Reference-Based Scheme Combined With K-SVD for Scene Image Categorization

Qun Li, *Student Member, IEEE*, Honggang Zhang, *Senior Member, IEEE*, Jun Guo, Bir Bhanu, *Fellow, IEEE*, and Le An

***Abstract*—A reference-based algorithm for scene image categorization is presented in this letter. In addition to using a reference-set for images representation, we also associate the reference-set with training data in sparse codes during the dictionary learning process. The reference-set is combined with the reconstruction error to form a unified objective function. The optimal solution is efficiently obtained using the K-SVD algorithm. After dictionaries are constructed, Locality-constrained Linear Coding (LLC) features of images are extracted. Then, we represent each image feature vector using the similarities between the image and the reference-set, leading to a significant reduction of the dimensionality in the feature space. Experimental results demonstrate that our method achieves outstanding performance.**

***Index Terms*—Image analysis, image classification, dictionary learning, pattern recognition.**

I. Introduction

BAG-OF-WORDS (BoW) representation combined with spatial pyramid matching (SPM) [1] has become one of the most popular methods for representing image content and has been successfully applied to object categorization. To improve the scalability, researchers have focused on obtaining nonlinear feature representations that work better with linear classifiers, e.g., [2], [3]. In particular, Yang *et al.* [3] proposed the sparse coding (SC) method where SC was used instead of vector quantization (VQ) to obtain nonlinear codes. Yu *et al.* [4] proposed to encourage the SC to be local by using the local coordinate coding (LCC) mechanism. Wang proposed a simple but effective coding scheme called Locality-constrained Linear Coding (LLC) [5]. With linear classifier, the LLC approach performs significantly better than the traditional nonlinear SPM, achieving the *state-of-the-art* performance on several benchmarks.

The performance of the above methods relies on the quality of the dictionary (codebook). Traditionally, dictionary was usually constructed in an unsupervised manner such as k-means to cluster the descriptor vectors of patches sampled either densely or sparsely from a set of training images. Although this kind of methods works well for texture analysis on images containing only a few homogeneous regions, it is not guaranteed to obtain an optimal codebook for a complicated image category, such as natural scenes. Gemert *et al.* [6] showed that this kind of methods of codebook construction have two drawbacks: codeword uncertainty and codeword plausibility, and proposed the kernel codebooks to improve categorization performance. [7] employed the entire set of training samples as the dictionary for discriminative sparse coding, and achieved impressive performances on face recognition. Many algorithms [5], [8] have been proposed to efficiently learn an over-complete dictionary that enforces a discriminative criterion. In order to scale to large training sets, several small-sized dictionary learning methods have been developed by [7], [9], [10]–[12]. In [10], a dictionary learning algorithm, K-SVD, is introduced to learn an over-complete dictionary and this method has been applied to infill missing pixels and to image compression. Discriminative K-SVD algorithm (D-KSVD) presented in [11] unified the dictionary and classifier learning processes. [12] proposed a supervised algorithm to learn a compact and discriminative dictionary for sparse coding.

In this paper, we propose a novel reference-based descriptor to represent images. To support this representation, we use sparse coding to learn the dictionaries and adopt the K-SVD algorithm to efficiently obtain the optimal solutions. After the representation, the feature dimensionality is remarkably reduced, and the classification accuracy is significantly increased. The main contributions of our method compared to other methods are reference-combined dictionary learning and reference-based representation for image classification. The computational complexity is bounded by K-SVD [10].

The remainder of the letter is organized as follows: Section II presents the reference-combined objective function and the K-SVD algorithm for simultaneously learning a dictionary with a reconstructive criterion. Section III introduces the reference-based method for image classification. Experimental results and analysis on three widely used datasets are reported in Section IV. Finally Section V concludes the letter and discusses future work.

II. Reference-Combined Dictionary Learning

The overall classification process is illustrated in Fig. 1.

Manuscript received August 01, 2012; revised October 19, 2012; accepted November 14, 2012. Date of publication November 21, 2012; date of current version November 29, 2012. This work was supported in part by National Natural Science Foundation of China under Grants 61005004 and 61175011, the Chinese 111 program Advanced Intelligence and Network Service under Grant B08004 and by a key project of the Ministry of Science and Technology of China under Grant 2011ZX03002-005-01. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Qian Du.

