Research Article

Adaptive Multilevel Kernel Machine for Scene Classification

Junlin Hu,¹ Liang Wang,² Fuqing Duan,³ and Ping Guo¹

¹ Image Processing and Pattern Recognition Laboratory, Beijing Normal University, Beijing 100875, China

² College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China

³ College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

Correspondence should be addressed to Fuqing Duan; fqduan@bnu.edu.cn and Ping Guo; pguo@ieee.org

Received 23 January 2013; Accepted 24 February 2013

Academic Editor: Carlo Cattani

Copyright © 2013 Junlin Hu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Scene classification is a challenging problem in computer vision applications and can be used to model and analyze a special complex system, the internet community. The spatial PACT (Principal component Analysis of Census Transform histograms) is a promising representation for recognizing instances and categories of scenes. However, since the original spatial PACT only simply concatenates compact census transform histograms at all levels together, all levels have the same contribution, which ignores the difference among various levels. In order to ameliorate this point, we propose an adaptive multilevel kernel machine method for scene classification. Firstly, it computes a set of basic kernels at each level. Secondly, an effective adaptive weight learning scheme is employed to find the optimal weights for best fusing all these base kernels. Finally, support vector machine with the optimal kernel is used for scene classification. Experiments on two popular benchmark datasets demonstrate that the proposed adaptive multilevel kernel machine method outperforms the original spatial PACT. Moreover, the proposed method is simple and easy to implement.

1. Introduction

With the rapid development of the smart phone and internet, millions of users can now share their photos or videos taken by these smart phones on the internet and then construct a special complexity system, an internet community. On one hand, these photos and videos of each internet community can be directly used to model the real world by 3D reconstruction, scene retrieval, navigation, and augmented reality [1–7]. On the other hand, they can provide information to model and analyze the complexity system by scene classification and face recognition [4–6]. Great advances have been made on the face recognition, which is widely used in real applications. However, scene classification still is a great challenging problem due to the clutter and the occlusion. Related techniques for associating scenes with semantic labels have a high potential for improving the performance of other computer vision applications such as image browsing, retrieval, and categorization [2, 4, 8]. Motivated by these applications, many image representation models have been proposed for addressing this problem, such as probabilistic latent semantic analysis model [9], part-based model [10], and BoW (bag of

visual words) model [3, 6]. Among these existing models, the BoW model has shown an excellent performance and been widely used in many real applications due to its robustness to scale, translation, and rotation transform, including image classification, image annotation [11, 12].

Generally, the BoW model treats an image as a collection of unordered appearance descriptors extracted from local image patches, quantizes them into discrete visual words, and computes a compact histogram representation for semantic image classification. However, the BoW method discards the spatial information of local descriptors, which makes the representation power of the descriptors severely limited. To include information about the spatial layout of the descriptors, Lazebnik et al. [13] propose an extension of BoW model, named as SPM (spatial pyramid matching), which has achieved excellent results on several challenging image classification tasks. For each image, spatial pyramid partitions an image into increasingly finer patches at different levels, and a pyramid match kernel is used to estimate the overall perceptual similarity between images. In their work, the k-means technique is used to generate the codebook, and each local descriptor is only quantized to its nearest

visual word. However, such a hard assignment may result in severe quantization error. To reduce quantization error, Yang et al. [14] propose an improved version of SPM called ScSPM (sparse coding SPM), which computes a spatial pyramid image representation based on sparse codes of the SIFT (scale invariant feature transform) [15]. In this way, the ScSPM can automatically learn the optimal codebook by replacing k-means with sparse coding, and concurrently search for the optimal weights to be assigned to the visual words for each local descriptor. By adopting sparse coding to generate the codebook, their model achieves an excellent performance in image classification [14, 16].

