Randomized Spatial Partition for Scene Recognition

Yuning Jiang, Junsong Yuan*, and Gang Yu

School of Electrical and Electronics Engineering
Nanyang Technological University, Singapore 639798
{ynjiang,jsyuan}@ntu.edu.sg, gyu1@e.ntu.edu.sg

Abstract. The spatial layout of images plays a critical role in natural scene analysis. Despite previous work, e.g., spatial pyramid matching, how to design optimal spatial layout for scene classification remains an open problem due to the large variations of scene categories. This paper presents a novel image representation method, with the objective to characterize the image layout by various patterns, in the form of randomized spatial partition (RSP). The RSP-based image representation makes it possible to mine the most descriptive image layout pattern for each category of scenes, and then combine them by training a discriminative classifier, i.e., the proposed ORSP classifier. Besides RSP image representation, another powerful classifier, called the BRSP classifier, is also proposed. By weighting a sequence of various partition patterns via boosting, the BRSP classifier is more robust to the intra-class variations hence leads to a more accurate classification. Both RSP-based classifiers are tested on three publicly available scene datasets. The experimental results highlight the effectiveness of the proposed methods.

Key words: Random Partition, Spatial Layout, Scene Recognition

1 Introduction

Images are different from other information carriers such as texts and audios since image patterns convey rich spatial information. Such a difference makes a number of techniques which have been successful in the text-based applications less effectiveness when applied to images, e.g., the popular bag-of-visual-word (BoVW) model [2–5], because the spatial information among the visual primitives in images is usually ignored. A lot of previous work [6–10] has shown that without considering the spatial information, the discriminative power of BoVW model is severely limited. This is especially true when the images are composed by several semantic components with clear spatial layout, e.g., the natural scenes in Figure 1, where the spatial configuration among the semantic components becomes essential in describing these scenes. Therefore, how to make use of the spatial layout information plays a critical role in natural scene recognition.

The most straightforward way to incorporate the spatial layout information is to quantize the image space. By a pre-defined partition pattern, the two-dimensional im-

* This work is supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Dr. Junsong Yuan.

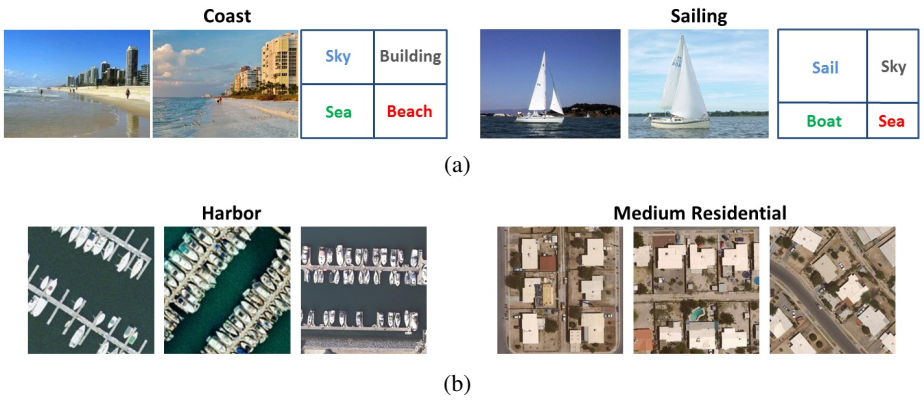


Fig. 1. (a) Examples of inter-class variations of the image layout. The left two pictures in each category are the original images while the right one is the optimal partition pattern to describe their spatial layout. We can see that sharing the same partition pattern for all the categories is not an optimal solution; (b) Examples of intra-class variations of the image layout (from 21-land-use dataset [1]). Even for the images containing the same components, their layout may vary quite a lot due to the changes in offset, scale, viewpoint and rotation.

age space is divided into several sub-regions. Then each visual word is encoded according to the sub-regions it belongs to. By doing so, the spatial layout of the images is characterized by the pre-defined partition pattern. The spatial pyramid matching (SPM) algorithm [11] is one representative of these methods, which symmetrically partitions the image into uniform cells at different levels of resolution. With the help of spatial layout information to improve the discriminative power, the SPM algorithm achieves a significant better performance, and its partition pattern, namely the spatial pyramid, is employed in many other work [1, 12].

Despite previous success, the spatial pyramid still has several limitations in capturing the spatial layout information. The first challenge comes from the inter-class variations of the spatial layout of the images. As shown in Figure 1(a), the spatial layout of the images in different scene categories differs a lot from each other, hence sharing one partition pattern for all scene categories is not the optimal solution to characterize the image layout, just as done in [11]. On the contrary, each category should have its own partition pattern that is most descriptive, such that its spatial layout can be optimally characterized. The second challenge comes from the large intra-class variations of the image layout. From Figure 1(b) we can see that even for the images in the same category that contain similar semantic components, their spatial layout may vary quite a lot due to the changes in offset, scale, viewpoint and rotation. It means even if we could find one descriptive partition pattern for each category, there will be also many outliers in the same category which cannot be described satisfactorily.

