A Hierarchical Scheme of Multiple Feature Fusion for High-Resolution Satellite Scene Categorization

Wen Shao[1], Wen Yang[1,2], Gui-Song Xia[2], and Gang Liu[1]

[1] School of Electronic Information, Wuhan University, Wuhan, 430072, China
[2] Key State Laboratory LIESMARS, Wuhan University, Wuhan, 430079, China
{shaowen1989,yangwen94111,gsxia.lhi,liugang.spl}@gmail.com

Abstract. Scene categorization in high-resolution satellite images has attracted much attention in recent years. However, high intra-class variations, illuminations and occlusions make the task very challenging. In this paper, we propose a classification model based on a hierarchical fusion of multiple features. Highlights of our work are threefold: (1) we use four discriminative image features; (2) we employ support vector machine with histogram intersection kernel (HIK-SVM) and L1-regularization logistic regression classifier (L1R-LRC) in different classification stages, respectively. The soft probabilities of different features obtained by the HIK-SVM are discriminatively fused and fed into the L1R-LRC to obtain the final results; (3) we conduct an extensive evaluation of different configurations, including different feature fusion schemes and different kernel functions. Experimental analysis show that our method leads to state-of-the-art classification performance on the satellite scenes.

Keywords: Scene Categorization, Hierarchical Fusion, Histogram Intersection Kernel, Logistic Regression.

1 Introduction

Scene categorization is a challenging problem in remote sensing image interpretation. It is difficult due to the high intra-class variability and low inter-class disparity in remote-sensing images. Other factors, such as changes of viewpoint, illuminations and shadows, partial occlusions, background clutter and multiple instances, further complicate these problems. Simultaneously, with the substantial increase of the resolution of images, details of the targets become more clear, and a multitude of cues also become more distinctive, such as structure, shape, texture and color. Thus, it is essential to design highly discriminative image features and to reasonably combine them based on different aspects. In the past few years, tremendous efforts have been made to develop advanced image features and classification techniques for boosting the classification accuracy[1,2].

Designing comprehensive and complementary image features is one of recent trends in image classification domain. Unlike the case of low-resolution satellite images, where texture and intensity cues have been proved to be effective and

M. Chen, B. Leibe, and B. Neumann (Eds.): ICVS 2013, LNCS 7963, pp. 324–333, 2013.

efficient enough for recognition [3], structure, shape and color information also play important roles in analyzing high-resolution satellite images. For example, Xia *et al.* [4] have confirmed the applicability of their shape-based image indexing scheme to satellite scene categorization, named tree of shapes.

Another trend is how to properly combine different features. Many approaches have been reported in the literature. A prominent kernel level fusion instance is the Multiple Kernel Learning (MKL). MKL linearly combines similarity functions between images, which yields good results on the application of object classification. However, although the MKL solution is sparse for every class separately, it is not sparse jointly in the multi-class setup. Based on the observation that higher computation efficiency and classification accuracy can be achieved by simple feature combination strategies than that by MKL, Gehler and Nowozin [5] concluded that the performance of MKL has been overestimated in the past. Recently, a frequently used approach is score level fusion, where scores from different feature channels are combined. Sheng *et al.* [6] designed a high-resolution satellite scene classification using a sparse coding based multiple feature fusion, simplified as SCMF. SCMF set the final fused result as the concatenation of the probabilities obtained by the individual feature channel, the approach turned out to work surprisingly well with the linear SVM classifier, while dramatically reducing the computational complexity. Furthermore, many breakthroughs have been made by discriminative feature fusion. Fernando *et al.* [7] presented a new logistic regression-based fusion method, called LRFF. LRFF first created a visual dictionary for each feature, and then used a Logistic Regression (LR) method to deduce the most class-specific discriminative visual words from the multiple dictionaries, finally applied the LR outputs to design an efficient marginalized kernel for the purpose of learning a new SVM classifier. Experimental results have demonstrated the effectiveness of their approach.

