Discriminative Reference-Based Scene Image Categorization

Qun LI\(^{(a)}\), Nonmember, Ding XU\(^{(††)}\), Member, and Le AN\(^{(†††)}\), Nonmember

A discriminative reference-based method for scene image categorization is presented in this letter. Reference-based image classification approach combined with K-SVD is approved to be a simple, efficient, and effective method for scene image categorization. It learns a subspace as a means of randomly selecting a reference-set and uses it to represent images. A good reference-set should be both representative and discriminative. More specifically, the reference-set subspace should well span the data space while maintaining low redundancy. To automatically select reference images, we adapt affinity propagation algorithm based on data similarity to gather a reference-set that is both representative and discriminative. We apply the discriminative reference-based method to the task of scene categorization on some benchmark datasets. Extensive experiment results demonstrate that the proposed scene categorization method with selected reference set achieves better performance and higher efficiency compared to the state-of-the-art methods.

**key words:** scene categorization, reference-based scheme, affinity propagation algorithm, discriminative reference-set

1. Introduction

Scene image categorization not only has a noteworthy impact on scene image management, but it also can offer immeasurable assistance to other computer vision problems, such as image retrieval, image completion, face recognition, human activity analysis, object recognition, etc. Bag-of-words (BoW) representation combined with spatial pyramid matching (SPM)\(^{(1)}\) such as the sparse coding (SC) method\(^{(2)}\), local coordinate coding (LCC)\(^{(3)}\) and locality-constrained linear coding (LLC)\(^{(4)}\) has become popular choices for representing image content and has been successfully applied to scene image categorization. In particular, LLC replaces vector quantization (VQ) and acquires nonlinear codes. With a linear classifier, the presented approach performs significantly better than the traditional nonlinear SPM. As studied previously\(^{(5)}\), a well designed BoW or Fisher Vector\(^{(6)}\) model trained on a single feature can beat all other approaches.

The problem of scene categorization has been approached in other papers from a variety of different angles. For example, Sadeghi and Tappen\(^{(7)}\) address this problem with a representation based on discriminative scene regions. Juneja et al.\(^{(5)}\) address this problem by learning distinctive parts incrementally, starting from a single part occurrence with an exemplar SVM. Although it has attracted a lot of interest and efforts in recent years, scene image categorization remains very challenging mainly due to the complexity and variability of the scene image itself, including significant appearance or pose variation of objects, multiple objects, complex background, frequent occlusion, changing illumination, etc.

In our previous work, we proposed a novel reference-based scheme combined with K-SVD\(^{(8)}\) for scene image categorization. The method is simple, efficient, and effective. However, a reference-set is selected randomly to form a set of basis, in which the image features are represented, leading to a significant reduction of the dimensionality of represented feature and achieve outstanding performance. To achieve better performance two key issues must be addressed\(^{(9)}\). The first is to construct a representative reference-set space. This difficulty is bypassed by defining ‘discriminative’. The second issue is to select a proper similarity measure between the probe image and the reference-set. In this work, we show that a more descriptive and discriminative reference-set should further improve the classification accuracy. More specifically, the reference-set should well span the data space while maintaining low redundancy. As Frey and Dueck reported\(^{(10)}\), affinity propagation algorithm (APA) outperforms other clustering methods in clustering images of faces, detecting genes in microarray data, identifying representative sentences in manuscript, and identifying cities that are efficiently accessed by airline travel. APA found clusters with much lower error than other methods, and it did so in less than one-hundredth the amount of time. What is more, it is able to avoid many of the poor solutions caused by unlucky initializations and hard decisions. So we adapt APA based on data similarity to construct a representative reference-set. Originally designed for data clustering, affinity propagation selects exemplars by message passing between the data points. Here the reference-set is constructed by exemplars selected from each image class.

The remainder of the paper is organized as follows: Section 2 presents details about the discriminative reference-set selection scheme based on the adapted affin-
ity propagation algorithm and how we incorporate it in the classification stage. Experimental results and analysis on three benchmark datasets are reported in Sect. 3. Section 4 concludes the paper.

