A BENCHMARK FOR SCENE CLASSIFICATION OF HIGH SPATIAL RESOLUTION REMOTE SENSING IMAGERY

Jingwen Hu[1,2]*, Tianbi Jiang*[1]*, Xinyi Tong*[1]*, Gui-Song Xia*[1]*, Liangpei Zhang*[1]

[1] State Key Laboratory of LIESMARS, Wuhan University, Wuhan, 430079, China
[2] Electronic Information School, Wuhan University, Wuhan, 430079, China

ABSTRACT

Scene classification for high-resolution remotely sensed imagery have been widely investigated in recent years. However, there is few public, widely accepted and large scale dataset for benchmarking different methods. This paper presents a new and large dataset consisting of 5000 high-resolution remote sensing images which is manually labeled in 20 semantic classes for scene classification. Each class includes more than 200 image samples with different appearances. Some classic classification algorithms are compared on this dataset. To our knowledge, this work is the first time to give a public benchmark dataset at this size on the problem of scene classification in high-resolution remote sensing imagery, and give comparative results and analysis of various classic classification algorithms.

Index Terms— benchmark, dataset, scene classification, remote sensing imagery

1. INTRODUCTION

With the development of satellite sensors, the quality of remote sensing images increase rapidly, and large amounts of remote sensing images with high spatial resolution are available, which makes the in-depth study of land-use/land-cover possible. Scene classification of remote sensing images can cross the barrier between low-level visual features and high-level semantic information and be used to automatically label an image from a set of semantic categories. It is a significant task in the interpretation of high resolution remote sensing imagery, which can be widely applied to urban planning, environmental monitoring, etc.

In recent years, tremendous scene classification approaches have been proposed and reach higher and higher classification accuracy on remote sensing datasets, such as bag-of-visual-words (BOVW) [1, 2], topic models [3, 4, 5] and unsupervised feature learning approaches [6, 7]. However, one main difficulty of this task lies in the fact that there are few available datasets of remote sensing images for testing, evaluating and comparing different scene classification methods. Moreover, the few existing datasets of remote sensing images, e.g. [2, 8], are actually relatively small to test the scalability of scene classification algorithms. These points finally limit the development and application of scene classification methods. In order to solve this problem, we reorganize and expand the dataset of [8] from 950 images to 5000 ones and build a benchmark for comparing different methods. The new dataset consists 20 semantic classes and each class contains more than 200 sample images with varied appearances. With the proposed benchmark, the classification methods can be comprehensively evaluated and the results are more instructive for real aplications. In our paper, we adapt five classification methods for benchmarking and analysis.

In the rest of the paper, we first describe the new dataset and detail the benchmark protocols in Section 2, then show the results of different methods with the new benchmark in Section 3, and finally draw some conclusion remarks in Section 4.

2. DATASET AND BENCHMARK

2.1. Dataset Description

Good datasets can help to evaluate the performance of the algorithms more precisely, add challenges and promote the development of new methods. In this section, we organize a new and large dataset of high-resolution remote sensing imagery for benchmarking various existing scene classification methods, called WHU20 dataset in the rest of this paper. Fig. 1 shows two examples for each scene class of WHU20. The example images of WHU20 are obtained from screen capture of Google earth satellite imagery (without any watermarks). We collect images of different countries and regions around the world at different time and seasons with different level of resolutions. The resolution levels change from level 14 (7.44m) to the highest level 19 (0.26m), and the size of each image is fixed at 600×600 pixels to cover a scene with various resolutions. These sample images are labeled into 20 classes with different numbers in each class (see Table. 1) by the specialists in the field of remote sensing image interpretation. The dataset has a number of 5000 images in all.

This work was partially founded by NSFC project No.91338113.

Fig. 1. Samples of WHU20: two examples of each semantic scene class are shown. There are 20 classes and more than 200 samples per class. Observe the changes of scales and geometries in the samples of the same class.

Table 1. The different semantic scene classes and the number of images in each class of the WHU20 dataset.

Types	#images	Types	#images
airport	220	meadow	250
bare land	250	mountain	250
beach	250	park	205
bridge	270	parking	250
commercial	250	pond	255
desert	250	port	250
farmland	280	railway station	215
football field	260	residential	275
forest	255	river	250
industrial	250	viaduct	265

2.2. Dataset Properties

Compared with the existing remote sensing image datasets [2, 8], WHU20 has the following properties:

- *Large scale:* The number of images contained in WHU20 is the largest among all the existing remote sensing datasets, which means to cover a wider range of remote sensing images that may be a challenge to the scene classification task. Meanwhile, the number of sample images in each class are also larger than others which can help to evaluate the performance more precisely. For common-used datasets [2, 8], there are no more than 50 images in each class for testing, and the performance will be seriously affected by even one image is predicted correctly or not. Thus, our dataset can provide a benchmark to compare different existing methods and evaluate their performances more precisely, which can help us to focus on the key problems in scene classification algorithms.