Inspired by the above two trends, this paper describes a robust two-level classification model by discriminatively fusing multiple features. The remainder of this paper is organized as follows. First, multiple feature extraction procedure is presented in Section 2. Then, Section 3 gives a detailed description of our hierarchical feature fusion method. Further, Section 4 shows the experimental results and performance evaluation. Finally, Section 5 ends up with conclusions and future work**.**

2 The Hierarchical Categorization Framework

In this section, we build a hierarchical categorization framework by discriminatively fusing multiple features, shown in Figure 1. Observe that our method is also flexible to other features and classifiers.

2.1 Multi-feature Extraction

Since a high-resolution satellite image usually consists of several kinds of information cues, capturing these cues is very helpful in recognizing and distinguishing categories. Thus, we exploit four channels of features to characterize the images. For structural cues, we extract SIFT gradient orientation histograms within the support region and

quantize local feature descriptors using a bag of features (BoF) model [8]. For shape cues, we incorporate color information into the shape-based invariant image indexing framework of Xia *et al.* [4], and the resulting output is the concatenation of R, G and B three channeled edge information, termed tree of colored shapes (tree of c-shapes). For textural cues, a supervised three-layered model is adopted, and the discriminative CLBP feature (disCLBP) [9] is obtained by concatenating pattern occurrence histograms of sign and magnitude operators. For color cues, we use the simple yet effective bag-of-colors [10] signature as an additional global color descriptor, which brings in the idea of a Lab-color-palette.

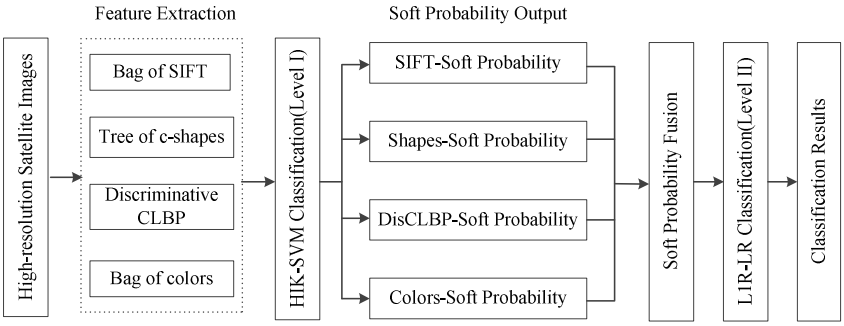

Fig. 1. The two-level categorization framework

2.2 Level-I Classification by HIK-SVM

Support vector machine (SVM) with histogram intersection kernel (HIK-SVM) [11] remains an extremely popular choice for histogram feature classification. Moreover, the HIK SVM is able to provide probabilistic output for the test image. After feature extraction, we therefore serve HIK-SVM as the base classifier for Level-I classification.

Let $\mathbf{x}_I^k$ and $\mathbf{x}_J^k$ denote the k^{th} ($k = 1,2,3,4$) feature vector extracted in image I and J, respectively. The similarity between $\mathbf{x}_I^k$ and $\mathbf{x}_J^k$ can be defined by the histogram intersection kernel K_{int} as

$$K_{\text{int}}(\mathbf{x}_I^k, \mathbf{x}_J^k) = \sum_{i=1}^{M} \min\{\mathbf{x}_I^k(i), \mathbf{x}_J^k(i)\} \tag{1}$$

where M is the dimension of the feature vectors. The histogram intersection can be used as a similarity measure for histogram-based representations of images. Integrating histogram intersection kernel into a SVM framework, a classification can be obtained by:

$$h(\mathbf{x}^k) = \sum_{l=1}^{m} \alpha_l^k y_l K(\mathbf{x}^k, \mathbf{x}_l^k) + b^k \tag{2}$$

where $K(\mathbf{x}^k, \mathbf{x}_l^k)$ represents the value of a kernel function for the l^{th} training image $\mathbf{x}_l^k$ and the test image $\mathbf{x}^k$, α_l^k and y_l are weight and class label c $(c = 1, 2, ..., C)$ of $\mathbf{x}_l^k$, and b^k is the learned threshold.

