. So we would implicitly make the assumption that all visual words have the same discriminative power. Then the representation has no capability of capturing the aspects of the data that are most useful for classification. And thirdly, the dictionary was shared by all categories and its size was manually fixed, so the discriminative power would depend heavily on the right setting of this parameter.

In this paper, we propose a framework for automatic learning a discriminative dictionary to address the aforementioned problems. This framework should be adapted for incremental learning, so when an additional category is considered, the system can still distinguish it with low-

This project was supported by the grant from Shenzhen municipal government for basic research on Information Technologies (No. JCYJ20130331144751105), and by Shenzhen Peacock Plan.

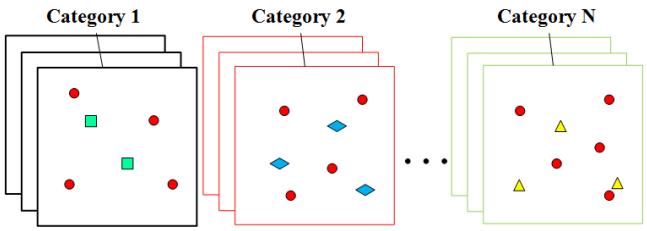

Fig. 1. Toy example of discriminative visual vocabulary learning. The images has four types of visual words, indicated by circles, squares, diamonds and triangles. The circle word exists in all categories, so it should be discarded. The words represented by squares, diamonds and triangles are words with great discriminative power.

er computational cost. The visual dictionary is trained in a semi-supervised manner by combining the information of cluster with labels of categories. Then we select a universal and adapted dictionary for scene categorization task.

The remainder of the paper is organized as follows: Section 2 introduces the related discriminative dictionary training work and the basic idea of our proposed framework; Section 3 describes the theory behind our framework; In Section 4, a novel training algorithm is employed to construct the discriminative visual vocabulary; Section 5 shows experimental results and corresponding analysis. Finally conclusions are drawn in Section 6.

2. RELATED WORK AND PROPOSED FRAMEWORK

There is semantic gap between the high level semantic labels and low level local descriptors, so it is not easy to strengthen the discriminative power of a system. In previous researches the discriminative power is enhanced by means of following. [15] learned a compact visual dictionary by pair-wise merging of visual words from an initial large dictionary and described the final visual words by GMMs. [16] proposed to generate image signature based on a universal vocabulary which can be adapted to class-specific data. [14] designed a new image signature by using generative models for classification. The idea behind [17] and [18] is a little similar to ours. [17] proposed a label consistent K-SVD (LC-KSVD) algorithm to learn a discriminative dictionary for sparse coding and [18] learning a weakly supervised visual dictionary by harnessing the semantic labels of images or regions.

The idea to enhance the BoF discrimination is selecting the "best visual word" into the dictionary, as illustrated by Figure 1. Meanwhile we add a dictionary retraining step compared to BoW in text domain, which could help to find an optimal center of the local features. The dimension of dictionary can be reduced when some useless visual words are discarded. It also leads to reduce the dimension of the final image

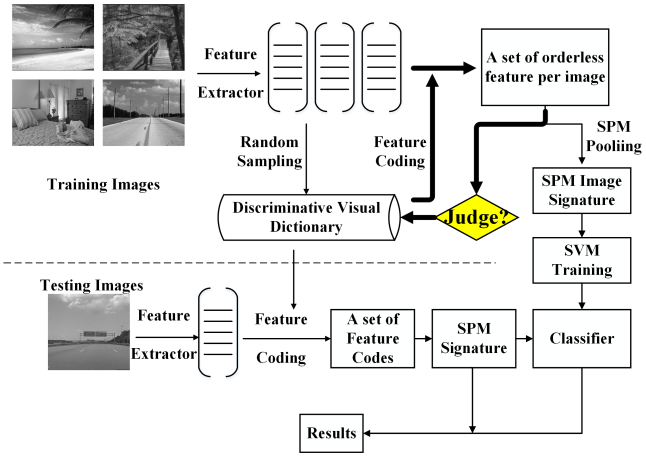

Fig. 2. Proposed framework

representation. Then the image signature will be generated through SPM method. The pipeline is showed in Figure 2. To evaluate the condition of discrimination, the framework will select the best visual word into dictionary and automatic set it to a appropriate size through small cycles which is shown by the bold arrows.

Under the proposed framework, there are several advantages: 1) In real world application, new categories can by added to the system through adding some visual words adapted to that category just like BoW in text domain; 2) By reinforcing the discriminative power of traditional BoF pipeline instead of abandoning it, the framework overcomes the BoF weakness while keeping the outstanding features, then the state-of-the-art SPM image representation is employed in the framework; and 3) the dimension of image signature can be set to an efficient one through the dictionary relearning step.

