Satellite Image Scene Classification Using Spatial Information

Weiwei Song[1], Dunwei Wen[2], Ke Wang[1], Tong Liu[1] and Mujun Zang[1]
[1]College of Communication Engineering, Jilin University, Changchun, Jilin, China
[2]School of Computing and Information Systems, Athabasca University, Alberta, Canada

ABSTRACT

In order to enhance the local feature's describing capacity and improve the classification performance of high-resolution (HR) satellite images, we present an HR satellite image scene classification method that make use of spatial information of local feature. First, the spatial pyramid matching model (SPMM) is adopted to encode spatial information of local feature. Then, images are represented by the local feature descriptors and encoding information. Finally, the support vector machine (SVM) classifier is employed to classify image scenes. The experiment results on a real satellite image dataset show that our method can classify the scene classes with an 82.6% accuracy, which indicates that the method can work well on describing HR satellite images and classifying different scenes.

Keywords: Satellite image, spatial pyramid model, spatial information, support vector machine.

1. INTRODUCTION

In scene classification, data mining and pattern recognition, categorization of high-resolution (HR) satellite image scene faces significant challenges owing to its two characteristics [1]. First, with the improvement of spatial resolution, the objects show more details in a satellite scene. It is necessary to find some effective features for representing detailed information of images. Second, objects may appear at different orientations and scales in the same category scene of HR satellite images and different scenes may contain the same object. For instance, the cars in parking areas may have different orientations and scales, and so do the buildings in commercial areas. Meanwhile, the brightness of the same scene is influenced by lighting under different weather conditions. These characteristics put big obstacle to represent images for scene classification. Therefore, we should employ some invariant properties, such as orientation invariance, scale invariance and contrast invariance.

Image representation is important to scene classification, unlike low-resolution satellite images, which are described through texture and intensity cues effectively [2]. Xu *et al.* [3] presented a scene classification method which employs feature combination based on probabilistic latent semantic analysis (pLSA) to describe the detailed information of a scene. Considering that the object contains a wealth of information in HR satellite images, Xia *et al.* used texture and structure features which are robust to orientation, scale and contrast for image indexing [1]. Another strategy is to take local features and semantic concepts into consideration, where image patches with the same semantic concept are assigned to the same class in a big scene. The strategy also combines semantic concepts with topic model, e.g., LDA (latent Dirichlet allocation), for large image scene semantic annotation [4]. Most of the existing methods have focused on the feature level to achieve HR satellite image scene classification, but the spatial information between local features plays a key role in describing the scene and thus improving the performance of local features representation.

Spatial pyramid matching model (SPMM) [5] encodes local feature descriptors, during which spatial information is integrated into the local features to better describe images. The main idea of SPMM is to repeatedly divide levels at increasingly fine resolution. Each level is divided into some regular grids, and local features have greater probability to assign to the same grid if the relations of spatial geometry are closer. In this paper, we will introduce spatial pyramid coding method to HR satellite image scene classification. Experimental results on a real satellite image dataset confirm that the local features of SPMM encoding can improve classification performance.

2. APPROACH

One contribution of this paper is the introduction of spatial information for representing images to HR satellite image scene classification. Spatial information is vital for geographic data composition and its analysis. As early as 1970's,

Sixth International Conference on Graphic and Image Processing (ICGIP 2014),
edited by Yulin Wang, Xudong Jiang, David Zhang, Proc. of SPIE Vol. 9443, 94431K

Walter Tobler [6] stated in his "first law of geography" that "everything is related to everything else, but near things are more related than distant things". Inspired by this law, when describing a HR satellite image scene we take into consideration the spatial distribution relationship between features of the scene. Another contribution is the choice of a proper scene feature for classification. Feature selecting is important to classify of HR satellite image scene, and in this paper we choose SIFT (scale-invariant feature transform) feature [7] for image description for two reasons. One is that SIFT descriptors have some invariant properties, such as affine, rotation, brightness and scale invariant, which can well describe some prominent regions, such as ground, road and river. The other reason is that SIFT descriptor performs image matching better than other local descriptors. The flow chart of the proposed method is shown in Fig.1.