Multi-class categorization is implemented using a set of binary classifiers and taking the majority vote. In terms of four feature channels of the image, we obtain four kinds of soft posteriori probabilities $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4 \in \mathrm{R}^C$ in parallel using HIK-SVM, their lengths are equivalent to the number of categories *C* within a dataset.

2.3 Level-II Classification by L1R-LR and Other Kernel Classifiers

Several data fusion methods have been reported in the literature [5]. In what follows, four representative fusion strategies are employed to combine the intermediate soft probabilities,

$$\mathbf{Y}_1 = \max\left(\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4\right) \in \mathrm{R}^C \tag{3}$$

$$\mathbf{Y}_2 = \mathrm{sum}\left(\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4\right) \in \mathrm{R}^C \tag{4}$$

$$\mathbf{Y}_3 = \mathrm{cat}\left(\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4\right) \in \mathrm{R}^{4C} \tag{5}$$

$$\mathbf{Y}_4 = \mathrm{multiply}\left(\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4\right) \in \mathrm{R}^C \tag{6}$$

In order to keep the consistency of data, most of the multi-stage classification tasks use the same classifier in each stage [6]. However, different classifiers may be more capable of helping each other by focusing on different aspects of the data, and not making the same mistakes. Here, we use an L1-regularization logistic regression (L1R-LR) model [7] to train the Level-II classifier, by taking $\mathbf{Y}_t$ $(t = 1, 2, 3, 4)$ as the input feature vectors.

Suppose an image *I* can be described as a normalized feature vector $\mathbf{x}$, multivariate logistic regression models the probability that the image label *y* belongs to class *c* as follows:

$$P\left(y = c \mid \mathbf{x}; \mathbf{b}, \mathbf{W}\right) = \frac{\exp\left(\mathbf{b}_c + \mathbf{w}_c^T \mathbf{x}\right)}{\boldsymbol{N}} \tag{7}$$

where $\mathbf{b}$ is a class-bias vector**,** $\mathbf{W}$ is a weight vector-matrix with columns $\mathbf{w}_n$, and $\boldsymbol{N} = \boldsymbol{N}\left(\mathbf{x}; \mathbf{b}, \mathbf{W}\right) = \sum_c \exp(\mathbf{b}_c + \mathbf{w}_c^T \mathbf{x})$ is a probability normalization term. L1-regularized logistic regression-based training [12] actually solves following unconstrained optimization problem,

$$\arg\min \sum_c \| \mathbf{w}_c \|_1 - A \sum_i \log\left(\frac{\exp(\mathbf{b}_{c_i} + \mathbf{w}_{c_i}^T \mathbf{x}_i)}{\boldsymbol{N}_i}\right) \tag{8}$$

where $\| \cdot \|_1$ denotes the L1-regularization, *A* is a regularization parameter and *i* runs over the training samples with features $\mathbf{x}_i$, labels c_i, and normalizations $\boldsymbol{N}_i$.

Unlike the L2-regularization that only restricts large values, the L1-regularization term penalizes all factors equally, which can create sparse answers.

For comparisons, we further consider other kernel functions, such as HIK, RBF kernel, χ^2 kernel, and especially logistic regression marginalized kernel (LRMK) [7], which is a new marginalized kernel designed by making use of the output weight and condition probability of logistic regression model. More details are discussed in the subsequent experiments.

3 Experimental Evaluation and Analysis

In the experiments, we report the experimental results on two datasets: a 19-class satellite scene[1], and a 21-class land-use dataset[2].

