Weak Metric Learning for Feature Fusion towards Perception-Inspired Object Recognition

Xiong Li[1], Xu Zhao[1], Yun Fu[2], and Yuncai Liu[1]

[1] Institute of Image Processing & Pattern Recognition
Shanghai Jiao Tong University, Shanghai 200240, China
{lixiong,zhaoxu,whomliu}@sjtu.edu.cn
[2] Department of CSE, University at Buffalo (SUNY), NY 14260, USA
raymondyunfu@gmail.com

Abstract. With extracted local features of a given image, computing its global feature under perceptual framework has shown promising performance in object recognition. However, under some tough applications with large intra-class variance, using only one kind of local feature is inadequate to build a robust classification system. To integrate the discriminability of complementary local features, in this paper, we extend the efficacy of perceptual framework to adapt to heterogeneous features. Given multiple raw global features, we propose a fusion strategy through metric learning, which is called weak metric learning in this work, for fusing high dimensional features. The fusion model is solved with the maximal kernel canonical correlation formulation with the multiple global features as outputs. Experimental results show that our method achieves significant improvements about 5% to 11% than the benchmark perceptual framework system, HMAX, on several difficult categories of object recognition with much less training samples and feature elements.

Keywords: Object recognition, feature fusion, weak metric learning, perceptual distance.

1 Introduction

Object recognition has seen rapid progress in recent years, motivated by innovative studies in relative fields such as statistical learning and cognition science. However, it is still in a long arduous travel for machine to approach human being's vision capability which can distinguish about 30,000 categories with very few training samples [1]. As a highlight of current researches, some human perception-inspired models [2,3] reach state-of-the-art performance.

Studies of human perception construct a basic framework for object recognition. Rosch [4] argued that categories are not defined by lists of features but by similarity to prototypes. Similarities defined on prototype examples, or equivalently perceptual distances rather than feature spaces attract the focus of researches. In this framework, scaling to a large number of categories just requires enough prototypes instead of adding new features. It is also possible to train the

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 273–283, 2010.

Fig. 1. Four images from scorpion category of the Caltech 101 dataset [1], which show large appearance variations. The main variations include mixed scorpion species with different biological morphology, texture blur, pose change, and cluttered background.

model with very few samples because the invariance to certain transformations or intra-class variation can be built into the perceptual distance function.

Serre and Poggio [5,3] modeled the ventral stream of primate visual cortex as a hierarchical structure (HMAX) for object recognition. The model is composed of S_1, C_1, S_2 and C_2 layers, of which C_1 produces local feature invariant to scaling and rotation and C_2 computes global features by defined perceptual distance (or similarity function). Corresponding to visual cortex, S layers improve invariance while C layers improve selectivity. The tradeoff between invariance and selectivity is achieved through alternate procedures. Frome [2] chose to learn a perceptual distance function for each example with metric learning algorithms [6], which determine weights for elements of all global features. In nature, these algorithms learn a transformation for the entire sample space.

These models and most perceptual inspired models follow the insight of Rosch [4] and share a basic outline: (1) For a test or training image, select a set of interest regions and extract patches from them. (2) Compute a local feature for each patch, which gives a set of local features for each given image. (3) For image pairs, return a value of distance by defining a distance function on their feature sets [7,2]. Or, given a local feature set from a image and the learned prototypes, return a set of distances as the global feature by defining a distance function between image and prototypes [3]. (4) Assign a category label to the image using the distance function or global feature. In step (3), both distance functions, known as perceptual distance, are defined on the local feature space.

However, local features developed for tasks like image registration can lead to a problem that they tend to fail under extreme lighting and pose conditions (for instance, SIFT [8] will be failure on binary images), and therefore could not provide enough discriminative information to classify complex objects, where even images from the same category show large intra-class variance. See Fig. 1.

Recently, some multiple local feature representations [9,10] were proposed to attack the above problem under "distance function learning" framework [2](learn distance functions instead of computing global feature with local features). Motivated by the capability of perceptual inspired models, we explore to integrate multiple local features with global feature computation models, such as HMAX.