Q. Li was with the Pattern Recognition and Intelligent System Laboratory (PRIS), Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China. She is now with the Center for Research in Intelligent Systems (VISLab), University of California, Riverside, CA 92521 USA (e-mail: liqun@bupt.edu.cn).

H.G. Zhang and J. Guo are with Pattern Recognition and Intelligent System Laboratory (PRIS), Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China (e-mail: zhhg@bupt.edu.cn; guojun@bupt.edu.cn).

B. Bhanu and L. An are with VISLab, University of California, Riverside, CA 92521 USA (e-mail: bhanu@ee.ucr.edu, lan004@ucr.edu).

Digital Object Identifier 10.1109/LSP.2012.2228852

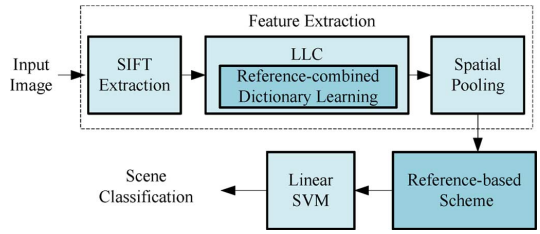

Fig. 1. An overview of the proposed scene categorization algorithm. This letter concentrates on the dictionary learning and reference-based processing.

A. Dictionary Learning for Reconstruction and Sparse Coding

Let X be a set of P-dimensional signals, i.e., $X = [x_1, \ldots, x_N \in \mathbb{R}^{P\times N}]$. Learning a reconstructive dictionary with M entries for sparse representation of X can be accomplished by solving the following constrained least square fitting problem:

$$\langle B, C\rangle = \arg\min_{B,C} \|X - BC\|_2^2, \quad s.t. \|c_i\|_0 \le T, \forall i, \quad (1)$$

where $B = [b_1, \ldots, b_M] \in \mathbb{R}^{P\times M}$ ($M > P$, making the dictionary over-complete) is the learned dictionary, $C = [c_1, \ldots, c_N] \in \mathbb{R}^{M\times N}$ is the sparse codes of X, and T is a sparsity constraint factor. The term $\|X - BC\|_2^2$ calculates the reconstruction error.

The K-SVD algorithm [10] is an iterative approach to minimize the energy in (1) and it learns a reconstructive dictionary for sparse representations of signals. It is highly efficient and works well in applications such as image restoration and compression. Given B, sparse coding computes the sparse representation C of X by solving:

$$C = \arg\min_{C} \|X - BC\|_2^2, \quad s.t. \|c_i\|_0 \le T, \forall i. \quad (2)$$

B. Reference-Combined Dictionary Learning

In this paper, a reference-based method using a reference-set for image representation is proposed and the reference-set is also coded by the dictionary, so we combine its construction error with the original object function to form a unified objective function. The objective function for dictionary construction is defined as:

$$\langle B, C, S\rangle = \arg\min_{B,C,S} \|X - BC\|_2^2 + \mu\|R - BS\|_2^2, \quad s.t. \ \|c_i\|_0 \le T_1, \|s_i\|_0 \le T_2, \forall i, \quad (3)$$

where μ controls the relative contribution between training data reconstruction and reference-set reconstruction, and $R \in \mathbb{R}^{P\times L}$ is the reference-set signals, $S \in \mathbb{R}^{M\times L}$ is the sparse codes of R. The reference-set is explained in Section III.

C. Optimization

We adapt the efficient K-SVD algorithm to find the optimal solution for all parameters simultaneously of our method, and we denote it as R-KSVD. Equation (3) can be rewritten as:

$$\langle B, C, S\rangle = \arg\min_{B,C,S} \|(X, \sqrt{\mu}R) - B(C, \sqrt{\mu}S)\|_2^2, \quad s.t. \ \|c_i\|_0 \le T_1, \|s_i\|_0 \le T_2, \forall i. \quad (4)$$

Fig. 2. The flowchart of the reference-based classification scheme.

