Pyramid of Spatial Relatons for Scene Level Land Use Classification

Shizhi Chen and YingLi Tian, *Senior Member, IEEE*

***Abstract*—Local feature with Bag of Words (BOW) representation has become one of the most popular approaches in object classification and image retrieval applications in the computer vision community. The recent efforts in the remote sensing community demonstrate that the BOW approach can also effectively apply to geographic images for the applications of classification and retrieval. However, the BOW representation discards spatial information, which is critical for the remotely sensed land use classification. Several algorithms have incorporated spatial information into the BOW representation by hard encoding coordinates of local features. Such rigid spatial encoding is not robust to translation and rotation variations, which are common characteristics of geographic images. To effectively incorporate spatial information into the BOW model for the land use classification, we propose a Pyramid of Spatial Relatons (PSR) model to capture both absolute and relative spatial relationship of local features. Unlike the conventional co-occurrence approach to describe pairwise spatial relationships between local features, the PSR model employs a novel concept of spatial relaton to describe relative spatial relationship of a group of local features. As the result, the storage cost of the PSR model only linearly increases with the visual word codebook size instead of the quadratic relationship as in the co-occurrence approach. The PSR model is robust to translation and rotation variations, and demonstrates excellent performance for the application of remotely sensed land use classification. On the Land Use and Land Cover image database, the PSR achieves 8% higher in the classification accuracy than the state of the art. If using only gray images, it outperforms the state of the art by more than 11%.**

***Index Terms*—Bag of Words, Spatial Pyramid Matching, Pyramid of Spatial Relatons, Geographical Image Classification, Land Use Classification**

I. Introduction

THE local features [2, 4, 8, 9, 18, 27] have been successfully applied to many computer vision applications, including image retrieval, object classification, and scene understanding etc. They also begin to gain popularity in remote sensing community due to the robustness to rotation, scale changes, and occlusion [3, 19, 29, 32, 33, 34, 40, 41, 42, 43, 44].

One of the most popular approaches to group local features in an image is the Bag of Words (BOW) model [8]. By simply counting occurrences of local features in an image without modeling their spatial relationships, the BOW demonstrates good performance in both computer vision and remote sensing applications.

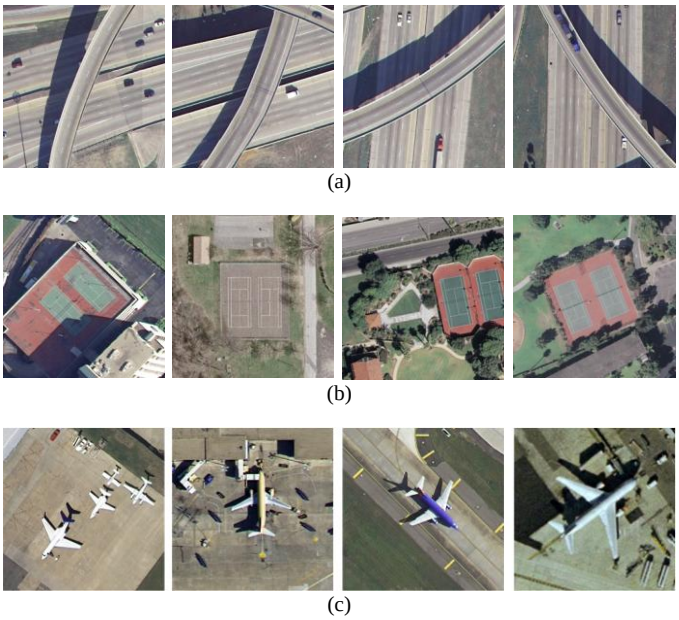


Figure 1: Sample images in three different classes of the Land Use Land Cover (LULC) dataset: (a) Overpass; (b) Tennis court; (c) Airplane.

Nevertheless, several researchers [7, 21, 28, 31, 43] have confirmed that the spatial relationship among local features improves performance over many applications, e.g., image classification and retrieval etc. Most of the proposed models hard code image coordinates of local features [7, 21, 28], which captures the absolute spatial information of local features based on their image coordinates. The absolute spatial information improves some computer vision applications, in which camera views are usually fixed in the upright orientation and the point of interest is approximately in the center of an image.

One of the most popular absolute spatial models is the Spatial Pyramid Matching (SPM) [21]. The SPM model hierarchically divides an image into several sub-regions. A Bag of Words (BOW) histogram is constructed for each sub-region. Then the BOW histograms of all sub-regions are concatenated together based on the absolute spatial order of sub-regions in the image space. The SPM model demonstrates excellent

Manuscript received May 15th 2014, revised July 12th 2014, accepted August 10th 2014. This work was supported in part by ONR grant N000141310450 and ARO grant W911NF-09-1-0565.

Shizhi Chen is with Naval Undersea Warfare Center. This work was done when he was with the City College of New York, CUNY, New York, NY 10031 USA (e-mail: shizhi.chen@navy.mil).

YingLi Tian is with the City College and Graduate Center, City University of New York, New York, NY 10031 USA (phone: 212-650-7046; fax: 212-650-8249; e-mail: ytian@ccny.cuny.edu)

performance and has been a key component in many computer vision applications [10, 17, 37, 38], including object recognition, scene understanding, and object detection etc.

However, the camera view of geographic images or aerial images can be freely rotated. The point of interest can also appear anywhere in the image. Figure 1 shows some sample images from three different classes in the Land Use and Land Cover (LULC) dataset, i.e., Overpass, Tennis court and Airplane. All three classes exhibit rotation and translation variations. Therefore, the absolute spatial information may not improve the classification accuracy for these geographic images. Actually, some researchers [43] have reported that the absolute spatial information degrades the classification performance as compared with the order-less approach.

Instead of absolute spatial information, some researchers [15, 16, 31, 43] explore relative spatial relationship among local features. Relative spatial information describes locations of local features relative to each other in an image, independent of their absolute image coordinates. The researchers generally model relative spatial relationship through co-occurrence of local features, which satisfies certain spatial constraints, such as distance or angular directions. The co-occurrence matrix models spatial relationships of local features within a local region, which achieves invariance to rotation and translation. By combining the co-occurrence matrix with the order-less Bag of Words model, the performance of land use classification improves [43]. However, the storage cost of the co-occurrence matrix has quadratic relationship with the visual word codebook size [43]. For a codebook size of 1000, which is common for the geographical image classification [43], the size of a co-occurrence matrix can reach a million.

Inspired by the success of Spatial Pyramid Matching (SPM) model for the computer vision applications and the co-occurrence approach for the remote sensing applications, we propose a Pyramid of Spatial Relatons (PSR) model to incorporate both absolute and relative spatial information into the Bag of Words framework. Similar to the texton and correlaton [31], we define relaton as a basic relationship unit. The PSR divides an image into successively finer sub-regions as in the SPM model. Then a collection of spatial relationships prototypes and a histogram of local features are extracted from each of the sub-regions.

The Pyramid of Spatial Relatons (PSR) model achieves excellent performance for land use classification on the Land Use and Land Cover (LULC) geographical image database. It achieves 8% higher in classification accuracy than the state of the art. If using only gray images, it outperforms the state of the art by more than 11%. As compared with the BOW model, the PSR model improves classification accuracy across all codebook sizes we evaluated in our experimental setup. The performance improvement is especially significant when the codebook is compact, e.g., the performance gain of the PSR is more than 11% for the codebook size smaller than or equal to 1000. Our extensive experiments demonstrate the effectiveness of the PSR model, which incorporates both relative and absolute spatial relationship into the BOW model.