2. Discriminative Reference-Based Scheme for Classification

The overall discriminative 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. In the offline process, we adapt APA to select the discriminative reference-set in a fully automated manner. After the discriminative reference-set gained, we project gallery images into the reference-set subspace and get the final image feature called reference descriptor (RD). The RD of the probe is generated in a similar manner and we train a linear SVM as classifier.

2.1 Affinity Propagation Algorithm (APA)

Originally designed for data clustering, affinity propagation selects exemplars by message passing between data points. Two kinds of message, “responsibility” $r(i, k)$ and “availability” $α(i, k)$, are exchanged between data points $i$ and $k$. $r(i, k)$ reflects the accumulated evidence for how well-suited point $k$ is to serve as the exemplar for point $i$, and $α(i, k)$ reflects the accumulated evidence for how appropriate it would be for point $i$ to choose point $k$ as its exemplar. After initialization, such as $α(i, k) = 0$, the responsibility and availability are computed in iteration as follows:

$$r(i, k) ← s(i, k) − \max_{k′, i′, k′ \neq k} \{α(i, k′) + s(i, k′)\},$$

(1)

$$α(i, k) ← \min\left\{0, r(k, k) + \sum_{i′ \in \mathcal{I} \setminus \{i\}} \max\{0, r(i′′, k)\}\right\},$$

(2)

where $s(i, k)$ denotes the similarity between point $i$ and point $k$. When the goal is to minimize squared error, each similarity is set to a negative squared error (Euclidean distance):

$$s(i, k) = −||x_i − x_k||^2.$$  

The algorithm stops once the exemplar decisions remain unchanged or the maximum allowed iteration number is achieved.

In our problem, we define the similarity as the negative chi-square distance as the following:

$$s(i, k) = 1 − F(d(i, k); \varepsilon) = 1 − \frac{\gamma(\frac{d(i, k)^2}{2}, \varepsilon)}{\Gamma(\frac{\varepsilon}{2})},$$

(3)

where $d(i, k)$ indicates the $χ^2$ distance of image $i$ and image $k$, $F(d(i, k); \varepsilon)$ presents cumulative distribution function, and $\varepsilon$ is a positive integer that specifies the number of degrees of freedom. $\gamma(\varepsilon/2, d(i, k)/2)$ denotes the lower incomplete Gamma function, $\Gamma(\varepsilon/2)$ is the Gamma function. To compute the similarity between images, LLC features are extracted from each image.

2.2 Discriminative Reference-Based Scheme

As shown in Fig. 1, after the discriminative reference-set generated, given an image, the similarity between it and each image in the reference set is given by

$$S(p, r_i) = 1 − F(d(p, r_i); \varepsilon) = 1 − \frac{\gamma(\frac{\varepsilon}{2}, \frac{d(p, r_i)}{2})}{\Gamma(\frac{\varepsilon}{2})},$$

(4)

$$\quad (r \in \alpha, \beta, \ldots; i = 1, \ldots, n).$$

Similar to Eq. (3), $d(p, r_i)$ is the $χ^2$ distance of the image $p$ and the $i$-th exemplar of the reference-subset (subclass) $r$ ($r \in \alpha, \beta, \ldots$ denoting different classes of the reference-set). $n$ gives the number of exemplars in each reference-subset. Then, the dimensionality of the similarity matrix is reduced by averaging per class to generate the final representation of the image denoted as $RD(p, r)$ for classification according to Eq. (5). Finally, we normalize the represented feature and use linear SVM as the classifier.

$$RD(p, r) = \frac{\sum_{i=1}^{n} S(p, r_i)}{n},(r \in \alpha, \beta, \ldots).$$

(5)

3. Experiments

Our approach is evaluated on three benchmark databases: Caltech-101[11], fifteen scene categories[12], and Pascal VOC2007[13]. The proposed method is compared with several state-of-the-art methods including the original reference-based method[8]. We use only a single descriptor, with the SIFT descriptors of $16 \times 16$ pixel patches computed over a grid with a spacing of 8 pixels, and $4 \times 4, 2 \times 2, 1 \times 1$ sub-regions for LLC, throughout all the experiments. Dictionary size for Caltech101 and VOC 2007 is 1024, and for fifteen scene categories the sizes are 200 and 400. 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 from all images in 392
different classes by APA in fifteen scene categories, Caltech101, Caltech-256 [14], and Pascal VOC2007. So the dimension of the final image feature is reduced significantly to 392. Figure 2 gives some examples of accepted and rejected images in the reference-set. From Fig. 2, we can see that: (1) it takes the more representative ones from similar images (as shown in the first and the last rows). (2) it takes the more discriminative ones from images in a same class (as shown in the second row). (3) it takes images with greatly illumination variation for the same face (compare the third row and the forth row) and several images for different faces without a doubt. With a randomly selected reference-set, we can not be sure all of the above.