This paper contributes to addressing the two challenges mentioned above: 1) we propose a novel image representation approach based on randomized spatial partition (RSP). Instead of partitioning the images by the pre-defined pattern, e.g., the symmetric spatial pyramid, we randomly partition the images multiple times and obtain a pool of

independent partition patterns that can be selected later. As a result, the image layout can be characterized by various partition patterns; 2) to address the inter-class variations, an effective RSP-based classifier via optimal selection (ORSP) is proposed to mine the most discriminative partition pattern for each category. By representing each category of images in its own partition pattern, the spatial layout of different image categories is better described hence the ORSP classifier improves the classification accuracy; 3) furthermore, a more powerful RSP-based classifier via boosting (BRSP) is proposed, in which a sequence of partition patterns are weighted according to their discriminative abilities and finally boosted into the strong classifier. Since it allows to characterize the spatial layout of one image using multiple patterns, this classifier is more robust to the large intra-class variations of the image layout. We apply both RSP-based classifiers for scene recognition on three publicly available datasets: the 15-scene dataset [13], the 8-event dataset [14], and the 21-land-use dataset [1]. The comparison with the state-of-the-art methods highlights the effectiveness of our RSP-based image classifiers.

2 Related Work

The bag-of-visual-word (BoVW) model [3, 4, 15, 13, 14] has been widely adopted in visual recognition although it has an obvious drawback of quantizing high-dimensional descriptors into visual words. In general, there are two ways to address the quantization error incurred by the BoVW model. One is to match individual descriptors in the feature space directly, e.g., the Naive-Bayes Nearest Neighbor (NBNN) classifier proposed in [16, 17]. However, the NBNN-based algorithms are under the Naive-Bayes assumption that each feature point is independent from the others, therefore they can fail when the assumption is violated. Besides, without considering the spatial information, matching individual features can not provide satisfactory recognition results.

Taking advantage of the spatial information is another way to mitigate the quantization error. By bundling the co-occurred visual words within a constrained spatial distance into a visual phrase or feature group as the basic unit for matching, the spatial context information is incorporated to enhance the discriminative power of visual words and leads to a better performance in object recognition [18–20]. To group local features, spatial random partition has been used in [21, 22] for common object discovery and visual object search. However, these methods only characterize the relative location of the visual words rather than their absolute location, hence their descriptive abilities are limited on the images which have clear spatial layout, such as natural scenes. As one of the most popular methods in scene recognition, the spatial pyramid matching algorithm [11, 1, 12] divides the image space into uniform cells at different levels of resolution, and quantizes the continues coordinates of the visual words into discrete cells. Thus, the location of the visual words is encoded and the spatial layout of the images is characterized by the spatial pyramid. However, the pre-defined spatial pyramid could not be the optimal pattern to describe the image layout of all the categories. In [23], a specific partition pattern is learnt for each category by a series of recursive axis aligned splits of cells, but it is still under the *one-pattern-per-category* restriction, that is, the images in the same category share the same pattern to describe their image layout.

3 Image Representation

In this section, we first briefly describe the original spatial pyramid matching (SPM) algorithm [11], then introduce a novel randomized spatial partition-based (RSP-based) image representation scheme.

3.1 Original SPM Image Representation

Given an image I, we denote by $\{f\}$ all the local features extracted from it. In general each local feature f is represented as $f = (x, y, \mathbf{d})$, where (x, y) is the location coordinates and $\mathbf{d}$ is the continuous high-dimensional descriptor, e.g., 128-dimensional SIFT [24]. Then in the BoVW framework, each local descriptor $\mathbf{d}$ is quantized into a discrete visual word using a vocabulary of V words, while the location coordinates (x, y) are discarded. Finally, each image I can be represented as a histogram with V bins which records its word-occurrence frequency.

In the SPM algorithm, the spatial locations are integrated to enhance the descriptive power of BoVW model. As in Figure 2, the two-dimensional image space is divided symmetrically into uniform cells at different levels of resolution, forming the spatial pyramid. The higher level of the pyramid generates the smaller cells. Let us consider the current level l $(0 \leq l \leq L-1)$, at which all the local features are assigned to 4^l cells according to their location coordinates. Essentially, for each local feature $f = (x, y, \mathbf{d})$, we quantize its location coordinates (x, y) into these discrete cells, just like quantizing the continuous descriptor $\mathbf{d}$ into discrete visual words. Therefore, the image I can be represented as a histogram h^l with $4^l \times V$ bins at the l level, and finally represented as a long histogram with $\sum_{l=0}^{L-1} 4^l V$ bins combining h^l at all levels. Note that the weights associated with different levels are inversely proportional to the cell sizes. Intuitively, the histogram bin associated with a larger cell is penalized because it corresponds to a coarser quantization.