In the following of this paper, Section II discusses the related work on both computer vision and remote sensing applications. Section III describes the details of the proposed Pyramid of Spatial Relatons (PSR) framework. Section IV presents experimental setup and results. Finally, Section V concludes the paper.

II. Related Work

The Pyramid of Spatial Relatons (PSR) is most closely related to the bag of words (BOW) model [8, 22, 35] and its spatial variants. The bag of words model remains one of the most popular approaches in image classification applications due to its simplicity and excellent performance. Over the past few years, several spatial variants [6, 7, 21, 23, 24, 28, 31, 43, 44, 46] of the BOW model are proposed to incorporate the spatial information of local features. These spatial variants utilize either absolute or relative spatial context of local features. They have demonstrated performance improvements in many computer vision applications. However, direct adaptation of these spatial models to the remotely sensed land use classification still remains challenging [43].

A. Order-less Approach: Bag of Words

The Bag of Words (BOW) is commonly used in document classification or text retrieval application [1]. A document is represented by a histogram of words, where each element in the histogram represents frequency of occurrences of a word in the document.

Inspired by the success of the BOW model in text retrieval literature and the robustness of local features as image representation [2, 13, 20, 27], computer vision community adapts the BOW to represent an image by treating local features as visual words [8, 35].

Different from words in text, there are much more variations in local features. Hence, similar features in feature space are grouped together to form a cluster, which is represented by a prototype feature, i.e., a visual word. The collection of visual words from each cluster in training data forms a visual dictionary or codebook.

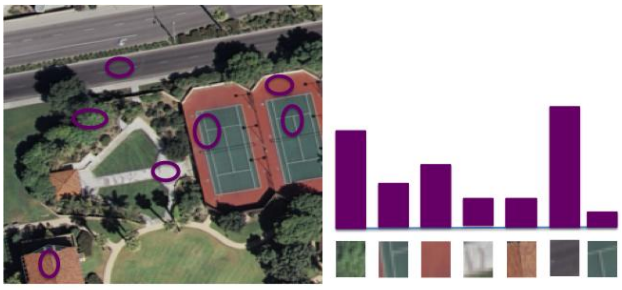


Figure 2: (a) Illustrate local features in an image; (b) Illustrate the corresponding Bag of Words representation of the image based on the statistics of local features.

Figure 2 illustrates the Bag of Words (BOW) representation for a geographical image. The local features are illustrated with purple ellipses, which are encoded based on visual words in a codebook. Then we form a histogram of visual words by proper pooling techniques [4]. The histogram is the final BOW

representation of the image.

In the coding step, local features are encoded using visual words in a codebook. The simplest coding method is vector quantization, i.e., hard assignment. The hard assignment of a local feature is to assign the weight of 1 to the nearest neighbor visual word, while all other visual words in the codebook are assigned with 0 weight. Other more sophisticated coding techniques include soft assignment [25] and sparse coding [45], which generally improve classification accuracy at the expense of increased complexity.

Pooling is to aggregate all codes of local features into a single vector as the image's BOW histogram. Two of the most popular pooling operators are average pooling [8] and maximum pooling [25]. Maximum pooling usually achieves higher accuracy and is employed in most state-of-the-art systems for image classification.

A major limitation of the Bag of Words (BOW) model is that it only represents an image as an order-less collection of local features without considering features' spatial information in the image. As proven by many researchers, the spatial relationship of local features can be very important for image classification [6, 7, 21, 31, 43].

B. Incorporating Spatial Context

To improve classification accuracy, both absolute spatial information [7, 21, 28] and relative spatial information [31, 43] have been incorporated into the bag of words representation by researchers.

1) Absolute Spatial Context

Absolute spatial context is the location information of local features with reference to the absolute image coordinates. Despite the fact that absolute spatial information of local features is not invariant to translation or rotation of cameras, it has achieved the state of the art performances on several benchmark datasets for image classification [7, 21, 28].

Lazebnik *et al*. [21] propose Spatial Pyramid Matching (SPM) as feature representation, by partitioning an image into successively smaller sub-regions and calculating a BOW histogram for each sub-region. Then, the model concatenates all BOW histograms of the sub-regions together to form the SPM representation of an image. With the finer regions assigned with larger weights, the intersection kernel is employed for the classification, as shown in Eq. (1).

$$K(I_1, I_2) = \sum_{l=0}^{L-1} \left\{ \frac{1}{2^{(L-l-\delta(l)+1)}} \sum_{r=0}^{S_l-1} \sum_{n=0}^{N-1} min\left(H^n_{I1,l,r}, H^n_{I2,l,r}\right) \right\}, \quad (1)$$

where L is the total number of levels. S_l is the total number of sub-regions at level l. N is the codebook size of visual words. Note that $\delta(l)$ is 1 when l is 0, otherwise $\delta(l)$ equals to 0.

Conventionally, SPM is calculated at three levels. At the first level, the whole image is a sub-region. At the second and the third level, the image is divided into 2 x 2, and 4 x 4 sub-regions respectively. As compared with the order-less BOW model, the SPM achieves better accuracy in several challenge datasets [11, 12, 14, 39] with slightly increase of computational cost.

Recently, McCann and Lowe [28] proposed Spatial Local Coding (SLC), which is also based on absolute spatial information of local features. SLC augments Scale Invariant Feature Transform (SIFT) features with their absolute location (x, y) in the image coordinate space. Then a dictionary is constructed based on these augmented SIFT features. Local soft assignment and maximum pooling are employed to build the final spatially encoded BOW representation.

Our previous work [7] employs a collection of EigenMaps as image representation to incorporate both appearance and absolute spatial information. An EigenMap is location likelihood of a visual word obtained from kernel density estimation. As compared with other spatial variants of the BOW model, the EigenMap is more computationally efficient while remaining comparable classification accuracy.

2) Relative Spatial Context

On the other hand, relative spatial context captures spatial information relative to some of local features in an image. It is invariant to translation and rotation.

Savarese *et al*. [31] borrow the idea of color correlograms [16] to develop visual word correlograms, which incorporate relative spatial information. Correlograms capture spatial correlation between all possible pairs of visual words by forming a co-occurrence matrix of visual words as a function of distance.

The correlograms matrix requires expensive computation and memory cost [31]. Savarese *et al*. utilize integral histogram techniques to improve computational efficiency. However, the technique becomes less effective as the codebook size increases. Furthermore, the storage cost for the integral histogram also increases dramatically for an image even at a moderate codebook size, e.g., 1000.

C. Related Remote Sensing Applications

Bag of Words (BOW) representation and its spatial variants have also demonstrated the effectiveness for the remotely sensed land use classification [43, 44]. The paper [43] describes the performance of the BOW and two other spatial variants for the application of large-scale land use image classification. They demonstrate that the BOW approach achieves comparable performance with the best of standard approaches, e.g. color histogram.

However, the absolute spatial information in the Spatial Pyramid Matching (SPM) model degrades the classification performance of the land use images according to their experiments [43]. The authors argue that land use images exhibit significant translation and rotation variation. Motivated by the importance of spatial structure of geographical images [36], they [43] proposed to incorporate relative spatial information into the BOW using spatial co-occurrence kernel (SCK), which is similar to correlograms [16, 31] approach with an intersection kernel.