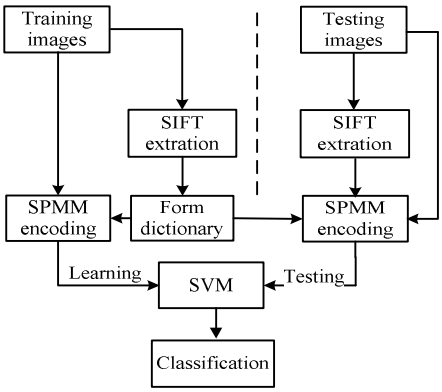


Figure 1. Flow chart of the method

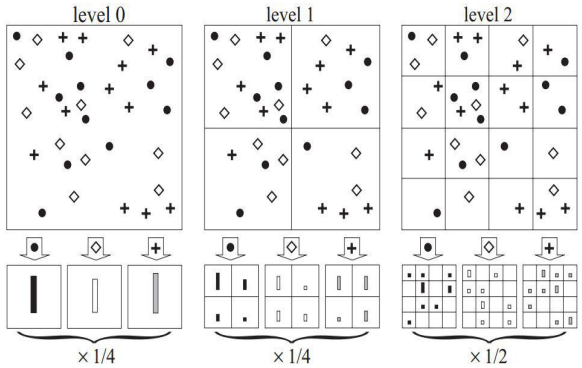


Figure 2. Example of constructing a three-level pyramid(adopted from [5])

2.1 Feature Extraction and Visual Dictionary Construction

It has been proved that dense features work better for image classification [8], and thus we decide to use a dense regular grid instead of interest point to sample SIFT descriptors in HR satellite image scene classification. Two schemes can be implemented to extract features for constructing visual dictionary. In the first scheme each image in the dataset is divided into some regular patches and all of these patches form a set, from it a subset of patches is selected randomly. A visual dictionary is constructed based on the features extracted from the subset. In the second scheme is each image from the training set is divided into some regular patches, and then features are extracted from the subset of these patches. We take into account the fact that images to be classified are unknown and they do not participate in constructing the visual dictionary in actual experiment. In order to make the experimental result more convinced, we employ the second scheme to learn visual words by performing k-means cluster algorithm on the extracted features. Each learned cluster is a visual word, and the number of clusters is the size of the visual dictionary.

2.2 Spatial Coding for HR Satellite Images

The spatial pyramid matching model (SPMM) is introduced for spatial coding and matching for HR satellite images. As the pyramid matching model (PMM) [9] is not suitable for high-dimensional features matching, Lazebnik *et al.* proposed SPMM, a pyramid matching model applied in two-dimensional space of the image. The approximate spatial correspondence can be found from the visual words of two images. The process of Spatial encoding for HR satellite images includes three steps, i.e., local features quantization, building features histograms and modeling spatial pyramid. In PMM, the feature space of image is divided by a series of increasingly finer grids. As shown in Fig.2, there are $2^l \times 2^l$ grids at the l th $(l = 0,\ldots, L)$ level. We calculate histograms of visual words for each grid of an image; the combination of all histograms is the pyramid representation of the image. Let X and Y denote two images, the similarity between them is obtained by weighted match values for each level, where the match value of the l th level can be calculated by the histogram intersection function [10]:

$$I(H_X^l, H_Y^l) = \sum_{i=1}^{D} \min(H_X^l(i), H_Y^l(i)) \tag{1}$$

We abbreviate $I(H_x^l, H_y^l)$ to I^l . H_x^l and H_y^l denote histograms of visual words of two images at level l. $H_x^l(i)$ and $H_y^l(i)$ denote the number of visual words in X and Y that fall in the i th grid at the l th level. D denote the number

of grids. It should be pointed out that the matching points exist in finer level $l+1$ also exist in level l, the number of new matching points at level l is $I^l - I^{l+1}$ for $l = 0,\ldots,L-1$. The finer the feature space is divided, the more attention the match value of similarity points in the grid is obtained. Therefore, the weight associated with level l is set to $1/2^{L-l}$. Putting all levels together forms the pyramid match kernel:

$$k^L(X,Y) = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(I^l - I^{l+1}) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \tag{2}$$