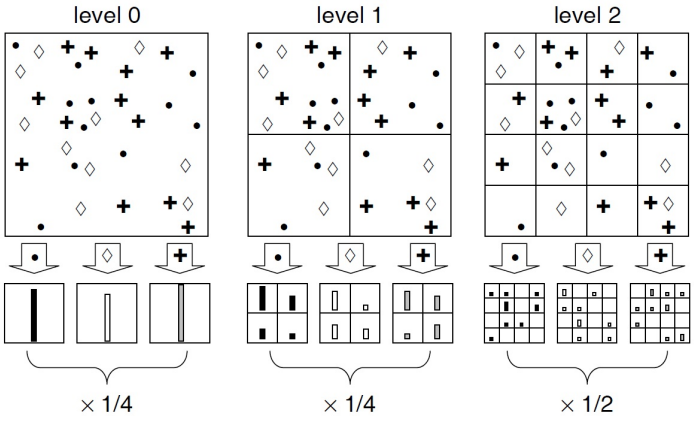


Fig. 2. Toy example of a three-level SPM image representation method (adapted from [11]). Essentially, the SPM method quantizes the location coordinates (x, y) into discrete cells using symmetric patterns.

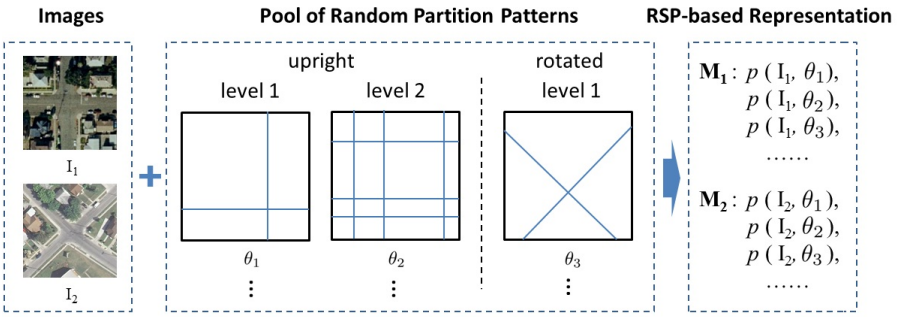


Fig. 3. Illustration of our RSP-based image representation.

3.2 RSP-based Image Representation

As reported in [11], with the help of spatial layout information, the SPM algorithm outperforms the BoVW model. However, quantizing the image space using the pre-defined spatial pyramid is empirical and arbitrary, and may not be optimal to describe the spatial layout information of an image. Therefore, we propose an image representation method based on randomized spatial partition, with the objective to better characterize the spatial layout of the images.

Let us consider a single level l first. Instead of symmetrically dividing the image space into uniform cells, we randomly partition the image space into $2^l \times 2^l$ sub-regions of various sizes and shapes. Such randomized partition is performed K times independently. In this way, the image space is quantized by these K random patterns, denoted by $\Theta^l = \{\theta^{l,k}\}_{k=1}^K$, where $\theta^{l,k}$ is a single $2^l \times 2^l$ partition pattern. Now for any image I_i, it will be partitioned into 4^l image patches by each $\theta^{l,k}$, and be represented as a histogram $h_i^{l,k}$ with $4^l \times V$ bins, denoted as: $h_i^{l,k} = p(I_i, \theta^{l,k})$. Therefore, in total we form a partition pattern pool $\Theta = \bigcup_{l=0}^{L-1} \Theta^l$ combining all levels, by which the image I_i is represented as a histogram collection M_i:

$$M_i = \hat{p}(I_i, \Theta) = \bigcup_{\theta \in \Theta} \{p(I_i, \theta)\}. \tag{1}$$

In this paper we not only use the upright partition patterns as in [11], but also introduce the rotated ones to enrich the variety of our pattern pool. Figure 3 gives an illustration of the RSP-based image representation method.

The benefits of the RSP-based image representation are two-fold: 1) from the image layout point of view, we randomly generate many spatial partitions so that can discover the descriptive partition patterns to better present the spatial configuration of the semantic components of a scene category; 2) from the local feature point of view, similar to the original SPM algorithm, the proposed RSP-based method also encodes the local features by quantizing the two-dimensional image space. However, different from SPM where each local feature is *hard-quantized* into the unique sub-region at each level, the RSP-based method provides many partition patterns in the same level, such that each local feature can be *soft-quantized* into multiple sub-regions. This will make the image representation more robust as it is less sensitive to the spatial quantization error.

4 Image Classification

Now given a collection of labeled images, denoted by $\Phi = \{(I_i, c_i)\}$, where $c_i \in \{1, 2, \ldots, C\}$ is the category label of the image I_i, and a pool of independent partition patterns Θ, we propose two approaches to obtain the robust classifier $F(\cdot)$ using the labeled training set.