We repeat the experiments 10 times with 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 personal computer with 16G memory and 4.00 GHz CPU.

3.1 Caltech-101

The first dataset is the Caltech-101 database [11] consisting of 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.

The result is compared with other state-of-the-art approaches [1], [2], [4], [15]–[18] and the original reference-based scheme [8]. The size of the reference-set is reduced significantly compared to original randomly selected as presented in Table 1. The first row corresponds to fifteen scene categories with 200 bases, and the next accords to fifteen scene categories with 400 bases. As can be seen from Table 2, our approach achieves a 1.6% increase in terms of accuracy over the original reference-based method, while remarkably outperforming all the competing approaches with nearly 10% accuracy increase compared to the next best result. Moreover, the classification accuracy with 5 training images per class is still comparable with the other methods.

3.2 Scene Category Recognition

The second experiment is on fifteen scene categories [12]. 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 1 The size of reference-set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Original</th>
<th>Discriminative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene15-200</td>
<td>30 × 392</td>
<td>18 × 392</td>
</tr>
<tr>
<td>Scene15-400</td>
<td>30 × 392</td>
<td>20 × 392</td>
</tr>
<tr>
<td>Caltech101</td>
<td>30 × 392</td>
<td>19 × 392</td>
</tr>
<tr>
<td>VOC2007</td>
<td>30 × 392</td>
<td>20 × 392</td>
</tr>
</tbody>
</table>

Table 2 Image classification results on caltech101 database.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Method</th>
<th>Classification Accuracy (%)</th>
<th>Classification</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lazebnik [1]</td>
<td>5</td>
<td>56.4</td>
<td>10</td>
<td>64.6</td>
</tr>
<tr>
<td>Gemert [16]</td>
<td>15</td>
<td>-</td>
<td>20</td>
<td>64.16</td>
</tr>
<tr>
<td>LC-KSVD [17]</td>
<td>49.8</td>
<td>59.8</td>
<td>59.8</td>
<td>70.7</td>
</tr>
<tr>
<td>D-KSVD [18]</td>
<td>49.6</td>
<td>59.5</td>
<td>65.1</td>
<td>71.1</td>
</tr>
<tr>
<td>Reference-based [8]</td>
<td>52.5</td>
<td>72.5</td>
<td>Ours</td>
<td>73.3</td>
</tr>
</tbody>
</table>

Table 3 Image classification results on scene15 database.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Method 1</th>
<th>Classification Accuracy (%)</th>
<th>Classification</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lazebnik [1]</td>
<td>200</td>
<td>74.5</td>
<td>400</td>
<td>74.8</td>
</tr>
<tr>
<td>Gemert [16]</td>
<td>200</td>
<td>74.3</td>
<td>400</td>
<td>76.7</td>
</tr>
<tr>
<td>Wang [4]</td>
<td>200</td>
<td>78.5</td>
<td>400</td>
<td>80.2</td>
</tr>
<tr>
<td>Reference-based [8]</td>
<td>82.8</td>
<td>83.8</td>
<td>Ours</td>
<td>84.4</td>
</tr>
</tbody>
</table>

Fig. 3 Confusion table of Scene15 dataset using 200 dictionary, the grid detector and patch based representation. The average performance is 83.8%.
to the state-of-the-art methods.

The classification results show that the proposed method increases the accuracy efficiently compared to the state-of-the-art methods. Experimental results were performed on three widely used image datasets to verify the benefits of the proposed method. Experimental results show that the proposed method increases the accuracy of classification while obtaining higher efficiency compared to the state-of-the-art methods.

