Boosting Relative Spaces for Categorizing Objects with Large Intra-Class Variation

Yi Ouyang, Ming Tang, Jinqiao Wang, Hanqing Lu, Songde Ma
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{youyang, tangm, jqwang, luhq, masd}@nlpr.ia.ac.cn

ABSTRACT
In this paper, a novel method for object categorization is proposed. We first analyze the phenomenon of large intra-class variation and attribute it to the "subcategory" problem. To reveal the local and distinct properties of the different subcategories, *relative spaces* are constructed. Then the weighted FLDs (Fisher Linear Discriminant) as weak learners trained in relative spaces are integrated with the boosting framework to form the final classifier. Experiments on 8 categories from Caltech database show the effectiveness of our algorithm.

Categories and Subject Descriptors
I.4.8 [Scene Analysis]: Object Recognition

General Terms
Algorithms, Experimentation, Performance.

Keywords
Object categorization, Adaboost, relative space, Geometric Blur.

1. INTRODUCTION
Object categorization is one of the most challenging problems in computer vision, and has been investigated intensively in decades. The main difficulties lie in the large intra-class variation, such as viewpoint change, scale change, shape deformation, occlusion, and so on.

Recently, a great many methods have been proposed to address the problem based on local features. Local features have been proved effective for object and scene classification because they are robust to rotation, scale change and geometric deformation to some extent.

Discarding the spatial information, a comprehensive study on object classification with local features has been accomplished by Zhang et al. [1], in which local features are clustered to visual words. Thus an image can be represented as the frequency of occurrence of the visual words. This is a simple and efficient representation, but quantification errors are introduced during the clustering process.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

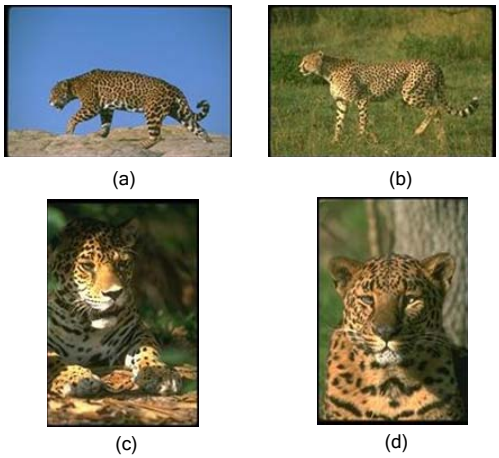

Figure 1. Leopards images with large intra-class variation.

Grauman et al. [2] represent an image as a set of local features and propose Pyramid Match Kernel (PMK) to calculate the distance between images. This distance measures how well the two sets' feature may be put into correspondence. Zhang et al. [3] propose to use the mean nearest distances to measure the distance between two sets of points. Different from Grauman et al. [2], the problem of the correspondence is ignored.

The above two methods employ some heuristic methods to define metrics. Being separated from the final classification task, they may fail to provide the maximal discriminativity. For this reason, some researchers try to learn the metric with data-driven methods.

Schutlz and Joachims [4] propose a method for learning a distance metric based on relative comparison such as "A" is closer to "B" than to "C". They try to learn the importance weight for each dimension of the feature space. Frome et al. [5] propose to learn local distances for each image also based on relative comparison. They try to learn the weight for each feature point, presuming that the focal image should be closer to the images from the same category than those from different categories. Such local distance owns some global properties.

Although many properties are shared among instances, it is difficult to find properties that are globally discriminative enough to all instances in the same category. As shown in Fig.1, it is often difficult to learn a unique global distance function at a time to classify positive samples, "a", "b", "c", and "d", against negative ones correctly. It is reasonable to consider that a category consists of many subcategories. Those subcategories may not be clearly separated from each other, but each has its own local and distinct properties (e.g., different viewpoints, shapes, and scales, etc.).

Therefore, Frome et al. [5] propose to learn local distance functions for focal images. Unfortunately, Frome has found that using hard positive samples is potentially hazardous in experiments. The reason is that hard positive samples may influence the result negatively seriously. Thus, they reduced hard ones heuristically before the training stage of experiments. In addition, their method needs to perform a second round of training or use heuristics to classify query images. In Frome's work [6], a hard positive sample is one that is very different from the focal image in its features. For example, in Fig.1, "c" and "d" could be seen as hard ones when "a" or "b" is focal image. In general, hard positive examples are those that are very close to negative ones in feature space, whereas they belong to positive set in semantics.

In this paper, we propose an alternative approach to object categorization. To deal with the large intra-class variation, weak classifiers are learnt for each subcategory. Specifically, a relative space is constructed for each positive training sample (belonging to one or several subcategories). Then a weak learner, weighted Fisher Linear Discriminant, is learnt in each relative space based on its local and distinct properties. Adaboost framework is employed to adaptively change the weights of points in relative spaces and produce a boosted classifier.