In this paper, we propose an integrated solution that extends feature computation model and fusion strategy of global features, as illustrated in Fig. 2. It extends the HMAX model to adapt multiple local features, however, in general

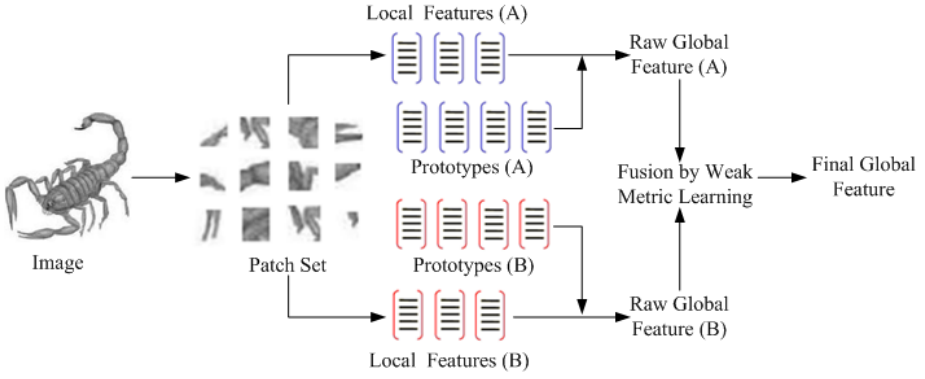

Fig. 2. Illustration of our framework

the extension to other models [7,2] is straightforward. The model could adapt to multiple kinds of local features. For each kind of feature, corresponding raw global feature is computed by measuring its distances from the pre-computed prototypes with the same feature representation[1]. Then an algorithm, namely *weak metric learning*, is developed to fuse these raw global features for object recognition. In nature, it aligns features at the metric level. For the feature fusion task, a criterion, maximal kernel canonical correlation [11,12], is used to solve the weak metric learning model. In sum, our main contributions are two folds.

1. Extend the global feature computation model, HMAX, to adapt to multiple complementary local features. It greatly improves the capability of the system to recognize ambiguous categories.
2. Introduce the kernel canonical correlation to learn the metrics to fuse different global features. Each set of metric weights is derived from the same template function with few free parameters so that the fusion model could be solved with few training samples.

For the improved features, experiments performed on Caltech 101 [1] show consistently significant improvement about 5 $\sim$ 11% than the benchmark model HMAX [3] with robust performance.

2 Model Extensions

The original HMAX model is composed of three steps: (1) compute C_1 response for the given image, (2) learn prototypes from C_1 responses of images, (3) for each prototype, compute the maximal response between the prototype and the C_1 response, which produces an element of the global feature C_2. To extend it for multiple local features, the point is to represent an image with a set of

[1] The prototypes are extracted from a set of randomly selected images, such as natural images.

patches and define the C_1 response on these image patches instead of the whole image (the patch based C_1 response is called C_1 descriptor in the paper). Then the following steps are updated accordingly and other local features could be introduced into the model by replacing C_1 descriptor.

In the extended model, the final global feature is computed according to the following four steps.

1. Extract patches from experimental images and arbitrary natural images. The natural images are used for learning prototypes.
2. Consider a type of local feature, compute a set of such features for the extracted patches.
3. Learn prototypes from the local feature set of natural images, and compute the raw global feature for an image with its local feature set and the learnt prototypes.
4. Multiple kinds of raw global features could be computed by replacing the feature type at step 2. Fuse the raw global features as the final global feature.

In this section, we describe how to compute the raw global features from local features. The fusion scheme will be introduced in Section 3.

Given image I and its local patch set $\mathcal{P}(I) = \{\mathbf{p}_i\}_{i=1}^n$ with varying sizes at detected interest regions, the local feature $\mathbf{c}$ is computed for each patch. Let $\mathcal{C}(I) = \{\mathbf{c}_i\}_{i=1}^n$ represent the local feature set of image I. At learning step, prototypes are extracted from local feature set $\cup_k \mathcal{C}(I_k)$ of natural images randomly and the learnt prototypes set is represented as $\mathcal{T} = \{\mathbf{c}_i^*\}_{i=1}^m$.