Yang and Newsam [44] further propose Spatial Pyramid of Co-occurrence Kernel (SPCK) model, which incorporates both absolute and relative spatial context for the geographical image classification. Similar to the Spatial Pyramid Matching model,

the SPCK method partitions an image into sub-regions successively. Within each sub-region, it computes the co-occurrence matrix instead of the bag of words histogram. Then different sub-regions are concatenated together with appropriate weights to form the final image representation. The method demonstrated improved performance over the order-less BOW model and the model with the absolute spatial context, i.e., the Spatial Pyramid Matching model.

III. Pyramid of Spatial Relatons

A. Overview

Motivated by both spatial pyramid matching (SPM) model and spatial co-occurrence kernel approach, we propose a new algorithm, i.e., the Pyramid of Spatial Relatons (PSR). The PSR captures both absolute and relative spatial relationships of local features. We evaluate the proposed PSR model for the remotely sensed land use classification.

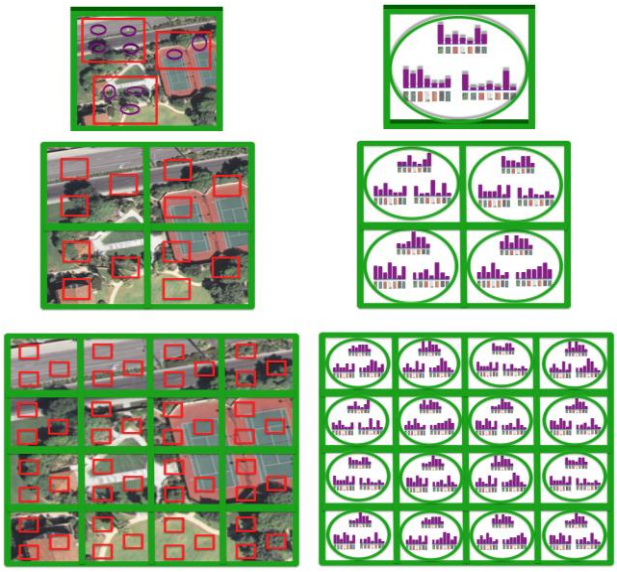


Figure 3: Illustrate Pyramid of Spatial Relatons model. An image is divided into sub-regions (green rectangles) hierarchically; An order-less collection of spatial relatons, which support regions are illustrated by the red rectangles, captures the relative spatial relationship of local features within each sub-region. A spatial relaton is represented by a quantized histogram of local features within a support region (red rectangle).

Similar to the SPM model, the PSR divides an image into sub-regions at multiple levels successively, as illustrated in Figure 3. Each sub-region (green rectangle) is represented by an order-less collection of spatial relatons (i.e., spatial relaton histogram). The spatial relaton is a quantized local features distribution within a support region (red rectangle). The concept of relaton, which is inspired by the texton and correlaton [31], means the basic relationship unit. By combining both spatial relaton histogram and Bag of Words (BOW) histogram in each sub-region, the PSR model incorporates both absolute and relative spatial relationship of local features. Hence, the PSR is more robust to rotation and translation, which are the main challenges for remotely sensed land use classification.

Unlike the co-occurrence matrix approach, the PSR describes spatial relationship of a group of local features in a support region without modeling detailed pair-wise relationship. Therefore, the PSR is more efficient in terms of computational and memory cost as compared with the conventional co-occurrence approach.

B. Spatial Relaton

As in the Bag of Words (BOW) approach for generic object recognition, we represent a geographical structure by the distribution over its structure parts. Such distribution implicitly represents the spatial relationship of the structure parts relative to the geographical structure.

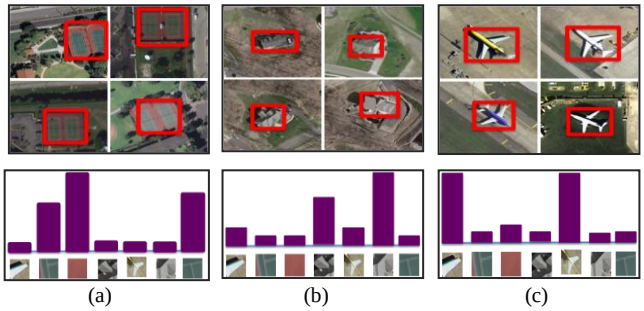


Figure 4: Illustrate structure part distribution for the geographical structure of (a) tennis court, (b) house, and (c) airplane.

Figure 4(a)-(c) illustrate structure part distribution for the geographical structures of tennis court, house, and airplane respectively. We introduce a novel concept of spatial relaton to represent spatial arrangement of structure parts relative to a geographical structure (i.e., structure part distribution). The concept of spatial relaton is analogous to the visual words of Bag of Words model. An arbitrary spatial relationship can be generated by linearly combining finite number of prototype spatial relationships, i.e., the spatial relatons.

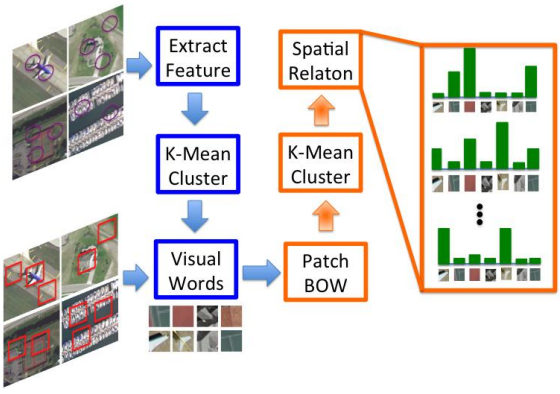


Figure 5: Illustrate the flowchart to generate a dictionary of spatial relatons.

Figure 5 illustrates the flowchart to generate a dictionary of spatial relatons. Local features, which represent structure parts, are first extracted from training images. Then we generate a codebook of N visual words by clustering local features with K-mean clustering algorithm [26]. Each training image samples a set of patches or support regions. For each of the support regions, a patch BOW histogram is generated to capture its local feature distribution.

These patch BOW histograms represent the spatial relationships of local features (i.e., structure parts), which spatial affinity constraint is defined by the size of patches. There can be infinite number of different spatial relationships of local features. Therefore, we cluster local features' spatial relationships (i.e., patch BOW histograms) into *M* prototype spatial relationships by the K-mean algorithm again. The prototype spatial relationships are spatial relatons, which provide a compact representation of relative spatial relationship of local features.

C. Pyramid of Spatial Relatons (PSR)

Motivated by the state of the art performance of the Spatial Pyramid Matching (SPM) kernel, the PSR model also partitions an image into successively fine sub-regions, as illustrated in Figure 3. In each sub-region, we extract both BOW histogram and spatial relaton histogram using local soft assignment coding and maximum pooling [25].

To perform local soft assignment coding, we introduce *metric to calculate* distance between two vectors, x_i and x_j, as shown in Equation (2).

$$D_{knn}(x_i, x_j) = \begin{cases} \|x_i - x_j\| & if\ x_i \in NN_k(x_j) \\ \infty & Otherwise \end{cases}, \quad (2)$$

where $NN_k(x_j)$ is *K* nearest neighbors of x_j. If two vectors are far from each other, the distance is simply infinite.

The local soft assignment coding of local featurex_i to visual word V_n is shown in Equation (3).

$$\psi_{i,n} = \frac{e^{-\beta * D_{knn}(x_i, V_n)}}{\sum_{j=1}^{N} e^{-\beta * D_{knn}(x_i, V_j)}}, \quad (3)$$

where β is a parameter to control the sensitivity of coding over distance. If β reaches infinitive, local soft assignment coding becomes hard assignment coding, which assigns *1* to the feature's nearest neighbor visual word, and *0* to the other visual words.