Let $X_{\text{new}} = (X, \sqrt{\mu}R), C_{\text{new}} = (C, \sqrt{\mu}S)$. We relax the optimization of (4) as

$$\langle B, C_{\text{new}}\rangle = \arg\min_{B,C_{\text{new}}} \|X_{\text{new}} - BC_{\text{new}}\|_2^2, \quad s.t. \ \|c_{\text{new}_i}\|_0 \le T, \forall i. \quad (5)$$

This is exactly the problem that K-SVD [10] solves. Following K-SVD, b_m and its corresponding coefficients, the m-th row in C, denoted as c_R^m, are updated at a time. Let $E_m = (X - \sum_{j\neq m} b_j c_R^j)$, and $\tilde{c}_R^j, \tilde{E}_m$ denote the result of discarding the zero entries in c_R^j and E_m, respectively. b_m and $\tilde{c}_R^m$ can be computed by solving the following problem:

$$\langle b_m, \tilde{c}_R^m\rangle = \arg\min_{b_m, \tilde{c}_R^m} \|\tilde{E}_m - b_m\tilde{c}_R^m\|_F^2. \quad (6)$$

An SVD operation is performed for $\tilde{E}_m$, i.e., $U\Sigma V^t = \text{SVD}(\tilde{E}_m)$. Then b_m and $\tilde{c}_R^m$ are computed as:

$$b_m = U(:, 1), \quad \tilde{c}_R^m = \Sigma(1,1)V(:, 1). \quad (7)$$

Finally, $\tilde{c}_R^m$ is used to replace the non-zero values in c_R^m.

The parameter B_0 for R-KSVD needs to be initialized. We employ several iterations of K-SVD within each class and then combine all the outputs (i.e., dictionary items learning from each class) of each K-SVD. The label of each dictionary item b_m is then initialized based on the class it corresponds to and will remain fixed during the entire dictionary learning process, although b_m is updated during the learning process. Dictionary elements are uniformly allocated to each class with the number of the elements proportional to the dictionary size [12].

III. Reference-Based Scheme for Classification

Fig. 2 shows the flowchart of the reference-based scheme. All the image features including the reference-set used for reference-based scheme are LLC features which are generated using the dictionary trained by the proposed method introduced in the last section. The LLC [5] coding method uses the following criteria:

$$\arg\min_{C} \|X - BC\|_2^2 + \lambda\|D \odot C\|_2^2, \quad s.t. \ 1^T c_i = 1, \forall i, \quad (8)$$

where $\odot$ is the element-wise multiplication, and $D \in \mathbb{R}^{M\times N}$ denotes the locality adaptor matrix. Specifically,

$$D = \exp\left(\frac{\text{dist}(X, B)}{\sigma}\right), \quad (9)$$

where $\text{dist}(X, B)$ is the Euclidean distance. LLC is easy to compute and gives superior image classification performance than

many existing approaches. The LLC features are saved for the following reference-based scheme:

1) In the first step, we select $n = 30$ images per class randomly from different datasets to assemble a reference-set.
2) Then given the probe image, the similarity between it and each image in the reference-set is calculated by

$$S_p^{r_i} = 1 - F\left(d_p^{r_i}; k\right) = 1 - \frac{\gamma\left(\frac{k}{2}, \frac{d_p^{r_i}}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}, \quad (r = \alpha, \beta, \ldots) \tag{10}$$

where $d_p^{r_i}$ is the χ^2 distance of the probe image p and the reference-set image r_i (r_i is the i-th image of reference-subset r), $F(d_p^{r_i}; k)$ is its cumulative distribution function, and k is a positive integer that specifies the number of degrees of freedom. $\Gamma(k/2)$ denotes the Gamma function, $\gamma(k/2, d_p^{r_i}/2)$ is the lower incomplete Gamma function. From this step, a similarity matrix is obtained as shown in Fig. 2, $S_p^{\alpha_i}$ or $S_p^{\beta_i}$ denotes the similarity between the probe image p and the i-th image in class α or β of reference-set.
3) In this step, the dimensionality of the similarity matrix is reduced by averaging in row to generate the final representation of the image denoted as $[F_p^{\alpha}, F_p^{\beta}, \ldots]^T$ for classification according to (11). So if the number of the reference-subsets is 300, the dimension of the final feature should be 300.

$$F_p^r = \text{mean}\left(\sum_{i=1}^{n} S_p^{r_i}\right), \quad (r = \alpha, \beta, \ldots). \tag{11}$$

4) Finally, we normalize the represented feature according to (12) and use linear SVM as the classifier.

$$F_p = F_p / \text{norm2}(F_p). \tag{12}$$

This algorithm is based on reference-set, and more importantly, the same reference-set is used on various image databases and does not need to be replaced.