For candidate image I with its local feature set $\{\mathbf{c}_i\}_{i=1}^n$ and learnt prototype set $\{\mathbf{c}_i^*\}_{i=1}^m$, the element of raw global feature $\mathbf{x}(I) \in \mathbb{R}^m$ which corresponds to the local feature set, namely, the perceptual distance, is defined as

$$x_i \stackrel{def}{=} \min_{\mathbf{c} \in \mathcal{C}(I)} d(\mathbf{c}_i^*, \mathbf{c}), \tag{1}$$

where function d is a distance measurement of local features $\mathbf{c}_i^*$ and $\mathbf{c}$. We employ Euclidean distance and normalized inner product to measure C_1 and SIFT based perceptual distance respectively in this work. Further, the minimum distance could be regarded as an implementation of the maximal neural response. For C_1 feature, the simulated neural response corresponds to the shape tuning process of visual cortex. On the other hand, x_i could be interpreted as the baseline representation of patch $\mathbf{c}$, based on the prototype set $\mathcal{T}$.

Some descriptors such as SIFT [8], shape context [13] and geometric blur [14], can be used in the extended model. Most of them follow the scale space theory [15] and are invariant to rotation, scaling, or affine translation. In our solution, two complementary descriptors, C_1 and SIFT, are introduced into the extended model because C_1 encodes rich contour and shape information and the complementary SIFT encodes rich gradient information. In the following sections, the C_1 based raw global feature and the SIFT based raw global feature are called as C_2 and SIFT_2 respectively. What should be mentioned here is that only two descriptors from patches with the same size could be used for computing the perceptual distance.

3 Weak Metric Learning for Feature Fusion

With previous steps, two raw global features are computed. However, global features derived from different local features have different metrics even though they share the same perceptual distance function. Common schemes suggest to learn two weights for them. Moreover, recent work in [9] shows that multiple features fusion could benefit from subspace learning. However, it is hard to merge the metric difference of features in this task with these methods. In this section, a novel fusion scheme towards eliminating metric difference through metric learning is proposed. We also develop a novel metric learning method called as *weak metric learning* to deal with high dimensional feature. A criterion, maximal canonical correlation is used to solve the metric weights.

3.1 Formulation

Metric learning [2,6] is originally proposed to learn distance or similarity function by weighting each feature dimension. In [16], a correlation metric for feature extraction and similarity measurement is proposed. The technique can eliminate metric difference between feature dimensions implicitly. However, the metric learning scheme has to determine large number of independent weights therefore the scheme tend to fail for high dimensional feature and relative few training samples. In the weak metric learning scheme, a set of nonlinearly dependent weights are assigned to feature dimensions, and only few function parameters, instead of large numbers of weights, have to be determined.

For a given image, suppose similar feature elements correspond to the similar prototypes therefore they have similar metrics with similar weights. The continuous function $h \in H$ is used for assigning weights $w_i = h(x_i)/x_i$ to the global feature $\mathbf{x}(I) \in \mathbb{R}^m$. Then the weighed feature $\mathbf{x}'(I)$ could be formulated as

$$\begin{aligned} \mathbf{x}'(I) &= diag(w_1, \ldots, w_m)\mathbf{x}(I) \\ &= (h(x_1), \ldots, h(x_m))^T \\ &= h \circ \mathbf{x}(I). \end{aligned} \tag{2}$$

It suggests that weighting feature with template derived weights equals to applying a nonlinear transformation on the feature. Because weights for a raw global feature are derived from the same template function h, the task of determining weight set $\{w_i\}_{i=1}^m$ is converted to determine the parameter set of the template function h. The weights are nonlinearly dependent because the number of free parameters of $\{w_i\}_{i=1}^m$ (equals to the parameter number of h) is much smaller than m. It leads to a weak learning scheme. However, with capacity increasing of the template function h, the weak metric learning scheme will approach the general metric learning.

