Boosted Multiple Kernel Learning for Scene Category Recognition

I-Hong Jhuo[1,2] and D. T. Lee[1,2]
[1]*Dept. of CSIE, National Taiwan University, Taipei, Taiwan*
[2]*Institute of Information Science, Academia Sinica, Taipei, Taiwan*
Email: ihjhuo@gmail.com, dtlee@ieee.org

***Abstract*—Scene images typically include diverse and distinctive properties. It is reasonable to consider different features in establishing a scene category recognition system with a promising performance. We propose an adaptive model to represent various features in a unified domain, i.e., a set of *kernels*, and transform the discriminant information contained in each kernel into a set of weak learners, called *dyadic hypercuts*. Based on this model, we present a novel approach to carrying out incremental *multiple kernel learning* for feature fusion by applying AdaBoost to the union of the sets of weak learners. We further evaluate the performance of this approach by a benchmark dataset for scene category recognition. Experimental results show a significantly improved performance in both accuracy and efficiency.**

***Keywords*-Dyadic hypercut; Multiple kernel learning**

I. Introduction

Visual recognition tasks for images of multiple categories have gained significant advances. The state-of-the-art methods have been developed by following some common strategies, such as designing a set of salient features based on which we derive a statistical model to learn optimal classifiers from training data, or clustering a number of semantic classes based on interested properties. However, major obstacles, e.g., lighting conditions, occlusions, and large intra-class variations, should be overcome in order to handle the recognition tasks more successfully.

Visual recognition models for image categorization usually utilize different and diverse properties to discriminate categories, and most methods learn promising recognition models by exploring these properties in training images. For example, images are represented by significant features and then operated by various scene classification methods [3], [11], [17], [8]. Further, methods considering the intra-class relationships between sub-image entities or co-occurrence of different objects were proposed [21], [13], [15], in which the bag-of-features and region-based methods were chosen. Furthermore, the performance of the method based on local features with co-occurrence property was significantly improved by including spatial relations [13].

Previous efforts of devising robust visual features and the corresponding distances have obtained significant results. Nevertheless, it is well-known that no single visual feature can be sufficient for recognition tasks. For example, the images in the first row of Fig. 1 should have a high recognition rate based on keypoint-based features [2], since those images consist of commonly shared patches of features. In contrast, for the images in the second and third rows, gist feature [3] might provide more recognition power. Thus, it is reasonable to make use of combined information cues from different features to improve recognition performance. Furthermore, to account for the fact that the optimal features for classification vary from category to category, we consider the use of *category-dependent* adaptive classifiers.

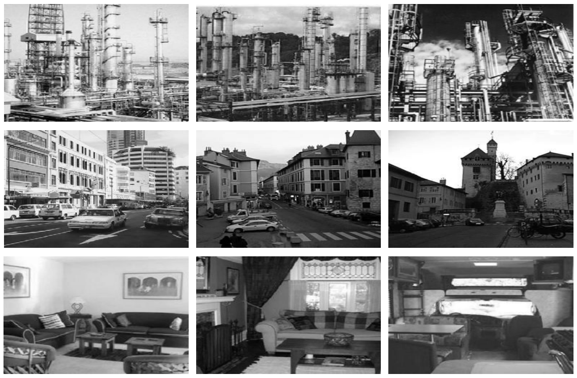

Figure 1: Example of images for scene category recognition. The three rows of images show scenes of category `industrial site`, `street scene` and `living room` respectively. Due to the diverse properties, recognizing these categories requires *category-dependent* combinations of several visual features.

Inspired by the good performance of multiple kernel learning (MKL) [20], [22], [14], [16], we adopt an adaptive learning approach to designing kernel machines for scene recognition. The use of multiple kernels not only provides richer information for recognition, but also gives an effective platform of feature fusion. In our approach, adaptive classifiers are learned from multiple kernels via a boosting algorithm. Empirically our method demonstrates a significantly improved performance.

II. MKL for feature fusion

We describe some key concepts about multiple kernel learning and its use for feature fusion in this section.

A. Conjoining kernel representation

Cognitive discoveries in prior work have demonstrated the importance of the use of multiple features, such as shape, texture, or color, to obtain a promising performance. However, the forms of the diverse visual features, e.g., histograms, feature vectors, or bag-of-words, may be different. To consider various visual features concurrently, we establish a kernel matrix to describe the pair-wise relationships among images under each visual feature. More specifically, suppose we have a training data set of C categories, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, where $\mathbf{x}_i$ is a training data, containing M kinds of different visual features, $\mathbf{x}_{i,m}$, $m = 1, 2, \ldots, M$. i.e., $\mathbf{x}_i = (\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,M})$, $y_i \in \{1, 2, ..., C\}$, and l is the number of training data. For each visual feature m, we convert pairwise distances between data, measured by Euclidean distance, to the m^{th} kernel matrix based on *radial basis function* (RBF) kernel function:

$$K_m(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma_m \|\mathbf{x}_{i,m} - \mathbf{x}_{j,m}\|^2}, \text{ for } m = 1, 2, \ldots, M \tag{1}$$

where γ_m is a positive constant. After that, we obtain a set of kernels $\{K_m\}_{m=1}^{M}$, and each kernel is derived by conjoining each visual feature and its corresponding distance function.