We employ maximum pooling to aggregate all codes together to form a BOW histogram as shown in Equation 4.

$$H_I^n = \max_{i=1\ldots|x|} \psi_{i,n}, \quad (4)$$

where $|x|$ is the total number of features in image *I*. Note that Equation (3) and (4) can also be used to calculate patch BOW histogram for a support region of spatial relaton except that the image *I* now becomes an image patch.

Similarly, we also compute spatial relaton histogram using local soft assignment coding and maximum pooling, as shown in Equation (5) and (6) respectively.

$$\phi_{i,m} = \frac{e^{-\beta * D_{knn}(H_{pi}, R_m)}}{\sum_{j=1}^{M} e^{-\beta * D_{knn}(H_{pi}, R_j)}}, \quad (5)$$

$$\bar{\bar{H}}_I^m = \max_{i=1\ldots|p|} \phi_{i,m}, \quad (6)$$

where H_{pi} is i^{th} patch BOW histogram in an image *I*. R_m is m^{th} spatial relaton in the relaton dictionary formed in last section. $|p|$ is the total number of patches in the image *I*.

Patches (i.e., support regions of spatial relatons) are illustrated by the blue and green bounding boxes in Figure 6(a). Each patch generates a patch BOW histogram H_{pi}, as illustrated in Figure 6(b) and 6(c) for the blue and green patch respectively. In this paper, we apply uniform grid method to extract $|p|$ patches in an image. Finally a spatial relaton histogram is the order-less collection of all patch BOW histograms in an image, as illustrated in Figure 6(d).

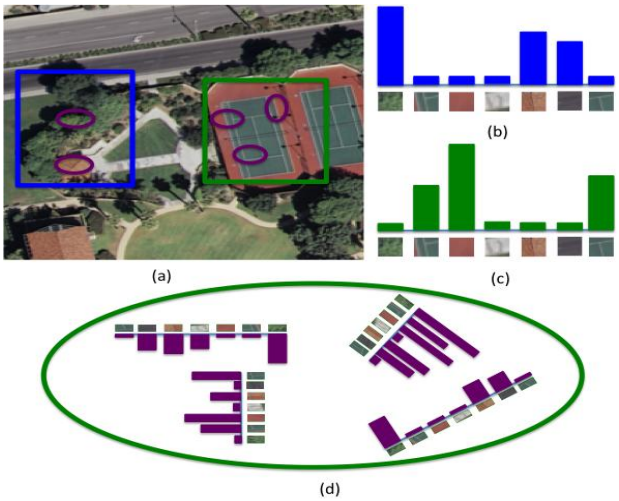


Figure 6: Illustrate spatial relaton histogram, which captures relative spatial relationship of local features. (a) A geographic image with two support regions marked by blue and green bounding boxes respectively; (b) Patch BOW histogram of the blue bounding box; (c) Patch BOW histogram of the green bounding box; (d) An order-less collection of spatial relatons to form a spatial relaton histogram.

While a spatial relaton histogram captures relative spatial relationship of local features, a BOW histogram represents an image's appearance information. To take the advantage of their complementarities, we combine both BOW histogram and spatial relaton histogram for each sub-region. The linear kernel matrix of the PSR is shown in Equation (7).

$$K(I_i, I_j) = \sum_{l=0}^{L-1} \sum_{r=0}^{S_l-1} \left(\sum_{n=0}^{N-1} H_{Il,l,r}^n H_{Ij,l,r}^n + \sum_{m=0}^{M-1} \bar{\bar{H}}_{Il,l,r}^m \bar{\bar{H}}_{Ij,l,r}^m \right), \quad (7)$$

where *N* is the codebook size of the BOW, and *M* is the dictionary size of the spatial Relatons. *H* and $\bar{\bar{H}}$ are the BOW histogram and the spatial relaton histogram respectively for each sub-region. *L* is the total number of levels in the pyramid, and S_l is the total number of sub-regions at l^{th} level.

During training step of the support vector machine (SVM), we maximize the dual form of the objective function in Equation (8).

$$L_D = \sum_{i=1}^{|I|} \partial_i - \frac{1}{2} \sum_{i=1}^{|I|} \sum_{j=1}^{|I|} \partial_i \partial_j y_i y_j K(I_i, I_j), \quad (8)$$

subjects to the following constraints.

$$0 \le \partial_i \le C, \quad (9)$$

$$\sum_{i=1}^{|I|} \partial_i y_i = 0, \quad (10)$$

where *C* is a constant to tradeoff classifier's margin with

training error, and $|I|$ is the total number of training images.

In testing phase, we can predict the class of an unknown image $I_t$with the following equation.

$$y_t = \sum_{i=1}^{|I|_s} \partial_i y_i K(I_i, I_t) + b, \quad (11)$$

where

$$b = \frac{1}{|\{i|0<\partial_i<C\}|} \sum_{\{i|0<\partial_i<C\}} \left[y_i - \sum_{j=1}^{|I|_s} \partial_j y_j K(I_i, I_j) \right], \quad (12)$$

where $|I|_s$ is the total number of support vectors obtained during the training. By substituting Equation (7) into Equation (11), we have:

$$y_t = \sum_{i=1}^{|I|_s} \partial_i y_i \sum_{l=0}^{L-1} \sum_{r=0}^{S_l-1} \left(\sum_{n=0}^{N-1} H_{Ii,l,r}^n H_{It,l,r}^n + \sum_{m=0}^{M-1} \bar{\bar{H}}_{Ii,l,r}^m \bar{\bar{H}}_{It,l,r}^m \right) + b \quad (13)$$

Simplifying equation (13), we have

$$y_t = \sum_{l=0}^{L-1} \sum_{r=0}^{S_l-1} \left(\sum_{n=0}^{N-1} H_{It,l,r}^n w_{l,r}^n + \sum_{m=0}^{M-1} \bar{\bar{H}}_{It,l,r}^m \bar{\bar{w}}_{l,r}^m \right) + b, \quad (14)$$

with

$$w_{l,r}^n = \sum_{i=1}^{|I|_s} \partial_i y_i H_{Ii,l,r}^n, \quad (15)$$

$$\bar{\bar{w}}_{l,r}^m = \sum_{i=1}^{|I|_s} \partial_i y_i \bar{\bar{H}}_{Ii,l,r}^m. \quad (16)$$

D. Computation and Memory Complexity

Since we employ linear kernel for the PSR model, the computational complexity to predict a new image is linearly proportional to the codebook size of visual words and the dictionary size of spatial relatons, as shown in Equations (14) to (16). Therefore it is much more efficient in computation as compared with the intersection kernel of the original Spatial Pyramid Matching model [21].

Similar to the co-occurrence matrix, spatial relaton histogram models relative spatial relationship between local features. However, spatial relatons capture relative spatial arrangement of a group of local features instead of detailed pairwise spatial relationship as in the co-occurrence matrix. As the result, the construction cost in computation and the storage cost in memory are reduced significantly.

If using brute force approach, it requires $O(n^4)$computation complexity to construct a co-occurrence matrix, where *n* is the number of rows or columns in an image [31]. Savarese *et al.* [31] employs integral histogram to reduce the computation cost to $O(n^2)$. However, memory cost of an integral histogram can become prohibitively expensive, especially when the codebook size increases. For a typical 1000 by 1000 image with moderate codebook size of 1000 visual words, the memory of its integral histogram is 4GB, assuming 4 bytes for a float value.

