A NOVEL CONTEXTUAL DESCRIPTORS FOR CATEGORY RECOGNITION

Yi Ouyang, Ming Tang, Jian Cheng, Jinqiao Wang, Hanqing Lu, Songde Ma

National Laboratory of Pattern Recognition,
Institute of Automation,Chinese Academy of Sciences, Beijing 100080, China
{youyang, tangm, jcheng, jqwang, luhq, masd}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we propose a novel contextual descriptor which combines the contextual information and local appearance. Based on Gibbs distribution, a local descriptor is designed. By assembling the contextual information and local descriptors, a new partial contextual descriptor (PCD) is finally presented. Combining Pyramid Match Kernel (PMK) and SVM, we test our new descriptor and obtain higher average precision of classification than using local appearance descriptor.

Index Terms— object classification, PMK, category recognition, contextual descriptor

1. INTRODUCTION

This paper investigates the problem of object classification, which is useful in many fields, such as video surveillance, image indexing and retrieval, etc. Research on this topic could also promote the development of image understanding. Object classification has attracted much attention in the past several years, but it still remains a challenging problem because of the large variance of objects in the same class. Such variance may be due to the change in the viewpoint, deformable object shape, different image scale, and occlusion, etc.

A great deal of algorithms have been proposed for category classification. Among them, bag-of-words model achieves dramatic success. Zhang et al. [1] conduct a comprehensive study using the occurrence distribution over the visual words as features.

The main drawback of this method is that it just counts the occurrences of visual words, that is, only the local appearance of spatial points is utilized. However, it is obvious that the context is also important to determine the categories. The key difficulty is how to describe the context in some way, whereas not to increase the computational complexity greatly. Many literatures focus on this problem and propose to consider not only the local appearance but the local geometric relationship between spatial points.

Savarese et al. [2] use correlograms to capture spatial correlations between visual words, and the elements of the correlograms are further clustered into correlatons. Those correlatons reflect some intrinsic shape pattern, and the occurrence histogram over the correlatons describes the global geometric information of an image. Ling et al. [3] also use correlograms of visual words. Unlike [2], [3] defines a "Proximity Distribution Kernel" directly on the correlograms.

Correlograms contain the information on how many times a visual word "A" occurs in some local neighborhood of word "B". They describe the global geometric pattern or layout. Because correlograms are just statistics of spatial co-occurrence of the visual words and do not involve any explicit model of the spatial relationships among points, they are efficient in computation and robust to some basic geometric transformations.

Another non-negligible drawback of bag-of-words is that the characteristic of each feature point is ignored to some extent, since the feature space is quantized to discrete visual words. This problem is also inherited by correlograms.

In a way different from bag-of-words model, many researchers investigate the similarity of two images (two feature sets). In their methods, no quantization is needed. Grauman et al. [4] propose PMK as an approximation to the Earth Mover's Distance (EMD) and its computational complexity is linear in the number of features. However, the local appearance of the spatial interesting points is used while the context is ignored. Lazebnik et al. [5] adapt the pyramid matching scheme [4] to compute rough geometric correspondence on a global scale. This work can be regarded as a development of [4], as the spatial relationship is considered implicitly.

However, in [5], only the absolute positions of the points are used. We believe the information of relative positions could also be important, as has been demonstrated in [2, 3].

In this paper, emphasizing on the characteristic of each spatial point, we propose a new descriptor to formulate the spatial context in category recognition. This descriptor describes both the appearance of one spatial point and contextual information around it.