IV. Experimental Results

In this section, we evaluate our approach on three widely used databases: Caltech-101 [13], fifteen scene categories [14], and Pascal VOC2007 [15]. The proposed method is compared with several *state-of-the art* methods. We use only a single descriptor, the SIFT descriptors of 16×16 pixel patches computed over a grid with a spacing of 8 pixels, and $4\times4, 2\times2, 1\times1$ sub-regions for LLC, throughout all the experiments. We refer to the publicly available software packages of [5], [12] to set $\mu = 4$ and $\lambda = 1 \times 10^{-4}$.

Dictionary size for Caltech101 and VOC 2007 is 1024, and for fifteen scene categories the sizes are 200 and 400. In our setup, we use linear SVM as the classifier. We partition the whole dataset of Caltech-101 into 30 training images per class and the rest for testing images, and 100 training images per class for the Scene 15. The reference-set is collected by 30 images per class from 392 different classes by randomly selected in fifteen scene categories, Caltech101, Caltech-256 [16], and Pascal VOC2007. The dimension of the final image feature is reduced significantly compared to original features as presented in Table I.

TABLE I
Feature Dimensionality Compare With 200, 400, and 1024 Bases Dictionaries

Feature	200	400	1024
LLC [5]	4200	8400	21504
Ours	**392**	**392**	**392**

TABLE II
Image Classification Results on Caltech101 Database

Classification Method	Classification Accuracy(%) 5	10	15	20	25	30
Lazebnik [1]	-	-	56.4	-	-	64.6
Gemert [6]	-	-	-	-	-	64.16
Yang [3]	-	-	67.0	-	-	73.2
Wang [5]	51.15	59.77	65.43	67.74	70.16	73.44
K-SVD [10]	49.8	59.8	65.2	68.7	71.0	73.2
D-KSVD [11]	49.6	59.5	65.1	68.6	71.1	73.0
LC-KSVD1 [12]	53.5	61.9	66.8	70.3	72.1	73.4
LC-KSVD2 [12]	54.0	63.1	67.7	70.5	72.3	73.6
Ours	**72.5**	**77.9**	**79.7**	**81.4**	**82.3**	**83.0**

TABLE III
Image Classification Results on Scene15 Database

Classification Method	Accuracy(%) 200	400	Classification Method	Accuracy(%) 200	400
Lazebnik [1]	74.5	74.8	Yang [3]	-	80.28
Gemert [6]	74.3	76.67	Wang [5]	78.5	80.2
Ours	**82.8**	**83.2**			

We repeat the experiments 10 times with different random selections of the reference-set and different random splits of the training and testing images to obtain stable results. The final recognition rates are reported as the average of each run. All experiments are conducted on a Dell D01X computer with 6 G memory and 3.2 Ghz Quad Core CPU.

A. Caltech-101

Our first set of experiments is on the Caltech-101 database [13], which contains 9144 images in 101 classes. Each category has 31 to 800 images, and most images are of medium resolution, i.e., about 300×300 pixels.

We compare our result with K-SVD [10], D-SVD [11], LC-KSVD [12] and other *state-of-art* approaches [1], [6], [3], [5]. As shown in Table II, our approach remarkably outperforms all the competing approaches with nearly 10% increase compared to the next best result. Moreover, the classification accuracy with 5 training images per class is still comparable with the other methods. The average computation time of classifying one test images is 0.77 ms. In our evaluation, totally 16 classes achieve 100% classification accuracy with 30 training images per class.

B. Scene Category Recognition

The second dataset is composed of fifteen scene categories [14]. The number of images per category varies from 200 to 400, and the average image size is 300×250 pixels. The major sources of the pictures in the dataset include the COREL collection, personal photographs, and Google image search. It is one of the most complete scene category dataset used in the literature.

TABLE IV
IMAGE CLASSIFICATION RESULTS ON PASCAL VOC 2007 DATABASE

Object Class	aero	bicyc	bird	boat	bottle	bus	car	cat	chair	cow
LLC [5]	74.8	65.2	50.7	70.9	28.7	68.8	78.5	61.7	54.3	48.6
Best PASCAL'07 [15]	77.5	63.6	56.1	71.9	**33.1**	60.6	78.0	58.8	53.5	42.6
Ours	**79.0**	**72.8**	**57.9**	**72.6**	29.9	**71.8**	**81.9**	**65.1**	**61.6**	**53.5**