Although the BoW approach has achieved good results for scene classification, the most important procedures in this method, codebook generation and feature quantization, have a high complexity both in time and space. Recently, Wu and Rehg [17] introduce PACT, an effective representation that fulfills the need for recognizing both instances and categories of places and scenes. The CT (Census Transform) encodes local shape information with significant relevance between neighboring CT values, while PCA (Principal Component Analysis) can extract the global shape information in an image patch to represent image compactly. To capture spatial information, they also propose spatial PACT scheme, spatial pyramid representation of PACT, which encodes roughly global spatial arrangement of subblocks in an image and tries to seek the tradeoff between discriminative power and invariance for scene recognition tasks. Furthermore, they conclude that the spatial PACT has several important advantages in comparison to previous state-of-the-art feature representations for scene classification, such as superior recognition performance, no parameter to tune, and extremely fast evaluation speed.

The spatial PACT, however, only concatenates compact histograms at all levels together and ignores the difference among different levels. In this paper, to alleviate the impact of this problem, we propose an adaptive multilevel kernel machine approach, which firstly computes a set of base kernels corresponding to each level in the pyramid of PACT, then a promising adaptive weight learning scheme is used to seek a set of optimal weights for best fusing all base kernels for scene categorization. Lastly, support vector machine (SVM) classifier [18] with the optimal kernel is employed to perform the scene classification task.

The remainder of this paper is organized as follows. Section 2 briefly reviews some related works, mainly including Census Transform, CENTRIST (CENSus TRansform hISTogram), PACT, and spatial PACT, respectively. Section 3 presents the proposed multilevel kernel machine approach and introduces an adaptive weight learning scheme. Section 4 presents experiment results on two public datasets. Finally, we conclude the paper in Section 5.

2. Related Works

In this section, we describe Census Transform, PACT, and spatial PACT in detail, respectively.

FIGURE 1: The CT operation.

2.1. Census Transform and CENTRIST. According to descriptions in [17, 19], Census Transform is a nonparametric local transform, which is originally designed to establish the correspondence between image patches. Census Transform compares the intensity value of a pixel with its eight neighboring pixels (see illustration in Figure 1). If the center pixel is bigger than (or equal to) one of its neighbors, a bit 1 is set at the corresponding location. Otherwise a bit 0 is set.

Usually the eight bits generated from intensity comparisons can be put together in any order. In our work, we adopt the same order as that in [17] and collect these bits from left to right and top to bottom. Consequently the eight bits are converted to a base-10 number, which is the CT value for this center pixel. Another important property of the transform is that CT values of neighboring pixels are highly correlated. What is more, similar to other nonparametric local transforms which are based on intensity comparisons, Census Transform is robust to illumination changes, gamma variations.

As an intuitive representation, a Census Transformed image is created by replacing a pixel with its CT value. An example in Figure 2 shows that the Census Transform retains global structures of the image (e.g., especially discontinuities), meanwhile, captures the local structures as what is expected. Thus, a histogram of the CT values in an image (or image patch) encodes both local and global information.

2.2. PACT and Spatial PACT. CT values, denoted in the base-2 format, represent homogeneous regions with small variations (e.g., (00001000)₂). That is, there exist strong correlations between pairs of CT values. In order to remove these correlation effects and get a more compact image representation, Wu and Rehg [17] propose PACT by performing the PCA operation on the CT histograms. When computing histograms and PCA, they remove two bins with CT = 0 and CT = 255 and further normalize the CT histograms and eigenvectors such that they have zero mean and unit norm.

Since PACT only encodes global shape structure in a small image patch [17], to capture the global structure of an image in larger scales, Wu and Rehg [17] also propose a spatial PACT representation based on the SPM approach [20]. The spatial pyramid scheme divides an image into patches and concatenates corresponding PACT features in these patches. Since it encodes roughly spatial structure of an image, the recognition accuracy usually can be improved. As described in [20], the level 2 split in a spatial pyramid divides the image into $2^2 \times 2^2 = 16$ blocks. We also shift the division (dash line blocks) to avoid artifacts produced by the nonoverlapping division, which makes a total of 25 blocks in level 2. Similarly, level 1 and level 0 have 5 blocks and 1 block, respectively. It is noted that the image is resized at different levels so that all blocks contain the same number of pixels. These blocks

FIGURE 2: An example of Census Transformed image.

