A Novel Distribution of Local Invariant Features for Classification of Scene and Object Categories

LiJun Guo[1], JieYu Zhao[2], and Rong Zhang[2]

[1] Institute of Computer Technology, CAS
Graduate University of Chinese Academy of Sciences
Faculty of Information Science & Engineering NingBo University
818, Fenghua, Ningbo City,
Zhejiang, China
guolijun@nbu.edu.cn
[2] NingBo University
Faculty of Information Science & Engineering
818, Fenghua, Ningbo City,
Zhejiang, China
Zhao_jieyu@nbu.edu.cn,
zhangrong@nbu.edu.cn

Abstract. A new image representation based on distribution of local invariant features to be used in a discriminative approach to image categorization is presented. The representation which is called Probability Signature (PS) is combined with character of two distribution models Probability Density Function and standard signatures. The PS representation retains high discriminative power of PDF model, and is suited for measuring dissimilarity of images with Earth Mover's Distance (EMD), which allows for partial matches of compared distributions. It is evaluated on whole-image classification tasks from the scene and category image datasets. The comparative experiments show that the proposed algorithm has inspiring performance.

Keywords: Image classification, distribution representation, probability signature, kernel method.

1 Introduction

Image Categorization is one of the most challenging problems in computer vision, especially in the presence of scale variation, view variation, intra-class variation, clutter, occlusion, and pose changes. Generally, performance of an image categorization system depends mainly on two ingredients, the image representation and the classification algorithm. Ideally these two should be well matched so that the classification algorithm works well with the given image representation.

Local features [3, 4] are very powerful and efficient image representation for categorization problems, as seen by the state of the art performance of [1, 4]. However, the image representation produced by local features is an unordered set of feature vectors, one for each interest point found in the image. Although this kind of

Z. Shi et al. (Eds.): IIP 2010, IFIP AICT 340, pp. 308–315, 2010.

representation can be used for image categorization directly by comparing image similarity with voting method and gets nice performance [5], there exist two problems aroused by the representation: The first question is that most machine learning algorithms expect a fixed dimensional feature vector as input; The other is efficiency problem, because an unordered set of feature vectors from a image includes thousands of points, each of which, i.e. local features, is a high dimension vectors usually.

Distributions of local invariant features are more constringent image representation relative to modeling image with local features directly. It can solve the problems aroused by the unordered set of feature vectors effectively. Histograms, signatures and PDF (Probability Density Function) are three mainly manners to model the representation of images with distributions of local features for classification. They all suit for discriminative classification algorithm such as Support Vector Machines (SVM), but with different matching kernel respectively.

Histograms representation with distribution of local invariant features can be got by Vector-quantized method (called as bag-of-keypoints [1]), the simplest and most popular methods in text classification and image classification, which corresponds to a histogram of the number of occurrences of particular image patterns in a given image. Here the particular image patterns are seen as keypoints which are found using a simple k-means unsupervised learning procedure over the local invariant features of the train set. Because the histogram is just a kind of coarse form of distribution description, the process of histogram quantification must lose a lot of discriminative information from local features. At the same time, just as earlier global methods based on color or gradient histograms, it cannot achieve a good balance between expressiveness and efficiency because of fixed-size bin structures [6].

A signature $\{S_j = (p_j, w_j)\}$ represents a set of feature clusters. Each cluster is represented by its mean (or mode) p_j, and by the fraction w_j of features that belong to that cluster. Since the definition of cluster is open, a histogram can be viewed as a special signature with a fixed priori partitioning of the underlying space. In contrast with histogram representation with Vector-quantized method, in which all images are limited to have the same Bin structures, the number of clusters in the signatures can vary with the complexity of images. It means that the signature is a more flexible representation of distributions. Signature feature's good performance in image categorization benefits from its categorization method with EMD (Earth Mover's Distance) kernels[6][11]。

The EMD is a cross-bin dissimilarity measure and can handle variable-length representation of distribu-tions [6][13]. It allows partial matches in a very natural way, which is important, for instance, in dealing with occlusions and clutter in image categorization applications, and in matching only parts of an image. In addition, if the ground distance is a metric and the total weights of two signatures are equal, the EMD is a true metric, which allows endowing image spaces with a metric structure. Meanwhile, the EMD can be computed efficiently by a streamlined simplex algorithm Mathematical Programming [8]. However, the signature is still not enough to retain more discriminative information for categorization task.