Similar to [2,6], the scheme can also be used to learn the distance function and solved with maximal margin formulation on the triplets training set. For fusion tasks, however, we develop a different model and solving scheme.

3.2 Metric Solving and Feature Fusion

For raw global features introduced in Section 2, we try to fuse them by weighting their elements. We next focus on fusing two features and the way to fuse multiple features is similar. The weight set for a raw global feature in Eq. (2) is derived from the same template function. As to fuse two features, two independent weight sets, equally two template functions have to be determined.

In [17], the canonical correlations of within-class sets and between-class sets for discriminative learning is explored. [18] uses canonical correlation analysis for feature fusion by determining pairs of projective matrices, given two candidate features. Different from above works, we employ canonical correlation to determine template functions (or the derived weight set equally) instead of projective matrices, though a weight set could be regarded as a special projective matrix.

The kernel version of canonical correlation [11,12] is used in the process of feature fusion, in our work, because it increases the flexibility of the feature selection through kernel trick. For the training image set $\mathcal{I}_N$, raw global features $X_{N\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$, $Y_{N\times q} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^T$ are computed from two different kinds of local features respectively with Eq. (1). Given nonlinear transformations $g, h \in H$, the kernel canonical correlation of two weighted global features is defined as

$$\phi(g, h, \alpha, \beta) = \mathrm{corr}_{\mathrm{ker}}(\alpha^T(g \circ X), \beta^T(h \circ Y)), \tag{3}$$

where $g \circ X$ represents applying transformation g on feature matrix X, as Eq. (2) formulated, and vectors $\alpha, \beta \in \mathbb{R}^N$ represent the combination coefficients of canonical correlation. We choose optimum nonlinear transformations by maximizing Eq. (3) stepwise

$$(g^*, h^*) = \arg \max_{g,h\in H} \widehat{\max}_{\alpha,\beta\in\mathbb{R}^N} \phi(g, h, \alpha, \beta), \tag{4}$$

where $\widehat{\max}_{\alpha,\beta}$ is a constrained maximizing process. We maximize Eq. (4) by enumerating g, h in function space H firstly. After g and h are given, we then further maximize $\phi(h, g, \alpha, \beta)$ in space $\mathbb{R}^N$. That is to maximize kernel canonical correlation

$$\begin{aligned} &\widehat{\max}_{\alpha,\beta\in\mathbb{R}^N} \mathrm{corr}_{\mathrm{ker}}(\alpha^T(g \circ X), \beta^T(h \circ Y)) \\ &\texttt{s.t.} : \mathrm{var}(\alpha^T(g \circ X)) = \mathrm{var}(\beta^T(h \circ Y)) = 1. \end{aligned}$$

It can be solved using Lagrange method which leads to an eigenvalue decomposition problem. Then $\phi_{\max}(g, h)$ could be substituted into Eq. (4) to continue maximizing in function space. It is time consuming to enumerate function space H. A specific yet effective solving procedure is to solve the optimization problem in the parameter space of a certain function instead of in the function space. Specially, let H be a function family parameterized by $\theta \in \mathbb{R}^S$. Eq. (4) can be formulated as

$$(\theta_g^*, \theta_h^*) = \arg \max_{\theta_g,\theta_h\in R^S} \widehat{\max}_{\alpha,\beta\in R^N} \phi(\theta_g, \theta_h, \alpha, \beta). \tag{5}$$

According to our experiments, enumerating θ_g and θ_h on an experiential range can satisfy this problem. To ensure optimization, for each candidate image, two local feature sets should be derived from the same patch set.

After parameter sets θ_g^* and θ_h^* are determined, two weighted global features could be given by Eq. (2), leading to the final global feature $(\mathbf{x}'(I)^T, \mathbf{y}'(I)^T)^T$. In the metric learning based fusion scheme, weighting on each feature element can be regarded as the adjusting process with feedback signals in visual cortex.