3.1 19-Class Satellite Scene

Data Description. Our first dataset is composed of 19 classes of scenes, including airport, beach, bridge, commercial area, desert, farmland, football field, forest, industrial area, meadow, mountain, park, parking, pond, port, railway station, residential area, river and viaduct. Each class has 50 images, with the size of 600×600 pixels. To make the results as robust and quantitative as possible, we randomly divided the dataset into five equal sets, three of which were chosen for training and the remainder for test the same as [6]. The classification accuracy is the mean and standard deviation over five evaluations.

Implementation Details and Results. To obtain the optimal parameter configuration, SIFT descriptors from 16×16 pixel patches were densely extracted from each image on a grid with a spacing of 8 pixels. During BoF processing, a visual dictionary with a size of 1024 was used. In addition, codebook sizes of 1400 and 512 were set separately for the tree of c-shapes and bag-of-colors methods to achieve satisfactory classification accuracies. Since different training sets produced different global dominant pattern sets, the dimension of disCLBP feature over five runs were unfixed (4135, 3955, 4144, 4173, 4243) but converged to a constant 4000. Table 1 gives the dimensions and classification rates of four features independently using HIK-SVM. We can find that bag of SIFT accounts for the largest importance and bag of colors the lowest on the 19-class satellite scene. An empirical combination of feature concatenation was also realized to classification, the accuracy increases at the expense of the reduced computational speed due to the high-dimensional vector.

In order to further improve the performance, four data fusion strategies were proposed to combine the intermediate results from four feature channels, five kernels were also considered to perform classification tasks. Table 2 shows the classification rates of different fusion methods using different kernels. Experimental results show

[1] http://dsp.whu.edu.cn/cn/staff/yw/HRSscene.html
[2] http://vision.ucmerced.edu/datasets

that all kernels under maximum, sum and concatenation fusion can dramatically improve the classification accuracy as opposed to the above single feature channels and simple feature concatenation, and their performances almost keep pace with each other. However, L1R-LRC shows a significant advantage over other kernels under multiplication fusion and explicitly performs the best in all experiments. The superiority can be attributed to at least two reasons. On the one hand, the four features we chose are relatively independent, which exactly meets the requirement of multiplication fusion. On the other hand, the choice of L1R-LRC is important since L1-regularization provides for robustness in the case that no categories matching all feature attributes of the image. Furthermore, multiplying probabilities from all feature channels would tend to favor the category where all features are somewhat likely, but none is particularly low.

Table 1. Performance comparison with different features using HIK-SVM on 19-class

Feature	Dimension	Accuracy (%)
Bag of SIFT	1024	**85.52 ± 1.23**
Tree of c-shapes	1400	80.42 ± 1.80
DisCLBP	~ 4000	80.42 ± 1.37
Bag of colors	512	70.63 ± 1.44
concatenation	~ 6936	90.79 ± 0.65

Table 2. Performance comparison of different fusions using different kernels on 19-class

Fusion method	Maximum	Sum	Concatenation	Multiplication
L1R-LR Accuracy (%)	90.11 ± 1.05	93.21 ± 1.02	92.37 ± 0.51	**94.53 ± 1.01**
LRMK Accuracy (%)	90.16 ± 0.76	93.23 ± 0.74	92.68 ± 1.08	53.46 ± 0.68
HIK Accuracy (%)	90.21 ± 1.02	93.26 ± 1.06	93.16 ± 0.97	58.74 ± 0.53
RBF Accuracy (%)	**90.21 ± 0.63**	93.42 ± 0.43	**93.26 ± 0.82**	50.36 ± 0.75
χ^2 Accuracy (%)	89.82 ± 0.33	**93.59 ± 0.96**	93.15 ± 0.70	60.18 ± 0.59

Table 3. Classification results using L1R-LRC under multiplication fusion with two features

Two features	Accuracy (%)
Bag of SIFT, Tree of c-shapes	89.05 ± 0.97
Bag of SIFT, DisCLBP	90.16 ± 0.71
Bag of SIFT, Bag of colors	**90.37 ± 0.72**
Tree of c-shapes, DisCLBP	87.89 ± 0.78
Tree of c-shapes, Bag of colors	88.58 ± 0.35
DisCLBP, Bag of colors	88.46 ± 0.46