- *Large intra-class geometrical diversity :* The variances in each class of the dataset can help to evaluate the robustness of scene classification methods. The resolution of the images in WHU20 varies from 0.26m to 7.44m which can help to test the scale invariance, and the scene images collected from different directions and positions can help to evaluate the rotational invariance as well as the translation invariance. Thus, our dataset add more challenges to the invariances of the classification methods.

- *Large intra-class radiometric diversity :* The images are obtained in different seasons and time accompanied with different weather and illumination conditions, which means that the scene images are taken under different kinds of imaging situations, such as the shadow of buildings, cloud cover, seasonal variation, illumination change, etc. These facts are also the challenges to scene classification task which need to be solved. Thus, our dataset can evaluate the robustness of the various algorithms under different imaging situations.

- *Variation of class size :* The existing datasets [2, 8] are both uniformly distributed with the same number of images per class, which makes us difficult to judge

whether the classification methods are biased to some specific classes for the Kappa coefficient is proportional to the overall accuracy (OA). However, WHU20 can make up for it with different number of images in each class. This enables one to compare the Kappa coefficients of different methods.

2.3. Tested Methods

The methods evaluated on our dataset are as follows :

BoVW: Bag-of-visual-words (BoVW) ignores the spatial information and represents the image by the frequencies of its visual words. It firstly describe the local image patches by a descriptor as visual words and clustering them to form the dictionary using k-means algorithm. Thus the global image feature is obtained by quantifying the frequency of the visual words in the dictionary using nearest neighbors, and support vector machine (SVM) [9] is used for classification.

SPM: Spatial pyramid matching (SPM) is different from BoVW in that it considers the spatial information of the images. It uses a sequence of increasingly coarser grids to build a spatial pyramid over the image plain to divide it into subregions and concatenate the weighted local BoVW features in each subregion at each level of resolution.

pLSA: Probabilistic latent semantic analysis (pLSA) is a topic model which introduce a latent variable called topic representing the conditional probability distribution of visual words on the dictionary. It connects the visual words to images and uses the distribution of topics to describe images in order to solve the influence of synonyms and polysomes. As a result,it depicts better similarity among images meanwhile reduce the feature dimension. It is used between BoVW feature extraction and SVM classification.

LDA: Latent Dirichlet allocation (LDA) is a generative probabilistic model evolved from pLSA with the main difference that it uses Dirichlet distribution to describe the latent variable topic so as to increase the robustness.

IFK: The improved Fisher kernel method combines the benefits of generative and discriminative approaches. It uses Gaussian mixture model (GMM) to construct a visual word dictionary and describe an image by Fisher vector encoding method using mean and covariance deviation vectors.

2.4. Evaluating Protocols

Parameter Settings : For fair comparisons, we use SIFT descriptor [10] to describe the local image patches for all the classification methods. The local image patches are fixed at 16×16 pixels and densely sampled by the spacing at 8 pixels. The size of the dictionary is set to be the optimal one which is different between methods, so as the topic number in pLSA and LDA. The libsvm classifier [9] is used to test the performance for all the methods with different kernels: for IFK, we adopt rbf kernel while the rest are HIK kernel [11], both are used with optimal parameters.

Evaluation Measures : To compare the classification quantitatively, we compute two common-used measures: overall accuracy (OA) and Kappa coefficient. OA is defined as the number of correct predicted images divided by the total number of predicted images, which is a direct measure to reveal the classification performance. Kappa coefficient is a more robust measure which takes into account the agreement occurring by chance. It considers the classification accuracy of each class to judge whether the classification methods are biased to some specific classes, which is defined as follows:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

where, $Pr(a)$ is the probability of observed agreement, and $Pr(e)$ is the probability of random agreement.

3. RESULTS AND ANALYSIS

Table. 2 illustrates the means and standard variances of OA using the five methods with randomly choosing different numbers of the training set for 100 times. Table. 3 shows the corresponding Kappa coefficients.

From the two tables, we can see the consistent trends between the two evaluation measures: the higher the OA, the higher the Kappa coefficient. We can also come to the conclusion that the IFK method has obviously better performances on this dataset: with the number of test images increasing from 50 to 100, and to 150, the average of OA increases from 80.03% to 84.91%, and to 87.08%, which are about 2% higher than the second highest OA, and the corresponding Kappa coefficient increases from 0.7889 to 0.8413, and to 0.8651, which are about 0.02 higher than the second highest Kappa coefficient. The other tested methods have similar results.