On the other hand, spatial relaton histogram requires $O(|p| * (n^2 + M * N))$, where $|p|$ is total number of patches used in an image, and *M* is the dictionary size of spatial relatons. *N* is the codebook size of visual words. $|p|$ is much smaller than *n*. The number of patches used in our experiment is 4. Based on our experiments, it only slightly improves the classification accuracy when the number of patches continues increasing. *M*N* is comparable to n^2 . Therefore, the computation cost to construct spatial relaton histogram is still $O(n^2)$. And it only requires memory to store a dictionary of spatial relatons, which is about the size of an image. Furthermore, the dictionary is shared among all images.

The storage cost of a co-occurrence matrix is $O(N^2)$, where *N* is the total number of visual words in the codebook. For a moderate codebook size of 1000, the storage cost can reach 4MB for a co-occurrence matrix. On the other hand, the storage cost for a spatial relaton histogram is only $O(M)$, where *M* is the total number of spatial relatons in the relaton dictionary. The size of the relaton dictionary is 300 in most of our experiments. Hence, the storage cost of a spatial relaton histogram is only about 1 KB.

E. Local Features

We adopt Scale Invariant Feature Transformation (SIFT) [27] as the local feature in our experiments. SIFT features have shown great success on both computer vision and remote sensing applications [5, 27, 28, 42, 43, 44].

In this paper, we apply SIFT feature on gray images with densely sampled interest points, which is the most popular approach in both computer vision and remote sensing communities [28, 30, 42, 43, 44,]. The sampling rate used in this paper is one interest point every 8 pixels in both *x* and *y* directions, as suggested in the paper [28].

IV. Experiment

To evaluate the effectiveness of the proposed PSR framework for the remotely sensed land use classification, we conduct experiments using the Land Use or Land Cover (LULC) high-resolution aerial image database [42], which is one of the largest geographical image databases with ground truth labeling.

We first compare classification accuracy of the proposed PSR with the state of the art performance reported on the LULC dataset. Our proposed PSR achieves 8% higher in accuracy than the state of the art result under similar experimental setting. If using gray image only, we outperform the state of the art by more than 11% on the LULC database.

Then we compare the PSR algorithm with the Bag of Words model and the Spatial Pyramid Matching model under our experimental setup. We find that the PSR model exceeds these state of the art methods by a significant margin across a range of visual words codebook size.

We further investigate the effect of several important parameters related to the PSR model, such as the dictionary size of the spatial relatons, and the number of hierarchical levels.

Figure 7: Sample geographical images from each of 21 categories in the Land Use or Land Cover (LULC) database.

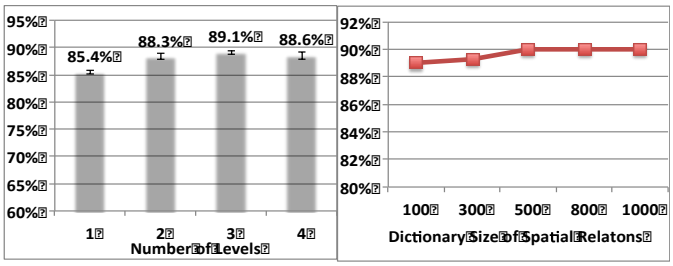
A. *Dataset*

The Land Use or Land Cover (LULC) dataset [42] is one of the largest geographical image databases with ground truth, which are publically available. The images are downloaded from the United States Geographical Survey (USGS) national map.

There are total of 21 categories, including agricultural, beach, buildings, chaparral, dense residential, and forest etc. Sample geographical images of each land use category are shown in the Figure 7. Each class has 100 images with same size, i.e., 256 by 256 pixels. The pixel resolutions of all images are 30cm per pixel.

B. *Experimental Setup*

To be consistent with other researchers' experimental setting on the LULC dataset [43, 44], we randomly partition the database into five subsets, with each subset contains 20 images from each land use category. Four subsets are used as the training data, and the remaining subset is used in the testing.



(a) (b)

Figure 8: Evaluate the effect on the classification accuracy for the parameters of (a) number of hierarchical levels; (b) and spatial relatons' dictionary size.

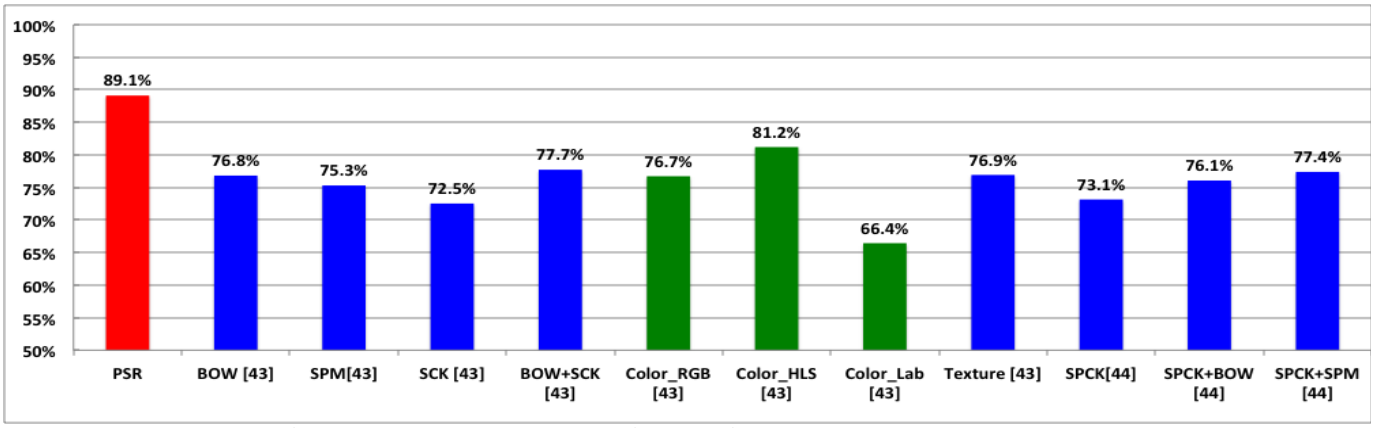


Figure 9: Compare the Pyramid of Spatial Relatons (PSR) with the state of the art performance reported in the literature on the LULC database. The blue bars are the performance based on gray images. The green bars are based on color images. The performance reported in the PSR is only based on gray images.

The experiments are repeated five times by selecting one of the five subsets as the testing data. The average classification accuracies and the standard errors are reported in this paper, unless otherwise noted.

As the codebook size increases, it can take several days to generate just one codebook from four subsets of training data. In order to facilitate the experiments, we generate a codebook of visual words by selecting four subsets of training data, and reuse the same visual word codebook for the other experiments as we rotate through the testing subset.

In the experiments, we use the dictionary size of 300 for spatial relatons, as it provides a good tradeoff between the accuracy and the efficiency. In Section C, we will evaluate how the dictionary size of spatial relatons affects the classification accuracy.

C. Effect of PSR Parameters

The Pyramid of Spatial Relatons (PSR) model has two important parameters, i.e., the number of pyramid levels and the dictionary size of the spatial relatons. In the following experiments, we will use the same experimental setup as in the previous section except changing one of the parameters.