The rest of the paper is organized as follows. The details to construct relative spaces are described in Section 2. In Section 3, we present a boosting framework to obtain a strong classifier with relative space and weighted FLD. Section 4 shows the experimental results on 8 classes from Caltech database. The conclusions and future work is in Section 5.

2. RELATIVE SPACE

Zhang et al. [3] and Frome et al. [5] both represent images as sets of features. They evaluate the distance between images based on the measure between a feature and an image (i.e., a set of features). To better express our method, we will construct the relative space which is based on the same measure.

Given a training set $X = \{X_p, X_n\}$, where X_p is composed of N_p positive samples, and X_n is composed of N_n negative ones. Each sample x_i in X is represented as a set of m_i local features:

$$\vec{f}_{i,j} \quad , \quad j = 1, 2, \cdots, m_i$$

For each positive sample x_i, a relative space is constructed as follows. X is represented as a $(N_p + N_n) \times m_i$ matrix R_i, where $R_i(j,k)$ is defined as the distance from the *k*-th point in x_i to its nearest point in x_j:

$$R_i(j,k) = \min_{\vec{f}_{j,s} \in x_j} \| \vec{f}_{i,k} - \vec{f}_{j,s} \| .$$

The *j*-th row vector of R_i is regarded as the feature of the *j*-th image, which is denoted as $r_{i,j}$. All row vectors of R_i constitute the m_i -dimensional relative space of x_i. Since there are N_p positive samples, N_p relative spaces are constructed. Each sample appears in every relative space.

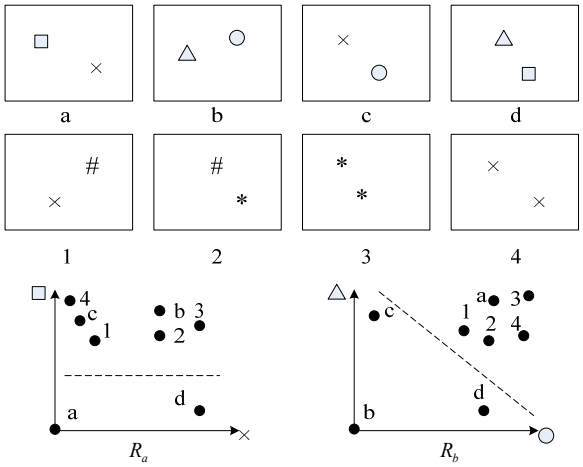

Figure 2. The construction of relative spaces.

Fig. 2 illustrates the construction of relative space. The first row includes four positive images, "a", "b", "c", and "d", and the second row negative images, "1", "2", "3", and "4". For better visualization, each image is assumed to consist of only two key points. {□, △, ○} represents the set of key points from the object, and {×, *, #} from the background. It should be noted that all these symbols do not represent visual words from clustering, they are just features. The same symbol, e.g., "□", in different images represents the similar part of the object, and the distances between them are small. And the distances between the different symbols are always large. The left graph of the third row is the relative space of image "a", i.e. R_a , which is constructed with two features: "×" and "□". Now, for example, we illustrate how to project the image "b" into R_a . The horizontal coordinate of point "b" is the minimum distance among ones from "×" in image "a" to the key pints ("○" and "△") in image "b" , and its vertical coordinate is the minimum distance among ones from "□" in image "a" to the key points ("○" and "△") in image "b". Other points are projected into relative space R_a in the same way. Relative space R_b is constructed with "○" and "△" , as shown in the third row, right.

It should be noticed that each relative space just reveals some perspectives (i.e., some subcategories) of the object in dataset. As illustrated in Fig.2, while "b" and "c" are indistinguishable from the negative samples in the relative space of image "a" with a linear classifier, they can be perfectly divided from the negative ones in that of "b". But even so, classifying with the relative space of "b" is still imperfect, because any linear classifier will misclassify "a". Subcategories phenomenon mentioned in Section 1 is the essential reason.

3. BOOSTING RELATIVE SPACES FOR OBJECT CATEGORIZATION

Since each relative space is only linearly discriminative for some parts of the positive samples against negative ones, it is necessary to integrate several relative spaces to improve the final

discriminativity. We propose to use weighted FLDs, trained in relative spaces, as weak learners, and then integrate them with the boosting framework.