4.1 RSP-based Classifier via Optimal Selection

Our intention is straightforward: for each category we seek for an optimal pattern with best discriminative power to separate this category from the others, and then combine these optimal patterns together to train the final classifier. We summarize this procedure in three steps as follows:

Step 1, training and validation. First we divide the entire training set Φ into two subsets Φ_t and Φ_v. Then for each pattern θ, the images in both Φ_t and Φ_v are represented as the corresponding histograms, and C binary classifiers $\{f_\theta^c(\cdot)\}_{c=1}^C$ are trained on Φ_t, which is done with the support vector machine (SVM). After that, the classifier $f_\theta^c(\cdot)$ is validated on Φ_v and its classification error err_θ^c is recorded:

$$err_\theta^c = \sum_{c_v = c} \mathbf{I}(f_\theta^c(I_v) \neq c) + \sum_{c_v \neq c} \mathbf{I}(f_\theta^c(I_v) = c), \tag{2}$$

where (I_v, c_v) is the sample in Φ_v, and $\mathbf{I}(\sigma) = 1$ if σ is true; otherwise $\mathbf{I}(\sigma) = 0$;

Step 2, pattern selection. The best pattern with minimum validation error is selected as the optimal description of spatial layout information for the images of category c:

$$\theta_{best}^c = \arg\min_{\theta \in \Theta} err_\theta^c. \tag{3}$$

We finally select C best patterns which are supposed to have the strongest discriminative power for each category;

Step 3, classifier recasting. After obtaining the best patterns $\{\theta_{best}^c\}_{l=1}^C$, we re-train the binary classifiers $f_{best}^c(\cdot)$ on the entire set Φ for each θ_{best}^c. Finally the multi-class classification is implemented using the C binary classifiers and taking the class of highest classification score.

Compared with SPM algorithm that shares the empirical and fixed partition pattern for all the categories, this approach selects one optimal partition pattern for each category, which can describe this category discriminatively. Therefore, it is more in accord with the fact that each image category has its own image layout pattern. Though the idea is straightforward, it contributes to a considerable improvement of classification accuracy in the experiment.

4.2 RSP-based Classifier via Boosting

In the above, we present how to train a RSP-based classifier via optimal selection (ORSP classifier for short). Despite the advantages in describing spatial layout of each category, the ORSP classifier is still under the *one-pattern-per-category* restriction hence it is not robust to the intra-class variations of image layout. Therefore, in the following we adopt the data-driven weighting strategy to train another RSP-based

Algorithm 1 Training RSP-based Classifier via Boosting

Input:

A collection of labeled images: $\Phi = \{(I_i, c_i)\}_{i=1}^{N}$.

A pool of independent partition patterns: $\Theta = \{\theta\}$.

The target classification accuracy: σ_{target}.

Output:

A robust image classifier $F(\cdot)$. Give an unlabeled image I, we have $c = F(I)$, where $c \in \{1, \ldots, C\}$ is the predicted label of I.

1. **For all** $\theta \in \Theta$:
   - Randomly sample a subset $\Phi_\theta \subset \Phi$, and represent the images in Φ_θ in pattern θ.
   - Train a multi-class classifier $f_\theta(\cdot)$ on the random subset Φ_θ using SVM.
2. Initialize the weight $w_i = \frac{1}{CN_{c_i}}$ for each images I_i, where N_{c_i} is the number of the images with label c_i; the current iteration number $j = 0$; the current accuracy $\sigma^{(0)} = 0$.
3. **While** $(\sigma^{(j)} < \sigma_{target})$
   - $\forall i = 1, \ldots, N$, $w_i \leftarrow \frac{w_i}{\sum_{i=1}^{N} w_i}$; $j \leftarrow j + 1$.
   - $\forall \theta \in \Theta$, calculate its classification error on Φ: $err_\theta = \sum_{I_i \in \Phi} w_i \cdot \mathbf{I}(f_\theta(I_i) \neq c_i)$.
   - Select the pattern $\theta^{(j)}$ with minimum error $err^{(j)}$, and then calculate the weight for $\theta^{(j)}$ as:
   $$\alpha^{(j)} = \log \frac{1 - err^{(j)}}{err^{(j)}} + \log(C - 1).$$
   - $\forall i = 1, \ldots, N$, $w_i \leftarrow w_i \cdot \exp(\alpha^{(j)} \cdot \mathbf{I}(f_{\theta^{(j)}}(I_i) \neq c_i))$.
   - Generate the strong classifier as:
   $$F(I) = \arg\max_c \sum_{m=1}^{j} \alpha^{(m)} \cdot \mathbf{I}(f_{\theta^{(m)}}(I) = c),$$
   and calculate its classification accuracy on Φ: $\sigma^{(j)} = \sum_{I_i \in \Phi} \mathbf{I}(F(I_i) = c_i)/N$.

classifier via boosting (BRSP classifier for short), in which a sequence of patterns are weighted in proportional to their discriminative power.

Since Algorithm 1 has illustrated our BRSP in detail, here we only discuss it briefly: first, to promote the independence among the weak classifiers, the bootstrapping is adopted in Step 1, i.e., each weak classifier $f_\theta(\cdot)$ is trained on a random subset $\Phi_\theta \subset \Phi$; second, in Step 2, the weight of each image is initialized inversely proportional to its category size to prevent the problem of unbalanced sample sizes; third, the *SAMME* Adaboost algorithm [25] is employed to address the multi-class cases; finally, we set the stop condition of the loop as a target training accuracy σ_{target}, rather than the number of iterations. The reason is that in the experiment, we find only a few patterns (generally less than 30) will be enough to make a good classification. Therefore, to avoid overfitting, we will stop the iterations before the training error becomes 0.