FIGURE 3: An intuitive illustration of the level 2, 1, and 0 of spatial pyramid split of an image, respectively.

in different levels are shown in Figure 3. PACT features in all blocks are then concatenated to form an overall long feature vector. For example, if 40 eigenvectors are selected in PACT, the $L = 3$ levels spatial pyramid will generate a long feature vector whose dimension is $40 \times (25 + 5 + 1) = 1240$.

After spatial PACT feature vectors are extracted from images, SVM classifier is used to perform scene image classification task. In addition, the CT values are only based on pixel intensity comparisons, so it might be helpful to include a few image statistics, such as average value and standard deviation of pixels in an image block. This statistical information is appended to spatial PACT in the input to SVM classifiers for scene classification problem.

3. Proposed Approach

The original spatial PACT, however, only simply concatenates compact histograms at all levels together and discards the difference among different pyramid levels. In this section, we will introduce an adaptive multilevel kernel machine scheme. It firstly computes a set of base kernel matrixes corresponding to each level of the pyramid of PACT, and then a promising adaptive weight learning scheme is used to seek a set of optimal weights for best fusing all base kernels for scene image categorization.

3.1. Multilevel Kernel Machine. Let $\mathbf{x} = [\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{L-1}]$ and $\mathbf{y} = [\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{L-1}]$ denote two spatial PACT feature vectors, where L is the total number of pyramid levels and $\mathbf{x}_l$ (or $\mathbf{y}_l$) is PACT feature vector located at the level l ($l = 0, \dots, L-1$). Generally speaking, PACT features from various levels have different contributions to image representation. In order to fully exploit the descriptive capacity of each level, we compute a set of base kernels corresponding to each pyramid level of PACT. The aim is to learn a SVM classifier [18, 21], where

FIGURE 4: The proposed multilevel kernel machine scheme.

the kernel is learnt to be a linear combination of given base kernels [22].

Here there are several ways to estimate similarity between $\mathbf{x}_l$ and $\mathbf{y}_l$, such as RBF kernel and linear kernel. By coupling each representation with its corresponding distance function, we obtain a set of similarity-based kernel matrices $\{\mathbf{K}_l\}_{l=0}^{L-1}$, for RBF kernel:

$$\mathbf{K}_l(\mathbf{x}_l, \mathbf{y}_l) = k_l(\mathbf{x}_l, \mathbf{y}_l) = \exp(-\gamma \|\mathbf{x}_l - \mathbf{y}_l\|^2) \quad (1)$$

or for linear kernel:

$$\mathbf{K}_l(\mathbf{x}_l, \mathbf{y}_l) = k_l(\mathbf{x}_l, \mathbf{y}_l) = \mathbf{x}_l \mathbf{y}_l^T. \quad (2)$$

Here, $\gamma > 0$ and $\mathbf{K}_l$ is the similarity kernel corresponding to level l . Our multilevel kernel machine approach is to find an optimal linear combination of given base kernels. Suppose we have obtained a set of base kernel functions $\{k_l\}_{l=0}^{L-1}$ (or base kernel matrices $\{\mathbf{K}_l\}_{l=0}^{L-1}$), an ensemble kernel function k (or an ensemble kernel matrix $\mathbf{K}$) is then defined by

$$k(\mathbf{x}_i, \mathbf{y}_i) = \sum_{l=0}^{L-1} \beta_l k_l(\mathbf{x}_i, \mathbf{y}_i) \quad \left(\sum_{l=0}^{L-1} \beta_l = 1, \beta_l \geq 0 \right), \quad (3)$$

$$\mathbf{K} = \sum_{l=0}^{L-1} \beta_l \mathbf{K}_l \quad \left(\sum_{l=0}^{L-1} \beta_l = 1, \beta_l \geq 0 \right).$$

With (1), (2), and (3), determining a set of optimal coefficients $\boldsymbol{\beta} = \{\beta_0, \beta_1, \dots, \beta_{L-1}\}$ can now be interpreted as finding appropriate weights for best fusing these spatial pyramid levels of PACT feature representations. Lastly, SVM classifier with kernel matrix $\mathbf{K}$ can be used to perform scene image classification task.

