**Notice of Violation of IEEE Publication Principles**

**"Generic Object Recognition Via Integrating Distinct Features with SVM"**
by Tong-Cheng Huang and You-Dong Ding
in Proceedings of 2006 International Conference on Machine Learning and Cybernetics, pp 3897-3902.

After careful and considered review of the content and authorship of this paper by a duly constituted expert committee, this paper has been found to be in violation of IEEE's Publication Principles.

This paper is a near duplication of the original text from the papers cited below. The original text was copied without attribution and without permission.

Due to the nature of this violation, reasonable effort should be made to remove all past references to this paper, and future references should be made to the following articles:

**"Generic Object Recognition by Combining Distinct Features in Machine Learning,"**
by Hongying Meng, David R. Hardoon, John Shawe-Taylor, Sandor Szedmak,
in the Proceedings of the 17th Annual Symposium on Electronic Imaging, January 2005, SPIE

Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006

# GENERIC OBJECT RECOGNITION VIA INTEGRATING DISTINCT FEATURES WITH SVM

**TONG-CHENG HUANG[1, 2], YOU-DONG DING[1]**

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China
[2] Department of Information and Electrical Engineering, Shaoyang University, Shaoyang 422000, China
E-MAIL: huang_tongcheng@163.com, ydding@yc.shu.edu.cn

**Abstract:**

**In a genetic image object recognition or categorization system, the relevant features or descriptors from a characteristic point, patch or region of an image are often obtained by different approaches. And these features are often separately selected and learned by machine learning methods. In this paper, the relation between distinct features obtained by different feature extraction approaches and that for the same original images were studied by Kernel Canonical Correlation Analysis (KCCA). We apply a Support Vector Machine (SVM) classifier in the learnt semantic space of the combined features and compare against SVM on the raw data and previously published state-of-the-art results. Experiments show that significant improvement is achieved with the SVM in the semantic space in comparison with direct SVM classification on the raw data.**

**Keywords:**

**KCCA; SVM; Data Fusion; Image Recognition; Feature Selection**

## 1. Introduction

The capacity to categorize objects plays a crucial role for a cognitive and autonomous visual system in order to compartmentalize the huge numbers of objects it has to handle into manageable categories. Generic object detection and recognition has recently gained a lot of attention in computer vision (e.g.[1-4]). For a generic object recognition system, there are mainly three parts. At first, features like points or regions have to be detected. These features should be flexible enough to accommodate a wide variety of object categories. Secondly, these features should be normalized or represented to be compared or learned. Finally, a suitable classifier or recognition algorithm should be provided. In most of these systems, distinct features were handled separately into the classification.

The increase of multimedia data during the past years has raised the issue of having efficient methods to analyze the data. In Ref. [5], it has been shown that the combination of different types of data is able to give a more accurate result than each component separately. In previous work[6] we follow this motive using KCCA where we combine image and text extracted from the web for a web page classification task. KCCA has been successfully applied in information retrieval application such as of cross-lingual[6] and content-based image retrieval[7,8] where one of two views is used to retrieve the other. In this paper we follow the idea of combining different components for a generic object classification task, where KCCA was used to learn the semantic feature space between interest points and key point features from the same image and produce a new kernel function for SVM.

In the section 2, the related research is summarized. In section 3, our system is introduced including feature extraction and whole system. In section 4, KCCA method was used to combine both the features obtained from interest points and key points. Finally both individual and combinative features were used into the generic object recognition by SVM classifier.

## 2. Related work

Agarwal and Roth[1] used sparse network of Winnows as the learning algorithm for the recognition of cars from side views. For this purpose images were represented as binary feature vectors that included a sparse, part-based representation of objects and spatial relations between them. These features were obtained by moving a window in the whole image and sensitive for the image with wide variety in scale. Complexity of the learning algorithm grows linearly with the number of relevant features and logarithmically with the total number of features. A different approach to object class recognition was presented by Fergus, Perona, and Zisserman[3]. They present a method of learning and recognizing object class models from unlabeled and un-segmented cluttered scenes in a scale invariant manner. Objects are modeled as flexible

constellations of parts. A probabilistic model was used for the representation of the object within the image. Using an EM-type learning algorithm they achieved very good recognition performance.