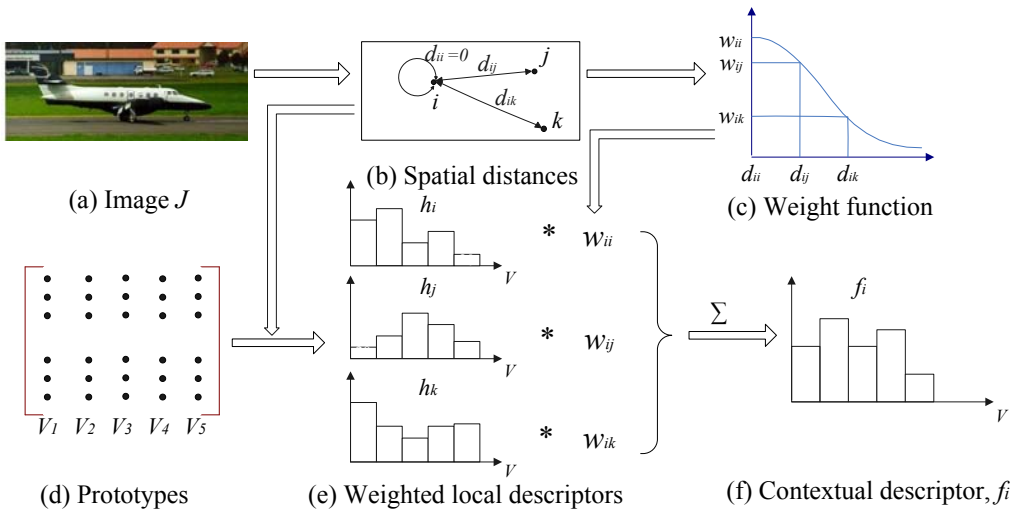

Fig. 1. Demonstration of the construction of contextual descriptor for point i. (a) An input image. (b) Relative distances from other points to i. (c) Weight function. (d) 5 prototypes. (e) Weighted local descriptors. (f) Contextual descriptor for point i.

The rest of the paper is organized as follows. In section 2, the detailed procedure of constructing contextual descriptors is described. Section 3 presents the experimental results on 7 classes from Caltech 101. Section 4 gives the conclusions and future work.

2. CONTEXTUAL DESCRIPTORS

In this section, we describe the construction of contextual descriptors. The procedure is illustrated in Fig. 2(a).

2.1 Construction of contextual descriptors

Three stages are needed to construct contextual descriptors which are S1, S2 and S3. We will discuss them in detail respectively.

2.1.1 Local feature extractions (S1)

We uses SIFT [6] to extract local stable and salient points. For each point the standard SIFT feature is a 128-dimensional vector. To get a more compact and robust representation, PCA-SIFT [7] is employed to reduce the dimension to 36.

2.1.2 Construction of local descriptors (S2)

K-means algorithm is utilized to cluster those PCA-SIFT features and currently K=100. To reduce the loss of quantification, each feature point is not merely represented by its nearest prototype but by a Gibbs distribution over all prototypes. We extend the representation in [8] slightly and construct the elements of local descriptor h_i as follows

$$h_{iv} = \exp(-\lambda \times dist_{iv}) / \sum_{v} \exp(-\lambda \times dist_{iv}) \quad v = 1, \cdots, 100 \qquad (1)$$

where $dist_{iv}$ is the Euclidean distance between the feature point i and prototype v. λ controls the shapes of the Gibbs distributions.

For each spatial interesting point i, h_i describes its local appearance more informatively than only with PCA-SIFT features, because it relates to the density of the original PCA-SIFT features.

2.1.3 Construction of contextual descriptors (S3)

Besides the local appearance h_i, incorporating i's contextual information will improve the ability of description further. There are many approaches to describe the context around one spatial point. Lyu et al. [9] use the neighboring angles in the constellation which is centered at a point and spanned by the k-nearest neighbors. Hence surrounding region is represented by two components: the appearances and the neighboring angles. Sivic et al. [10] ignore the relationship between points and only counts the occurrence of "visual words" in the neighborhood. Amores et al. [11] propose to use constellation of contextual descriptors.