The flowchart of our proposed multilevel kernel machine scheme is shown in Figure 4 for three levels. Note that SP means spatial pyramid scheme.

FIGURE 5: Some example images from the 15-class scene category dataset (two images per category).

3.2. Adaptive Weight Learning Scheme. In our previous work [23], the weight parameters $\beta = \{\beta_0, \beta_1, \dots, \beta_{L-1}\}$ are empirically set by a large number of experiments, which is a time consuming procedure and does not always achieve reasonable results. To address this problem, in this paper, we adopt an adaptive weight learning scheme; our aim is to find the optimum weights to best fuse the entire various PACT features at different levels by a promising method.

According to the descriptions in [24], their basic idea of weight learning is that finding optimal weight is actually equivalent to the problem of fusing predefined kernels according to the label information, which is a standard kernel matrix learning problem.

As for kernel matrix learning, it is always believed that the target similarity should be close to the label similarity. Therefore, we need to seek $\beta = \{\beta_0, \beta_1, \dots, \beta_{L-1}\}$, such that the matrix $\mathbf{K}$ is close to the label similarity matrix $\mathbf{C}$. Here, the label similarity is defined as follows:

$$C_{i,j} = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where c_i is the corresponding label information of target images $\mathbf{I}_i$.

Then, the following criterion is used to minimize the distance between $\mathbf{K}$ and $\mathbf{C}$, which is essentially a quadratic programming problem:

$$\|\mathbf{K} - \mathbf{C}\|_F^2 = \sum_i \sum_j (K_{i,j} - C_{i,j})^2. \quad (5)$$

Here $\|\mathbf{X}\|_F^2 = \text{tr}(\mathbf{X}\mathbf{X}^T)$, where $\text{tr}(\cdot)$ means trace operation. More intuitively, $\|\mathbf{K} - \mathbf{C}\|_F^2$ is the sum of element-wise distance between $\mathbf{K}$ and $\mathbf{C}$, where the subscript “F” denotes Frobenius norm operation.

In practice, if we use (5) directly, we may suffer a scale problem. As the elements in $\mathbf{K}$ are not scaled in $[0, 1]$, minimizing $\|\mathbf{K} - \mathbf{C}\|_F^2$ may not guarantee a reasonable result. In order to address the scaling problem, all the elements in $\mathbf{K}$ should be constrained within $[0, 1]$ by following method:

$$\mathbf{K} = \sum_{l=0}^{L-1} \beta_l \overline{\mathbf{K}}_l, \quad m_l = \max(|\mathbf{K}_l|), \quad (6)$$