Table 4. Classification results using L1R-LRC under multiplication fusion with three features

Three features	Accuracy (%)
Bag of SIFT, Tree of c-shapes, DisCLBP	91.89 ± 0.51
Bag of SIFT, Tree of c-shapes, Bag of colors	**93.84 ± 0.47**
Bag of SIFT, DisCLBP ,Bag of colors	93.79 ± 0.60
Tree of c-shapes, DisCLBP, Bag of colors	92.68 ± 0.79

Meanwhile, we have redone the two-level classification experiments using L1R-LRC under multiplication fusion with two and three features, and the corresponding classification results are shown in Table 3 and Table 4. It is clear that the overall performance with two features is worse than three features, and the performance with three features is inferior to four features. The results show the benefits and complementarities of the four features we chose.

For comparison, we also give results for two competitive methods in Table 5. In the "sparse coding-based multiple feature combination (SCMF)" method [6], *SIFT, CH, LTP-HF* were processed with the same dictionary size 512. In the "logistic regression-based feature fusion (LRFF)" method [7], the descriptors *SIFT+Hue+CN+Opp.SIFT* were with sizes of 1024+300+300+2000. The comparative results demonstrate the effectiveness of our hierarchical multiple feature fusion method.

Table 5. Performance comparison with state-of -the-art methods on 19-class

Method	SCMF	LRFF	Ours
Accuracy (%)	92.75 ± 0.64	91.26 ± 0.47	**94.53 ± 1.01**

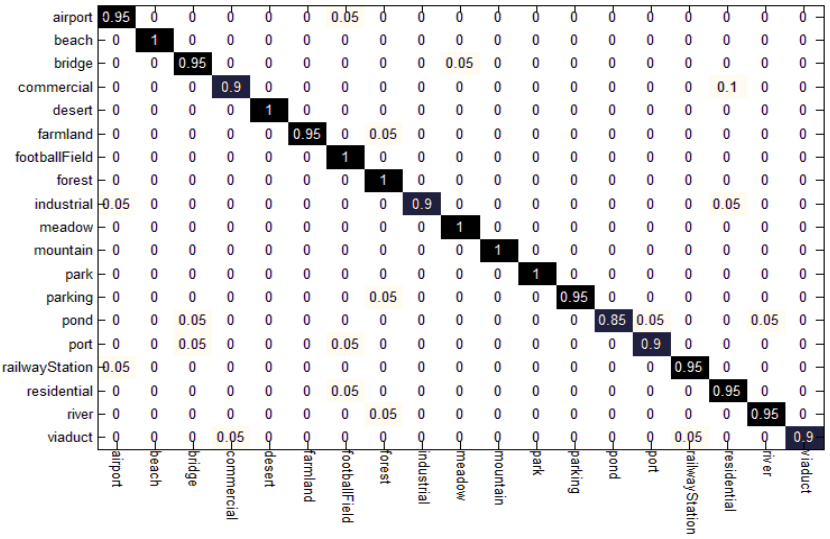

Fig. 2. Confusion matrix of the 19-class satellite scene using our classification method

In particular, using L1R-LRC under multiplication fusion, our method can even achieve a overwhelming accuracy of 95.26%, which exceeds the highest accuracy 93.62% previously obtained in [6]. An overview of the best performance from one run of our approach for all 19 categories is given by the confusion matrix presented in Figure 2. Performance is measured as the average classification accuracy per class. Totally 7 classes of satellite scenes achieve 100% classification accuracy using our fusion method. The visually complex classes appear to be the main source of confusion. Especially some scenes that should belong to commercial area are misclassified into residential area, perhaps because commercial area images often

contain patterns such as dense houses, horizontal and vertical lines which are also characteristics of residential area images. In addition, it is not surprising that pond areas filled with waters are easily classified into ports and rivers.