Among the above three modes, PDF is a kind of the most direct description of distribution and can encode more discriminative features during modeling representation

of image for categorization or recognition. However, generally, the complex form of PDFs can lead to heavy computation when applied to image categorization [2][9].

Herein we propose a novel image representation combined with characters of two distribution models, Probability Density Function and standard signatures. We call it as Probability Signatures (PS). Images categorization is completed by learning a SVM classifier with EMD (Earth Mover's Distance) kernels [6][11] based on PS. The paper evaluates the classification method by scene recognition and image categorization on different image databases. Our experimental results demonstrate that the proposed approach in this paper is superior to vector-quantized and standard signature method.

2 Improved Distribution Representation

The PS is an improving to standard signature distribution by introducing generation model. First, in the PS, the initial distribution models based on local features for each image are created by Gaussian Mixture Models. Second, the mean vector of every single model of GMMs is viewed as the center of a cluster and the summation of posteriori probability reflecting all the local features to the same single model as the weights of corresponding cluster in the PS. Therefore, the PS combines the merits of PDF with that of signature. On one hand, compared with standard signature, as each component has its own covariance structure, a point is not based solely on the Euclidean distance to the clusters but upon some local measure of the importance of different feature components. Thus different clusters can emphasize different feature components depending on the structure they are trying to represent. Finally, we obtained a much smoother approximation to the input sets density model. On the other hand, by using the probability, this approach can encode more discriminative information and capture more perceptual similarity between distributions as a local feature is allowed to respond to multi clusters. Consequently, compared with PDF, the PS retain the same discriminative information with PDF for categorization, moreover, it allows for partial matches that the SVM categorizing with PDF kernel does not possess.

2.1 Local Invariant Feature Selection

Local invariant features include detector and descriptor. Some researches[4][11] have shown that the discriminative power of local features for classification can be raised by combining multi types of detector and descriptor efficiently. This kind of integration must have some complementary in invariance such as scale and affine or in patch types such as salience or texture. We use two complementary local region detector types to extract salient image structures: The Harris-Laplace detector responds to corner-like regions, while the Laplacian detector extracts blob-like regions. In order to raise the efficiency to generate probability model, we employ the low-dimensional gradient-based descriptor called PCA-SIFT [12] as a descriptor for patches extracted at these interest regions.

2.2 Probability Signature Generation

In order to get the PB representation of image, first, we need to establish the initial distribution models based on distribution of PCA-SIFT features for each image by PDF model. We use GMMs model and its maximum likelihood parameters are estimated by EM algorithm. Given an image, its PCA-SIFT feature vectors set $\mathrm{X} = \{x_1, x_2, ...x_n\}$ is extracted from detected regions, and GMMs model:

$$p(x \mid \theta) = \sum_{i=1}^{m} k_i N(x \mid u_i, \Sigma_i) \tag{3.1}$$

is estimated by EM, where $(k_i, \mu_i, \Sigma_i)_{i=1}^{m}$ are parameter vectors, $N(x \mid \mu_i, \Sigma_i)$ means a normal distribution and $k_i \geq 0$, $\sum_{i=1}^{m} k_i = 1$. Then, we generate the initial PS representations of the image:

$$S = \{(p_1, w_1), ..., (p_i, w_i), ..., (p_m, w_m)\} \tag{3.2}$$

where p_i is the mean vector of i th single mode of GMMs, w_i means the weights of i th mode, $w_i = \sum_{j=1}^{n} p(x_j) N(\mu_i, \Sigma_i)$, $p(x_j) N(\mu_i, \Sigma_i) > \alpha$, and α is called as correlation threshold to filter some noises from local features which is a little relation with the mode. The initial PS's length is m. The final PS is formed with compression process by a compression threshold. It is noted that different images have their PSes of different length. Two thresholds and compression process will be introduced in section 3.2.

2.3 EMD Kernel-Based Classification

Supposed $S_1 = \{(p_1, w_{p_1}), ..., (p_i, w_{p_i}), ..., (p_m, w_{p_m})\}$ and $S_2 = \{(q_1, w_{q_1}), ..., (q_j, w_{q_j}), ..., (q_n, w_{q_n})\}$ are two image Probability Signatures (having the same form with standard signature). The EMD is defined as follows:

$$EMD(S_1, S_2) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{3.3}$$

where f_{ij} is a flow value that can be determined by solving a linear programming problem[6], and d_{ij} is the Euclidean distance between cluster centers p_i and q_j. As the EMD is a measure of dissimilarity of two signatures, to incorporate EMD into the SVM framework, we use extended Gaussian kernels[7]:

$$K(S_i, S_j) = \exp(-\frac{1}{A} EMD(S_i, S_j)) \tag{3.4}$$

The $K(S_i, S_j)$ is called the EMD kernel. A is a scaling parameter which is set to the mean value of the EMD distances between all training images to reduce the computational cost[11].