We can see several benefits of the BRSP classifier: first of all, a data-driven weighting strategy is adopted instead of the uniform weighting strategy. That is, the partition patterns are weighted according to their recognition performances on the training dataset. Next, it breaks the *one-pattern-per-category* restriction and allows to describe the spatial layout of one image in multiple partition patterns. Therefore, it can better deal with the categories with large intra-class variations. Finally, taking the advantages

of boosting algorithm, these confusing images which are likely to be classified mistakenly will be picked out hence the final classifier is more robust to those outliers.

5 Experiment

In this section, we report our experimental results on three publicly available datasets: the 15-scene dataset [13], the 8-event dataset [14], and the 21-land-use dataset [1]. All experiments are repeated 3 times with different randomly selected training and test sets, and the average recognition rates are recorded as the final results. The SVMs are implemented using the LIBSVM package.[1]

5.1 The 15-Scene Dataset

First, we perform our algorithm for natural scene recognition on the 15-scene datasets, which is one of the most complete scene category datasets collected gradually by several research groups [13, 11, 26]. The 15-scene dataset is composed of fifteen natural scene categories: *bedroom, suburb, industrial, kitchen, livingroom, coast, forest, highway, insidecity, mountain, opencountry, street, tallbuilding, office* and *store*. Each category has 216 to 400 images with resolution around 300×250. As in [11], from each image the dense SIFT descriptors of 16×16 pixel patches computed over a grid are extracted with spacing of 8 pixels. Then, 100 images per category, 1500 images in total, are randomly selected out as the training set Φ, while the rest is prepared as testing set. We randomly sample 450 images from the training set Φ, and perform k-means clustering on all the SIFT descriptors of these 450 images to generate a vocabulary with $V = 400$ visual words, by which all the SIFT descriptors from the 15-scene dataset are quantized into discrete words.

We set the randomized spatial partition parameters as follows: the highest level L is set to 3, i.e., $l = 0, 1, 2$. We choose $L = 3$ because [11] has shown that a higher level ($l \geqslant 3$) may lead to a decrease in accuracy due to over subdivision. For both $l = 1, 2$ levels, the randomized partition number K is set to 100. For $l = 0$ level the partition number is $K = 1$ since in fact no division is made at this level. To make a fair comparison with the SPM algorithm [11], in which only upright spatial pyramid is used, we do not introduce the rotated partition patterns for this dataset. Therefore, a pool of upright patterns Θ_u is formed which contains 201 random partition patterns in total.

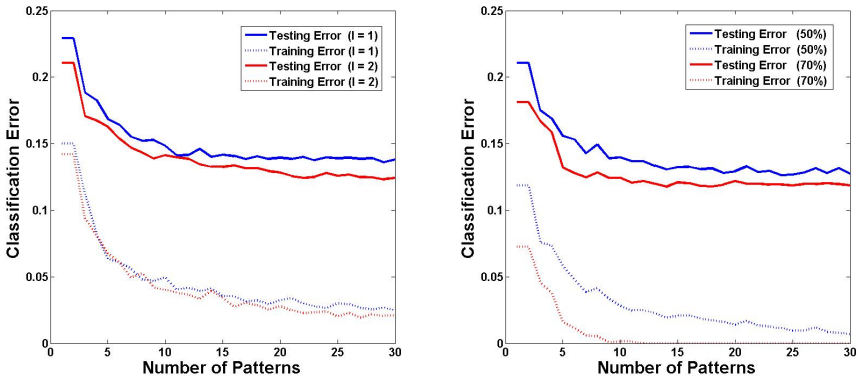
After the preparation step on image representation, now both two types of RSP-based classifiers proposed in Section 4 are trained. Linear kernel is incorporated for all SVMs as in [11]. For the ORSP classifier, we set two subsets of the Φ: Φ_t for training and Φ_v for validation, respectively, which are in the same size, $|\Phi_t| = |\Phi_v|$. For the BRSP classifier, we also keep $|\Phi_\theta| = 50\% \times |\Phi|$ for all the patterns. The target training error is set as $\sigma_{target} = 98\%$. Table 1 reports the classification accuracy of these two types of RSP-based classifiers and compares with the results of the SPM algorithm.

The results shown in Table 1 is analyzed as follows: first, by comparing the performance of the original SPM algorithm (2_{nd} column) and the ORSP classifier (3_{rd}

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

level	Original SPM [11]	ORSP Classifier	BRSP Classifier
$l = 0$	74.8%	-	-
$l = 1$	78.8%	80.7%	86.4%
$l = 2$	79.7%	82.6%	87.1%
$l = 0, 1, 2$	81.4%	83.9%	87.2%

Table 1. Comparison with SPM at different levels on the 15-scene dataset.



