REPRESENTATIVE REFERENCE-SET AND BETWEENNESS CENTRALITY FOR SCENE IMAGE CATEGORIZATION

Qun Li[1,2], *Zhen Qin*[2*], *Lunshao Chai*[1,2], *Honggang Zhang*[1], *Jun Guo*[1], *Bir Bhanu*[2]

[1]Pattern Recognition and Intelligent System Laboratory,
Beijing University of Posts and Telecommunications, Beijing, China
[2]University of California, Riverside, CA, USA
{ liqun, zhhg, guojun, chailunshao }@bupt.edu.cn, zqin001@cs.ucr.edu, bhanu@ee.ucr.edu

ABSTRACT

Reference-based image classification approach introduces a reference-set for both image representation and dictionary learning. It significantly reduces the dimensionality of represented images and shows outstanding performance even with randomly selected reference images and simple distance measure. In this paper, we improve upon existing work with two major contributions. First, we show that a more representative reference-set contributes to better classification accuracy. To this end, we carefully adapt the K-means clustering algorithm in the feature space to select a distinguished reference-set. Second, in the image classification process, we propose to represent each image by measuring its betweenness centrality in a social network composed of the representative reference-set in each class, leading to a more coherent distance measure that considers the overall connectivity between the probe image and the reference-set. Extensive experiment results demonstrate that our proposed scheme achieves better performance than existing methods.

Index Terms— Scene categorization, reference-based scheme, K-means, social network, betweenness

1. INTRODUCTION

Image classification is a fundamental problem in computer vision and has attracted a lot of attention in recent years. Current research converges on leveraging bag-of-words (BoW) representation combined with spatial pyramid matching (SPM) [1]. Such scheme provides an effective way of capturing image statistics for natural scene classification and reports *state-of-the-art* performance. Recently, it has been shown that combining the BOW modules with sparse representation is very effective and has been successfully applied to object categorization by many researchers, *e.g.* the sparse coding (SC) method [2] and local coordinate coding (LCC) [3]. In particular, Wang *et al.* [4] present a simple but useful coding scheme called Locality-constrained Linear Coding (LLC) which replaces vector quantization (VQ) and acquires nonlinear codes. With linear classifier, the presented approach performs significantly better than the traditional nonlinear SPM. Meanwhile, Li *et al.* [5] propose a novel reference-based scheme for scene image categorization. They select a set of images to form a reference-set and associate them with training data in sparse codes during the dictionary learning process. In the classification stage, they further represent each image feature vector using the similarities between the image and the reference-set, leading to a significant reduction of the dimensionality of represented feature and achieve outstanding performance even with a randomly selected reference-set and the simple Chi-square distance between an image and the reference-set.

In this work, we show that a more representative reference-set and a more carefully selected similarity measure between the probe image and the reference-set can further improve the classification accuracy. We optimize the reference-set selection procedure by carefully adopting K-means clustering in the feature space, and combine the betweenness centrality from social network with existing similarity measure to measure the overall connectivity between the probe image and the reference-set. Betweenness centrality is a measure of a node's centrality in a network, which equals to the number of

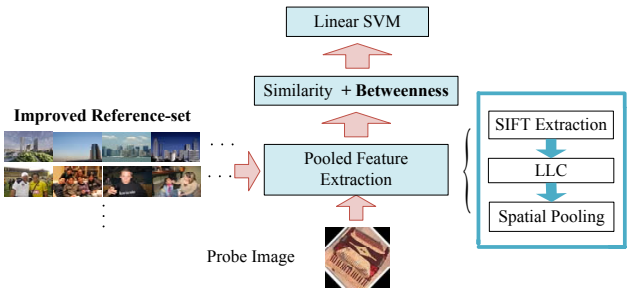

Fig. 1: An overview of improved reference-based scheme

*Zhen and Qun contributed equally for this paper. This work was partially supported by National Natural Science Foundation of China under Grant No.61273217, 61175011 and 61171193, the 111 project under Grant No.B08004, and the Fundamental Research Funds for the Central Universities.

shortest paths from all vertices to all others that pass through that node [6]. Betweenness centrality can be regarded as a measure of the extent to which a node has control over the information flowing among others.

The remainder of the paper is organized as follows: Section 2 presents details about the representative reference-set selection scheme based on the adapted K-means and how we incorporate betweenness centrality in the classification stage. Experimental results and analysis on three widely used datasets are reported in Section 3. Section 4 concludes the paper.

2. IMPROVED REFERENCE-BASED SCHEME FOR SCENE CATEGORIZATION

2.1. An overview of our improved reference-based scheme

The overall improved reference-based classification process is illustrated in Fig. 1. All the image features including the reference-set used for our reference-based scheme are LLC features, which are generated using the dictionary trained by the reference-combined dictionary learning method [5].