4 Experiments

Object recognition experiments with the fused global feature are performed on Caltech 8 to (1) show the advantage of the extended model and the fusion scheme; (2) examine the stability of fused features under varying number of samples and feature elements. The HMAX is chosen as the benchmark system because it provides the basic framework for our method. The SVM is used as classifier.

4.1 Dataset and Experimental Setup

A subset of Caltech 101 is chosen in the experiments. Although some categories in Caltech 101 are relative easy to classify, many categories with images taken under extreme lighting and large variations on view and pose are hard to be recognized. The same difficult may also come up in several sub categories with large intra-class variance. To validate the efficacy of our model, we deliberately select 8 difficult categories and the background category with the size of samples ranging from 80 to 800 for test. To speed up feature computation, all the images are normalized to gray images with 140 pixels high and a fixed aspect ratio.

We extract patches from interest regions. Several interest region detectors such as MSER [19], Harris-Affine, and Hessian-Affine [20] can be embedded into our framework. According to the comparison studies in [21], we select the Hessian-Affine as the detector of the interest regions. For a candidate image, patches with the sizes of 4×4, 8×8, 12×12, and 16×16 are extracted from all the interest regions respectively. The C_1 descriptors are constructed for each patch while SIFT descriptors are constructed for 12×12 and 16×16 patches. For prototype learning, patch extraction and descriptor construction are similar to the candidate images, except that for 500 patches per size are randomly extracted from interest points. Although descriptors could be constructed for all size of patches, descriptors from 12×12 and 16×16 patches work well. Then two prototype sets are learnt from natural images for C_1 and SIFT respectively.

At feature fusion step, two independent Gaussian functions are chosen for the weak metric learning procedure, with scale factor ranging from 6.5 to 10, variance ranging from 0.4 to 1.1 and mean ranging from 1.6 to 2.8. Larger range might improve the performance with more expensive time cost. Given the set sizes, the training set and testing set are sampled randomly from corresponding categories, as the same scheme for raw global features. For each setting, we sample data set and raw features about 20 to 40 rounds respectively. Then the average performance and its variance are reported in the final results.

Table 1. Performance comparison of three global feature settings: the feature with 2000 C_2 elements, the combinational feature of 1600 C_2 and 400 SIFT_2 elements, and the fused feature of 1600 C_2 and 400 SIFT_2 elements using our proposed method. Experiments are conducted under a configuration that the number of positive training samples, negative training samples, positive testing samples and negative testing samples are 30, 50, 50, and 50 respectively.

Data set	C_2	Combination of C_2 and SIFT_2	Fusion of C_2 and SIFT_2
Butterfly	0.8092	0.8515	0.8879
Brain	0.8112	0.8458	0.8833
Bonsai	0.7969	0.8130	0.8681
Chandelier	0.7783	0.7891	0.8281
Car-side	0.9737	0.9791	0.9929
Airplanes	0.9674	0.9735	0.9800
Buddha	0.7947	0.8349	0.8729
Scorpion	0.7754	0.8058	0.8438

4.2 Results

To test the performance under different configurations, the size of the positive training set and the length of the global feature are varying in our experiments. The experimental setting is as follows: the sizes of negative training set, positive testing set and negative testing set are taken as 50 respectively.

We run a series of experiments using 30 positive training images per category and 2000 elements (corresponding to 2000 prototypes) per global feature on the 9 categories dataset, with 30 random training sets (also 30 testing sets) and 20 random subsets of global feature (30×20 rounds overall). To evaluate performance, three global feature settings which share the same feature dimension but different element configurations, the feature with 2000 C_2 elements, the combinational feature of 1600 C_2 elements and 400 SIFT_2 elements, and the fused features of 1600 C_2 elements and 400 SIFT_2 elements, are compared.