(a) (b)

Fig. 4. The curves of training error and testing error of the BRSP classifiers in different conditions, with the number of boosted patterns increasing.

column), we can see that the latter one leads to a significant increase in classification accuracy while the only difference between these two methods is the way to divide the image space. This comparison supports our claim that quantizing the image space by symmetric divisions is not optimal to describe the spatial layout information, and it can be improved by the RSP-based image representation. Second, compared with the ORSP classifier, the BRSP classifier (4_{th} column) has shown more discriminative power, especially at the low levels (3_{rd} row and 5_{th} row). In fact, it is the main advantage of the BRSP classifier: the final decision can be made more robustly after combining the votes from several patterns, even though the descriptive power of each single pattern is limited.

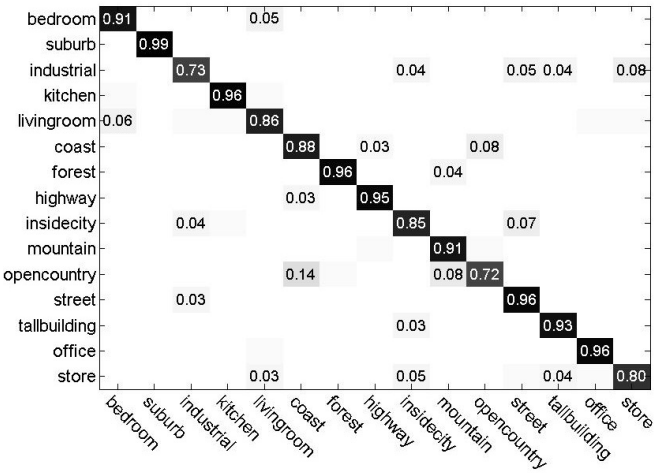
Next, we study how the number of boosted patterns, as well as the effectiveness of each single weak classifier, affects the overall performance of the final BRSP classifier. Here we change the conditions when training the weak classifiers, and make two comparisons: in the first comparison, the weak classifiers are trained using only the level 1 or the level 2 patterns, corresponding to the 3_{rd} and 4_{th} rows in Table 1, respectively; in the second comparison, we increase the size of the random subset $|\Phi_\theta|$ for weak classifier training, from $50\% \times |\Phi|$ to $70\% \times |\Phi|$, hence obtain weak classifiers with different powers. The comparison results are given in Figure 4, in which both of the testing error and training error are plotted with the number of weak classifiers increasing. We can see that in each comparison, the gap of discriminative power between the two BRSP classifiers narrows after convergence, despite comparably big difference existing at the very beginning. Moreover, both curves show a sharp decrease in the first ten iterations,

nearly 9% in testing error, which means more than 250 initially mis-classified images are corrected by the latter weak classifiers. It once again validates the advantage of the BRSP classifier.

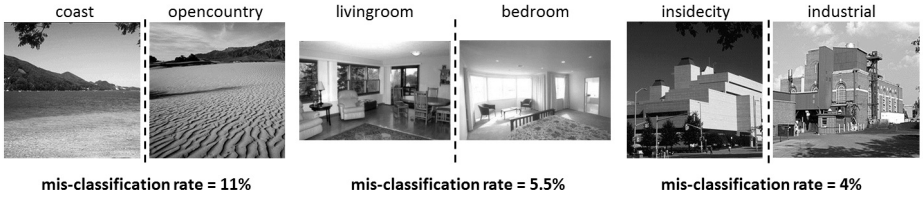
Finally, the proposed algorithms are compared with the state-of-the-art techniques, as shown in Table 2. Though only the simplest SIFT feature and linear-SVM are adopted, the recognition accuracy of our RSP-based algorithms is already better than the best of the previous work [12, 27], which validates the effectiveness of the RSP-based image representation. Also a confusion matrix is given in Figure 5(a). Similar to the results in [11], the top three confusing pairs are: coast/opencounty, bedroom/livingroom and insidecity/industrial. Several mis-classified examples are shown in Figure 5(b).

Method	Avg. Accuracy
SPM + SIFT with 400 clusters [11]	81.4%
SPM + SIFT with 400 concepts [28]	83.3%
DSP + SIFT with 1000 clusters [23]	80.7%
SP-pLSA + SIFT with 1200 topics [29]	83.7%
CENTRIST + RBF-SVM [12]	83.9%
CENTRIST + LCC + Boosting [27]	87.8%
RSP + Optimal Selection	83.9%
RSP + Boosting ($\|\Phi_\theta\|/\|\Phi\| = 50\%$)	87.2%
RSP + Boosting ($\|\Phi_\theta\|/\|\Phi\| = 70\%$)	**88.1**%

Table 2. Comparison with the state-of-the-art methods on the 15-scene dataset.



(a)



(b)

Fig. 5. (a) The confusion matrix of 15-scene recognition by BRSP classifier. Only rates higher than 3% are shown; (b) Examples of the top 3 confusing pairs in the 15-scene dataset.

Fig. 6. Examples of the images in the 8-event dataset.

Method	Avg. Accuracy
Scene Model + SIFT [14]	$\approx 60\%$
Scene Model + Object Model + SIFT [14]	73.4%
PACT + RBF-SVM [12]	78.2%
SPM + RBF-SVM	74.0%
RSP + Optimal Selection	77.9%
RSP + Boosting	**79.6**%

Table 3. Comparison with the state-of-the-art methods on the 8-event dataset.