A reference-based scheme for scene image categorization adopts a reference-set composed by some images for images representation by using the similarities between the probe image and the reference-set. The reference-set is also associated to generate the dictionary by following the reference-combined dictionary learning method [5] in sparse coding, which is defined as Eq. (1),

$$< B, C, S >= \underset{B,C,S}{\operatorname{argmin}} ||X - BC||_2^2 + \mu||R - BS||_2^2, \quad s.t. ||c_i||_0 \leq T_1, ||s_i||_0 \leq T_2, \forall i, \quad (1)$$

where X denotes a set of input signal, $X = [x_1, \ldots, x_N] \in \mathbb{R}^{P\times N}$, $B = [b_1, \ldots, b_M] \in \mathbb{R}^{P\times M}$ denotes the learned dictionary, $C = [c_1, \ldots, c_N] \in \mathbb{R}^{M\times N}$ denotes the sparse code of X, and T_i is the sparsity constraint factor, $R \in \mathbb{R}^{P\times L}$ is the reference-set signal, $S \in \mathbb{R}^{M\times L}$ is the sparse code of R, and μ balances the weights between training data reconstruction and reference-set reconstruction. The optimal solution is efficiently obtained by the K-SVD algorithm with its effectiveness verified in [5]. Given the learned dictionary, our improved reference-based scheme is described as follows:
(1) Select m images per class from different datasets to assemble a reference-set using adapted K-means clustering, which will be introduced in section 2.2.
(2) Given the probe image, calculate the similarity between it and each image in the reference-set as Eq. (2),

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

where $d_p^{r_i}$ presents the χ^2 distance of the probe image p and the i-th image of the reference-subset (class) 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. $\gamma(k/2, d_p^{r_i}/2)$ denotes the lower incomplete Gamma function, $\Gamma(k/2)$ is the Gamma function.
(3) Calculate the betweenness centrality b_p^r after adding the probe image to a social network composed of images in reference-subset r, as detailed in section 2.3.
(4) Concatenate the betweenness centrality calculated in each class with the average similarity per class to generate the final representation of the probe image for classification, accoring to Eq. (3) - Eq. (4).

$$f_p^r = mean(\sum_{i=1}^{m} S_p^{r_i}), (r = \alpha, \beta, \ldots). \quad (3)$$

$$F_p = (f_p; b_p)/norm2(f_p; b_p). \quad (4)$$

(5) Finally, perform classification using the linear SVM.

Algorithm 1 Adapted K-means based reference-set optimization algorithm

Input: I, K. % I denotes the input images from different classes, K is the number of images per class of the reference-set.
Output: R. % R is the improved reference-set.

```
nr ← 0.
for each i ∈ [1, n] do
    % n is the number of image classes.
    K-means clustering in I_i;
    sort clusters according to the number of images in per cluster in descending order;
    while nr < K, from high to low do
        if n_cluster ≥ 2 then
            find N and F;   % n_cluster presents the number of images per cluster, N is the nearest neighbor from the cluster center, F is the farthest one.
            R_i(nr) = N; nr = nr + 1;
            if nr < K then
                R_i(nr) = F; nr = nr + 1;
            end if
        else
            R_i(nr) = I_c;   % I_c is the cluster center.
        end if
    end while
end for
```

2.2. Adapted K-means based representative reference-set selection method

Unlike previous works that generate the reference-set randomly, we further optimize the reference-set selection procedure. More specifically, we adapt the K-means clustering method to assemble the representative reference-set. We observe that a more diverse reference-set leads to a more discriminative representation, so we keep both the center image

and the one farthest to the center image in each cluster, if they are available, to form the reference-set. The reference-set optimization method is illustrated in Alg. 1.

2.3. Betweenness centrality as a distance measure

Betweenness centrality is a measure of the centrality of a node in social networks. It reflects the influence a node has over the spread of information through the network. Recently, various methods have been proposed to calculate betweenness and closeness centralities [6]. We propose to use betweenness centrality instead of a direct distance measure for the following reasons: (1) it considers the overall connectivity by measuring how well the probe image fits in a social network composed of the reference images of one class. In this way, it is also an image-to-class measure, instead of a image-to-image measure used before (where each reference image played equal importance). (2) Betweenness centrality can be calculated very efficiently when given the small size of each social network involved (reference images in one class plus the probe image) and *state-of-art* algorithms developed in the social network.