$$\overline{\mathbf{K}}_l = \frac{\mathbf{K}_l}{m_l}, \quad l = 0, \dots, L-1.$$

Suburb	99.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0
Coast	0.4	85.0	0.4	1.9	0.0	1.9	9.6	0.0	0.0	0.4	0.4	0.0	0.0	0.0
Forest	0.0	0.0	92.5	0.0	0.0	4.8	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.4
Highway	0.0	2.5	0.0	87.5	1.9	1.3	3.1	1.3	0.0	0.0	0.6	0.0	0.6	0.6
Inside city	1.9	0.0	0.0	0.0	83.2	0.0	0.0	3.4	7.7	0.0	0.0	1.4	0.0	0.5
Mountain	0.0	1.8	2.9	0.4	0.0	87.2	6.9	0.0	0.4	0.4	0.0	0.0	0.0	0.0
Open country	0.3	9.7	3.9	2.3	0.0	6.8	76.5	0.6	0.0	0.0	0.0	0.0	0.0	0.0
Street	1.0	0.0	0.0	1.0	4.2	0.0	0.0	90.1	2.1	0.0	0.0	0.5	0.0	1.0
Tall building	0.0	0.0	0.0	0.4	5.9	2.3	0.0	0.4	86.7	0.0	0.4	2.7	0.0	0.4
Office	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	93.3	0.9	0.0	2.6	1.7
Bedroom	0.0	0.0	0.0	0.0	0.9	1.7	0.0	0.9	0.0	2.6	69.8	0.9	5.2	16.4
Industrial	0.0	0.5	0.9	0.5	2.8	0.5	0.5	0.5	1.4	0.0	1.9	74.9	1.9	0.0
Kitchen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	4.5	3.6	2.7	77.3	0.9	10.0
Living room	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.5	1.1	3.2	21.2	1.1	1.6	68.8
Store	0.0	0.0	0.5	0.0	1.4	0.0	0.0	0.9	0.5	1.4	3.3	2.8	0.0	89.3
	Suburb	Coast	Forest	Highway	Inside city	Mountain	Open country	Street	Tall building	Office	Bedroom	Industrial	Kitchen	Living room
	Suburb	Coast	Forest	Highway	Inside city	Mountain	Open country	Street	Tall building	Office	Bedroom	Industrial	Kitchen	Living room

FIGURE 6: The Confusion Matrix on the 15-class scene category dataset (%). Average classification rates for individual classes are listed along the diagonal. The entry in the i th row and j th column is the percentage of images from class i that are misidentified as class j . Here, average classification accuracy is 84.09%.

FIGURE 7: Some misclassified instances in the 15-class scene category dataset. The images in the first row are from bedroom, but they are misclassified as living room (the misclassified rate is about 16.40 percent), and the images in the second row are from class living room, but they are misclassified as bedroom (the misclassified rate is about 21.20 percent).

Here $\bar{\mathbf{K}}_l$ is a normalized variant of $\mathbf{K}_l$, “ $|\cdot|$ ” is the absolute operation, and m_l is the largest absolute value in $\mathbf{K}_l$.

To avoid overfitting in the learning procedure, a regularization term $\|\boldsymbol{\beta}\|^2$ can be added to (5). Then, we can find weights $\boldsymbol{\beta} = \{\beta_0, \beta_1, \dots, \beta_{L-1}\}$ by solving the optimal problem:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{K} - \mathbf{C}\|_F^2 + \lambda \|\boldsymbol{\beta}\|^2 \\ \text{s.t.} \quad & \sum_{l=0}^{L-1} \beta_l = 1, \quad \beta_l \geq 0, \end{aligned} \quad (7)$$

where λ is a tradeoff parameter to prevent overfitting problem. In our experiments, the parameter λ is set to $0.1 \times \text{tr}(\mathbf{K}_0 \mathbf{K}_0^T)$, and,

$$\mathbf{K}_0 = \frac{1}{L} \sum_{l=0}^{L-1} \bar{\mathbf{K}}_l. \quad (8)$$

More intuitively, we may reformulate the first term in (7) as following:

$$\begin{aligned} \|\mathbf{K} - \mathbf{C}\|_F^2 &= \left\| \sum_{l=0}^{L-1} \beta_l \bar{\mathbf{K}}_l - \mathbf{C} \right\|_F^2 \\ &= \text{tr} \left(\left(\sum_{l=0}^{L-1} \beta_l \bar{\mathbf{K}}_l \right)^T \left(\sum_{l=0}^{L-1} \beta_l \bar{\mathbf{K}}_l \right) - 2\mathbf{C}^T \sum_{l=0}^{L-1} \beta_l \bar{\mathbf{K}}_l + \mathbf{C}^T \mathbf{C} \right) \\ &= \boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} - 2\mathbf{b}^T \boldsymbol{\beta} + d, \end{aligned} \quad (9)$$

FIGURE 8: Images from 8 sports event dataset (one image per category).