Figure 8(a) shows the classification accuracy over different number of hierarchical levels in the PSR model. As the hierarchical level increases, the performance continues improving until the number of levels is 3.

The performance decreases slightly, when the number of hierarchical levels reaches 4. As the hierarchical level increases, the influence of absolute spatial information in the model also increases since the partitioned sub-regions becomes finer. At the fourth hierarchical level, an image is divided into 8 by 8 sub-regions.

Figure 8(b) shows the effect of spatial relatons' dictionary size on the classification accuracy. We run this experiment on the first subset of testing data. The performance improves only slightly when the dictionary size increase from 100 to 500. After that, the performance remains constant. Hence, the dictionary size of 300 provides a good tradeoff between the accuracy and efficiency.

D. Comparison with the state of the art

To prove the effectiveness of the proposed Pyramid of Spatial Relatons (PSR), we first compare its classification performance on the LULC dataset with the state of the art performance reported in the literatures under similar experimental setup (*i.e.*, 80% of images from each category are used as training, and the remaining 20% images are used as testing). As shown in Figure 9, the PSR model achieves around 8% higher in accuracy, as compared with the best performance reported using HLS color histogram features [43]. Note that the PSR does not use color information in the model, although the accuracy can potentially be higher if the color information is also modeled.

If considering gray level image only, the PSR model achieves more than 11% higher in accuracy than the best performance of the Spatial Co-occurrence Kernel with the Bag of Words (BOW+SCK). Yang and Newsam [43] also employed SIFT features for the BOW+SCK model. Their experimental setup is also similar to ours, *i.e.*, 80% of the images in each category are used as training data, and the remaining 20% of the images are used as testing data.

The superior performance, as comparing with the current state of the art results on the LULC dataset, demonstrates the effectiveness of the proposed PSR model for remotely sensed land use classification.

To further evaluate the performance of the PSR model with the state of the art algorithms, we implement two of the most successful algorithms on the geographical image classification, *i.e.*, the Bag of Words (BOW) model and the Spatial Pyramid Matching (SPM) model. We then directly compare the PSR with those two algorithms under our experimental setup. We found that the PSR outperforms the two state of the art algorithms by a significant margin over various codebook sizes. The detailed comparisons are shown in Section E and F below.

E. Comparison with Bag of Words (BOW)

The performance of the Bag of Words (BOW) model can vary significantly with the codebook size. Hence, we compare the performance of the Pyramid of Spatial Relatons (PSR) model with the BOW over different codebook sizes, as shown in Figure 10.

The PSR outperforms the BOW model over all codebook sizes, *i.e.*, 5000, 1000, 500, 300, 100, 50. The performance improvement is especially prominent when the codebook size is small. At the codebook size of 50, the improvement is more than 30%.

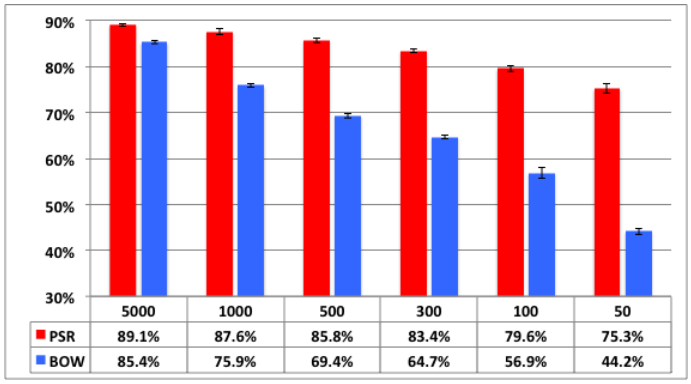


Figure 10: Compare classification accuracy of the Pyramid of Spatial Relatons (PSR) with the Bag of Words (BOW) model over codebook sizes (x-axis).

Even at the codebook size of 1000, which are typically used by other researchers on the LULC dataset [43], the PSR model exceeds the BOW model by more than 11%. At the codebook size of 5000, the PSR model achieves the average classification accuracy of 89.1%, which is almost 4% higher than the BOW model.

As shown in Figure 10, the standard errors are very small over all codebook sizes, which confirm that the improvement of the PSR is statistically significant across all codebook sizes we evaluated.

For the PSR model, the most confusion occurs in the "building", "dense residential", "mobile home park" and "medium residential" categories. By observing the sample images in Figure 7 on these four categories, we find the great similarity between the "mobile home park" and the "medium residential" categories. The "building" and "dense residential" categories also show some similarity.

(a)

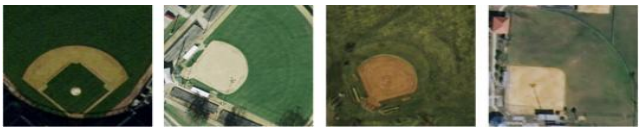
(b)

Figure 11: (a) Building images misclassified as dense residential by the PSR model; (b) Mobile home park images misclassified as medium residential by the PSR model.

Figure 11(a) shows some of the building images, which are misclassified as dense residential. The first two images have large shadow in the image, which cause the recognition to become very challenged even by human. The last two images in Figure 11(a) share some similarity with the dense residential category. Note that the first three building images are also misclassified as dense residential by the BOW model. And the last image in Figure 11(a) is misclassified as tennis court by the BOW model.

Figure 11(b) shows some misclassified images of mobile home park. All four images are incorrectly predicted as medium residential area, as these two categories have very similar appearance.

Except for the Mobile home park and the Medium Residential categories, the PSR has better or comparable performance over the BOW model in all other categories. The performance improvement is especially profound over the baseball diamond and sparse residential categories, which is around 25% and 35% respectively as shown in Figure 15. The BOW model confuses the baseball diamond with storage tanks and sparse residential. For the sparse residential, the BOW model confuse it with many other categories including intersection, baseball diamond and storage tanks etc.



(a)

(b)

Figure 12: (a) Baseball diamond images, (b) and Sparse Residential images, are predicted correctly by the PSR model, but not by the BOW model.

Figure 12 shows the geographical images from those two categories, which are predicted correctly by the PSR model, but not by the BOW model. The BOW model misclassifies the second and the fourth image in Figure 12(a) as storage tanks, and the third image of the baseball diamond as beach. By incorporating both relative and absolute spatial information of baseball diamond and its surroundings, the PSR model is able to correctly predict them as the baseball diamond. Similarly, without the spatial order of sparse residential area and its surroundings, the BOW misclassifies the last two geographical images in Figure 12(b) as intersection.

F. Comparing with Spatial Pyramid Matching (SPM)

We implement the original Spatial Pyramid Matching (SPM) model, which employs intersection kernel [21]. Figure 13 compares the classification accuracy of the PSR with the SPM model over same set of codebook sizes in Figure 10.

As shown in Figure 13, the best performance of the PSR model is more than 3% higher in classification accuracy than that of the SPM model. The PSR model has better performance when the codebook size is large, which is at least 500 for the LULC database in our experiments. The SPM model has slightly better performance than that of the PSR model when the codebook size is small.