3 Experiments

We have applied our method to two domains which belong to whole image categorization: scene recognition, object categorization.

3.1 Methodology

For each categorization task, we compare our algorithm's performance with two other techniques: Vector-quantized method using linear SVM classifier in[1] and standard signature using the same classifier based on EMD kernel in[11]. All three methods share the idea of representing images based on their distribution of local invariant features and discriminative classification algorithm SVM, but they vary in distribution form and corresponding kernel in SVM. Multi-class classification is done with a SVM trained by using the one-versus-all rule.

For the Vector-quantized method, considering that classification effect is sensitivity to the size of a Bin in histogram distribution representation as image content, Bins with two sizes are selected in our experiment. We call them fine Vector-quantized(1000)and coarse Vector-quantized(200) respectively. For the standard signature scheme, we use signatures of fixed length by extracting 40 clusters with k-means for each image, although EMD can handle variable-length representation of distributions. The ground distance d_{ij} of EMD is computed by Euclidean distance in standard signature and PS.

3.2 Compression Probability Signature

A simply and flexible method to determine the length of PS is used in our experiment. The initial PS with uniform length 50 is generated according to the steps in section 2.2. Then the PS is compressed in compression procedure by setting a compression threshold which depends on the correlation threshold to some extent. If the ratio of the number of local features which posteriori probability responding a component from PS is larger than the correlation threshold to the number of total local features from the image is larger than the compression threshold, this component will be deleted from the PS of the image. The compression procedure can improve not only the performance of categorization but also efficiency of computing EMD distance between signatures.

We learn the two thresholds of PS from the same databases with scene recognition experiment introduced as next section. The correlation threshold and the compression threshold are determined respectively in two phrases: First, without executing compression procedure (i.e. under no compression threshold), determine the change curve corresponding to the correlation threshold and the recognition rate, as shown in Fig. 1. In the following phrase, according to the result of Fig.1, draw the change curve between compression threshold and recognition rate under the correlation threshold 0.4, as shown in Fig.2. By comparing the recognition rates in Fig.1 and Fig.2, it is indicated that when the compression threshold is 0.02, the recognition rate is 0.83 which has exceeded the highest recognition rate in Fig.1 and when the compression

threshold is 0.03, the recognition rate reaches its peak. However, if the compression threshold is too large, over compression of PS can cut down the recognition rate.

In the above experiments, we select 50 images per class for the training set and 10 images from the remaining in each class for the test set. The results are reported as the average recognition rate.

Experiments show when correlation threshold and compression threshold are taken with 0.4 and 0.03 respectively, the average length of all category images reduce to 29 and the performance of recognition increase 4 percentage points. Because in PS, one local feature can simultaneously contribute to multi components by probability, the compression will not lead to delete local feature directly. However, it reduces noise affection in computing EMD. so the correlation and compression threshold of PS will be set as 0.4 and 0.03 respectively in our next experiments.

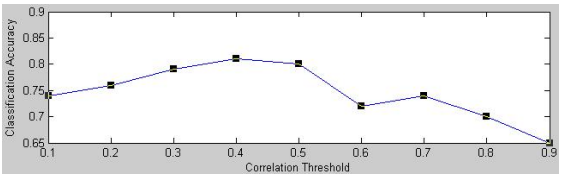

Fig. 1. Recognition accuracy with different choices of correlation threshold

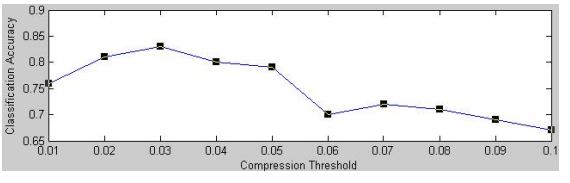

Fig. 2. Recognition accuracy with different choices of compression threshold

3.3 Scene Recognition

Our recognition task dataset is composed of eight scene categories provided by Oliva and Torralba[10]. Each category has about 300 images, and each image size is 256×256 pixels. Figure 3 shows the average recognition accuracy for a varying number of training examples per class, over 10 runs with randomly selected training examples. These are recognition rates that have been normalized according to the number of test examples per class. We can observe that our method works best among the four methods, while fine Vector-quantized approach works better than coarse and standard signature. Overall, the improved performance of our method over standard signature and two kind of Vector-quantized shows that more discriminative information are learn and more perceptual similarity between distributions are captured in our method.