5.2 The 8-Event Dataset

The 8-event dataset [14] is composed of eight sport classes: *badminton, bocce, croquet, polo, rock climbing, rowing, sailing, and snowboarding* (see Figure 6). Each class has 137 to 250 high-resolution images (from 800×600 to thousands of pixels per dimension). Following [14, 12], we randomly select 70 images per class for training, and 60 for testing. Although more complex PACT features are used in [12], here we only extract the same features as in [14], i.e., the SIFT descriptors of 12×12 pixel patches computed over a grid with spacing of 10 pixels, and then cluster them into a vocabulary of $V = 300$ visual words. As in [12], the RBF kernel replaces the linear kernel in SVM training, and kernel parameters are chosen by a three-fold cross validation on the training set. The other experimental conditions, e.g., the partition parameters, are the same as those in the 15-scene dataset.

Table 3 compares our RSP-based approaches with other state-of-the-art methods. From this table we can see that: 1) spatial layout information plays a critical role in visual recognition. Although a high-level probability model is used in [14] (3_{rd} row), the original SPM algorithm (5_{th} row) still can reach a comparable accuracy with the help of spatial layout information; 2) it once again validates that the symmetric pyramid is not optimal to describe the spatial layout of all the categories of images, since the ORSP classifier (6_{th} row) has a remarkable improvement over the original SPM algorithm, and obtains a comparable results with [12] in which the complex PACT features are used; 3) the BRSP classifier (7_{th} row) has a more discriminative power than the ORSP classifier, and outperforms all the state-of-the-arts in this dataset.

5.3 The 21-Land-Use Dataset

The 21-land-use dataset [1] contains 21 classes of aerial orthoimagery downloaded from the United States Geological Survey (USGS) National Map: *agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts.* Each class has 100 images with resolution 256×256. SIFT features are computed over the 16×16 patches with an 8-pixel grid spacing. As in [1], a vocabulary of $V = 100$ visual words is generated by applying k-means clustering on a random subset of the feature pool, and then used to label all the SIFT features extracted from this dataset. For each class, we split it into five equal sized sets. Four of the sets are used for training and the held-out set is for testing.

We introduce the rotated partition patterns in this dataset to better describe the spatial layout information of the rotated cases, as shown in Figure 1(b). The rotated partition is performed at $l = 1$ level and repeated independently for $K = 100$ times, which forms a pool of rotated partition patterns Θ_r with 100 random patterns. Therefore, the entire pool we used in this dataset is $\Theta = \Theta_u \cup \Theta_r$ consisting of all upright and rotated patterns. Besides, we set $|\Phi_\theta|/|\Phi| = 75\%$ for training the BRSP classifiers. The other parameters are set the same as the experiments above.

In Table 4 we compare our RSP-based approaches with other methods that use spatial pyramid to characterize the arrangement of visual words, namely spatial pyramid matching kernel (SPMK) in [11], spatial pyramid co-occurrence kernel (SPCK), and its extensions SPCK+ and SPCK++ in [1]. The comparison highlights the effectiveness of our RSP-based approaches. Moreover, Figure 7 explains how our BRSP classifier can handle the large variations of spatial layout of images, hence achieves a more accurate classification result.

	BoVW	SPMK [11]	SPCK [1]	SPCK+ [1]	SPCK++ [1]	ORSP	BRSP
Acc.	71.9	74.0	73.1	76.1	77.3	75.5	**77.8**

Table 4. Classification accuracy for the 21-land-use dataset.

6 Conclusion

This paper presents a novel image representation method based on randomized spatial partition, with the objective to optimally characterize the spatial layout of the images. In contrast to the pre-defined spatial pyramid, in the RSP-based image representation method the spatial layout of the images is characterized by the randomized partition patterns. Furthermore, two discriminative image classifiers, the ORSP classifier and the BRSP classifier, are proposed based on the RSP image representation. The ORSP classifier discovers the most descriptive pattern for each category of images to address the inter-class variations of image layout, while the BRSP classifier can handle the large intra-class variations by boosting a sequence of various patterns together. The comparison with the state-of-the-art methods on three publicly available datasets validates the effectiveness of our RSP-based methods.