As shown in Table 1, C_2 feature achieves high performance about 96.7% to 97.4% on Car-side and Airplanes categories and relative low performance about 77.5% to 81.1% on other categories. Similar situation appeared in other two settings. This is because that images of Car-side or Airplanes have small variance or similar appearance even though they are taken from different lighting and pose conditions. For all 8 categories, the combinational feature of C_2 and SIFT_2 elements outperforms C_2 feature about 0.6% to 4.2%. On the other hand, experiments in [3] indicate that increasing the number of feature elements is hard to improve the performance when the number is more than 1000. These evidences suggest that only appending other complementary descriptors based feature elements may be helpful. Compared with C_2 feature, our fusion scheme reaches an improvement about 5.0% to 7.9% on categories except Car-side and Airplanes (improvement about 1.3% to 1.9%).

To validate the fusion scheme on scalable positive training images, varying number of positive training images are tested. Fig. 3 shows these results for 5

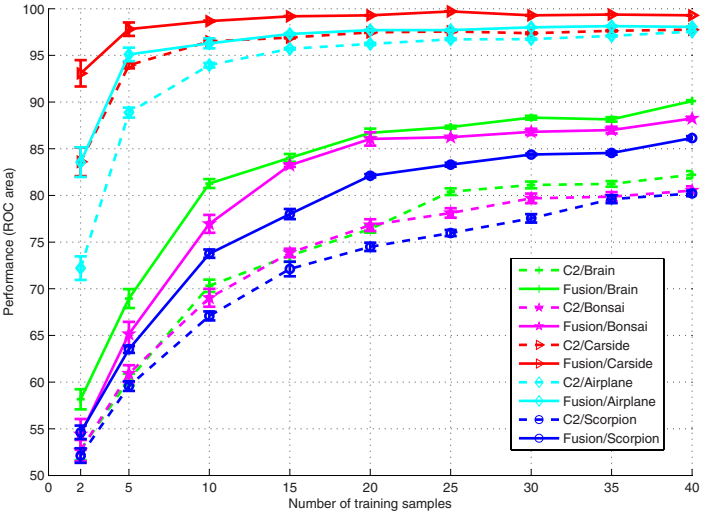

Fig. 3. Comparison between the feature of 2000 C_2 elements and the fused feature of 1600 C_2 and 400 SIFT_2 elements on Caltech 8 for varying number of training examples

categories in Caltech 101. Our fusion scheme outperforms C_2 on all tested categories. For easy categories Car-side and Airplanes, fusion scheme with 5 positive training images achieve satisfying performance about 97.8% and 95.1% with improvements about 3.9% and 9.2%. For other three categories, fusion scheme with 20 positive training images reach significant performance more than 83.1% while the performance of C_2 feature is under 77.2%, and it also outperforms C_2 about 5.1% to 7.9% when the number of positive training images is more than 20.

We also perform a series of experiments for varying number of feature elements from 2 to 2500 to test the fusion scheme. As shown in Fig. 4, the fusion scheme outperforms C_2 feature under all settings. For easy categories Car-side and Airplanes, the fused feature with 50 elements reach a satisfying performance about 97.5% and 95.5% with improvements about 4.0% and 4.5%. For other tested categories, the fused feature with 100 elements reaches significant performance exceeding 80% when the performance of C_2 feature is no more than 72.5%. Under settings of 50 or more feature elements, the fusion scheme outperforms C_2 feature at least 5.9%, especially 11.8% for Butterfly category.

In the fusion scheme, the optimization process of Eq. (5) consumes more time than other steps. Using Gaussian function as the template function to solve Eq. (5) by enumerating 448 parameter points on a normal computer takes about 110 seconds per 80 training images. When the size of dataset grows, the computation complexity mainly depends on the maximization in $\mathbb{R}^N$ and linearly depends on the enumeration number on $\mathbb{R}^S$. In our solution, the patch set that represents the candidate image is extracted from interest regions instead of overlapping regions [3] so that Eq. (1) takes only 1/70 time of it to compute global features.