The previously two described methods are based on the model of the object in the image. For image datasets that have a wide variation in scale, views and highly textured background, we find that the models used in Ref. [1, 3] are difficult to be well estimated.

Recently, Opelt A. et al.[4] provided a new framework for a generic object recognition system that is a model-free approach to allow flexibility. In this system, the characteristic regions were detected by the interest points and key point detector , these were one of successful methods used to detect the low-level feature of an image. For one interest point, different local descriptors were calculated as the feature vectors. Finally, Boosting algorithm was used in combining several weak classifiers based on arbitrary and inhomogeneous sets of image features into a final strong classifier. This method can provide very good performance on relatively difficult datasets.

For each image, there are a different number of points and for each point there are a set of feature vectors. Although in the standard Adaboost algorithm, the feature vectors of all samples should be of the same length. In Ref. [4], all the distance between every feature and every image were calculated in the preprocessing and the best weak learner among them is found as a weak hypothesis in every step of Adaboost algorithm. It is quite a huge computing burden during training.

The original idea for interest points detector comes from Harris corner detector[9]. It was originally used to capture the characteristic corner and edge points in an image. Later, it has been extended to several improved versions. Based on the evaluation of interest points by Schmid et al.[10] the improved Harris detector get better results. In application, the scale invariant Harris-Laplace detector[11] and the affine invariant interest point detector[12] both proposed by Mikolajczyk and Schmid are most common ones. Another one proposed by Lowe[13] is also very common.

## 3. Generic object recognition system

The outline of our method is similar with the generic object recognition framework in Ref. [4], but there are three main differences. Our whole system is showed in Figure.1. Different original feature vectors are constructed from the neighbor of extracted characteristic points in the grayscales images. The uniform feature vectors are created based on clustering on the training set. KCCA are used to build combined feature from distinct features. Finally, SVM is used as the classifier. The first difference is that we use clustering in order to reduce the number of features. The second is that we apply KCCA to study the relationship between two distinct features and obtain a new kernel mapping function. Finally, we use SVM instead of Adaboost algorithm for the classification. The advantage is that the new kernel mapping can be regarded as a new kernel function and can be easily implemented in SVM.

### 3.1. Feature extraction

Our feature extraction is similar to the one used in Ref. [4]. For all the images, the characteristic patches are detected by Harris corner detector[9] and key points detector[13]. For each image $I_i$, there are $N_i$ detected patches, this can range from hundreds to thousands. For each patch l, feature vectors $f_{i,l}$ are constructed by some local descriptors such as invariant moment[12] and Scale Invariant Feature Transform (SIFT) [13].

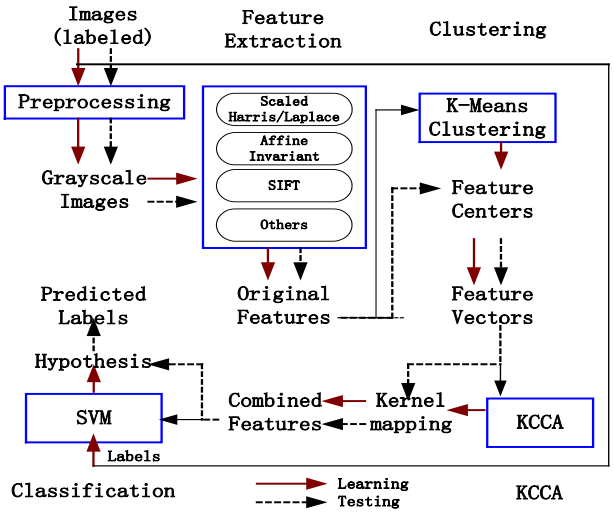


Figure 1. The proposed system framework for generic object recognition

### 3.2. Clustering

In order to obtain a simple feature vector for each image, K-means method was used to cluster the features into a uniform frame. In the training process, all the characteristic patch features were clustered by the K-means methods. K-mean method clustered all the training patches into K classes and their centers are in the set $O = \{o_k, k = 1,\cdots,K\}$ , where each center $o_k$ is a vector. Each center has the same length with the original feature vector for one

patch. Then the feature vector $X_i = \{x_{i,k}, k = 1,\cdots,K\}$ of an image $I_i$ is the minimum distance between $o_k$ and all features $f_{i,l}$ in $I_i$. It can be represented in the following way:

$$x_{i,k} = \min_{l=1,\cdots N_i} d(f_{i,l}, O_k), \tag{1}$$

where d(., .) is the Euclidean distance.