Object Class	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**Average**
LLC [5]	51.8	44.1	76.6	66.9	83.5	30.8	44.6	**53.4**	78.2	53.5	59.3
Best PASCAL'07 [15]	54.9	**45.8**	**77.5**	64.0	85.9	36.3	44.7	50.9	**79.2**	53.2	59.4
Ours	**64.6**	44.8	71.4	**69.7**	**88.8**	**38.9**	**45.3**	52.9	78.4	**59.3**	**63.0**

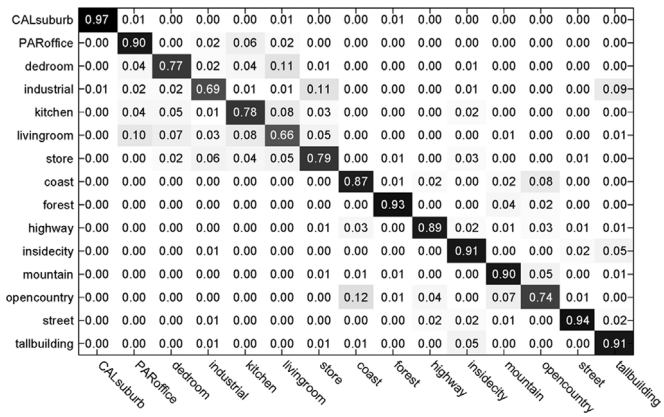

Fig. 3. Confusion table of Scene15 dataset using 400 dictionary, the grid detector and patch based representation. The average performance is 83.2%.

Table III shows that our method yields the best results with 200 bases and 400 bases. A closer look at the confusion table (Fig. 3) reveals that the highest block of errors occurs among the four categories: livingroom, industrial, opencountry and bedroom. The average computation time of classifying one test images is 0.31 ms.

C. Pascal VOC 2007

The PASCAL 2007 dataset consists of 9,963 images in 20 classes. The dataset is a challenging one because all the images are daily pictures got from Flicker where the size, viewing angle, illumination, appearances of objects and their poses vary greatly, with frequent occlusions. The classification performance criterion used is the standard metric used by PASCAL challenge [15]. It computes the area under the Precision/Recall curve, and the higher the score, the better the performance.

Table IV lists our scores for all 20 classes in comparison with the best performance of the 2007 challenge [15], as well as another recent result in [5]. As seen from Table IV, our reference-based method can achieve the best performance in most classes.

V. CONCLUSION

In this paper, we present a novel reference-based approach which combines reference-set with dictionary learning and image categorization. The approach uses a reference-set to represent the images. We perform experiments on various image databases to demonstrate the benefits of the proposed method. Experimental results show that the proposed method increases the classification accuracy significantly while obtains higher efficiency compared to the *state-of-the-art* methods. In the future work, we will combine the discriminative sparse-code error and the classification error with reconstruction error to form a unified objective function for dictionary leaning based on reference-set and explore the optimization methods of the reference-set.

ACKNOWLEDGMENT

The authors would like to thank Mehran Kafai and Zhen Qin for helpful discussions.

REFERENCES

[1] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006, pp. 2169–2178.
[2] X. Zhou, N. Cui, Z. Li, F. Liang, and T. Huang, "Hierarchical gaussianization for image classification," in *Proc. ICCV*, 2009, pp. 1971–1977.
[3] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. CVPR*, 2009, pp. 1794–1801.
[4] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. NIPS*, 2009.
[5] J. W. *et al.*, "Locality- constrained linear coding for image classification," in *Proc. CVPR*, 2010, pp. 3360–3367.
[6] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Proc. ECCV*, 2008, pp. 696–709.
[7] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 19–60, Jan. 2010.
[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. CVPR*, 2008, pp. 1–8.
[10] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: Design of dictionaries for sparse representation," in *Proc. SPARS*, 2005, pp. 9–12.
[11] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. CVPR*, 2010, pp. 2691–2698.
[12] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. CVPR*, 2011, pp. 1697–1704, IEEE.
[13] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *IEEE CVPR Workshop on Generative-Model Based Vision*, 2004, p. 178.
[14] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. CVPR*, 2005, pp. 524–531.
[15] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.
[16] G. Griffin, A. Holub, and P. Perona, Caltech-256 Object Category Dataset California Institute of Technology, , Tech. Rep. 7694, 2007 [Online]. Available: http://authors.library.caltech.edu/7694