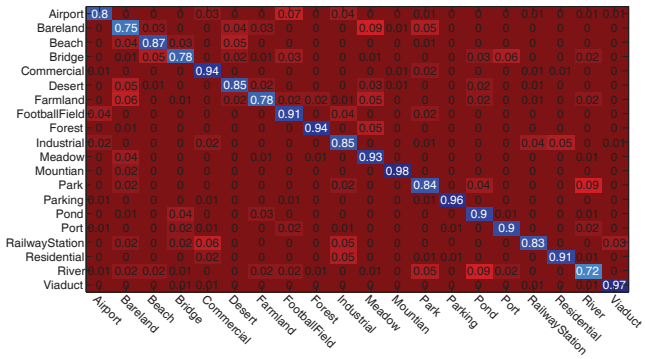

Fig. 2. Confusion mtrix of IFK with 150 testing images.

In order to further analyse the performance of IFK on our dataset, we have drawn the confusion matrix of the classification result using 150 testing images in Fig. 2. By further investigate the confusion matrix of four classes with lower classification accuracies, we can see that bare land is mainly mis-

Table 2. Overall accuracy (OA) of different methods compared in this paper.

Methods	50 test images	100 test images	150 test images
BoVW [2]	$77.26 \pm 0.68\%$	$82.39 \pm 0.54\%$	$84.52 \pm 0.69\%$
SPM [1]	$77.41 \pm 0.75\%$	$82.98 \pm 0.70\%$	$85.56 \pm 0.71\%$
pLSA [3]	$78.60 \pm 0.62\%$	$83.20 \pm 0.54\%$	$85.46 \pm 0.73\%$
LDA [4]	$77.59 \pm 0.85\%$	$82.44 \pm 0.60\%$	$84.86 \pm 0.68\%$
IFK [12]	$\mathbf{80.03 \pm 0.65\%}$	$\mathbf{84.91 \pm 0.61\%}$	$\mathbf{87.08 \pm 0.68\%}$

Table 3. Kappa coefficient of different methods compared in this paper.

Methods	50 test images	100 test images	150 test images
BoVW [2]	0.7605	0.8086	0.8439
SPM [1]	0.7623	0.8199	0.8535
pLSA [3]	0.7744	0.8231	0.8472
LDA [4]	0.7634	0.8157	0.8398
IFK [12]	**0.7889**	**0.8413**	**0.8651**

labeled into meadow; bridge is mainly mislabeled into port and beach; farmland is mainly mislabeled into bare land and meadow; river is mainly mislabeled into pond. The confusion between these classes can be explained by that these scene classes share similar textures such as bare land, farmland and meadow are fine grained textures, or have similar land-cover features such as river and pond mainly consists of water and grass. Thus, we should consider how to distinguish these confused classes in the future works, e.g. by taking the appropriate color information and spatial information into account.

4. CONCLUSION

To solve the current situation that the lack of public remote sensing datasets makes it difficult to benchmark different scene classification methods, this paper present a new public and large scale remote sensing dataset with manually labels for scene classification. Our benchmark dataset consists of 5000 images labeled with 20 classes. We describe the various properties of the new dataset, and shows the comparative results for some classic scene classification algorithms. From the comparative results, we can concludes that complementary local features and spatial information are important to scene classification, which are meaningful instructions to promote the development of classification methods on this field.

5. REFERENCES

[1] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 2169–2178.

[2] Yi Yang and Shawn Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Int. Symposium on Advances in Geographic Information Systems*, 2010, pp. 270–279.

[3] Thomas Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1/2, pp. 177–196, 2001.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," in *NIPS*, 2001, pp. 601–608.

[5] Retno Kusumaningrum, Hong Wei, Ruli Manurung, and Aniati Murni, "Integrated visual vocabulary in latent dirichlet allocation–based scene classification for ikonos image," *Journal of Applied Remote Sensing*, vol. 8, no. 1, pp. 083690–083690, 2014.

[6] Anil M Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 439–451, 2014.

[7] Fan Hu, Gui-Song Xia, Zifeng Wang, Liangpei Zhang, and Hong Sun, "Unsupervised feature coding on local patch manifold for satellite image scene classification," in *IGARSS*, 2014, pp. 1273–1276.

[8] Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, Henri Maître, et al., "Structural high-resolution satellite image indexing," in *ISPRS TC VII Symposium*, 2010, vol. 38, pp. 298–303.

[9] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27, 2011.

[10] David G Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[11] S. Maji, A.C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *CVPR*, 2008, pp. 1–8.

[12] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010, pp. 143–156.