### 3.3. Classifier

We use SVM for the classification in our system. This is due to SVM being an outstanding classifier that has shown very good performance on many real-world classification problems. Using arbitrary positive definite kernels provides a possibility to extend SVM capability to handle high or even infinite dimensional feature space.

If the binary labels are denoted as $y_i$, the norm-2 soft-margin SVM can be represented as a constrained optimization

$$\min_{a,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \tag{2}$$

s.t.

$$\langle x_i, w\rangle + b \geq 1 - \xi_i, y_i = 1,$$
$$\langle x_i, w\rangle + b \leq -1 + \xi_i, y_i = -1,$$
$$\xi_i \geq 0,$$

where C is a penalty parameter and $\xi_i$ are slack variables. It can be converted by applying Langrange multipliers into its Wolfe dual problem

$$\max_{\alpha_i} \quad L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle X_i, X_j \rangle \tag{3}$$

s.t.

$$0 \leq \alpha_i \leq C, \sum_i \alpha_i y_i = 0.$$

The primal optimum solution for w can be represented as

$$W = \sum_i \alpha_i y_i X_i. \tag{4}$$

The weight vector w can be expressed as a linear combination of the support vectors for which $\alpha_i > 0$. It can be solved by quadratic programming methods. The final hypothesis is:

$$h_{w,b} = sign(\langle w, x\rangle + b). \tag{5}$$

It should be mentioned here that only dot products of feature vectors appear in the dual of the optimization problem. If we define the dot products as:

$$K(X_i, X_j) = \langle X_i, X_j \rangle. \tag{6}$$

Then the dual problem can be represented by

$$\max_{\alpha_i} L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(X_i, X_j), \tag{7}$$

$$0 \leq \alpha_i \leq C, \sum_i \alpha_i y_i = 0.$$

Based on it, SVM can be generalized to the case where the decision function is not a linear function of the data. Now suppose we first mapped the data to some other (possibly infinite dimensional) Euclidean space H, using a mapping that we call $\Phi$ :

$$\Phi : R^d \mapsto S \tag{8}$$

Then of course the training algorithm would only depend on the data through dot products in S, i.e. on functions of the form $\langle \Phi(x_i), \Phi(x_j) \rangle$ . We only need to use $K(\Phi(x_i), \Phi(x_j))$ in the training algorithm and never need to explicitly even know what $\Phi$ is.

## 4. Kernel canonical correlation analysis

### 4.1. Brief introduction of CCA

Canonical Correlation Analysis (CCA) is a method of correlating linear relationships between two multi-dimensional variables. Proposed by H. Hotelling in 1936[14] , CCA can be viewed as the problem of finding basis vectors for two (or more) sets of variables such that the correlation between the projections of the variables on to these basis sets are mutually maximized. The advantage of canonical correlation over correlation is that it is invariant to affine transformations. Consider linear combination $s = a_u' u$ and $t = a_v' v$ , where $a'$ is the transpose of a matrix or vector *a. u, v* are two random variables from a multi-normal distribution, with zero mean. The correlation between s and t is given by the following

$$\max_{a_u, a_v} \rho = \frac{E[st]}{\sqrt{E[s^2]E[t^2]}} = \frac{a_u' C_{uv} a_v}{\sqrt{(a_u' C_{uu} a_u)(a_v' C_{vv} a_v)}}. \tag{9}$$

$C_{uu}$ and $C_{vv}$ are the non-singular within-set covariance matrices and $C_{uv}$ is the between-sets covariance matrix.