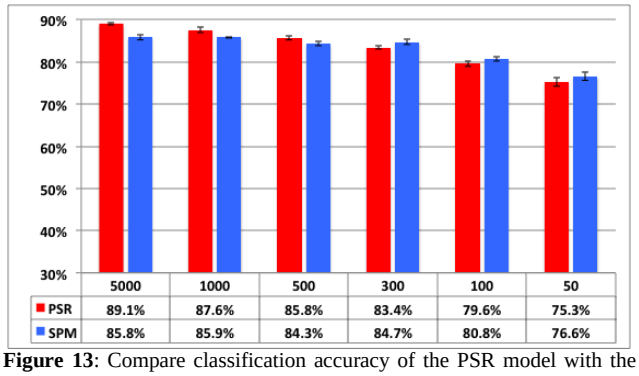


Figure 13: Compare classification accuracy of the PSR model with the Spatial Pyramid of Matching (SPM) model over codebook sizes (x-axis).

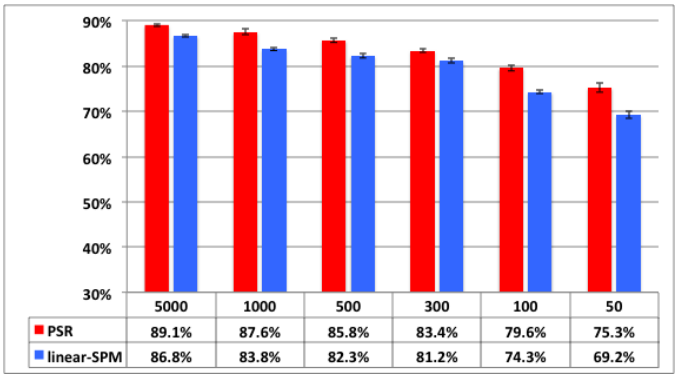


Figure 14: Compare classification accuracy of the Pyramid of Spatial Relatons (PSR) with the linear Spatial Pyramid of Matching (linear-SPM) model over different codebook sizes (x-axis).

To understand the performance advantage of the SPM at small codebook size, we replace the non-linear kernel, i.e., the intersection kernel, with the linear kernel as in the PSR model. We found that the performance of the PSR model improves across all codebook sizes. The detailed comparison is shown in Figure 14. Note that the best performance of the linear kernel SPM at the codebook size of 5000 is actually better than that of the original SPM model with the non-linear intersection kernel.

As compared with the PSR, the SPM model has lower performance particularly on the categories, which are not invariant to the rotation, such as the airplane and the baseball diamond. As an example, the baseball diamond is most confused with the golf course in the SPM model. By capturing relative spatial information, the PSR model is able to reduce the confusion, and achieves 95% on the baseball diamond category, as shown in Figure 15.

Figure 15 shows the performance of all the three models over each category of LULC database. The PSR model has better performance than the SPM in 13 out of 21 categories. In the remaining categories, they have equal performance. As compared with the BOW model, the PSR model has better performance in 9 out of 21 categories. The remaining categories have equal performance except 2 categories.

V. Conclusion

In this paper, we have proposed a Pyramid of Spatial Relatons (PSR) model to incorporate both relative and absolute spatial information into the Bag of Words (BOW) model for the remotely sensed land use classification. The novel spatial relaton captures quantized relative spatial relationship of local features using local feature distribution, which is more computationally efficient than the co-occurrence matrix approach. The proposed PSR model outperforms the state of the art performance by 8%. Comparing to the classification accuracy reported without color information, the PSR achieves more than 11% higher. We also implement two state of the art algorithms for the geographical image classification in our experimental setting. The side-by-side comparison with those two algorithms further demonstrates the advantages of the PSR model.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. ACM Press, ISBN: 020139829, 1999.

[2] H. Bay, A. Ess, T. Tuytelaars, L. Gool, “Speeded-Up Robust Features (SURF)”, Computer Vision and Image Understanding Vol. 110, Issues 3, PP. 346-359, June 2008

[3] J. Bordes, V. Prinet, “Mixture distributions for weakly supervised classification in remote sensing images”, BMVC 2008.

[4] Y. Boureau, N. Roux, F. Bach, J. Ponce, Y. LeCun, “Ask the locals: multi-way local pooling for image recognition”, ICCV 2011.

[5] G. Burghouts, J. Geusebroek, “Performance evaluation of local color invariants”, Computer Vision Image Understanding, 2009.

[6] Y. Cao, C. Wang, Z. Li, L. Zhang, L. Zhang, “Spatial-Bag-of-Features”, CVPR 2010.

[7] S. Chen and Y. Tian, “Describing Visual Scene through EigenMaps”, Journal of Computer Vision and Image Processing, Vol. 2, No. 1, March 2012.

[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, “Visual Categorization with Bag of Key points”, ECCV, 2004.

[9] N. Dalal, B. Triggs, “Histograms of Oriented Gradients for Human Detection”, CVPR 2005.

[10] J. Deng, A. Berg, K. Li, L. Fei-Fei, “What does classifying more than 10,000 image categories tell us”, ECCV, 2010.

[11] L. Fei-Fei and P. Perona,“A Bayesian hierarchy model for learning natural scene categories”, CVPR 2005.

[12] L. Fei-Fei, R. Fergus and P. Perona, “One-Shot learning of object categories”, IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2006.

[13] W. Freeman and E. Adelson, “The Design and Use of Steerable Filters,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 13, no. 9, pp. 891-906, Sept. 1991.

[14] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset”, Technical Report 7694, California Institute of Technology, 2007.

[15] R. Haralick, “Statistical and structural approaches to texture”, Proceedings of IEEE, Vol. 67, no. 5, pp. 786-804, May 1979.

[16] J. Huang, S. Kumar, M. Mitra, W. Zhu, R. Zabih, “Image indexing using color correlograms”, Computer Vision and Pattern Recognition (CVPR), 1997.

[17] P. Gehler, S. Nowozin, “On feature combination for multiclass object classification”, ICCV, 2009.

[18] J.Gemert, C. Veenman, A. Smeulders, J. Geusebroek,“Visual Word Ambiguity”, IEEE TPAMI, 99, 2009.

[19] H. Goncalves, L. Corte-Real, J. Goncalves, “Automatic image registration through image segmentation and SIFT”, IEEE Trans. Geosci. Remote Sens. Vol. 49, no. 7, pp. 2589-2600, Jul. 2011.

[20] Y. Ke and R. Sukthankar, “PCA-SIFT: A More Distinctive Representation for Local Image Descriptors,” Proc. Conf. Computer Vision and Pattern Recognition, pp. 511-517, 2004.

[21] S. Lazebnik, C. Schmid, J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”, Computer Vision and Pattern Recognition (CVPR), 2006.

[22] Z. Li, J. Imai, M. Kaneko, “Robust Face Recognition Using Block-based Bag of Words”, International Conference on Pattern Recognition, 2010