3. VISUAL VOCABULARY SELECTION

The semantic labels can not be connected to local features directly. We need to gather the statistics information between them, then use following techniques to select the appropriate features.

3.1. Information gain

Let C represents the label set for the image categories, $word$ is one item from the dictionary. The value set for w is W. Then the system entropy can be computed as follows,

$$H(C) = -\sum_{c \in C} P(c) \cdot \log P(c) \tag{1}$$

The condition entropy of $word$ for system is:

$$\begin{aligned} H(C|W) &= -\sum_{w \in W} P(w) \cdot H(C|W=w) \\ &= -\sum_{w \in W} P(w) \sum_{c \in C} P(c|w) \cdot logP(c|w) \end{aligned} \tag{2}$$

Table 1. χ^2 test

	label $l = c_i$	label $l \neq c_i$	Row total
$word$ in	a	b	$a+b$
$word$ not in	c	d	$c+d$
Column total	$a+c$	$b+d$	n

Information gained from $word$ can be formulated as:

$$IG(word) = H(C) - H(C|W) \tag{3}$$

The most discriminative words for the system can be exploited through sorting the IG for all words. But there is still weakness by employing the Shannons information theory to select features, the word selected is not applicable to every category. While chi-square test is another statistics technique that can overcome it.

3.2. Chi-square test

The chi-square test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in categories. When it is applied to test of independence, an "observation" consists of the values of two outcomes and the null hypothesis is that the occurrence of these outcomes is statistically independent. Let c_i be one of the semantic categories, l is the label for image and $word$ is one vocabulary in the visual dictionary V. Then the test is applied to a 2×2 contingency table as Table 1.

The original χ^2 test statistics is given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

In our case, it is used to test of independence between one high-level semantic category c and low-level visual descriptor $word$, the null and alternative hypotheses are:

H_0 : $word$ is independent of c_i

H_1 : $word$ is depend on c_i

When $i = 1$, E_1 means the expected times for $word$ to show in label c_i images:

$$E_1 = (a+c)\frac{(a+b)}{n} \tag{5}$$

The $i = 1$ variable in Equation 4 can be expressed as follow:

$$D_1 = \frac{(a - E_1)^2}{E_1} \tag{6}$$

Then χ^2 can be formulated as:

$$\chi^2(c_i, word) = \frac{n(ad-bc)^2}{(a+c)(a+b)(b+d)(c+d)} \tag{7}$$

When the χ^2 value between c_i and $word$ is larger than the threshold (ith value of the category), we will reject the null hypothesis H_0 and accept the alternative hypothesis H_1. By computing all the χ^2 value between one category and all visual vocabulary, we can select the best visual word that adapted to this category.

4. DISCRIMINATIVE DICTIONARY LEARNING

In all the above discussion, we have assumed that the dictionary is given first. A simple way to generate the dictionary is to use clustering based method such as K-Means algorithm. According to our experimental results in Section 5, the dictionary generated by K-Means can satisfy our requirement. Then we use the proposed small cycle in Figure 2 to relearning the dictionary and the condition in Section 3 to select the universal and adapted vocabularies. When a new category is added to system, the condition in Subsection 3.2 is used to select the adapted visual vocabulary for it. This enables our approach to gain the ability of incremental learning. To elaborate, the dictionary trained by K-Means clustering is used to initialize V. Then the IG and χ^2 test is computed on all $word \in V$ and $c_i \in C$. This step selects the discriminative words from V. Then the V is retrained on sampling descriptors F and get an optimal V. The above process is illustrated in Algorithm 1.

Algorithm 1 Discriminative dictionary learning

Input: $X \in \mathbb{R}^{K \times N}, L \in \mathbb{R}^{1 \times N}, V_{init} \in \mathbb{R}^{D \times K}, F \in \mathbb{R}^{D \times M}$ ▷ F : sampling local features

Output: $V_{opt} \in \mathbb{R}^{D \times K_{opt}}$

$IG \leftarrow 1 \times K$ zero vector
$C \leftarrow unique(L)$ ▷ C : number of classes
$CS \leftarrow K \times C$ zero matrix
for all $word \in V_{init}$ **do**
  $IG(word) \leftarrow H(L) - H(L|W)$
  for each $c_i \in C$ **do**
    $CS(word, c_i) = \chi^2(c, word)$
  end for
end for
$I_{uni} \leftarrow Find(IG > 0.1)$
for each $c_i \in C$ **do**
  $S_{idx} \leftarrow Find(CS(:, c_i) > threshold(ith))$
  $I_{adapt} \leftarrow [I_{adapt}; S_{idx}]$
end for
$idx \leftarrow unique([I_{adapt}; I_{uni}])$
$V_{opt} \leftarrow V_{init}(:, idx)$ ▷ V_{opt} : selected words
$V_{opt} \leftarrow$CLUSTER(F, V_{opt}) ▷ call clustering function
return V_{opt}

procedure CLUSTER(F, V_{opt}) ▷ retraining V_{opt}
  Clustering features F with initial V_{opt}
end procedure