As mentioned above, SPMM is the application of pyramid matching model in a two-dimensional image space. Spatially closer local features have greater probability of appearing in the same grid at a finer level, which means that the spatial geometric relationships between local features is integrated into the image matching, we call this process is encoding. Let M be the size of visual dictionary, X_m and Y_m be two sets of two-dimensional vectors, representing the coordinates of visual word of the image, the kernel of SPMM is given by:

$$K^L(X,Y) = \sum_{m=1}^{M} k^L(X_m, Y_m) \tag{3}$$

Our spatial pyramid encoding for HR satellite image is performed as follows. After dividing the HR satellite image dataset into training set and testing set, we extract SIFT descriptors from the two sets respectively and train visual dictionary on the training set according to the above method. An image is partitioned into a sequence of spatial grids at different resolution $0,\ldots,L$. We count histograms of visual words for every grid and combine all of these histograms at different resolution into a feature vector V for each image.

2.3 Training and Classification

Detail discussion and comparison for selecting the number of levels has been introduced in [5], which regards $L = 2$ as the optimal choice for spatial pyramid-based classification, considering the computational complexity and the classification performance. Thus we employ $L = 2$,i.e., a three-levels spatial pyramid for coding the HR satellite image. An image can be represented by the feature vector V describe in above section. The LIBSVM toolbox [11] is used to train an SVM classifier on the training set. We select histogram intersection function as the kernel function of the SVM classifier. We use 5-fold cross-validation to verify the validity of the classifier, and then test the images in the testing set by the trained classifier to realize the HR image scene classification.

3. EXPERIMENT

3.1 Experimental Dataset and Parameter Settings

In order to make the classification more reliable, we performed the experiments on a real HR satellite image dataset obtained from [1] and [3] (both were originally gathered from Google Earth). It contains 12 types of HR satellite image scenes, including airport, bridge, commercial, forest, industrial, meadow, parking, pond, port, residential, river, viaduct; three typical images of each scene are shown in Fig.3. For each class, there are 50 images with the size of 600×600 pixels. For each type of scene: 5, 10, 15, 20, 25 images are selected randomly from the dataset as training set and the rest as testing set. All the experiments are repeated five times, we calculate the average accuracy of the five experiments as the final performance for each group. The SIFT descriptors are extracted on a regular grid size of 16×16 pixels at a step of 8 pixels. The size of visual dictionary is 300.

3.2 Experimental Results and Analysis

In this part, we compare our method with other methods on the same dataset, and analyze the experimental results. Xu *et al.* adopted single feature (pLSA+SIFT) and multiple features combination (pLSA+Feature Combination) with pLSA topic model respectively [3], as shown in Fig.4 (a), we can see that, when the number of training images is small, the classification accuracy of our method (SPMM+SIFT) is relatively low, which is not sufficient to show the advantage of the model. This can be explained by the characteristics of HR satellite images, where the objects in the same scene tend to appear at different orientation, scales and brightness, which requires a relatively large number of training images to convey enough information about the scenes for classification. Xu *et al.* applied multiple features combination to

describe images and adopt a two-stage classifier for classification, which enhanced the ability to represent and classify images when the number of training images is small. As the number of training images increases, the advantage of SPMM for representing images is gradually revealed and the performance of classification is improved. When the number of training images is 25, the average accuracy of classification reaches 82.6%, almost 3 percentage points higher than Xu *et al.* (79.8%). It also shows that the classification accuracy base on the combination of single SIFT feature and topic model is not very satisfactory. Fig.4 (b) gives a comparison of standard deviations of the three methods. It shows that the standard deviation obtained by our method is relatively small; it also demonstrates that the results of each experiment are relatively stable and less affected by randomly chosen training set and testing set.