We propose to assemble all the local descriptors h_j in the context of point i to form new features, i.e., contextual descriptors (CD), as shown in Fig.1. In our method, every spatial interesting point is represented with contextual information, including both appearance and geometric structure. To reflect the geometric relation, each h_j is weighted by an exponential function of the distance

between j and i. Note that contextual descriptor f_i for each spatial interesting point i is also a histogram with 100 elements which are defined follows.

$$f_{iv} = \sum_j w_{ij} \times h_{jv} \quad i, j = 1 \cdots N, v = 1 \cdots 100 \quad (2)$$

$$w_{ij} = \exp(-\alpha \times d_{ij}) \quad (3)$$

where d_{ij} is the spatial distance between i and j, N is the number of the spatial interesting points in one image. α controls the decrement of weights so as to control the range of context. If $\alpha = 0$, all weights are the same, and the contextual descriptors of all points are identical. Consequently, the proposed descriptor is degenerated to the traditional bag-of-words [1]. If $\alpha \to +\infty$, only i's weight is nonzero, and contextual descriptor, f_i, regresses to its local descriptor, h_i.

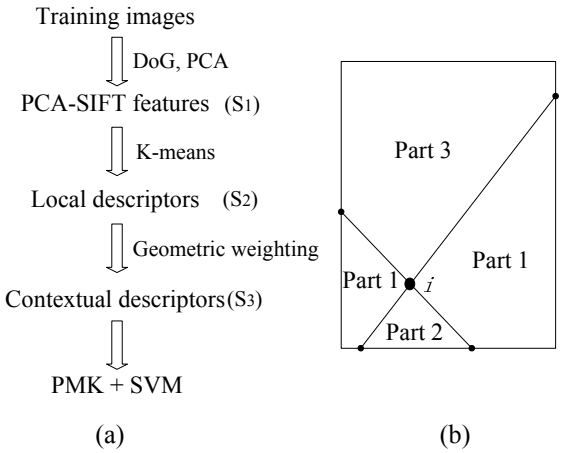

Fig. 2. (a) The flowchart of the proposed algorithm. (b) The division of i's contextual domain into three parts.

2.2 Extensions beyond relative distance information

The contextual descriptors above use only distance information, thus they are rotation invariant. But under some situations, relative position may also be important. For example, the windows of a car are always above its wheels while the wheels are always located on the same horizontal line. Based on the above consideration, we further extend the contextual descriptors by dividing the context domain into three parts as illustrated in Fig. 2(b). It can be observed that the relative position of "left" and "right" of an object usually does not have important influence on the perception of objects. Therefore "left" and "right" parts are merged into one, Part 1.

Now three partially contextual descriptors (PCD), one per part, are constructed by the method in 2.1.3. Note that the central point i is always included and its weight is 1. Consequently, the set of contextual descriptors of an image is divided into 3 sets of PCD. The similarity, K_p, of each pair of PCD sets of two images is calculated with PMK, and the final similarity is

$$K(I, J) = \sum_p K_p(I, J) \quad p = 1, 2, 3 \quad (4)$$

where (I, J) is an image pair.

3. EXPERIMENT

To evaluate the proposed algorithm, we select seven categories from the Caltech 101 database, which are airplanes, watch, leopards, motorbikes, faces, ketch and cars, as shown in Fig. 3. We use only the first 240 images from each category (use all images if the total number is less than 240) and adopt the one-vs-all strategy to train 7 classifiers.

Fig. 3. Seven categories used in the experiments

Firstly we randomly select 90 images per class to form prototypes and hierarchical tree used later by PMK.

And the numbers of samples used to train and test classifiers are shown in Table 1. All data are randomly selected.

Table 1. Data setting in experiments

	Training data		Testing data	
Categories	positive	negative	positive	negative
airplanes	120	20*6	120	20*6
watch	120	20*6	119	20*6
leopards	120	20*6	80	20*6
motorbikes	120	20*6	120	20*6
faces	120	20*6	120	20*6
ketch	90	20*6	24	20*6
cars	90	20*6	33	20*6

3.1 Parameters settings

There are 4 parameters needed to be set: λ in equation (1), the numbers of levels and branches of the hierarchical tree used by PKM, and α in equation (3). In our experiment, the former 3 parameters do not have impact on the results significantly. $\lambda = 5/a$, where a is the average of the distance of all pairs of prototypes. The number of levels is

set to 4, and that of branches to 11. α determines the effective range of a contextual domain. As discussed in section 2.1, too big or small α is not appropriate. Different values are tested using contextual descriptors, and the resulting average precision is shown in Fig. 4.