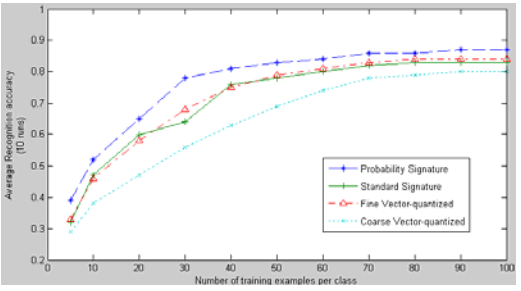

Fig. 3. Recognition results on the scene data set

3.4 Object Categorization

We evaluated four method on an object categorization task using two dataset with different styles. ETH-80[14] contains images of 80 objects from eight different classes in various poses against a simple background. All objects are almost full of the images. While Xerox7[1] includes total 1776 images which belongs 7 category object: faces, buildings, trees, cars , phones, bikes and books. These images are all of the objects in natural settings and thus the objects are in highly variable poses with substantial amounts of background clutter.

Table 1. Average classification accuracy rates in two objects datasets with four methods

Methods / datasets	vector-quantized		standard signature	probability signature
	fine-1000	coarse-200		
Eth-80	86.2	81.8	83.3	87.6
Xerox7	83.5	80.7	84.9	89.5

Table 1 shows that: Under object categorization tasks in images with simple background and object covering the whole image, although PS can obtain the best categorization rate, the categorization discrimina-tion power of signature is not superior to that of histogram features. However, under object category-zation tasks in natural images with complex background, signature representation has better categorization discrimination power than histogram representation, and PS categorization accuracy rate is higher than the standard Signature by 5 percentage points.

4 Conclusion

This paper proposes a novel representation of image: probability signature formed by improving the distribution of local features. The representation can capture more discriminative information for categorization in discriminative method with EMD kernel. We evaluate our method on three image databases in scene recognition and image categorization tasks. And our experiments demonstrate that the proposed approach in this paper is superior to vector quantization and standard signature method.

Acknowledgments

This work was supported by Scientific Research Fund of Zhejiang Provincial Education Department (Y200803738) and Ningbo Natural Science Foundation (2008A610027).

References

[1] Csurka, G., Dance, C., Fan, L., Williamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 59–74. Springer, Heidelberg (2004)

[2] Farquhar, J., Szedmak, S., Meng, H., Shawe-Taylor, J.: Improving Bag-of-Keypoints Image Categorisation: Generative Models and PDF-Kernels. LAVA report, 118, 141 (February 2005), http://www.ecs.soton.ac.uk/_jdrf99r/

[3] Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 2(60), 91–110 (2004)

[4] Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1615–1630 (2005)

[5] Gionis, A., Indyk, P., Motwani, R.: Similarity Search in High Dimensions via Hashing. In: Proc. of the 25th Intl Conf. on Very Large Data Bases, pp. 518–529 (1999)

[6] Rubner, Y., Tomasi, C., Guibas, L.: The Earth Mover's distance as a metric for image retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)

[7] Chapelle, O., Haffner, P., Vapnik, V.: Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks 10(5), 1055–1064 (1999)

[8] Hillier, F.S., Liberman, G.J.: Introduction to Mathematical Programming. McGraw-Hill, New York (1990)

[9] Moreno, P.J., Ho, P.P., Vasconcelos, N.: A kullback-leibler divergence based kernel for svm classification in multimedia applications. In: Neural Information Processing Systems, pp. 430–441 (2004)

[10] Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV 42(3), 145–175 (2001)

[11] Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. International Journal of Computer Vision 73(2), 213–238 (2007)

[12] Ke, Y., Sukthankar, R.: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In: Proc. CVPR, Washington, D.C., June 2004, vol. 2, pp. 506–513 (2004)

[13] Grauman, K., Darrell, T.: Efficient Image Matching with Distributions of Local Invariant Features. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 627–634 (2005)

[14] http://www.vision.ethz.ch/projects/categorization/