3.1 Weighted FLD

Fisher linear discriminant is designed to find an optimal direction of projection to separate the positive and negative samples. The projection function is defined as:

$$g = w^T r,$$

where $w = (S^1 + S^2)^{-1}(\mu^1 - \mu^2)$, μ^1, μ^2 are the means of the two classes, and S^1, S^2 are the covariance matrices [11],

$$\mu^t = \frac{1}{n_t \sum d_j} \sum_j^{n_t} d_j r_j$$

$$S^t = \frac{1}{(n_t - 1)\sum d_j^{\,2}} \sum_j^{n_t} d_j^{\,2} (r_j - \mu^r)(r_j - \mu^r)^T$$

where d_j is the weight of sample r_j, and t = 1, 2.

3.2 Boosting Relative Spaces

Instead of training weighted FLDs directly in the relative spaces, we first reduce their dimensionality to avoid the singularity of $S^1 + S^2$. The Adaboost algorithm we use is in Algorithm 1.

Algorithm 1. The Adaboost algorithm.

Given images labels and their features in the relative spaces with reduced dimensionality.

Initialize weight $d_j = 1/2N_p, 1/2N_n$ for positive samples and negative samples, respectively.

For $t = 1, \ldots, T$

1: Train weighted FLD in each relative space.

2: Select the pair of optimal weighted FLD and the relative space it performs in.

3: Update the weights and normalize them.

End

Output a strong classifier.

3.3 Related Work

Frome et al. [7] propose to learn a global distance for object categorization, which is an extensional work of [5]. As both of the papers are learning some global properties of the whole image set, to avoid the subcategory problem as mentioned in section 1, they all need to heuristically remove some "hard positive samples" from the constraints before training.

Opelt et al. [10] also adopt boosting framework. The difference from our work is that, [10] is actually constructing weak learners on every dimension of the relative spaces, while we construct only one weak learner in one relative space. As demonstrated in [11], based on vector-valued features, weak learners may have better generalization.

4. EXPERIMENT RESULTS

In our experiments, 400 key points are extracted in each image. The key point is represented with geometric blur descriptor [8], which can capture the local shape information and is robust to the change of lighting and viewpoint. Specifically, key points are first randomly sampled on the edges, and then each point is described by a 51 × 4-dimensional vector, which includes 4 oriented channels and 51 locations around. Half of the sampled points are described in the large scale which has a patch radius of 70 pixels, and the others are in the small scale of 42 pixels.

Principal component analysis is used to reduce the dimensionality of relative spaces. While higher dimensionality easily cause over-fitting and non-reversible problem when training a weight FLD, lower dimensionality is opt to lose the discriminative power. It is a trade-off to select the dimensionality. Through a grid search for dimensionality from 10 to 40, we find there is no significant change in the final results, so we simply set the dimensionality to be 20 in the following experiments.

4.1 Experiments on Object Classification

To compare the proposed algorithm with the approach in [9], we select the same 8 categories from the Caltech and Caltech 101 database, which are airplanes, watch, leopards, motorbikes, faces, ketch, cars-rear and background.

In each experiment, 60 images per class are randomly selected as training set and 40 images as testing set. The one-versus-all strategy is adopted to train 8 classifiers, and a test image is assigned to the label of the classifier with the highest response. We run the experiments for 10 times, and the average precision is reported in Table 1.

Table 1. Average precision of experiment results.

category	[9]	ours	category	[9]	ours
leopards	90.00	**92.25**	airplanes	92.54	**94.75**
motorbikes	75.47	**97.00**	faces	88.89	**98.00**
cars	75.74	**96.25**	watch	**100**	93.00
ketch	0.00	**93.25**	background	**88.14**	71.75

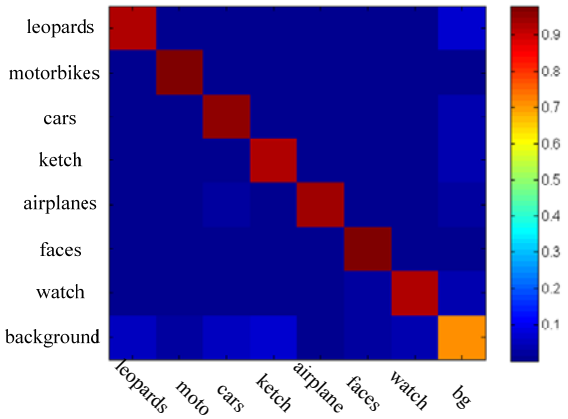

Figure 3. The confusion matrix.

From Table 1, the average precision of [9] is 76.35% while ours is 92.03%. The average improvement is 15.68%.

The confusion matrix is shown in Fig. 3. It can be observed that most misclassifications are caused by category "background". This is because the "background" is too diversified to be well modeled.