### 4.2. KCCA

It may be the case that due to the linearity of CCA, useful descriptors may not be extracted from the data. In Ref.[15- 16], a complete kernel representation of CCA is presented. Kernel CCA offers an alternative solution by first projecting into a higher dimensional feature space prior to performing the CCA. The KCCA mapping is represented by Equ. (8) and the kernel function by K ($\Phi$(u), $\Phi$(v)). Due to the curse of dimensionality, the flexibility of the feature projection causes over-fitting of the data. Therefore to avoid finding spurious correlations we introduce a regularization parameter k to control the flexibility of the

feature projection, for brevity we refer the reader to Ref.[16].

The principle of KCCA is

$$\max_{g_u, g_v} g'_u K_u K_v g_v \tag{10}$$

s.t.

$$\left(g'_u K_u^2 g_u + k g'_u K_u g_u\right) = 1,$$
$$\left(g'_v K_v^2 g_v + k g'_v K_v g_v\right) = 1,$$

where $K_u$ and $K_v$ are kernel matrices defined based on the samples. For the linear case, they can be defined as $K_u = UU'$ ,$K_v = VV'$ , where $U = (u_1, u_2, \cdots, u_N)'$ , $V = (v_1, v_2, \cdots, v_N)'$ , $a_u = U' g_u$ and $a_v = V' g_v$.

Therefore following Ref. [16-17], we rewrite Equ.(9) in the dual representation with regularization parameter k

$$\max_{\alpha, g_v} \rho = g'_u K_u g_v \tag{11}$$

s.t.

$$\left(g'_u K_u (1-k) K_u + kI\right) g_u = 1,$$
$$\left(g'_v K_v (1-k) K_v + kI\right) g_v = 1.$$

Solving Equ.(11) as an eigenvalue problem, as shown in Ref.[16,18]. Partial Gram-Schmidt orthogonalization (PGSO) was applied on the kernel matrices to reduce their dimensionality, please see Ref. 16 for details. We can obtain several solutions for $\rho$ and corresponding $g_u$s and $g_v$s. In practice, we only use a subset of the solutions based on all sample feature vectors learnt in Equ.(11) denoted as $G_u = \left(g_u^1, g_u^2, \cdots, g_u^{N_1}\right)'$ and $G_v = \left(g_v^1, g_v^2, \cdots, g_v^{N_1}\right)'$.

### 4.3. Kernel mapping

We choose two feature vectors $\hat{X}_i^1$ and $\hat{X}_i^2$ from one image $I_i$ in the training set. If there are $M_1$ training samples, we can define the kernel matrices as $\hat{K}_1 = \hat{X}_1 \hat{X}_1'$ and $\hat{K}_2 = \hat{X}_2 \hat{X}_2'$ where $\hat{X}_1 = \left(\hat{X}_1^1, \hat{X}_2^1 \cdots, \hat{X}_{M_1}^1\right)'$ and $\hat{X}_2 = \left(\hat{X}_1^2, \hat{X}_2^2 \cdots, \hat{X}_{M_1}^2\right)'$. This is a linear kernel, other kernel functions can be defined.

.For any image $I_i$, we also have two feature vectors $X_i^1$ and $X_i^2$. Based on the KCCA analysis on training samples, we can define the kernel mappings in the following way:

$$\Phi\left(X_i^1\right) = G_u \hat{X}_1 X_i^1, \qquad \Phi\left(X_i^2\right) = G_v \hat{X}_2 X_i^2,$$

The combined new feature vectors will be a linear combination of the two mappings:

$$\hat{\Phi}\left(X_i\right) = \delta G_u \hat{X}_1 X_i^1 + (1-\delta) G_v \hat{X}_2 X_i^2,$$

where $\delta$ is a combination factor satisfying $\delta \in [0, 1]$. These feature vectors can be used in the SVM classifier. Of course, based on the combined features, further kernels such as Gaussian kernel or polynomial kernel can be defined and used in the SVM classifier.