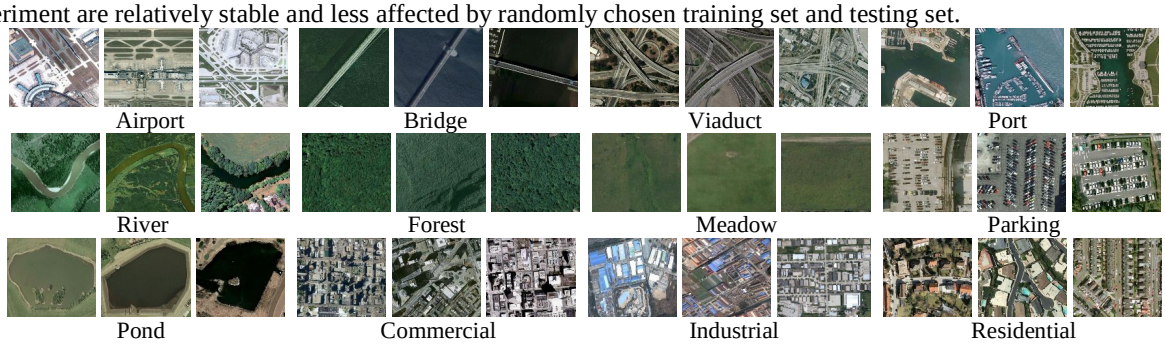
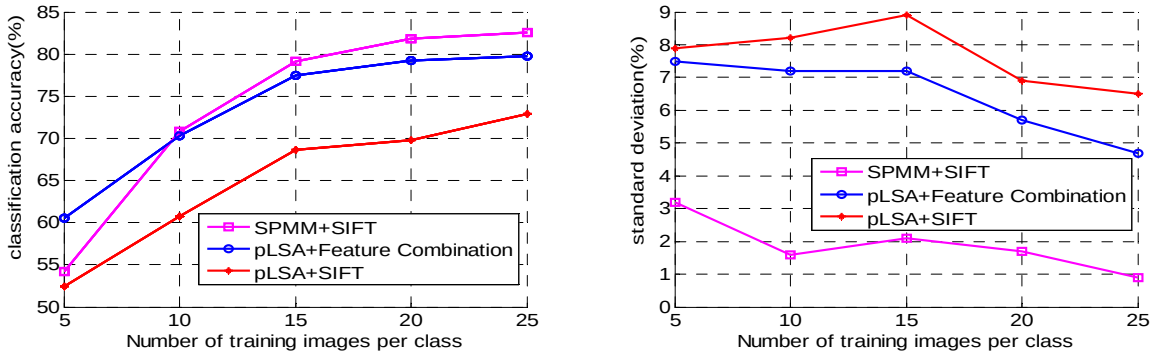


Figure3. Some samples of the high-resolution satellite image database, 3 of each class are shown here