B. Multiple kernel learning

Learning problems of classification or regression can be efficiently solved by kernel methods, such as support vector machines (SVMs). Generally, a kernel matrix is constructed from data, denoted as $K(\mathbf{x}_i, \mathbf{x}_j)$, and specifies the similarity between two data $\mathbf{x}_i$ and $\mathbf{x}_j$. For such learning problems, the formulation of the learned model is of the form

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{2}$$

where l is the number of training data, $\{\alpha_i\}_{i=1}^{l}$ and b are learned constants, and $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ are training data.

To enhance the interpretability of the resulting classifier, recent research efforts, e.g., [20] have shown a significant progress on learning SVMs with multiple kernels. In such methods, an ensemble kernel, which is a convex combination of multiple kernels, will be learned and is of the following form

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} c_m K_m(\mathbf{x}_i, \mathbf{x}_j), \tag{3}$$

$$c_m \geq 0, and \sum_{m=1}^{M} c_m = 1,$$

where M is the number of kernels. By substituting multiple kernels into (2), we obtain the formulation of learned models of MKL from binary-class data $\{(\mathbf{x}_i, y_i \in \pm 1)\}$:

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i \sum_{m=1}^{M} c_m K_m(\mathbf{x}, \mathbf{x}_i) + b \tag{4}$$

The multiple kernel learning problem is to optimize both the sample coefficients $\{\alpha_i\}_{i=1}^{l}$ and kernel weights $\{c_m\}_{m=1}^{M}$. However, the learning process of MKL is computationally expensive and difficult to implement.

III. Our approach

The proposed approach that carries out MKL in a boosting way is described in this section. We start by depicting the *dyadic hypercuts* which are constructed from kernels and serve as weak learners in the work. Then, we use AdaBoost to combine these dyadic hypercuts generated from different kernels. The resulting approach accomplishes MKL in an incremental manner, and can control the tradeoff between classification accuracy and efficiency via adjusting the number of selected weak learners.

A. Dyadic hypercut

It is known that the cost of classifying testing data by a kernel machine depends on the number of kernel function evaluations. Therefore, we exploit *dyadic hypercuts* proposed by Moghaddam and Shakhnarovich [19], which are generated from a kernel matrix, and can be combined to produce a strong classifier via a boosting algorithm. In this way, we can control the run-time computational cost of a learned kernel machine by adjusting the number of selected weak learners.

In our case of multiple kernels, a dyadic hypercut h can be constructed with a specific kernel K_m, positive sample $\mathbf{x}_i$ and negative sample $\mathbf{x}_j$. The constructed dyadic hypercut weak learner h has the following expression:

$$h(\mathbf{x}) = sign(K_m(\mathbf{x}, \mathbf{x}_i) - K_m(\mathbf{x}, \mathbf{x}_j) + \delta) \tag{5}$$

where δ is a threshold. Totally, the size of the generated weak learner pool is $|\{h\}| = M \times N_p \times N_n$, where M, N_p, and N_n are the numbers of kernels, positive training data, and negative training data respectively.

B. Incremental multiple kernel learning via boosting

With diverse visual features, we can transfer information embedded in each kernel m to a set of dyadic hypercuts through (5). The resulting weak learner pool, which is the union of the generated dyadic hypercuts from all the kernels, will contain useful information for classification from multiple visual features. We select and combine those weak learners to obtain an ensemble classifier via a practical framework, Adaboost [1]. AdaBoost iteratively selects discriminant weak learners via maintaining a weight distribution D_t over data. At each iteration t, AdaBoost selects dyadic hypercut h_t with the minimal weighted error ϵ_t, determines the weight α_t of h_t, and updates the next distribution of data weight D_{t+1}. The learned classifier $H(\mathbf{x})$, which is a linear combination of selected weak learners, is shown as follows

$$H(\mathbf{x}) = sign(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})) \tag{6}$$

We provide an incremental way of a learning classifier with multiple kernels based on AdaBoost. It carries out multiple kernel learning, and the tradeoff between classification accuracy and efficiency can be controlled via adjusting the number of boosted weak learners. In addition, for scene category recognition, we could learn a boosted classifier for each category, since the optimal feature combination for separating images of a category from the rest varies from category to category.