4.2 Image Set "Summarization"

Each relative space is constructed based on one positive image, and reveals some perspectives of the object. Therefore, boosting on relative spaces also selects out the most "representative" positive images for each subcategory. All these selected images could be considered as the whole view of the positive samples, or "summary" of the category. Different from traditional clustering algorithms, our method finds the "representative" samples in a discriminative manner.

It is hard to judge whether the selected images could serve as the "summary" without examining the whole dataset carefully. To make the judgment easier, we construct a positive dataset that includes three subcategories, i.e., leopards, motorbikes and ketch. And the negative dataset is the background category. We randomly select 150 positive (50 images per subcategory) and 150 negative images. The experiment is repeated for 10 times. In every experiment, the first 8 selected images are chosen as the "summary".

The expected "summary" should include the images from all three subcategories. Some results are shown in Fig. 4, where each row is the result from one experiment. In experiments, the "leopards" appears most frequently, because this subcategory is more diversified than others.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a novel approach for object categorization based on Adaboost framework. The key observation is that each category may be made up of several subcategories. By constructing a relative space for each positive sample, we propose to learn the local and distinct properties for each subcategory and then integrate them in Adaboost framework to gain higher precision of classification. At the same time, a summary of an image set is also obtained with our approach.

Experiments on larger datasets and the experimental comparison to Frome's work [7] are ongoing.

Though currently, only the geometric blur is used as the point descriptor, our algorithm is ready for combination of other kinds of descriptors, e.g. SIFT. Another promising direction is to take into consideration of the spatial locations of the key points when matching them.

6. ACKNOWLEDGMENTS

The research was supported by National High Technology Research and Development Program of China (grant No. 2006AA01Z315), Beijing Natural Science Foundation (grant No. 4072025) and NSFC (grant NO.60572057).

7. 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*, 2007.

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

[3] H. Zhang, A. Berg, M. Maire, and J. Malik, Svm-knn: Discriminative nearest neighbor classification for visual category recognition, in *CVPR*, 2006.

[4] M. Schultz and T. Joachims, Learning a distance metric from relative comparisons, in *NIPS*, 2003.

[5] A. Frome, Y. Singer, and J. Malik, Image retrieval and classification using local distance functions, in *NIPS*, 2006.

[6] Andrea Frome, Learning Distance Functions for Exemplar-Based Object Recognition, Ph.D. thesis, August 2007.

[7] A. Frome, Y. Singer, F. Sha, and J. Malik, Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification, in *ICCV* 2007.

[8] A. Berg and J. Malik, Geometric blur for template matching, in *CVPR*, pp. 607-614, 2001.

[9] L. Wu, M. J. Li, Z. Li, W. Y. Ma, and N. H. Yu, Visual language modeling for image classification, in *MIR*, 2007.

[10] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, Generic object recognition with boosting, in *PAMI*, 28(3), 2006.

[11] I. Laptev, Improvements of Object Detection Using Boosted Histograms, in *BMVC*, 2006.

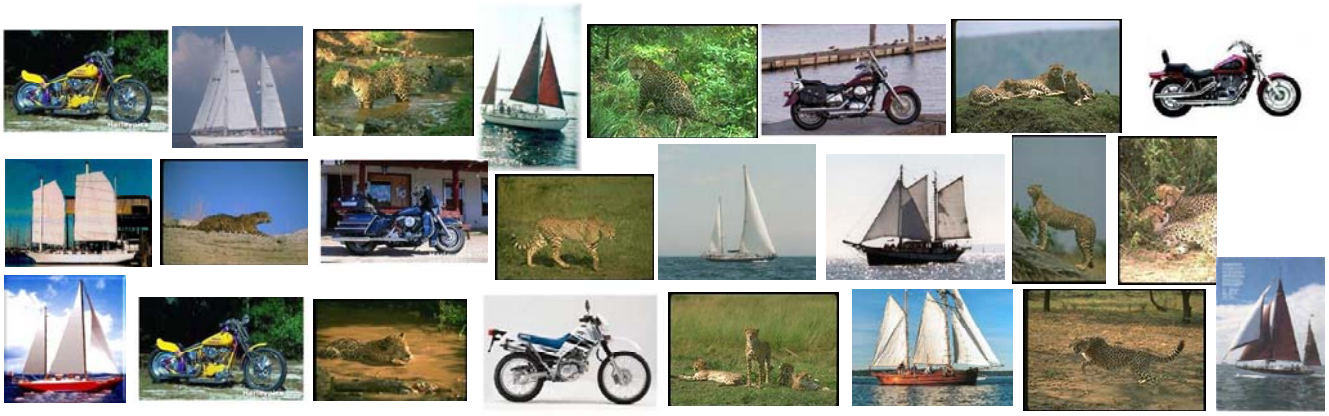

Figure 4. Image set summaries from 3 experiments.