（a）classification performance comparison　　（b）standard deviation comparison

Figure4. Performance comparison of classification and standard deviation

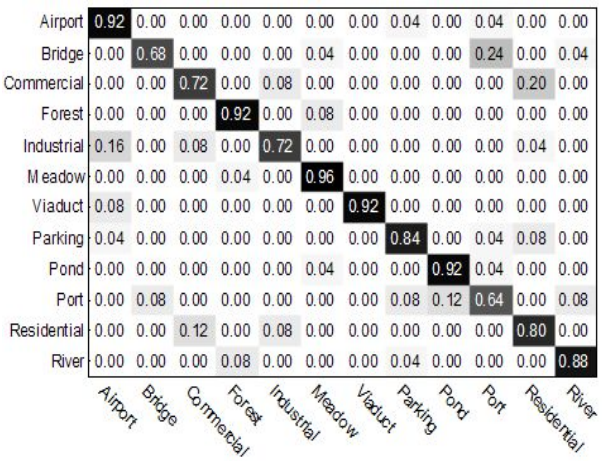


Figure5. Confusion matrix of the classification accuracy

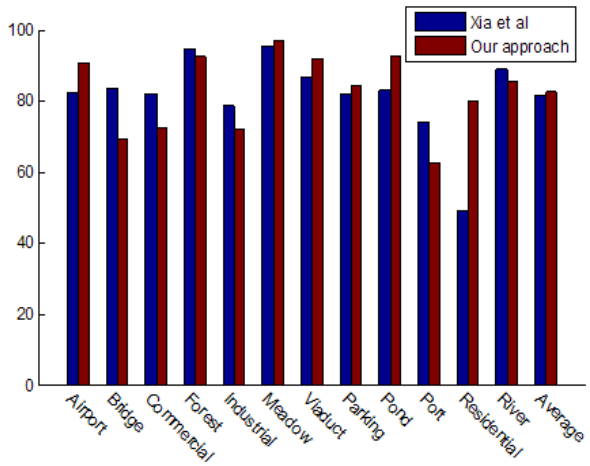


Figure6. Each class classification performance

Fig.5 gives a confusion matrix for a group of experiments performed when the number of training images is 25. We can see that the distinctiveness of objects contained in different scenes is greater, the classification accuracy is higher. For example, the airport, forest, and meadow types get an accuracy of 92%, 92% and 96% respectively. Different scenes containing similar object information are easily misclassified mutually. For instance, building is among the main objects appeared in commercial, industrial, and residential types of scenes, whose classification accuracy only reach 72%, 72%, and 80.2% respectively. Fig.6 shows that both Xia's method and our method obtain higher classification accuracy on distinctive scenes than that on similar scenes, and we have achieved higher classification accuracy than Xia's on most of these distinctive scenes. Xia *et al.* got superior performance on some types of scenes, such as bridge, pond, port, industrial and commercial. In these similar scenes, the combination of structure information and texture information may have stronger capacity of feature representation than our method. However, the accuracy of residential is only 48.87%, which is far below the accuracy of our method 80.2%. This confirms that our method is relatively stable in representing similar scenes. The average accuracy of our approach is 82.6%, about one percentage point higher than the method reported in [1]. From the above performances comparison and analysis, we conclude that, by using SIFT feature and utilizing the spatial information of local features through SPMM encoding, our approach can enhance the ability of describing HR satellite images and improve the classification performance.

4. CONCLUSION

In this paper, we have presented a method for HR satellite image scene classification by using spatial information of local features. Experimental comparison and analysis have proved that spatial coding of local features by SPMM plays a key role in satellite image scene classification. The average accuracy of classification is 82.6%, which is superior to the other two methods. The combination of spatial information with local features can significantly enhance the ability of description for representing images.

Although the proposed method has improved the classification performance, similar scenes may be misclassified mutually. In order to improve this situation, in our future work, we will try to use multiple features combination encoded by SPMM to represent images, and apply multiple stage classifiers for learning and testing.

REFERENCES

[1] Xia, G. S., Yang, W., Delon, J., Gousseau, Y., Sun, H., Maître, H., "Structural high-resolution satellite image indexing,"ISPRS TC VII Symp.-100 Years ISPRS. **38**, 298-303 (2010)

[2] Ruiz, L. A., Fdez-Sarría, A., Recio, J. A., "Texture feature extraction for classification of remote sensing data using wavelet decomposition: a comparative study,"20th ISPRS Congr., (2004)

[3] Xu, K., Yang, W., Chen, L. J., Sun, H., "Satellite image scene categorization based on topic models,"Geomatics and Inform. Sci. of Wuhan University, **36**(5), 540-543 (2011)

[4] Liénou, M., Maître, H., Datcu, M., "Semantic annotation of satellite images using latent Dirichlet allocation,"IEEE Geosci. Remote. Lett. **7**(1), 28-32 (2010)

[5] Lazebnik, S., Schmid, C., Ponce, J., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," IEEE Conf. **2**, 2169-2178 (2006)

[6] Tobler, W., "A computer movie simulating urban growth in the Detroit region,"Econ. Geography. **46**(2), 234-240 (1970)

[7] Lowe, D. G., "Distinctive image features from scale-invariant keypoints,"Int. J. Comput. Vision **60**(2), 91-110 (2004)

[8] Li, F. F., Perona, P., "A Bayesian hierarchical model for learning natural scene categories," IEEE Conf. **2**, 524-531 (2005)

[9] Grauman, K., Darrell, T., "The pyramid match kernel: Discriminative classification with sets of image features," 10th IEEE Int. Conf. **2**, 1458-1465 (2005)

[10] Swain, M. J., Ballard, D. H., "Color indexing," Int. J. Comput. Vision, **7**(1), 11-32 (1991)

[11] Chang, C. C., Lin, C.J., "LIBSVM: a library for support vector machines,"ACM Trans. Intell. Syst. Tech. **2**(3), 27 (2011)