## 5. Experimental results

### 5.1. Dataset

Two data sets were used in our experiment. The first one is a very difficult dataset used by Opelt et. al.[4] These images contain the objects at arbitrary scales and poses with highly textured background. There are two categories of objects, persons (P) and bikes (B), and images containing none of these objects (N). We tested the images containing the object (e.g. categories B and P) against non-object images from the database (e.g. category N). The performance was measured with the receiver-operating characteristic (ROC) corresponding error rate[3-4]. The training set contains 100 positive and 100 negative images. The tests are carried out on 100 new images, half belonging to the learned class and half not. For this dataset, there are several results published. In our experiment, the selected training and testing dataset are same with Ref.[4].

The second data set is a common one in the field of generic object recognition of images used by Fergus et al.[3] , Opelt et al. [4] and other papers. Motorbikes, airplane and faces are three object datasets and there is also a background dataset.

### 5.2. Experiment setup

For one image, two types of features were extracted by two different methods. One is from affine invariant interest point detector where moment invariant descriptor was calculated for each interest point. Another is the SIFT feature obtained from key point detector. These features were used in paper Ref.[4] as well. For any image in the two datasets, about from 10 to 3000 characteristic points were detected based on complexity of the images. And K = 400 was used as the number of centers for feature vectors in the clustering. So finally, for one image there are two uniform feature vectors with same length 400.

KCCA algorithm was used to produce the combined feature vectors. The selection method for regularization parameter k is the as in Ref. [16]. SVM was then used to classify the object categories against the no-object categories. The program SVMlight (Version 5.0) was used

in the experiment.

The results were compared to the one in which individual features are inputted to SVM. It is also compared with SVM on the mixed features in which two feature vectors were concatenated into one high dimensional vector. These results were also compared with the state-of-art performance obtained by other methods.

### 5.3. Performance

#### 5.3.1. Results on dataset 1

The state-of-art performance of the dataset 1 by using complex Adaboost algorithm was listed in the Tab. 1. The experiments results of our generic object recognition system based on the same two features were illustrated in the Figure.2. Experimental results show that significant improvement was achieved by the new combined features in comparison with classifications by using SVM on both individual features and mixed features. It is also a competitive results in comparison with the state-of-art performance on this dataset.

Table 1. Classification accuracy based on Adaboost algorithm4 according to ROC Equal Error Rate on dataset 1

| Dataset | Moment | SIFT |
|---|---|---|
| Bikes | 76.5 | 86.5 |
| Persons | 68.7 | 80.8 |

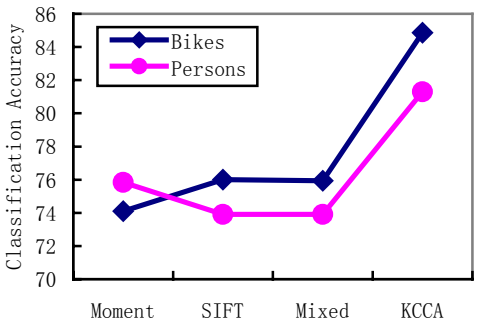


Figure 2. The chart of classification accuracy based on proposed generic object recognition system according to ROC Equal Error Rate on dataset 1.

#### 5.3.2. Results on dataset 2

The experimental results of our proposed generic object recognition system based on the same two features on dataset 2 were illustrated in the Figure.3. Experiment results show that very good performance is achieved by our generic object recognition system. There is a significant improvement in comparison with the state-of-art performance from previous papers.

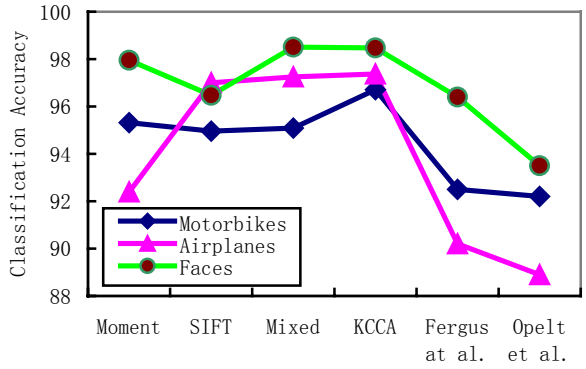


Figure 3. The chart of classification accuracy based on proposed generic object recognition system according to ROC Equal Error Rate on dataset 2 in comparison with the state-of-art performance from previous papers.