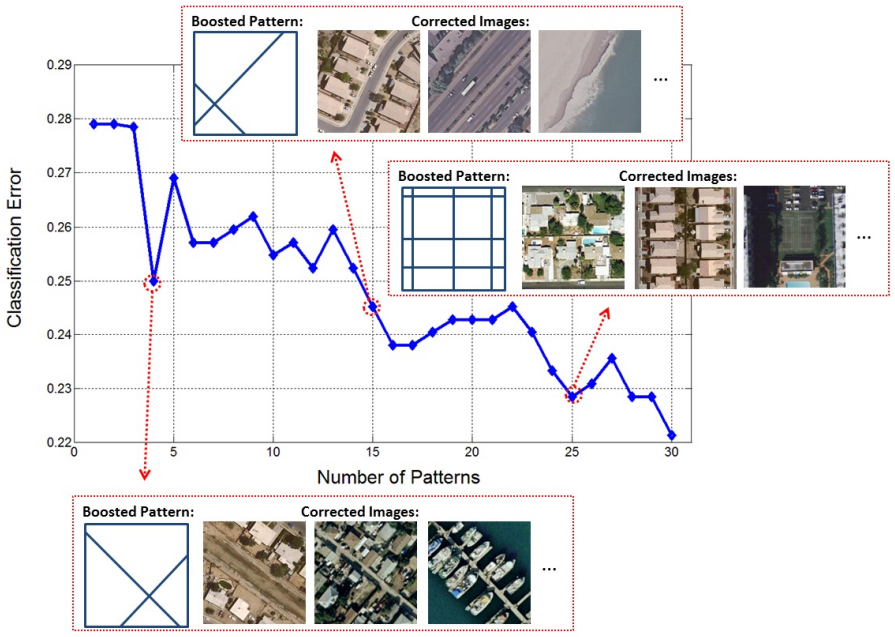


Fig. 7. The classification error curve of the BRSP classifier on the testing set, with the number of boosted patterns increasing. We show three examples of the selected patterns and the testing images which are mistakenly classified at the beginning but corrected by the current pattern in the red boxes. From it we can see that the BRSP classifier can well handle the large intra-class variations of image layout (e.g., the *dense residential*) by representing the images using multiple partition patterns. Note that although several bad patterns are boosted due to the bias between the training set and testing set, the classification accuracy generally tends to be improved when increasing the number of boosted patterns.

References

1. Yang, Y., Newsam, S.: Spatial pyramid co-occurrence for image classification. In: ICCV'11: IEEE International Conference on Computer Vision, Barcelona, Spain (2011)
2. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence (2009)
3. Winn, J., A.Criminisi, T.Minka: Object categorization by learned universal visual dictionary. In: Proc. IEEE Intl. Conf. on Computer Vision. (2005)
4. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: Proc. IEEE Intl. Conf. on Computer Vision. (2005)
5. Grauman, K., Darrel, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: Proc. IEEE Intl. Conf. on Computer Vision. (2005)
6. Li, T., Mei, T., Kweon, I., Hua, X.: Contextual bag-of-words for visual categorization. IEEE Transactions on Circuits and Systems for Video Technology. (2010)
7. Perdoch, M., Chum, O., Matas, J.: Efficient representation of local geometry for large scale object retrieval. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2009)

8. Wang, X., Yang, M., Cour, T., Zhu, S., Yu, K., Han, T.X.: Contextual weighting for vocabulary tree based image retrieval. In: ICCV'11: IEEE International Conference on Computer Vision, Barcelona, Spain (2011)
9. Zhang, Y., Jia, Z., Chen, T.: Image retrieval with geometry-preserving visual phrases. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2011)
10. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: Proc. IEEE Intl. Conf. on Computer Vision. (2005)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2006)
12. Wu, J., Rehg, J.M.: Centrist: A visual descriptor for scene categorization. IEEE Trans. on Pattern Analysis and Machine Intelligence (2011)
13. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2005)
14. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: Proc. IEEE Intl. Conf. on Computer Vision. (2007)
15. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bra, C.: Visual categorization with bags of keypoints. In: In Workshop on Statistical Learning in Computer Vision, ECCV. (2004)
16. Behmo, R., Marcombes, P., Dalalyan, A., Prinet, V.: Towards optimal naive bayes nearest neighbor. In: Proc. European Conf. on Computer Vision. (2010)
17. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2008)
18. Zhang, S., Tian, Q., Hua, G., Huang, Q., Li, S.: Descriptive visual words and visual phrases for image applications. In: Proc. ACM Multimedia. (2009)
19. Zhang, S., Huang, Q., Hua, G., Jiang, S., Gao, W., Tian, Q.: Building contextual visual vocabulary for large-scale image applications. In: Proc. ACM Multimedia. (2010)
20. Perina, A., Jojic, N.: Image analysis by counting on a grid. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2011)
21. Yuan, J., Wu, Y.: Spatial random partition for common visual pattern discovery. In: Proc. IEEE Intl. Conf. on Computer Vision. (2007)
22. Jiang, Y., Meng, J., Yuan, J.: Randomized visual phrases for object search. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2012)
23. G.Sharma, F.Jurie: Learning discriminative spatial representation for image classification. In: Proc. of BMVC. (2011)
24. Lowe, D.: Distinctive image features from scale-invariant keypoints. Intl. Journal of Computer Vision (2004)
25. Zhu, J., Rosset, S., Zou, H., Hastie, T.: Multi-class adaboost. In: Technical report, Stanford Univ. Available at http://www-stat.stanford.edu/ hastie/Papers/samme.pdf. (2005)
26. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision (2001)
27. Yuan, J., Yang, M., Wu, Y.: Mining discriminative co-occurrence patterns for visual recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2011)
28. Li, J., Shah, M.: Scene modeling using co-clustering. In: Proc. IEEE Intl. Conf. on Computer Vision. (2007)
29. Bosc, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Trans. on Pattern Analysis and Machine Intelligence (2008)