IV. Kernels for scene recognition

In this section, we introduce visual features and their associated distance functions used to characterize images. Each of them captures distinctive properties of images, such as shape, texture, or some sensitive properties. In the work, a kernel is constructed for each visual feature. We briefly describe them as follows:

Gist: We adopt the *gist* descriptor [4] as the first feature for its compactness and high performance. The gist descriptor performs Fourier transform analysis to each individual sub-region of an image, and the image is then summarized by a set of perceptual properties. Euclidean distance is used to estimate the distance between a pair images under gist descriptor.

Pyramid HOG (Histogram of Oriented Gradients): Shape-based features provide a strong evidence for image categorization, and this phenomenon has been reported in the literature of object recognition. To utilize the discriminant power of shape information, we adopt the *Pyramid HOG* descriptor [9] for shape feature extraction. We use χ^2 distance to measure the dissimilarity between two images under this representation.

SIFT: We apply *DoG* detector and use the SIFT (Scale Invariant Feature Transform) descriptor [2] to depict interest points for each image. We also transform all the descriptors by k-means clustering, following by vector quantization [17] to convert each image into a feature vector. In our experiment, we implement two kinds of settings: The first one sets the number of clusters to 200 (BoW-200) and the second sets it to 800 (BoW-800). We use the χ^2distance as the similarity measure.

V. Experimental results

Scene recognition/classification involves not only data of diverse categories but also large intraclass variations. Furthermore, problems like different scales, views, or lighting conditions also give challenges in designing recognition algorithms. Therefore, we propose an approach for tackling these problems by fusing multiple features. The proposed approach is compared with the codebook approaches [17], [11], [8] and other learning algorithms, especially with a multiple kernel learning software, simpleMKL [20], in which a classifier is learned by taking multiple kernels into account simultaneously. For scene category recognition, we implement the one-versus-all rule for multi-class classification: a classifier is learned for each category, and used to separate images of the category from the rest. We carry out our experiments on the set of fifteen natural scene categories, which is collected by Lazebnik et al [17]. We will provide the experimental results and some discussions in the following sections.

Table I: Accuracy rates for scene recognition.

Method	Dataset	Classifier	Accuracy%
our method	**NS15**	Boosted MKL	**88.6**
Rakotomamonjy [20]	”	simpleMKL	87.9
Rasiwasia [21]	”	Bayes	72.5
Lazebnik [17]	”	SVM	72.2
our method	**NS13**	Boosted MKL	**91.8**
Rakotomamonjy [20]	”	simpleMKL	90.2
Rasiwasia [21]	”	Bayes	76.2
Lazebnik [17]	”	SVM	74.7

A. Dataset

The dataset is composed of images from fifteen natural scene categories, *NS15*. Thirteen of these categories were provided by [11]. Eight among those were collected by [3]. Each category includes 200-400 images, where 100 of them are used for training, and the rest are for testing. All the experiments are repeated 5 times with different randomly selected training and testing images.

B. Classification results

Most of the previous methods[8], [17], applying to the database, use the discriminative classifiers, such as SVMs, to perform category recognition. We instead exploit the dyadic hypercuts with AdaBoost for classification. Fig. 2 shows the classification accuracy of each visual feature and fusion of diverse visual features. As we can see, using only single visual feature is not sufficient to obtain a satisfactory result (with recognition rates $50\% \sim 70\%$). In contrast, to transfer those visual features to multiple kernels could give a significant improvement (with recognition rate 90%).

Since each scene category includes distinct properties, our approach can select the useful combination of visual features via weak learners to best separate images of the category from the rest. Moreover, the tradeoff between classification accuracy and efficiency can be controlled by adjusting the number of boosted weak learners. In Fig. 3, we show different combinations of the four kinds of visual features (weighted weak learners) used for different categories, where each bar represents its corresponding proportion.

We report the results of our approach and compare to that of the state-of-the-art systems in Table **??**. The proposed method achieves recognition rates of 88.6% and 91.8% on NS15 and NS13 respectively, and outperforms the state-of-the-art systems. The results validate the successfulness of our method in fusing multiple features to boost classification accuracy.

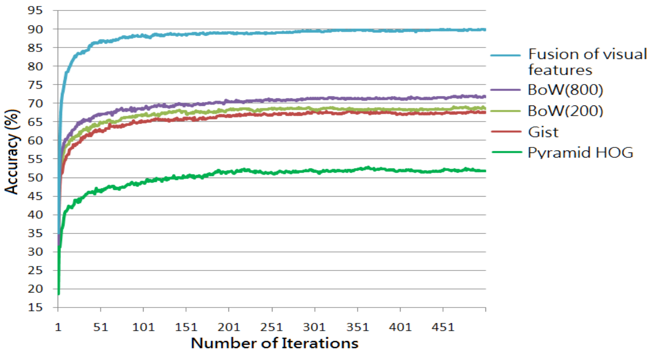

Figure 2: The recognition rates with each single and multiple kernels over 500 boosting iterations. The recognition rates are 52.86%, 68.04%, 69.15%, 72.16%, 88.94% for Pyramid HOG, Gist, BoW(200), BoW(800) and fusion features, respectively.