Betweenness centrality can be calculated as follows: suppose that g_i^{st} is the number of geodesic paths from node s to node t that pass through i, and n_{st} is the total number of geodesic paths from s to t. Normally, the betweenness of a node i is calculated as the fraction of shortest paths between node pairs in a network that pass through i. Then the betweenness of node i is

$$b_i = \frac{\sum_{s<t} g_i^{st}}{(1/2)n_{st}(n_{st}-1)}. \quad (5)$$

Betweenness can be calculated for all nodes in time $O(jk)$ for a network with j edges and k nodes [6].

In our case, we first construct a social network for the reference images in each class. Then given a probe image, we add it to existing networks and calculate its betweenness centrality to measure its connectivity to the corresponding class. For each image in the constructed network, we use the dimension of the biggest component in its LLC feature as the node label, and the next 30 biggest components as its friendship nomination data. Thus, we have a measurement of how the probe image fits into a class-specific network by letting it find "friends" that have similar LLC features in a global sense. In the improved reference-based scheme, the whole process is shown as follows:

(1) Construct a social network composed of reference images per class.

(2) Given the probe image, add it to each class-specific network and measure its betweenness centrality using the 31 biggest components in the LLC feature space.

(3) Concatenate betweenness centrality in all classes of the probe image to generate its final betweenness representation feature. In this way, the dimensionality of the feature is bounded by the number of classes in the reference-set.

Table 1: Feature Dimensionality Comparison With 200, 400, And 1024 Bases Dictionaries

Feature	200	400	1024
LLC[6]	4200	8400	21504
Ours	**784**	**784**	**784**

Table 2: Image Classification Results On Caltech101 Dataset

Classification Method	Classification Accuracy(%)					
	5	10	15	20	25	30
Lazebnik[1]	-	-	56.4	-	-	64.6
Gemert[12]	-	-	-	-	-	64.16
Yang[2]	-	-	67.0	-	-	73.2
Wang[4]	51.15	59.77	65.43	67.74	70.16	73.44
K-SVD[13]	49.8	59.8	65.2	68.7	71.0	73.2
D-KSVD[14]	49.6	59.5	65.1	68.6	71.1	73.0
LC-KSVD1[10]	53.5	61.9	66.8	70.3	72.1	73.4
LC-KSVD2[10]	54.0	63.1	67.7	70.5	72.3	73.6
Reference-based[5]	72.5	77.9	79.7	81.4	82.3	83.0
Ours	**72.9**	**78.6**	**80.7**	**82.3**	**83.5**	**84.2**

Table 3: Image Classification Results On Scene15 Dataset

Classification Method	Accuracy(%)		Classification Method	Accuracy(%)	
	200	400		200	400
Lazebnik[1]	74.5	74.8	Yang[2]	-	80.28
Gemert[12]	74.3	76.67	Wang[4]	78.5	80.2
Reference-based[5]	82.8	83.2	**Ours**	**83.4**	**84.2**

3. EXPERIMENTAL RESULTS

We evaluate our approach on three diverse datasets: Caltech-101 [7], fifteen scene categories [8], and Pascal VOC2007 [9], and compare our results with several *state-of-the-art* methods, including the original reference-based method. We use the dense SIFT descriptors of 16×16 pixel patches computed over a grid with a spacing of 8 pixels, and 4×4, 2×2, 1×1 sub-regions for LLC, throughout all the experiments. The dictionary size for Caltech101 and VOC 2007 is 1024, and for fifteen scene categories the size is 200 and 400. We refer to the publicly available software packages of [10] to set $\mu = 4$.

The reference-set is collected by gathering 30 images per class from 392 different classes by the adapted K-means clustering in fifteen scene categories, Caltech101, Caltech-256 [11], and Pascal VOC2007. The dimension of the final image feature is presented in Table 1.

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

Table 4: Image Classification Results On Pascal VOC 2007 Dataset

Object Class	aero	bicyc	bird	boat	bottle	bus	car	cat	chair	cow
LLC[4]	74.8	65.2	50.7	70.9	28.7	68.8	78.5	61.7	54.3	48.6
Best PASCAL'07[9]	77.5	63.6	56.1	71.9	**33.1**	60.6	78.0	58.8	53.5	42.6
Reference-based[5]	79.0	72.8	57.9	**72.6**	29.9	71.8	**81.9**	65.1	61.6	53.5
Ours	**79.8**	**73.4**	**58.0**	**72.6**	32.9	**72.5**	81.5	**67.1**	**61.8**	**54.6**

Object Class	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**Average**
LLC[4]	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[9]	54.9	45.8	**77.5**	64.0	85.9	36.3	44.7	50.9	**79.2**	53.2	59.4
Reference-based[5]	64.6	44.8	71.4	69.7	88.8	38.9	45.3	52.9	78.4	59.3	63.0
Ours	**64.8**	**46.0**	75.5	**71.2**	**89.7**	**40.7**	**48.2**	52.7	79.1	**59.6**	**64.1**