5. EXPERIMENT RESULTS

In this section, we report results on two widely used datasets: 15 Scenes [19] and Indoor Scenes [20]. We performed all processing in grayscale, even when color images are available. All experiments were repeated eight times with different

Table 2. Comparison of accuracy with K visual words

Dictionary Size	50	100	200	400
Lazebnik[10]	74.5(0.5)	75.4(0.5)	79.1(0.3)	81.4(0.5)
Van Gemert[13]	62.5(0.5)	67.5(0.4)	71.2(0.3)	72.5(0.2)
Zhen[14]	70.5(0.5)	71.4(0.4)	72.5(0.3)	73.0(0.3)
OURS	**75.0(0.5)**	**77.2(0.4)**	**80.0(0.3)**	**82.5(0.5)**

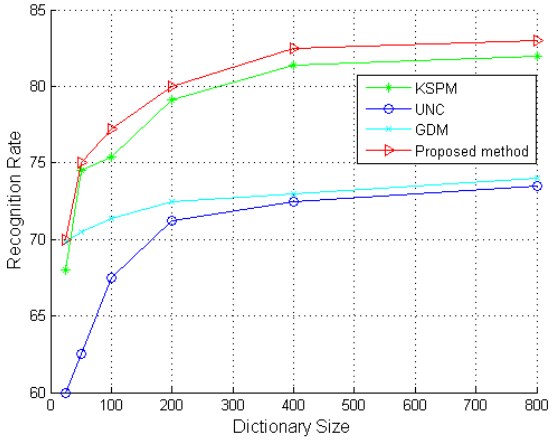

Fig. 3. Comparison for 15 Scenes

randomly selected train and test images, then the mean and standard deviation of per-class recognition rates were recorded for each run. The images were resized with maximum side 500 pixels. As for the image features, the image patches were densely sampled from each image with step size 8 pixels and side length 16 pixels, and SIFT descriptors were adopted with grid size 4×4 to form the final 128 dimensional feature vectors. The features used for training dictionary were randomly sampled from training images (10 images per category) and the amount of them was about $100,000$. The initial dictionary size was set to twice over the final one. K-Means clustering was run to get the initial dictionary. Spatial pyramid matching representation was embeded in the pooling step (the images were split into three scales $l = 0, 1, 2$, each of which had $1, 4, 16$ segments and the final histogram concatenated the 21 segments). One against all multi-class classification strategy was adopted. VLFeat [21] and LIBSVM package [22] were used.

The dictionary relearning step in this paper run with the parameters as follows: threshold for information gain was set to 0.1. Because of the changeful χ^2 value, the tenth highest χ^2 values for per category were set as threshold for adapted vocabulary selecting. Unlike the BoW in text domain, the visual words had many overlapping part between different categories. So the size of adapted vocabulary was about $0.5 \times i_{th} \times c_{num}$ for each loop. The times for running the dictionary relearning and index reassignment loop is depend on the initial value and final size of the visual dictionary. For the initial final value pair $(400, 200)$, the vocabulary selecting

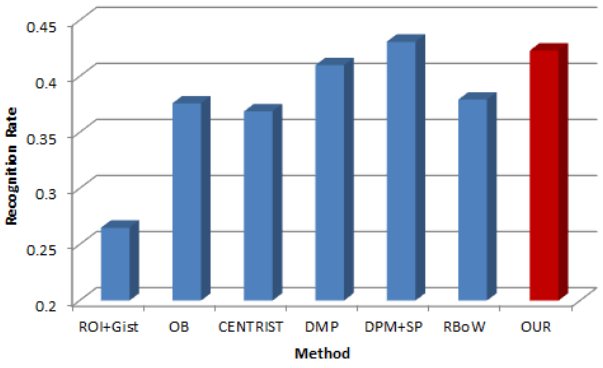

Fig. 4. Indoor Scenes Performance

and index reassignment step run for 5 times to achieve it.

Figure 3 shows the comparison of the results from proposed framework and three other methods: the state-of-the-art spatial pyramid matching (KSPM) method [10], the code word uncertainty (UNC) method [13] and the combining generative and discriminative model (GDM) method [14]. For a fair comparison, we plot the graphs of recognition rate versus dictionary size.