Table 4 Image classification results on pascal VOC 2007 database.

<table>
<thead>
<tr>
<th>Object Class</th>
<th>aero</th>
<th>bicy</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC[4]</td>
<td>74.8</td>
<td>64.5</td>
<td>50.7</td>
<td>70.9</td>
<td>28.7</td>
<td>68.8</td>
<td>78.5</td>
<td>61.7</td>
<td>54.3</td>
<td>48.6</td>
</tr>
<tr>
<td>Best PASCAL’07[13]</td>
<td>77.5</td>
<td>63.6</td>
<td>56.1</td>
<td>71.9</td>
<td>33.1</td>
<td>60.6</td>
<td>78.0</td>
<td>58.8</td>
<td>53.5</td>
<td>42.6</td>
</tr>
<tr>
<td>Reference-based</td>
<td>79.0</td>
<td>72.8</td>
<td>57.9</td>
<td>72.6</td>
<td>29.9</td>
<td>71.8</td>
<td>81.9</td>
<td>65.1</td>
<td>61.6</td>
<td>53.5</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>79.7</strong></td>
<td><strong>73.3</strong></td>
<td><strong>58.2</strong></td>
<td><strong>72.5</strong></td>
<td><strong>33.0</strong></td>
<td><strong>72.8</strong></td>
<td><strong>82.1</strong></td>
<td><strong>67.3</strong></td>
<td><strong>61.9</strong></td>
<td><strong>55.4</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Object Class</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC[4]</td>
<td>51.8</td>
<td>44.1</td>
<td>76.6</td>
<td>66.9</td>
<td>83.5</td>
<td>30.8</td>
<td>44.6</td>
<td>53.4</td>
<td>78.2</td>
<td>53.5</td>
</tr>
<tr>
<td>Best PASCAL’07[13]</td>
<td><strong>54.9</strong></td>
<td>45.8</td>
<td>77.5</td>
<td>64.0</td>
<td>85.9</td>
<td>36.3</td>
<td>44.7</td>
<td>50.9</td>
<td><strong>79.2</strong></td>
<td>53.2</td>
</tr>
<tr>
<td>Reference-based</td>
<td>64.6</td>
<td>44.8</td>
<td>71.4</td>
<td>69.7</td>
<td>88.8</td>
<td>38.9</td>
<td>45.3</td>
<td>52.9</td>
<td>78.4</td>
<td>59.3</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>64.9</strong></td>
<td>44.7</td>
<td>76.9</td>
<td><strong>72.2</strong></td>
<td><strong>89.8</strong></td>
<td><strong>40.9</strong></td>
<td><strong>46.5</strong></td>
<td><strong>54.9</strong></td>
<td><strong>78.4</strong></td>
<td><strong>60.3</strong></td>
</tr>
</tbody>
</table>

Reference-based method achieves the best performance in most classes.

3.3 Pascal VOC 2007

The last experiment is conducted on a most challenging dataset which holds of 9,963 images in 20 classes, called the PASCAL 2007 dataset. All the images in this dataset 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 evaluation criterion is the standard metric used by PASCAL challenge [13]. It computes the area under the Precision/Recall curve, and the higher the score, the better the performance.

Table 4 lists our scores for all 20 classes in comparison with the best performance of the 2007 challenge [13], as well as the reference-based method without reference set selection in [8]. As seen from Table 4, our discriminative reference-based method achieves the best performance in most classes.

4. Conclusions

In this paper, a new reference-based scene images categorization approach based on adapted APA for discriminative reference-set selection is proposed. In contrast to the previous methods in which either sophisticated features are developed or the distance metric is learned, a discriminative reference-set with images from different datasets is utilized. It is an image-to-class measure, instead of an image-to-image measure widely used in previous work. Experiments were performed on three widely used image datasets to verify the benefits of the proposed method. Experimental results show that the proposed method increases the accuracy of classification while obtaining higher efficiency compared to the state-of-the-art methods.

Acknowledgment

This work was partially supported by Nanjing University of Posts and Telecommunications Program under Grant No. NY213071 and NY213086, National Natural Science Foundation of China under Grant No.61373065, 61271334 and 61302158.

References