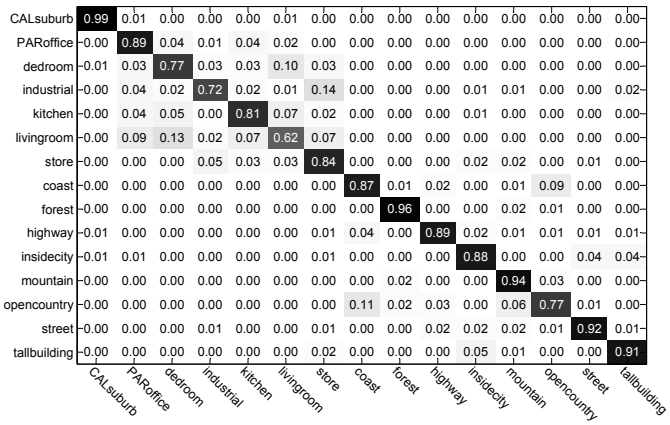

Fig. 2: Confusion table of Scene15 dataset using 400 bases dictionary, the grid detector and patch based representation. The average performance is 84.2%.

3.1. Caltech-101

The Caltech-101 dataset [7] contains 9144 images in 101 classes, the number of images per category varies from 31 to 800. Most images are of medium resolution, i.e., about 300×300 pixels. We partition the whole dataset of Caltech-101 into 5, 10, ..., 25, 30 training images per class and the rest for testing.

We compare our result with the original reference-based scheme and other *state-of-art* approaches [1, 2, 3, 4, 10, 12, 13, 14]. As we can see from Table 2, our approach achieves a 1.2% increase in terms of accuracy over the original reference-based method, while noticeably outperforming all the competing approaches with nearly 10% accuracy increase compared to the former best result.

3.2. Scene Category Recognition

Our second set of experiments is performed on the fifteen scene categories [8], including the COREL collection, personal photographs, and Google image search. Each category has 200 to 400 images, and the average image size is 300×250 pixels. We partition the whole dataset into 100 training images per class and the rest for testing.

As shown in Table 3, our method is superior to all the others with both 200 bases and 400 bases dictionaries. A closer look at the confusion table (Fig. 2) reveals that the highest block of errors occurs among the four categories: livingroom, industrial, opencountry and bedroom.

3.3. Pascal VOC 2007

The PASCAL 2007 dataset is an extremely challenging one which holds of 9,963 images in 20 classes. All the images are daily pictures got from Flicker with different sizes, viewing angles, illuminations. The appearances of objects and their poses vary greatly, with frequent occlusions. We use the standard metric used by PASCAL challenge [9] to evaluate the classification performance. It computes the area under the Precision/Recall curve, and the higher the score, the better the performance.

In Table 4, we list our scores for all 20 classes in comparison with one recent result in [4], as well as the best performance of the 2007 challenge. As seen from Table 4, our reference-based method can achieve the best performance in the most of classes and the highest average rate.

4. CONCLUSION

In this paper, we present a new reference-based scene images categorization approach which is based on adapted K-means clustering for representative reference-set selection and betweenness centrality for distance measure by treating images in the reference-set as social networks. We perform experiments on three widely used image datasets to verify the benefits of the proposed method.

5. REFERENCES

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

[2] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. of CVPR*, 2009, pp. 1794 – 1801.

[3] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. of NIPS*, 2009.

[4] J. Wang and *et al.*, "Locality- constrained linear coding for image classification," in *Proc. of CVPR*, 2010, pp. 3360 – 3367.

[5] Q. Li, H.G. Zhang, J. Guo, B. Bhanu, and L. An, "Reference-based scheme combined with k-svd for scene image categorization," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 67–70, 2013.

[6] A. Mantrach and *et al.*, "The sum-over-paths covariance kernel: A novel covariance measure between nodes of a directed graph.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1112–1126, 2010.

[7] 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.

[8] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. of CVPR*, 2005, pp. 524 – 531.

[9] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*.

[10] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD.," in *Proc. of CVPR*. 2011, pp. 1697–1704, IEEE.

[11] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep. 7694, California Institute of Technology, 2007.

[12] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Proc. of ECCV*, 2008, pp. 696–709.

[13] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: Design of dictionaries for sparse representation," in *Proc. of SPARS*, 2005, pp. 9–12.

[14] Q. Zhang and B.X. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. of CVPR*, 2010, pp. 2691–2698.