To prove the robustness of our framework, we also test it on Indoor Scenes, compared with 6 other methods: ROI+Gist [20], Object bank [23], CENTRIST [9], depth embedded pooling [24], DPM [25] and RBow [26] method. As seen in Figure 4, the proposed framework achieved competitive 42.31% recognition rate.

As shown in results, our approach outperforms several other methods. That was caused by selecting the right words into the dictionary. At first, the initial dictionary size is a large one, then the visual vocabularies which bring a lot of information were selected. With the dictionary relearning process run, the dimension of the dictionary is reduced without decaying the recognition rate. Meanwhile, the index reassignment step tries to capture all the information with the discriminative visual vocabularies. These framework help to achieve optimal performance under the same setting of dictionary size.

6. CONCLUSION

The proposed framework improves the discriminative power of the traditional Bag-of-Features model. Through discarding the useless vocabulary of the visual dictionary, the dimension of the dictionary can be reduced without decreasing the system accuracy. Meanwhile, the index reassignment step enable the framework to get a dictionary that is best fit for the classification task. When a new category is added, low computational cost is needed by selecting a few visual vocabularies for it. Future works may test our framework on a soft-assignment mechanism like sparse coding.

7. REFERENCES

[1] Wei Jiang, Guihua Er, Qionghai Dai, and Jinwei Gu, "Similarity-based online feature selection in content-based image retrieval," *IEEE Trans. IP*, vol. 15, no. 3, pp. 702–712, March 2006.

[2] Amirhossein Habibian, Koen EA van de Sande, and Cees GM Snoek, "Recommendations for video event recognition using concept vocabularies," in *Proc. of ICMR*. ACM, 2013, pp. 89–96.

[3] J. Sivic and A Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proc. of ICCV*. IEEE, Oct 2003, vol. 2, pp. 1470–1477.

[4] David G Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[5] Jianchao Yang, Kai Yu, Yihong Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. of CVPR*. IEEE, June 2009, pp. 1794–1801.

[6] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *Proc. of CVPR*. IEEE, June 2010, pp. 3360–3367.

[7] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. of CVPR*. IEEE, June 2010, pp. 2559–2566.

[8] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[9] Jianxin Wu and J.M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Trans. PAMI*, vol. 33, no. 8, pp. 1489–1501, Aug 2011.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of CVPR*. IEEE, 2006, vol. 2, pp. 2169–2178.

[11] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proc. of ICCV*. IEEE, Oct 2005, vol. 2, pp. 1458–1465.

[12] Bruno A Olshausen et al., "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[13] J.C. Van Gemert, C.J. Veenman, A W M Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. PAMI*, vol. 32, no. 7, pp. 1271–1283, July 2010.

[14] Zhen Li, Kim-Hui Yap, and Xiao-Ming Chen, "Beyond bag of words: Combining generative and discriminative models for natural scene categorization," in *Proc. of ICASSP*. IEEE, May 2011, pp. 965–968.

[15] J. Winn, A Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. of ICCV*. IEEE, Oct 2005, vol. 2, pp. 1800–1807.

[16] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Trans. PAMI*, vol. 30, no. 7, pp. 1243–1256, July 2008.

[17] Zhuolin Jiang, Zhe Lin, and L.S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Trans. PAMI*, vol. 35, no. 11, pp. 2651–2664, Nov 2013.

[18] Yue Gao, Rongrong Ji, Wei Liu, Qionghai Dai, and Gang Hua, "Weakly supervised visual dictionary learning by harnessing image attributes," *IEEE Trans. IP*, vol. 23, no. 12, pp. 5400–5411, Dec 2014.

[19] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. of CVPR*. IEEE, June 2005, vol. 2, pp. 524–531.

[20] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. of CVPR*, June 2009, pp. 413–420.

[21] Andrea Vedaldi and Brian Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. of ICMM*. ACM, 2010, pp. 1469–1472.

[22] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Trans. IST*, vol. 2, no. 3, pp. 27, 2011.

[23] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. of NIPS*, 2010, pp. 1378–1386.

[24] Zhen Zhou, Yongzhen Huang, Liang Wang, and Tieniu Tan, "Depth-embedded multiple pooling for image classification," in *Proc. of ICIP*, Sept 2013, pp. 4335–4339.

[25] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proc. of ICCV*, Nov 2011, pp. 1307–1314.

[26] S.N. Parizi, J.G. Oberlin, and P.F. Felzenszwalb, "Reconfigurable models for scene recognition," in *Proc. of CVPR*, June 2012, pp. 2775–2782.