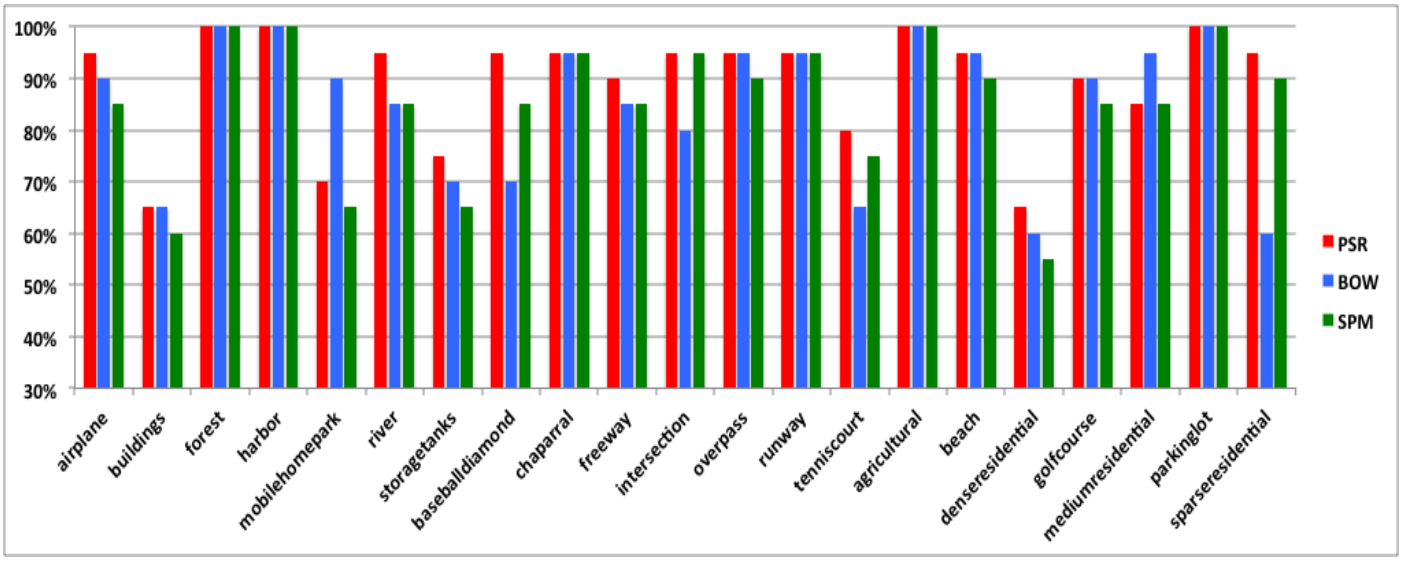


Figure 15: Compare category-wise performance of the Pyramid of Spatial Relatons (PSR) with the Bag of Words (BOW) and the Spatial Pyramid Of Matching with original intersection kernel (SPM) on the LULC database.

[23] H. Ling, S. Soatto, "Proximity distribution kernels for geometric context in category recognition", ICCV 2007.

[24] D. Liu, G. Hua, P. Viola, T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization", CVPR 2008.

[25] L. Liu, L. Wang, X. Liu, "In Defense of Soft-assignment Coding", International Conference on Computer Vision (ICCV), 2011.

[26] S. Lloyd, "Least Squares Quantization in PCM", IEEE Trans. Information Theory, Vol. 28, no. 2, pp. 129-137, Mar. 1982.

[27] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", IJCV, 2004.

[28] S. McCann, D. Lowe, "Spatially Local Coding for Object Recognition", Asian Conference on Computer Vision (ACCV), 2012.

[29] A. Mukherjee, M. Velez-Reyes, B. Roysam, "Interest points for hyperspectral image data", IEEE Trans. Geosci. Remote Sens. Vol. 47, no. 3, pp. 748-760, Mar. 2009.

[30] E. Nowak, F. Jurie, B. Triggs, "Sampling Strategies for Bag of Feature Image Classification", ECCV, 2006.

[31] S. Savarese, J. Winn, A. Criminisi, "Discriminative Object Class Models of Appearance and Shape by Correlatons", CVPR, 2006.

[32] A. Sedaghat, M. Mokhtarzade, H. Ebadi, "Uniform robust scale invariant feature matching for optical remote sensing images", IEEE Trans. Geosci. Remote Sens. Vol. 49, no. 11, pp. 4516-4527, Nov. 2011.

[33] B. Sirmacek, C. Unsalan, "Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory", IEEE Trans. Geoscience and Remote Sensing, Vol. 47, no. 4, pp. 1156-1167, Apr. 2009

[34] B. Sirmacek, C. Unsalan, "Urban area detection using local feature points and spatial voting", IEEE Trans. Geoscience and Remote Sensing, Vol. 7, no. 1, pp. 146-150, Jan. 2010

[35] J. Sivic, A. Zisserman, "Efficient visual search of videos cast as text retrieval", IEEE. Trans. Pattern Analysis and Machine Intelligence, Vol. 31, no. 4, pp. 591-606, Apr. 2009.

[36] W. Tobler, "A computer movie simulating urban growth in the Detroit region", Economy Geography, Vol. 46, pp. 234-240, 1970.

[37] M. Varma, D. Ray, "Learning The Discriminative Power-Invariance Trade-Off", ICCV 2007.

[38] A. Vedaldi, V. Gulshan, M. Varma, A. Zisserman, "Multiple kernels for object detection", ICCV, 2009.

[39] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, "SUN Database: Large-Scale Scene Recognition from Abbey to Zoo", Computer Vision and Pattern Recognition (CVPR), 2010.

[40] Z. Xiong, Y. Zhang, "A novel interest-point-matching algorithm for high-resolution satellite images", IEEE Trans. Geosci. Remote Sens., vol. 47, no. 12, pp. 4189-4200, Dec. 2009.

[41] S. Xu, T. Fang, D. Li, S. Wang, "Object Classification of Aerial Images with Bag-of-Visual Words", IEEE Trans. Geoscience and Remote Sensing, Vol. 7, no. 2, pp. 366-370, Apr. 2010.

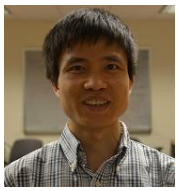
[42] Y. Yang, S. Newsam, "Geographic Image Retrieval Using Local Invariant Features", IEEE Trans. Geoscience and Remote Sensing, Vol. 51, no. 2, pp. 818-832, Feb. 2013.

[43] Y. Yang, S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification", ACM GIS, 2010.

[44] Y. Yang, S. Newsam, "Spatial Pyramid Co-occurrence for Image Classification", ICCV, 2011.

[45] J. Yang, K. Yu, Y. Gong, T. Huang, "Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification", Computer Vision and Pattern Recognition (CVPR), 2009.

[46] E. Zhang, M. Mayo, "Improving Bag-of-Words model with spatial information", International conference of Image and Vision Computing, New Zealand, 2010.



Shizhi Chen received his BS degree of Electrical Engineering from SUNY Binghamton, New York in 2004, and the MS degree of Electrical Engineering and Computer Science from UC Berkeley, California in 2006. He received his Phd in Electrical Engineering from the City College of New York in 2013. From 2006 to 2009, he worked as an engineer in several companies including Altera, Supertex Inc., and US Patent and Trademark Office. He is a member of Eta Kappa Nu (electrical engineering honor society), and a member of Tau Beta Pi (engineering honor society). He also received numerous scholarships and fellowships, including Achievement Rewards for College Scientists (ARCS) Fellowship, and NOAA CREST Fellowship. He is currently working in Naval Undersea Warfare Center (NUWC) as an engineer. His research interests include object recognition, detection and tracking, and machine learning.

YingLi Tian (M'99–SM'01) received her BS and MS from TianJin University, China in 1987 and 1990 and her PhD from the Chinese University of Hong Kong, Hong Kong, in 1996. After holding a faculty position at National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, she joined Carnegie Mellon University in 1998, where she was a postdoctoral fellow of the Robotics Institute. Then she worked as a research staff member in IBM T. J. Watson Research Center from 2001 to 2008. She is currently a professor in Department of Electrical Engineering at the City College of New York and Department of Computer Science at the Graduate Center, the City University of New York. Her current research focuses on a wide range of computer vision problems from motion detection and analysis, assistive technology, to human identification, facial expression analysis, and video surveillance.