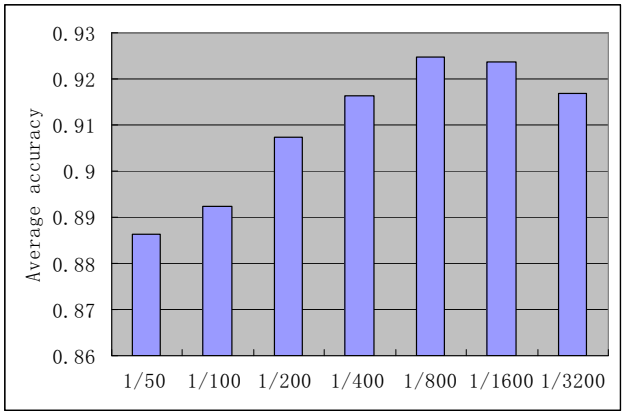

Fig. 4. Average accuracy with different α

Fig. 4 shows that the average accuracy is insensitive to α in quite a large range. In all experiments, α = 1/800.

3.2 Experimental results

To evaluate the effectiveness of contextual descriptor and its extension, PCD, the method in [4] is implemented in which PMK with local PCA-SIFT features is used. SVM [12] is employed as classifier. All precision is averaged over ten runs at equal-error rate. Table 2 shows the results. In our experiment, the average improvement with PCD over PCA-SIFT is 2.51%.

Table 2. Average precision

Categories	PCA-SIFT	CD	PCD
airplanes	88.43	89.44	**93.06**
watch	82.57	81.38	**83.05**
leopards	**99.75**	98.83	99.00
motorbikes	88.90	92.71	**93.06**
faces	94.44	97.78	**98.54**
ketch	88.19	90.05	**90.74**
cars	95.20	94.44	**96.62**
Ave.	90.93	92.09	**93.44**

It is noticed that the average precision of PCD is lower than that of PCA-SIFT in "leopards". The analysis of such phenomenon is one of our future works.

4. CONCLUSOIONS AND FURTURE WORK

Based on the analysis of existing problems in category recognition, a new contextual descriptor is proposed in this paper. An extended local descriptor is first presented appealing to Gibbs distribution. By assembling the contextual information and local descriptor into a partial contextual descriptor, we obtain higher average precision of classification than PCA-SIFT and contextual descriptor.

Currently, only one weight function is used. This may be unstable under scale variation. We are working on integrating multiple weight functions. Integrating other kinds of features is also under consideration.

5. ACKNOWLEDGEMENT

The research was supported by National Natural Science Foundation of China (Grant No. 60605004, Grant No. 60675003 and Grant No. 60572057)

6. REFERENCES

[1] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study", in *IJCV*, 73(2):213-238, 2007

[2] S. Savarese, et al. "Discriminative Object Class Models of Appearance and Shape by Correlatons", in *CVPR*, 2006

[3] H. Ling, S. Soatto, "Proximity Distribution Kernels for Geometric Context in Category Recognition", in *ICCV*, 2007

[4] K. Grauman and T. Darrell, "Approximate Correspondences in High Dimensions", in *NIPS,* 2007.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", in *CVPR*, II: 2169-2178, 2006.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints", in *IJCV*, vol. 20, pp. 91 – 110, 2003.

[7] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors", in *CVPR,* 2004

[8] B. Ommer, J. M. Buhmann, "Learning Compositional Categorization Models", in *ECCV,* 2006.

[9] S. Lyu, "Mercer Kernels for Object Recognition with Local Features", in *CVPR*, 2005

[10] J. Sivic, A. Zisserman, "Video data mining using configurations of viewpoint invariant regions", in *CVPR,* 2004

[11] J. Amores, N. Sebe, P. Radeva, "Fast Spatial Pattern Discovery Integrating Boosting with Constellations of Contextual Descriptors", in *CVPR,* 2005

[12] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm