IIEEJ **Paper**

Image Categorization Using Hierarchical Spatial Matching Kernel

Tam T. LE† , Yousun KANG††(*Member*) , Akihiro SUGIMOTO†††

† Kyoto University , †† Tokyo Polytechnic University , ††† National Institute of Informatics

<**Summary**> Spatial pyramid matching (SPM) has been an important approach to image categorization. This method partitions the image into increasingly fine sub-regions and computes histograms of local features at each sub-region. Although SPM is an efficient extension of an unordered bag-of-features image representation, it still measures the similarity between sub-regions by application of the bag-of-features model. Therefore, it is limited in its capacity to achieve optimal matching between sets of unordered features. To overcome this limitation, we propose a hierarchical spatial matching kernel (HSMK) that uses a coarse-to-fine model for the sub-regions to obtain better optimal matching approximations. Our proposed kernel can deal robustly with unordered feature sets as well as various cardinalities. In experiments, results of HSMK outperformed those of SPM and led to state-of-the-art performance on several well-known databases of benchmarks in image categorization, even though we use only a single type of image feature.

Keywords: kernel method, hierarchical spatial matching kernel, image categorization, coarse-to-fine model

1. Introduction

Image categorization is the task of classifying a given image into a suitable semantic category. The semantic category is definable as the depiction of a whole image such as a forest, a mountain or a beach, or of the presence of an interesting object such as an airplane, a chair or a strawberry. Among existing methods for image categorization, the bag-of-features (BoF) model is a popular and efficient one. It considers an image as a set of unordered features extracted from local patches. The features are quantized into discrete visual words, with sets of all visual words designated as a dictionary. A histogram of visual words is then computed to represent an image. A main weakness in this model is that it discards the spatial information of local features in the image. To overcome it, spatial pyramid matching (SPM)[1], an extension of the BoF model, uses aggregated statistics of local features on fixed sub-regions. It uses a sequence of grids at three levels of resolution to partition the image into sub-regions. Then it computes a BoF histogram for each sub-region at each level of resolution. Consequently, the representation of the whole image is the concatenation vector of all histograms.

Empirically, it is realized that to obtain good performance, the BoF model and SPM must be applied together with specific nonlinear Mercer kernels[2]such as the intersection kernel or χ^2 kernel. When the kernel function is proved to be positive definite, Mercer kernels guarantee the optimal solutions in learning algorithms. The intersection kernel for a BoF histogram is useful in Support Vector Machine (SVM) based image categorization and object recognition tasks. The Pyramid Match Kernel[3] is suitable for discriminative classification with unordered sets of local features.

Therefore, a kernel-based discriminative classifier is trained by calculating the similarity between each pair of sets of unordered features in whole images or in sub-regions. Numerous problems are well known to exist in image categorization such as the presence of heavy clutter, occlusion, different viewpoints, and intra-class variety.

In addition, the sets of features have various cardinalities and are lacking in the concept of spatial order. SPM embeds a part of the spatial information over the whole image by partitioning an image into a sequence of sub-regions, but to measure the optimal matching between corresponding sub-regions, it still applies the BoF model, which is known to be confined when dealing with sets of unordered features.

As described in this paper, we propose a new kernel function based on the coarse-to-fine approach. We call it a hierarchical spatial matching kernel (HSMK). HSMK enables not only capturing of the spatial order of local features, but also accurate measurement of the similarity between sets of unordered local features in sub-regions. In HSMK, a coarse-to-fine model on sub-regions is realized using multiple resolutions as shown in (b) of **Fig. 1**. Therefore, our feature descriptors capture not only local details from fine resolution sub-regions, but also global information from coarse resolution ones. In addition, matching based on our coarse-to-fine

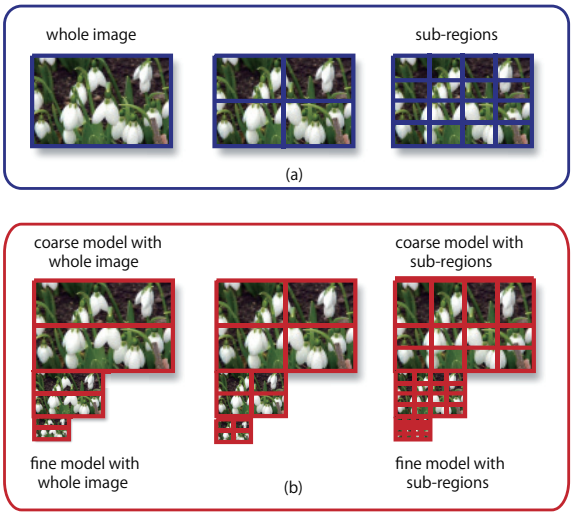

Fig. 1 Feature descriptor can be extracted local features from difference regions and resolutions by (a) and (b). (a) Spatial pyramid matching (SPM)[1]. (b) The proposed Hierarchical spatial matching kernel (HSMK).

model involves a hierarchical process, which indicates that a feature that does not find its correspondence in a fine resolution still presents the possibility of having its correspondence in a coarse resolution. Accordingly, our proposed kernel can achieve better optimal matching approximation between sub-regions than SPM.

2. Related work

The BoF model is a popular and powerful method for image categorization and generic object recognition. This framework functions by extracting local image features, quantizing them according to typical clustering method such as k-means vector quantization, accumulating histograms of the ”visual word” over the input image, and then classifying the histograms with simple classifiers such as an SVM and Boosting. However, the traditional BoF model discards the context information for spatial layout of an image.

Numerous methods have been proposed recently to improve the problems inherent in the traditional BoF model. Boiman *et al.*[4] presented no descriptor quantization for non-parametric Nearest-Neighbor(NN) classifier in image categorization. FeiFei *et al.*[5] proposed a generative probabilistic visual model based on Bayesian incremental algorithm. Moosmann *et al.*[6] introduced extremely randomized clustering forests to generate discriminative visual words using clustering decision trees. Yang *et al.*[7] unified codebook generation for object category recognition with classifier training. They proposed generative methods to model the co-occurrence of visual words, or discriminative visual words learning.

Lazebnik *et al.*[1] proposed an SPM method that can capture the spatial layout of features that are ignored in the BoF model. The SPM is particularly effective as well as being easy and simple to construct. It is used as a major part in many state-of-the-art frameworks in image categorization[8]. SPM is often applied with a nonlinear kernel such as the intersection kernel or χ^2 kernel. It requires high computation and large storage.

Grauman and Darrell[3] proposed a fast kernel function called the pyramid match using multi-resolution histograms. The pyramid match hierarchically measures the similarity between histograms which consist of sets of features extracted from the finest resolution to the coarsest one. The proposed kernel approximates the optimal partial matching by computing a weighted intersection over multi-resolution histograms for classification and regression tasks. Maji *et al.*[9] proposed an approximation to improve efficiency in building the histogram intersection kernel, but efficiency can be attained merely using pre-computed auxiliary tables, which are regarded as a type of pre-trained nonlinear support vector machine (SVM).

Mairal *et al.*[10] modeled data vectors as sparse linear combinations called sparse coding methods. They improved the visual dictionary in terms of discriminative ability or lower reconstruction error instead of using quantization by K-means clustering. To give SPM the linearity needed to address large datasets, Yang[11] proposed a linear SPM with spare coding (ScSPM), in which a linear kernel is chosen instead of a nonlinear kernel because of the more linearly separable property of sparse features.

Our proposed kernel emphasizes improvement of the similarity measurement between sub-regions using a coarse-to-fine model instead of the BoF model used in SPM. In recent works, some methods devoted to image categorization through multi-scale method have been described. Wang & Wang[12] proposed a multiple scale learning (MKL) framework in which multiple kernel learning (MKL) is used to learn the optimal weights instead of using predefined weights of SPM. The multiple scale learning method can determine the optimal combination of base kernels constructed in different image scales for visual categorization. However, we consider the sub-regions on a sequence of different resolutions as the pyramid matching kernel (PMK)[3]. Furthermore, instead of using the pre-defined weight vector for basic intersection kernels to penalize across different resolutions, we reformulate the problem into a uniform MKL to obtain it more effectively. In addition, our proposed kernel can deal with different cardinalities of sets of unordered features by application of square root diagonal normalization[13] for each intersection kernel, which is not considered in PMK.

3. Hierarchical Spatial Matching Kernel

In this section, we first describe the original formulation of SPM and then introduce our proposed HSMK, which uses a coarse-to-fine model as a basis for improving SPM.

3.1 Spatial Pyramid Matching

Each image is represented as a set of vectors in the D-dimensional feature space. Features are quantized into discrete types called visual words using K-means clustering or sparse coding. The matching between features turns into a comparison between discrete corresponding types. Therefore, they are matched if they are of the same type and unmatched otherwise.

SPM constructs a sequence of different scales with $l = 0, 1, 2..., L$ on an image. In each scale, it partitions the image into $2^l \times 2^l$ sub-regions and applies the BoF model to measure the similarity between sub-regions. Let X and Y be two sets of vectors in the D-dimensional feature space. The similarity between two sets at scale l is the sum of the similarity among all corresponding sub-regions:

$$\mathcal{K}_l(X, Y) = \sum_{i=1}^{2^{2l}} \mathcal{I}(X_i^l, Y_i^l), \tag{1}$$

where X_i^l is the set of feature descriptors in the i^{th} sub-region at scale l of the image vector set X. The intersection kernel $\mathcal{I}$ between X_i^l and Y_i^l is formulated as

$$\mathcal{I}(X_i^l, Y_i^l) = \sum_{j=1}^{V} \min(\mathcal{H}_{X_i^l}(j), \mathcal{H}_{Y_i^l}(j)), \tag{2}$$

where V is the total number of visual words and $\mathcal{H}_\alpha(j)$ is the number of occurrences of the j^{th} visual word which is obtained by quantizing feature descriptors in the set α. Finally, the SPM kernel (SPMK) is the sum of weighted similarity over the scale sequence:

$$\mathcal{K}(X, Y) = \frac{1}{2^L}\mathcal{K}_0(X, Y) + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\mathcal{K}_l(X, Y). \tag{3}$$

The weight $\frac{1}{2^{L-l+1}}$ associated with scale l is inversely proportional to the sub-region width at that scale. This weight is used to penalize the matching because it is easier to find the matches in the larger regions. All matches found at scale l are also included in a finer scale $l - \zeta$ with $\zeta > 0$.

3.2 Proposed kernel: Hierarchical Spatial Matching Kernel

To improve efficiency in achieving the similarity measurement between sub-regions, we use a coarse-to-fine model on sub-regions by mapping them into a sequence of different resolutions $2^{-r} \times 2^{-r}$ with $r = 0, 1, 2..., R$ as in[3].

X_i^l and Y_i^l respectively denote the sets of feature descriptors in the i^{th} sub-regions at scale l of image vector sets X, Y. At each resolution r, we apply the normalized intersection kernel $\mathscr{F}^r$ using the square root diagonal normalization method to measure the similarity as

$$\mathscr{F}^r(X_i^l, Y_i^l) = \frac{\mathcal{I}(X_i^l(r), Y_i^l(r))}{\sqrt{\mathcal{I}(X_i^l(r), X_i^l(r))\mathcal{I}(Y_i^l(r), Y_i^l(r))}}, \tag{4}$$

where $X_i^l(r)$, $Y_i^l(r)$ respectively denote the sets X_i^l, Y_i^l at resolution r. The histogram intersection between X and itself is equivalent with its cardinality. Consequently, letting $\mathcal{N}_{X_i^l(r)}$ and $\mathcal{N}_{Y_i^l(r)}$ be the cardinality of sets $X_i^l(r)$ and $Y_i^l(r)$, the equation (4) is rewritten as

$$\mathscr{F}^r(X_i^l, Y_i^l) = \frac{\mathcal{I}(X_i^l(r), Y_i^l(r))}{\sqrt{\mathcal{N}_{X_i^l(r)}\mathcal{N}_{Y_i^l(r)}}}. \tag{5}$$

The square root diagonal normalization of the intersection kernel not only satisfies Mercer's conditions[13], but also penalizes the difference in cardinality between sets as in equation (5).

To obtain the synthetic similarity measurement of the coarse-to-fine model, we define the linear combination over a sequence of local kernels, each term of which is calculated using equation (5) at each resolution. Accordingly, the kernel function $\mathscr{F}$ between two sets X_i^l and Y_i^l in the coarse-to-fine model is formulated as

$$\begin{aligned} &\mathscr{F}(X_i^l, Y_i^l) = \sum_{r=0}^{R} \theta_r \mathscr{F}^r(X_i^l, Y_i^l) \\ &\text{where} \quad \sum_{r=0}^{R} \theta_r = 1, \theta_r \geq 0, \forall r = 0, 1, 2, ..., R. \end{aligned} \tag{6}$$

Moreover, when the linear combination of local kernels is integrated with SVM, it can be reformulated as an MKL problem where basic local kernels are defined as equation (5) across the resolutions of the sub-region as

$$\begin{aligned} \min_{\vec{w_\alpha}, w_0, \vec{\xi}, \vec{\theta}} \quad & \frac{1}{2}(\sum_{\alpha=1}^{\Re} \theta_\alpha \left\|\vec{w_\alpha}\right\|_2)^2 + C\sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & y_i(\sum_{\alpha=1}^{\Re} \theta_\alpha \langle \vec{w_\alpha}, \Phi_\alpha(\vec{x_i})\rangle + w_0) \geq 1 - \xi_i \\ & \textstyle\sum_{\alpha=1}^{\Re} \theta_\alpha = 1, \vec{\theta} \geq \vec{0}, \vec{\xi} \geq \vec{0}, \end{aligned} \tag{7}$$

where $\vec{x_i}$ is an image sample, y_i is the category label for $\vec{x_i}$, N is the number of training samples, ($\vec{w_\alpha}$, w_0, and $\vec{\xi}$) are parameters of SVM, C is a soft margin parameter defined by users to penalize training errors in SVM, $\vec{\theta}$ is a weight vector for basic local kernels, $\Re$ is the number of the basic local kernels of the sub-region over the sequence of resolutions, $\vec{\theta} \geq \vec{0}$ means that any entry of vector $\vec{\theta}$ is nonnegative, $\Phi(\vec{x})$ is the function

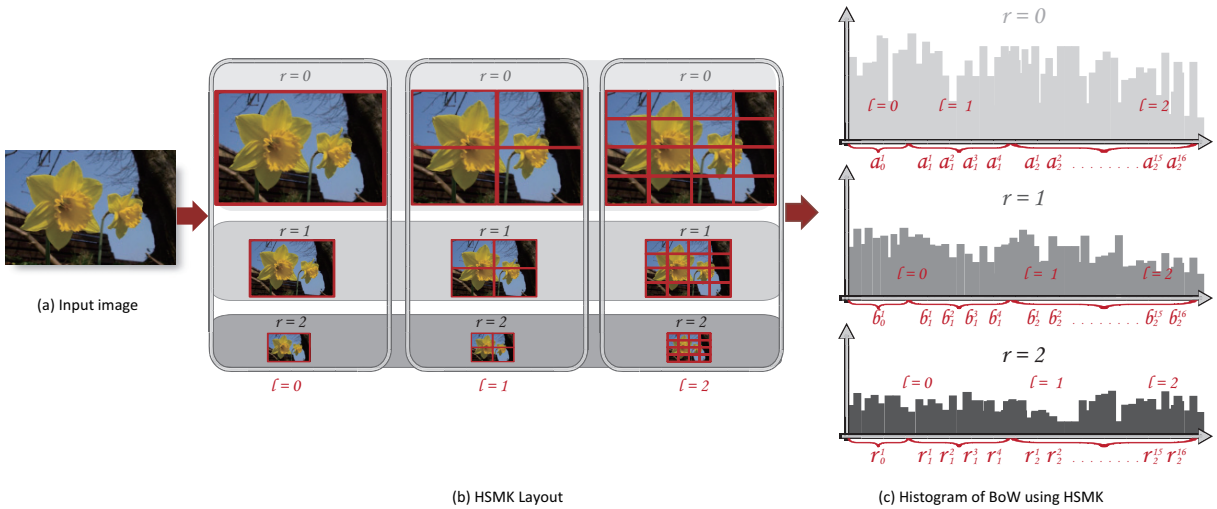

Fig. 2 Illustration for HSMK applied to images X and Y with $L = 2$ and $R = 2$ (a). HSMK first partitions the images into $2^l \times 2^l$ sub-regions with $l = 0, 1, 2$ as SPMK (b). However, HSMK applies a coarse-to-fine model for each sub-region by considering it on a sequence of different resolutions $2^{-r} \times 2^{-r}$ with $r = 0, 1, 2$ (c). The weight set notes (a_i^j, b_i^j, r_i^j), where i is $0, 1, 2$ and j is $1, 2, ..., 15, 16$. The equation (8) with the weight vector achieved from the uniform MKL is applied to obtain better optimal matching approximation between sub-regions instead of using the BoW model as in SPMK.

that maps the vector $\vec{x}$ into the reproducing Hilbert space, and $< \cdot, \cdot >$ denotes the inner product. MKL solves the parameters of SVM and the weight vector simultaneously for basic local kernels.

These basic local kernels are defined analogously across resolutions of the sub-region. Therefore, the redundant information between them is high. The experiments described by Gehler and Nowozin[8] and especially Kloft *et al.*[14] have shown that the uniform MKL, which is an approximation of MKL into traditional nonlinear kernel SVM, is the most efficient for this case in terms of both performance and complexity. Consequently, formulae (6) with linear combination coefficients obtained from the uniform MKL method become

$$\mathscr{F}(X_i^l, Y_i^l) = \frac{1}{R+1} \sum_{r=0}^{R} \mathscr{F}^r(X_i^l, Y_i^l). \tag{8}$$

Figure 2 illustrates an application of HSMK with $L = 2$ and $R = 2$. HSMK also maps the sub-regions into a sequence of different resolutions for PMK to obtain better measurement of similarity between them. However, the weight vector is achieved from the uniform MKL. Consequently, it is more efficient and theoretical than the predefined one in PMK. Furthermore, applying the square root diagonal normalization allows it to deal robustly with differences in cardinality that are not considered in PMK. HSMK is formulated based on SPM in the coarse-to-fine model, which is efficient with sets of unordered feature descriptors, even in the presence of differences in cardinality. Mathematically, the formulation of HSMK is the following:

$$\begin{aligned} \mathscr{K}(X, Y) &= \frac{1}{2^L} \mathscr{F}_0(X, Y) + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} \mathscr{F}_l(X, Y) \\ \text{with} \quad \mathscr{F}_l(X, Y) &= \sum_{i=1}^{2^{2l}} \mathscr{F}(X_i^l, Y_i^l) \\ &= \frac{1}{R+1} \sum_{i=1}^{2^{2l}} \sum_{r=0}^{R} \mathscr{F}^r(X_i^l, Y_i^l). \end{aligned} \tag{9}$$

Briefly, HSMK uses the *kd*-tree algorithm to map each feature descriptor into a discrete visual word; then the normalized intersection kernel by the square root diagonal method is applied to the histogram of V bins to measure the similarity. We have $\mathscr{N}$ feature descriptors in the D-dimension space, and the *kd*-tree algorithm costs $O(\log V)$ steps to map feature descriptors. Therefore, the complexity of HSMK is $O(DM \log V)$ with $M = \max(\mathscr{N}_X, \mathscr{N}_Y)$. In fact, the complexity of the optimal matching kernel[15] is $O(DM^3)$.

4. Experimental results

Most recent approaches use local invariant features as an effective means of representing images because they can well describe and match instances of objects or scenes under widely various viewpoints, illuminations, or even background clutter. Among them, SIFT[16] has robust and efficient features. To achieve better discriminative ability, we use the dense SIFT by operating a SIFT descriptor of 16×16 patches computed over each pixel of an image instead of key points[16] or a grid of points[1]. In addition to improving robustness, we convert images into gray scale ones before computing the dense SIFT. Dense features have the capability of capturing uniform regions such as sky, water or grass where key points usually do

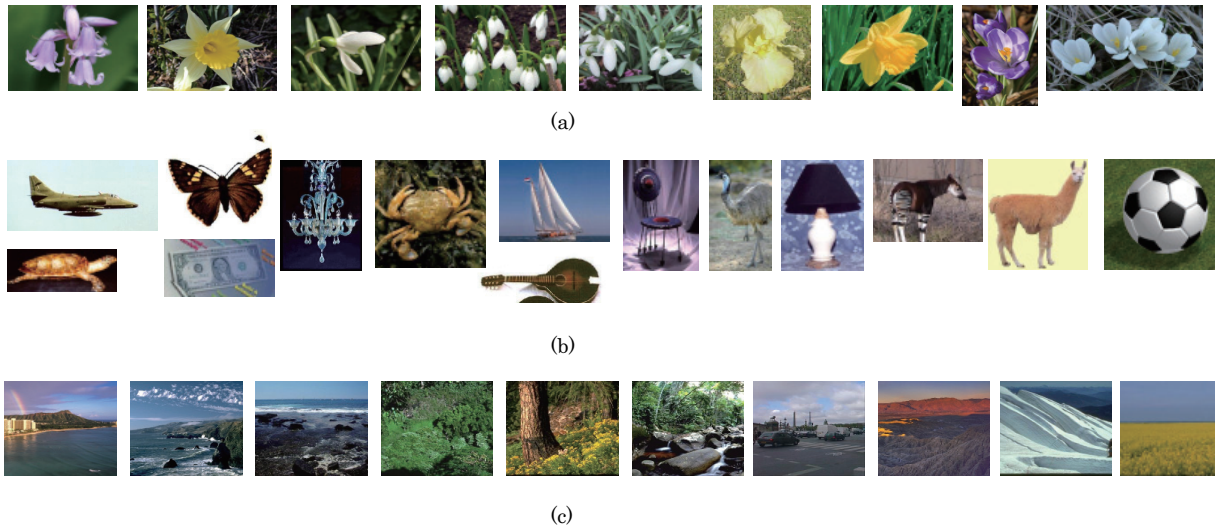

Fig. 3 Example images of various datasets used in experiments: (a) Oxford flower dataset, (b) Caltech 101 dataset, and (c) Scene Categorization dataset

Table 1 Classification rate (%) with a single feature comparison on Oxford Flower dataset (with NN that denotes the nearest neighbor algorithm)

Method	Accuracy (%)
HSV (NN)[20]	43.0
SIFT-Internal (NN)[20]	55.1
SIFT-Boundary (NN)[20]	32.0
HOG (NN)[20]	49.6
HSV (SVM)[8]	61.3
SIFT-Internal (SVM)[8]	70.6
SIFT-Boundary (SVM)[8]	59.4
HOG (SVM)[8]	58.5
SIFT (MSL)[12]	65.3
Dense SIFT (HSMK)	72.9

Table 2 Classification rate (%) comparison between SPMK and HSMK using the Oxford Flower dataset

Kernel	$M = 400$	$M = 800$
SPMK	68.09%	69.12%
HSMK	71.76%	72.94%

not exist. Moreover, the combination of dense features and the coarse-to-fine model allows images to be represented more exactly because feature descriptors achieve more neighbor information across many levels in resolution. We performed unsupervised K-means clustering on a random subset of SIFT descriptors to build visual words. Typically, we used two dictionary sizes M in our experiment: $M = 400$ and $M = 800$.

We conducted experiments for image categorization of two types: object categorization and scene categorization. For object categorization, we used the Oxford Flower dataset[17]. To show the efficiency and scalability of our proposed kernel, we also used the large scale object datasets such as CALTECH-101[5] and CALTECH-256[18]. For scene categorization, we evaluated the proposed kernel on the MIT scene[19] and UIUC scene[1] datasets. Example images of datasets used in experiments are presented in **Fig. 3**.

4.1 Object categorization

To assess the efficiency of the proposed HSMK for object categorization, we compared the classification accuracy with that of conventional SPM in Oxford Flowers dataset and Caltech datasets.

Oxford Flowers dataset: This dataset includes common flowers in the United Kingdom 17 classes, collected by Nilsback *et al.*[17]. Each class has 80 images with wide scale, pose, and light variations. Moreover, intra-class flowers such as irises, fritillaries, and pansies are widely diverse in their colors and shapes. Some cases show close similarity between flowers of different classes such as that between dandelion and Colts' Foot. In our experiments, we followed the set-up of Gehler and Nowozin[8], randomly choosing 40 samples from each class for training and using the rest for testing. We did not use a validation set as in[17),20] for choosing the optimal parameters.

Table 1 shows that our proposed kernel achieved a state-of-the-art results obtained using a single image feature. We use various classifiers for comparison results such as Nearest Neighbour (NN), SVM, and MSL[12] method. The HSMK using dense SIFT gives 72.9% that outperformed not only SIFT-Internal[20] of 70.6%, the best feature for this dataset computed on a segmented image, but also the same feature on SPMK with the optimal weights by MSL of 65.3%. **Table 2** shows that the performance of our HSMK outperformed that of conventional SPMK when using a single SIFT feature.

Caltech datasets: To show the efficiency and robustness of HSMK, we also evaluated its performance on large-scale object datasets, i.e., the CALTECH-101 and CALTECH-256 datasets. These datasets feature high intra-class variability,

Table 3 Classification rate (%) comparison using the CALTECH-101 dataset

	5 training	10 training	15 training	20 training	25 training	30 training
Grauman & Darrell[3]	34.8%	44%	50.0%	53.5%	55.5%	58.2%
Wang *et al.*[12]	-	-	61.4%	-	-	-
Lazebnik *et al.*[1]	-	-	56.4%	-	-	64.6%
Yang *et al.*[11]	-	-	67.0%	-	-	73.2%
Boimann *et al.*[4]	56.9%	-	72.8%	-	-	79.1%
Gehler & Nowozin (MKL)[8]	42.1%	55.1%	62.3%	67.1%	70.5%	73.7%
Gehler & Nowozin (LP-β)[8]	54.2%	65.0%	70.4%	73.6%	75.7%	77.8%
Gehler & Nowozin (LP-B)[8]	46.5%	59.7%	66.7%	71.1%	73.8%	77.2%
Our method (HSMK)	50.5%	62.2%	69.0%	72.3%	74.4%	77.3%

Table 4 Classification rate (%) comparison between SPMK and HSMK using the CALTECH-101 dataset

	5 training	10 training	15 training	20 training	25 training	30 training
SPMK (M = 400)	48.18%	58.86%	65.34%	69.35%	71.95%	73.46%
HSMK(M=400)	50.68%	61.97%	67.91%	71.35%	73.92%	75.59%
SPMK (M = 800)	48.11%	59.70%	66.84%	69.98%	72.62%	75.13%
HSMK(M=800)	50.48%	62.17%	68.95%	72.32%	74.36%	77.33%

Table 5 Classification rate (%) comparison on UIUC Scene (15 classes) dataset

Method	Accuracy (%)
Lazebnik *et al.* (SPMK)[1]	81.4
Yang *et al.* (ScSPM)[11]	80.3
SPMK	79.9
Our method (HSMK)	82.2

Table 6 Classification rate (%) comparison with the CALTECH-256 dataset

Kernel	15 training	30 training
Griffin *et al.* (SPMK)[18]	28.4%	34.2%
Yang *et al.* (ScSPM)[11]	27.7%	34.0%
Gehler & Nowozin (MKL)[8]	30.6%	35.6%
SPMK	25.3%	31.3%
Our method (HSMK)	27.2%	34.1%

Table 7 Classification rate (%) comparison on MIT Scene (8 classes) dataset

Method	Accuracy (%)
GIST[19]	83.7
Local features[21]	77.2
Dense SIFT (SPMK)	85.8
Dense SIFT (HSMK)	88.3

poses, and viewpoints. On CALTECH-101, we conducted experiments with 5, 10, 15, 20, 25, and 30 training samples for each class, including the background class, and used up to 50 samples per class for testing. **Table 3** compares the classification rate results of our approach with other ones. As shown, our approach obtained comparable results with those of state-of-the-art approaches even using only a single feature, whereas others used many types of features and complex learning algorithms such as MKL and linear programming boosting (LP-B)[8]. **Table 4** shows that the result of HSMK outperformed that of SPMK in this case as well. It is noteworthy that when the experiment was conducted without the background class, our approach achieved a classification rate of 78.4% for 30 training samples. This result shows that our approach is efficient in spite of its simplicity.

On the UIUC Scene dataset, we followed the experimental setup described in[1]. We randomly chose 100 training samples per class. The rest were used for testing. As shown in **Table 5**, the result of our proposed kernel also outperformed that of SPMK[1] as well as SPM based on sparse coding[11] for this dataset.

On CALTECH-256, we performed experiments with HSMK using 15 and 30 training samples per class, including the clutter class, and 25 samples of each class for testing. We also re-implemented SPMK[18] but used our dense SIFT to enable a fair comparison of SPMK and HSMK. As shown in **Table 6**, the HSMK classification rate was about 3 percent higher than that of SPMK.

4.2 Scene categorization

We also performed experiments using HSMK on the MIT Scene (8 classes) and UIUC Scene (1 5 classes) datasets. For them, we set M = 400 as the dictionary size. On the MIT Scene dataset, we randomly chose 100 samples per class for training and 100 other samples per class for testing. As shown in **Table 7**, the classification rate for HSMK was 2.5 percent higher than that of SPMK. Our approach also outperformed other local feature approaches[21] as well as local feature combinations[21] by more than 10 percent, and was better than the global feature GIST[19], an efficient feature in scene categoriza-

Table 8 Classification rate (%) comparison between HSMK with vector quantization and HSMK with sparse coding on an Oxford Flower dataset

HSMK	Vector Quantization	Sparse Coding
Linear kernel	63.53%	73.38%
Intersection kernel	72.94%	75.00%

Table 9 Classification rate (%) comparison between HSMK with vector quantization and HSMK with sparse coding on CALTECH-101 dataset with 30 training samples

HSMK	Vector Quantization	Sparse Coding
Linear kernel	65.28%	78.93%
Intersection kernel	77.33%	80.60%

tion.

5. Experiments revisited: HSMK with Sparse Coding

As in section 4., hierarchical spatial matching kernel is proved as an efficient and effective kernel. However, it is still a nonlinear kernel because of the fact that an intersection kernel is used as a basic kernel to build it. Therefore, it is difficult to apply HSMK to address large-scale datasets effectively in terms of time consumption. To help HSMK overcome this issue, we exploit a sparse coding approach and max spooling strategy to make data linear instead of using a vector quantization method by K-means. Therefore, we can replace the intersection kernel by a linear kernel as a basic kernel to construct HSMK based on the linear property of such data. It is worthwhile noting that the performance will become much greater when we apply the linear kernel as a basic kernel in HSMK in the case of using the vector quantization method.

We conducted the same configuration as that in section 4. for experiments of HSMK with sparse coding, but we set a dictionary size of $M = 800$. For sparse coding, we apply l_1 regularization instead of other regularization constraint like l_0 or l_2 norm because l_1 norm regularization is known as the best choice for an image categorization problem[11),22)]. Subsequently, we follow an efficient algorithm proposed by Lee *et al.*[23)] to achieve the solution for a sparse coding problem.

Table 8 and **Table 9** respectively show a comparison of application between vector quantization and sparse coding with HSMK on the Oxford Flower and CALTECH-101 datasets. They prove that sparse coding is an efficient method to make HSMK linear, it can maintain the performance of HSMK as in case of using an intersection kernel as basic kernels. The performance of HSMK with a linear kernel decreases about 1.62% and 1.07% on Oxford Flower and CALTECH-101 dataset respectively in comparison with HSMK with intersection kernel while it is about 10% in the case of using vector quantization.

We can explore from the results presented in Table 8 and Table 9 that the performance of HSMK with the intersection kernel is better than that of HSMK with the linear kernel for both vector quantization and sparse coding in Oxford Flower and CALTECH-101 datasets. Furthermore, it differs with the case of spatial pyramid kernel which in[11)], Yang *et al.* claimed that SPK with linear kernel was also better than SPK with nonlinear kernel when we used sparse coding.

The results of sparse coding for HSMK with the intersection kernel in Table 8 and Table 9 are, respectively, state-of-the art results for the Oxford Flower and CALTECH-101 datasets. Therefore, HSMK with sparse coding is an effective approach for image categorization. Especially the performance of HSMK with a linear kernel can achieve comparable results to those of HSMK with a nonlinear kernel in the case of using sparse coding.

6. Conclusion

As described in this paper, we propose an efficient and robust kernel that we call the hierarchical spatial matching kernel (HSMK). It uses a coarse-to-fine model for sub-regions to improve the spatial pyramid matching kernel (SPMK). Thereby, it obtains more neighbor information through a sequence of different resolutions. In addition, the kernel efficiently and robustly handles sets of unordered features as SPMK and pyramid matching kernel as well as sets having different cardinalities.

Combining the proposed kernel with a dense feature approach was found to be sufficiently effective and efficient. It enabled us to obtain at least comparable results with those by existing methods for datasets of many kinds. Moreover, our approach is simple because it is based solely on a single feature with nonlinear support vector machines, in contrast to other more complicated recent approaches based on multiple kernel learning or feature combinations. Additionally, it is more effective when we combine HSMK with sparse coding.

In most well-known datasets of object and scene categorization, the proposed kernel was also found to outperform SPMK, which is an important component as a basic kernel in multiple kernel learning. Therefore, we can replace SPMK with HSMK to improve the performance of frameworks based on basic kernels.

Acknowledgements

This work was supported in part by JST, CREST, and JSPS.

References

1) S. Lazebnik, C. Schmid, J. Ponce: “Bneyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”, Proc. of Computer Vision and Pattern Recognition(CVPR), pp.2169 – 2178 (2006).
2) S. Boughhorbel, J.P. Tarel, F. Fleuret: “Non-Mercer Kernels for Svm Object Recognition”, Proc. of British Machine Vision Conference (2004).
3) K. Grauman, T. Darrell: “The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features”, Proc. of International Conference on Computer Vision(ICCV), Vol.2, pp.1458 –1465 (2005).
4) O. Boiman, E. Shechtman, M. Irani: “In Defense of Nearest-Neighbor Based Image Classification”, Proc. of Computer Vision and Pattern Recognition(CVPR) (2008).
5) L. Fei-Fei, R. Fergus, P. Perona: “Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories.”, Proc. of Workshop on Generative-Model Based Vision (2004).
6) F. Moosmann, B. Triggs, F. Jurie: “Randomized Clustering Forests for Building Fast and Discriminative Visual Vocabularies”, Proc. of NIPS Workshop on Kernel Learning: Automatic Selection of Kernels (2008).
7) L. Yang, R. Jin, R. Sukthankar, F. Jurie: “Unifying Discriminative Visual Codebook Generation with Classifier Training for Object Category Recognition”, Proc. of Computer Vision and Pattern Recognition(CVPR), pp. 1–8 (2008).
8) P. Gehler, S. Nowozin: “On Feature Combination for Multiclass Object Classification”, Proc. of International Conference on Computer Vision(ICCV), pp.221 –228 (2009).
9) S. Maji, A. Berg, J. Malik: “Classification Using Intersection Kernel Support Vector Machines is Efficient”, Proc. of Computer Vision and Pattern Recognition(CVPR), pp. 1–8 (2008).
10) J. Mairal, F. Bach, J. Ponce, G. Sapiro: “Online Dictionary Learning for Sparse Coding”, Proc. of International Conference on Machine Learning(ICML), pp.689–696 (2009).
11) J. Yang, K. Yu, Y. Gong, T. Huang: “Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification”, Proc. of Computer Vision and Pattern Recognition(CVPR), pp.1794 –1801 (2009).
12) S.C. Wang, Y.C.F. Wang: “A Multi-Scale Learning Framework for Visual Categorization”, Proc. of Asian Conference on Computer Vision(ACCV) (2010).
13) B. Scholkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, USA (2001).
14) M. Kloft, U. Brefeld, P. Laskov, S. Sonnenburg: “Non-Sparse Multiple Kernel Learning”, Proc. of NIPS Workshop on Kernel Learning: Automatic Selection of Kernels (2008).
15) R.I. Kondor, T. Jebara: “A Kernel Between Sets of Vectors”, Proc. of International Conference on Machine Learning(ICML), pp.361–368 (2003).
16) D.G. Lowe: “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, Vol.60, No.2, pp.91–110 (2004).
17) M.E. Nilsback, A. Zisserman: “A Visual Vocabulary for Flower Classification”, Proc. of Computer Vision and Pattern Recognition(CVPR), pp.1447–1454 (2006).
18) G. Griffin, A. Holub, P. Perona: “Caltech-256 Object Category Dataset”, Technical Report No.7694, California Institute of Technology (2007).
19) A. Oliva, A. Torralba: “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope”, International Journal of Computer Vision, Vol.42, No.3, pp.145-175 (2001).
20) M.E. Nilsback, A. Zisserman: “Automated Flower Classification over a Large Number of Classes”, Proc. of Indian Conference on Computer Vision, Graphics and Image Processing(ICVGIP) (2008).
21) M. Johnson: “Semantic Segmentation and Image Search”, Ph.D. thesis, University of Cambridge (2008).
22) R. Raina, A. Battle, H. Lee, B. Packer, A.Y. Ng: “Self-Taught Learning: Transfer Learning from Unlabeled Data”, Proc. of International Conference on Machine Learning(ICML) (2007).
23) H. Lee, A. Battle, R. Raina, A.Y. Ng: “Efficient Sparse Coding Algorithms”, Proc. of NIPS (2006).
24) T.T. Le, Y. Kang, A. Sugimoto, S.T. Tran, T.D. Nguyen: “Hierarchical Spatial Matching Kernel for Image Categorization”, Proc. of International Conference on Image Processing and Recognition (ICIAR) (2011).

(Received November 30, 2012)

Tam T. LE
He received his B.S. degree in the honors program and M.S degree in Computer Science from the University of Science, Vietnam National University HCMC, Vietnam in 2008 and 2011, respectively. He was a Lecturer and Research Assistant at the University of Science, VNU-HCMC, Vietnam. He is currently a PhD student at Graduate school of Informatics, Kyoto University, Japan. His research interests include image categorization, feature representation, sparse coding, and kernel methods.

Yousun KANG (*Member*)
She received a Ph.D. degree from Tokyo Institute of Technology in 2010. She worked with Toyota Central R&D LABS., Inc. for three years from 2007. During 2010–2011, she was a researcher in the National Institute of Informatics, Japan. She is currently an Associate Professor at Tokyo Polytechnic University. Her research interests include texture analysis, scene understanding, pattern recognition, image processing, and computer vision. She is a member of the RSJ, IIEEJ and IEICE of Japan.

Akihiro SUGIMOTO
He received his B.S, M.S, and Dr. Eng. degrees in Mathematical Engineering from The University of Tokyo in 1987, 1989, and 1996, respectively. After working at Hitachi Advanced Research Laboratory, ATR, and Kyoto University, he joined the National Institute of Informatics, Japan, where he is currently a professor. During 2006–2007, he was a visiting professor at ESIEE, France. He received a Paper Award from the Information Processing Society in 2001. He is a member of IEEE. He is interested in mathematical methods in engineering. Particularly his current main research interests include discrete mathematics, approximation algorithm, vision geometry, and modeling of human vision.