Badminton	85.0	5.0	1.7	1.7	0.0	0.0	0.0	6.7
Bocce	5.0	65.0	11.7	1.7	6.7	3.3	0.0	6.7
Croquet	5.0	15.0	70.0	5.0	5.0	0.0	0.0	0.0
Polo	3.3	3.3	8.3	78.3	1.7	5.0	0.0	0.0
Rockclimbing	0.0	3.3	0.0	1.7	91.7	0.0	0.0	3.3
Rowing	0.0	1.7	0.0	3.3	0.0	90.0	3.3	1.7
Sailing	0.0	0.0	0.0	0.0	0.0	6.7	91.7	1.7
Snowboarding	0.0	6.7	5.0	0.0	3.3	3.3	0.0	81.7
	Badminton	Bocce	Croquet	Polo	Rockclimbing	Rowing	Sailing	Snowboarding

FIGURE 9: The Confusion Matrix on the 8 class event dataset (%). In this run, the average classification accuracy is 81.67%.

where

$$\begin{aligned}
 A_{i,j} &= \text{tr} \left(\left(\overline{\mathbf{K}}_l \right)^T \overline{\mathbf{K}}_l \right), \\
 b_j &= \text{tr} \left(\mathbf{C}^T \overline{\mathbf{K}}_l \right), \\
 d &= \text{tr} \left(\mathbf{C}^T \mathbf{C} \right).
 \end{aligned} \tag{10}$$

By combining (7) and (9) together, the optimal problem in (7) can be reformulated as the quadratic programming, and the optimal weights $\boldsymbol{\beta} = \{\beta_0, \beta_1, \dots, \beta_{L-1}\}$ can be obtained as following:

$$\begin{aligned}
 \min_{\boldsymbol{\beta}} \quad & \boldsymbol{\beta}^T (\mathbf{A} + \lambda \mathbf{E}) \boldsymbol{\beta} - 2\mathbf{b}^T \boldsymbol{\beta} \\
 \text{s.t.} \quad & \sum_{l=0}^{L-1} \beta_l = 1, \quad \beta_l \geq 0.
 \end{aligned} \tag{11}$$

Here, matrix $\mathbf{E}$ is the identity matrix with the same size of matrix $\mathbf{A}$.

4. Experimental Results

In this section, we implement and evaluate our methods on two widely used datasets: 15-class scene category dataset [13] and the 8-class event dataset [25].

4.1. Experimental Setup. In our setup, following the same experiment procedure of the spatial PACT [17], the images are all preprocessed into gray scale. For each dataset, experiments are repeated five times to obtain reliable results by randomly selecting training and testing images. The average of per-class classification accuracy is recorded for each run, and the final results are shown as the mean and standard deviation of the classification accuracy. For spatial PACT, CENTRIST descriptors are reduced to 40 dimensions by PCA, and the

TABLE 1: Classification accuracy comparison on the 15-class scene category dataset.

L	Method	Feature type	Classification accuracy (%)
2	SPM [13]	SIFT, 400 cluster centers	81.40 ± 0.50
2	ScSPM [14]	SIFT, 400 cluster centers	80.28 ± 0.93
2	Spatial PACT [17]	CENTRIST, 40 eigenvectors	83.38 ± 0.53
2	Our method	CENTRIST, 40 eigenvectors	83.45 ± 0.47

TABLE 2: Performance comparison using different kernels on the 15-class scene category dataset.