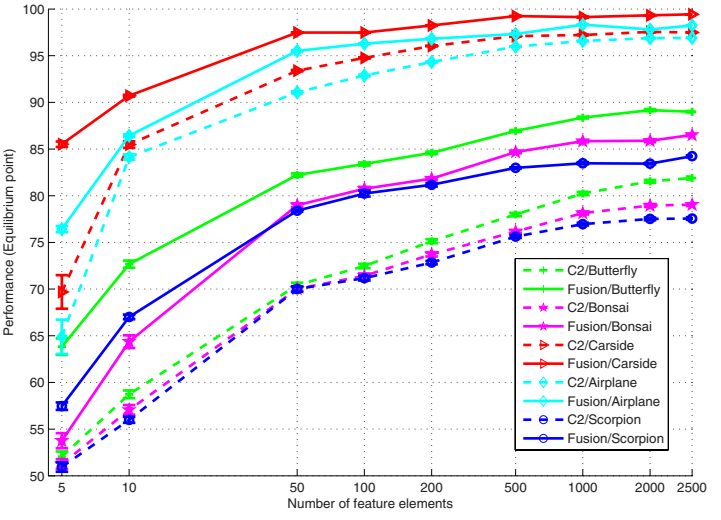

Fig. 4. Comparison between the feature with pure C_2 elements and the fused feature of 75% C_2 and 25% SIFT_2 elements on Caltech 8 for varying number of feature elements

5 Conclusions

In this paper, the perception inspired framework, HMAX, is extended to adapt multiple local features, producing multiple raw global features. A weak metric learning algorithm is developed for high dimensional features towards constructing the feature fusion model. The metric learning based model is solved through maximal canonical correlation formulation, giving the final global feature for object recognition towards difficult categories. Experiments on Caltech 8 show significant improvements under settings of varying number of training images and feature elements, which also confirms the validity and stability of our scheme. The fusion scheme, however, reaches the performance at the cost of much computing time, which will be the further research topic of this model.

Acknowledgment

Thanks to China National 973 Program 2006CB303103, China NSFC Key Program 60833009 and National 863 program 2009AA01Z330 for funding.

References

1. Fei-Fei, L., Fergus, R., Perona., P.: 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)

2. Frome, A., Singer, Y., Malik, J.: Image retrieval and classification using local distance functions. In: NIPS (2007)
3. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE TPAMI 29(3), 411–426 (2007)
4. Rosch, E.: Natural Categories. Cognitive Psychology 4(3), 328–350 (1973)
5. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience 2, 1019–1025 (1999)
6. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: NIPS (2004)
7. Zhang, H., Berg, A., Maire, M., Malik, J.: SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: IEEE CVPR (2006)
8. Lowe, D.: Object recognition from local scale-invariant features. In: IEEE ICCV (1999)
9. Fu, Y., Cao, L., Guo, G., Huang, T.S.: Multiple feature fusion by subspace learning. In: ACM CIVR, pp. 127–134 (2008)
10. Lin, Y., Liu, T., Fuh, C.: Dimensionality Reduction for Data in Multiple Feature Representations. In: NIPS (2008)
11. Lai, P., Fyfe, C.: Kernel and nonlinear canonical correlation analysis. International Journal of Neural Systems 10(5), 365–378 (2000)
12. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
13. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE TPAMI, 509–522 (2002)
14. Berg, A., Malik, J.: Geometric blur and template matching. In: IEEE CVPR (2001)
15. Lindeberg, T.: Scale-space: A framework for handling image structures at multiple scales. European Organization for Nuclear Research-Reports-CERN, 27–38 (1996)
16. Fu, Y., Yan, S., Huang, T.: Correlation metric for generalized feature extraction. IEEE TPAMI, 2229–2235 (2008)
17. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE TPAMI 29(6), 1005–1018 (2007)
18. Sun, Q., Zeng, S., Liu, Y., Heng, P., Xia, D.: A new method of feature fusion and its application in image recognition. Pattern Recognition 38(12), 2437–2448 (2005)
19. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22(10), 761–767 (2004)
20. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 128–142. Springer, Heidelberg (2002)
21. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.: A comparison of affine region detectors. International Journal of Computer Vision 65(1), 43–72 (2005)