VI. Conclusions

We have proposed a useful method by taking advantage of multiple features for classifying scene images. In addition, our approach not only provides a way of achieving incremental multiple kernel learning but also can control the tradeoff between classification accuracy and efficiency. Experimental results show that our proposed method can significantly improve the recognition performance.

References

[1] Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, In JCSS, 55(1), 119-139, 1997.

[2] D. G. Lowe, *Distinctive image features from scale-invariant keypoints*, In IJCV, 60, (2), 91-110, 2004.

[3] A. Oliva and A. Torralba, *Modeling the shape of the scene: A holistic representation of the spatial envelope*, In IJCV, 42, 145-175, 2001.

[4] A. Oliva and A. Torralba, *Building the gist of a scene: The role of global image features in recognition*, In Brain Research, 155, 23-26, 2006.

[5] B. Scholkopf and A. Smola, *Learning with kernels*, MIT Press, 2001.

[6] M. Varma and A. Zisserman, *A statistical approach to texture classification from single images*, In IJCV, 62, 61-81, 2005.

[7] A. C. Berg and T. L. Berg and J. Malik, *Shape Matching and Object Recognition using Low Distortion Correspondences*, In IEEE CVPR, 2005.

[8] A. Bosch and A. Zisserman and X. Munoz, *Scene Classification via pLSA*, In ECCV, 2006.

[9] A. Bosch and A. Zisserman and X. Munoz, *Image Classification using Random Forests and Ferns*, In ICCV, 2007.

[10] C. Domeniconi and D. Gunopulos, *Adaptive nearest neighbor classification using support vector machines*, In NIPS, 2001.

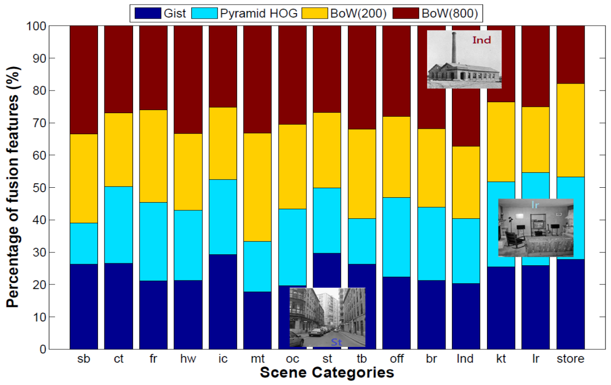

Figure 3: The proportion of four kinds of visual features used in learning each category-dependent classifier. Each of them is illustrated by a color bar. For example, in the `street scene`, `industrial site`, and `living room` categories, the highest proportion among the four visual features is Gist, BoW(800), and Pyramid HOG, respectively.

[11] Li, F.-F. and P. Perona, *A Bayesian Hierarchical Model for Learning Natural Scene Categories*, In CVPR, 2005.

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

[13] C. Galleguillos and A. Rabinovich and S. Belongie, *Object categorization using co-ocurrence, location and appearance*, In CVPR, 2008.

[14] P.V. Gehler and S. Nowozin , *On Feature Combination for Multiclass Object Classification*, In ICCV, 2009.

[15] J. C. Gemert and J. Geusebroek and C. J. Veenman and A. W.M. Smeulders , *Kernel Codebooks for Scene Categorization*, In ECCV, 2008.

[16] A. Kembhavi and B. Siddiquie and R. Miezianko and S. McCloskey and L. S. Davis, *Incremental Multiple Kernel Learning for Object Recognition*, In ICCV, 2009.

[17] S. Lazebnik and C. Schmid and J. Ponce, *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*, In CVPR, 2006.

[18] J. Liu and M. Shah , *Scene modeling using co-clustering*, In ICCV, 2007.

[19] B. Moghaddam and G. Shakhnarovich , *Boosted Dyadic Kernel Discriminants*, In NIPS, 2002.

[20] A. Rakotomamonjy and F. Bach and S. Canu and Y. Grandvalet , *More efficiency in multiple kernel learning*, In ICML, 2007.

[21] N. Rasiwasia and N. Vasconcelos , *Holistic Context Modeling using Semantic Co-occurrences*, In CVPR, 2009.

[22] Y.-Y. Lin and J.-F. Tsai and T.-L. Liu , *Efficient Discriminative Local Learning for Object Recognition*, In ICCV, 2009.

[23] H. Zhang and A. Berg and M. Maire and J. Malik , *SVM-KNN: Discriminative nearest neighbor classification for visual category recognition*, In CVPR, 2006.