L	Method	RBF kernel (1)	Linear kernel (2)
2	spatial PACT [17]	83.38 ± 0.53	81.92 ± 0.48
2	Our method	83.45 ± 0.47	82.51 ± 0.70

largest 40 eigenvalues accounted for about 90 percents and 94 percents of the sum of all eigenvalues in these datasets, respectively. In regard to computing CENTRIST descriptors and PCA eigenvectors, we remove the two bins with $CT = 0$, 255 and normalize the CENTRIST descriptors and PCA eigenvectors such that they have zero mean and unit norm. Also, average value and standard deviation of pixels in each image block are appended to spatial PACT for including some statistics information. In this paper, $L = 3$ is used in spatial PACT for all datasets, which has achieved the best performance in [17]. Thus, the dimension of the spatial PACT feature vector of $L = 3$ levels becomes $(40 + 2) \times (25 + 5 + 1) = 1302$. The dimension of PACT vector in pyramid level $l = 2$ is $(40 + 2) \times 25 = 1050$, level $l = 1$ is $(40 + 2) \times 5 = 210$, and level $l = 0$ is $(40 + 2) \times 1 = 42$, respectively. Lastly, multiclass classification task is done using an excellent toolbox LibSVM [26] which is trained by one-versus-one rule and parameters $(C, \gamma) = (8, 2^{-7})$ are recommended for RBF kernel [17].

4.2. The 15-Class Scene Category Dataset. The 15-class scene category dataset contains totally 4485 images with 15 categories (including coast, forest, kitchen, street, and office). Images are about 300×250 in average size, with 210 to 410 images in each category. This dataset contains a wide range of scene categories in both indoor and outdoor environments. In addition, this is one of the most complete scene category dataset used in the literature so far. Figure 5 lists several representative example images from this dataset.

Following the same experimental procedure of the spatial PACT [17], the first 100 images in each category are used to obtain a projection matrix by PCA, 100 images per category are chosen randomly for training and the remaining images for testing. The comparison results are shown in Table 1. From this table, we can see that our adaptive multilevel machine method achieves the highest accuracy and outperforms SPM about 2 percents, ScSPM, about 3 percents, respectively. To further illustrate the performance of our adaptive multilevel kernel machine scheme, Table 2 also shows the performance comparison using different kernels, respectively. We can find that linear kernel does not perform as well as RBF kernel, but the difference is not significant (about 1 percent). In addition, we can find that the linear kernel has more improvement

in classification accuracy than RBF kernel. To summary, all these results show that fusing different levels PACT can effectively improve classification performance.

Moreover, Figure 6 shows a confusion matrix from one run on this dataset using our approach. It can be observed that our method performs better for categories suburb, forest, street, and office (more than 90%). Meanwhile, we also notice that the categories bedroom and living room have a high percentage being classified wrongly. Not surprisingly, this may result from that they are visually similar to each other. Figure 7 lists some misclassified images.

4.3. The 8-Class Event Dataset. The event dataset [25] contains 1579 images of eight sport event classes: 200 badminton, 137 bocce, 236 croquet, 182 polo, 194 rock climbing, 250 rowing, 190 sailing, and 190 snowboarding, respectively. The images in this dataset are almost with small resolution variations and are high-resolution ones. Figure 8 lists some images from this dataset. In [25], this dataset is used to classify these events by integrating scene and object categorizations. In our work, we use this dataset for image scene classification purposes only.

Also, following the common benchmarking procedures on this dataset, we select 70 images per category for training and 60 for testing, randomly. The first 50 images in each class are used to produce the eigenvectors by PCA operation. The performance comparison results are shown in Table 3. From Table 3, we can see that our method achieves the best result on this dataset and outperforms spatial PACT by nearly 1.5 percent in classification accuracy, which shows the effectiveness of our adaptive multilevel kernel machine approach. Similar to the 15-class scene category dataset, we also list confusion matrix on this dataset, as shown in Figure 9. It can be observed that confusion occurs between bocce and croquet and also between polo and croquet, which is consistent with the result of spatial PACT. It is because that these two pairs of events share very similar scene or background.