## 6. Conclusion and future work

From experiments results in the Figure.2 and Figure.3, it is clear that combining SIFT with moment invariant feature can produce better results. Most of the results are even better than the state-of-the-art performances. The new kernel mapping can efficiently combine two distinctive features into a semantic feature space where significant improvement can be achieved in the SVM classification.

Due to the good results of the image generic object recognition by combining the interest points and key point features, it can be regarded that this methodology can be applied to a more wide generic field rather then just generic object recognition. It should be regarded as a general data fusion method of combining two or more sources for increasing the classification accuracy provided by only a single source.

Future work will include the study on how to choose a-priori the best subset of the KCCA solution for the projection into the semantic space. We would like to look into existing and new problems of multimedia application and data, where data fusion and multi-source handling is needed.

## Acknowledgements

This work is supported by a project supported by Scientific Research Fund of Hunan Provincial Education Department.

**References**


[1] Agarwal S. and Roth D. "Learning a sparse representation for object detection," ，Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark, pp. 113–130, 2002.

[2] Borenstein E. and Ullman S., "Class specific top down-segmentation,", Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark, pp. 109–124, 2002.

[3] Fergus R., Perona P. and Zisserman A. "Object class recognition by unsupervised scale-invariant learning," , Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003.

[4] Opelt A., Fussenegger M., Pinz A., and Aue r P., "Weak hypotheses and boosting for generic object detection and recognition", Proceedings of the 2004 European Conference on Computer vision, pp. 71–84, 2004.

[5] Kolenda T., Hansen L. K., Larsen J. and Winther O., "Independent component analysis for understanding multimedia content", Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII, pp. 757–766, IEEE Press, (Piscataway, New Jersey), Martigny, Valais, Switzerland, Sept.4-6, 2002.

[6] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini, "Inferring a semantic representation of text via cross-language correlation analysis" , Proceedings of Advances of Neural Information Processing Systems, 2002.

[7] D. R. Hardoon and J. Shawe-Taylor, "KCCA for different level precision in content-based image retrieval", Proceedings of Third International Workshop on Content-Based Multimedia Indexing, (IRISA, Rennes, France), 2003.

[8] Vinokourov A., Hardoon D. R., and Shawe-Taylor J. "Learning the semantics of multimedia content with application to web image retrieval and classification", Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Source Separation, (Nara, Japan), 2003.

[9] C. Harris and M. Stephens, "A combined corner and edge detector" ，Alvey Vision Conference, pp. 147–151, 1988.

[10] Schmid C., Mohr R. and Bauckhage C. "Evaluation of interest point detectors", International Journal of computer vision, No. 4, pp. 151–172, 2000.

[11] Mikolajczyk K. and Schmid C. "Indexing based on scale invariant interest points", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 525–531, 2001.

[12] Mikolajczyk K. and Schmid C. "An affine invariant interest point detector ", Proceedings of the 2002 European Conference on Computer vision, pp. 128–142, 2002.

[13] Lowe D. "Object recognition from local scale-invariant features ", Proceedings of the 7th IEEE International Conference on Computer vision, pp. 1150–1157, 1999.

[14] Hotelling H. "Relations between two sets of variates", Biometrika , No. 28, pp. 312–377, 1936..

[15] Hardoon D. R., Szedmak S. and Shawe-Taylor J. "Canonical correlation analysis: an overview with application to learning methods", Technical Report, CSD-TR-03-02, Royal Holloway University of London, 2003.

[16] Hardoon D. R., Szedmak S. and Shawe Taylor J. "Canonical correlation analysis: an overview with application to learning methods", Neural Computation，No. 16, pp. 2639–2664, 2004.

[17] Bach F. and Jordan M. "Kernel independent component analysis", Journal of Machine Leaning Research, No. 3, pp. 1–48, 2002.

[18] Borga M. "Learning Multidimensional Signal Processing", PhD thesis, Linkping Studies in Science and Technology, 1998.