5. Conclusions

In this paper, we propose an adaptive multilevel kernel machine method for scene classification. Unlike the original spatial PACT, which simply concatenates all pyramid levels of compact PACT histograms together and discards the difference between different pyramid levels, our proposed scheme computes a set of base kernels at each pyramid level of PACT and thus considers the difference existing in various levels. Since PACT features in each pyramid level play

TABLE 3: The performance comparison (%) on the 8-class event dataset.

L	Method	Feature type	RBF kernel (1)	Linear kernel (2)
2	Spatial PACT [17]	CENTRIST, 40 eigenvectors	79.08 $\pm$ 1.15	76.96 $\pm$ 1.16
2	Our method	CENTRIST, 40 eigenvectors	80.50 $\pm$ 1.49	78.38 $\pm$ 1.28

different roles in representing scene, the optimal kernel is learnt to be a linear combination of given base kernels; that is, it can be interpreted as finding appropriate weights for best fusing all pyramid levels PACT feature representations. To seek a set of optimal weights, a promising adaptive weight learning method is employed. Experiments on two popular benchmark datasets are present to demonstrate that our adaptive multilevel kernel machine method outperforms the original spatial PACT on image scene classification.

Acknowledgment

This research work was supported by the Grants from the National Natural Science Foundation of China (Nos. 90820010, 61101207).

References

- [1] N. Snavely, R. Garg, S. M. Seitz, and R. Szeliski, "Finding paths through the world's photos," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 11–21, 2008.
- [2] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.
- [3] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 883–890, October 2005.
- [4] A. Shabou and H. LeBorgne, "Locality-constrained and spatially regularized coding for scene categorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3618–3625, 2012.
- [5] J. H. Zhang, J. W. Zhang, S. Y. Chen, Y. Hu, and H. J. Guan, "Constructing dynamic category hierarchies for novel visual category discovery," in *Proceedings of the Intelligent Robots and Systems (IROS '12)*, pp. 2122–2127, 2012.
- [6] D. Gokalp and S. Aksoy, "Scene classification using bag-of-regions representations," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.
- [7] S. Y. Chen, J. H. Zhang, Y. F. Li, and J. W. Zhang, "A hierarchical model incorporating segmented regions and pixel descriptors for video background subtraction," *IEEE Trans. Industrial Informatics*, vol. 8, no. 1, pp. 118–127, 2012.
- [8] S. Y. Chen, H. Tong, and C. Cattani, "Markov models for image labeling," *Mathematical Problems in Engineering*, vol. 2012, Article ID 814356, 18 pages, 2012.
- [9] T. Hofmann, "Unsupervised learning by probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no. 1–2, pp. 177–196, 2001.
- [10] R. Fergus, P. Perona, and A. Zisserman, "A sparse object category model for efficient learning and exhaustive recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 380–387, June 2005.
- [11] J. Wu and J. M. Rehg, "Beyond the Euclidean distance: creating effective visual codebooks using the histogram intersection kernel," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 630–637, October 2009.
- [12] D. Xu and S. F. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1985–1997, 2008.
- [13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2169–2178, June 2006.
- [14] J. C. Yang, K. Yu, Y. H. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1794–1801, June 2009.
- [15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [16] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.
- [17] J. Wu and J. M. Rehg, "CENTRIST: a visual descriptor for scene categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.
- [18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [19] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proceedings of the European Conference on Computer Vision*, pp. 151–158, 1994.
- [20] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1458–1465, October 2005.
- [21] B. Scholkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.
- [22] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, October 2007.
- [23] J. Hu and P. Guo, "Multi-level kernel machine for scene image classification," in *Proceedings of the International Conference on Computational Intelligence and Security*, pp. 1169–1173, 2011.
- [24] J. He, S. F. Chang, and L. Xie, "Fast kernel learning for spatial pyramid matching," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–7, June 2008.

- [25] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV 07)*, October 2007.
- [26] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.