3.2 21-Class Land-Use Dataset

Data Description. To provide more comprehensive analysis, the 21-class land-use dataset was additionally introduced, and it has been quantitatively evaluated in the literature [8]. The dataset consists of 21 classes of images which was manually extracted from aerial orthoimagery with a pixel resolution of one foot, including 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 separately contains 100 images with 256×256 pixels. To make a quantitative comparison with representative methods [6,7], we chose 80 samples of each class for training and 20 for test. This setup was selected based on the comparative evaluation of [8], which just applied one feature and the best classification accuracy was only 76.81%. At the same time, the experiments were also repeated over five runs.

Implementation Details and Results. Similar to the case of 19-class satellite scene, the parameter configuration was also selected by a linear search process. Table 6 shows the performance comparison with different type of features indendently using HIK-SVM across all 21 classes of land-use scene. We can observe that all features work very well. Thus, there is a need to reasonably combine these four features to further improve the classification performance, certainly feature concatenation is a general alternative. As expected, using soft probabilities from four feature channels together, our proposed data fusion strategies drastically improve the performance, which is in consistence with the results of the 19-class satellite scene. Under maximum, sum and concatenation, the performance of L1R-LRC is almost on a par with other kernels, while far more surpasses others under multiplication fusion. From Table 7 we can draw up a same conclusion that our level-II implementation with L1R-LRC under multiplication fusion satisfies the optimal configuration.

In addition, we also compared our approach with SCMF (*SIFT+CH+LTP-HF* with sizes of 512+512+512) [6] and LRFF (*SIFT+Hue+CN+Opp.SIFT* with sizes of 512+300+300+1000) [7] on the 21-class dataset. Table 8 shows the overall results of these representative methods.

Table 6. Performance comparison with different features using HIK-SVM on 21-class

Feature	Dimension	Accuracy (%)
Bag of SIFT	512	83.33 ± 1.64
Tree of c-shapes	1400	**83.52 ± 0.94**
DisCLBP	~ 2000	82.52 ± 1.75
Bag of colors	512	83.46 ± 1.57
concatenation	~ 4424	89.48 ± 0.81

Table 7. Performance comparison of different fusions using different kernels on 21-class

Fusion method	Maximum	Sum	Concatenation	Multiplication
L1R-LR Accuracy (%)	89.43 ± 0.98	91.38 ± 0.69	91.19 ± 1.24	**92.38 ± 0.62**
LRMK Accuracy (%)	**89.68 ± 1.09**	91.48 ± 0.56	91.14 ± 0.86	73.33 ± 0.84
HIK Accuracy (%)	89.57 ± 0.83	**91.62 ± 0.83**	91.57 ± 1.06	77.14 ± 0.58
RBF Accuracy (%)	89.57 ± 0.83	91.52 ± 1.03	91.52 ± 0.61	71.67 ± 0.31
χ^2 Accuracy (%)	88.86 ± 0.37	91.33 ± 0.52	**91.57 ± 0.30**	80.81 ± 0.56

Table 8. Performance comparison with state-of -the-art methods on 21-class

Method	SCMF	LRFF	Ours
Accuracy (%)	91.03 ± 0.48	90.76 ± 0.59	**92.38 ± 0.62**

Compared to the results reported in the literature [6,7], our proposed method is capable of gaining the best performance of 92.62%. To get a convincing completion, confusion matrice from one run of our method demonstrating the highest classification rate on the 21-class dataset is shown in Figure 3. On the whole, 9 land-use classes are absolutely recognized by our proposed method. Of course, there are also a few explainable confusions between some classes. It is obvious that there exists certain resemblance in the structure and texture between baseball diamond and storage tanks, and many buildings are just next to the airport. In particular, the most difficult categories are dense residential and medium density residential. This can be explained by the fact that both categories share similar image components such as trees, buildings and roads. These are partial factors for misclassifications.

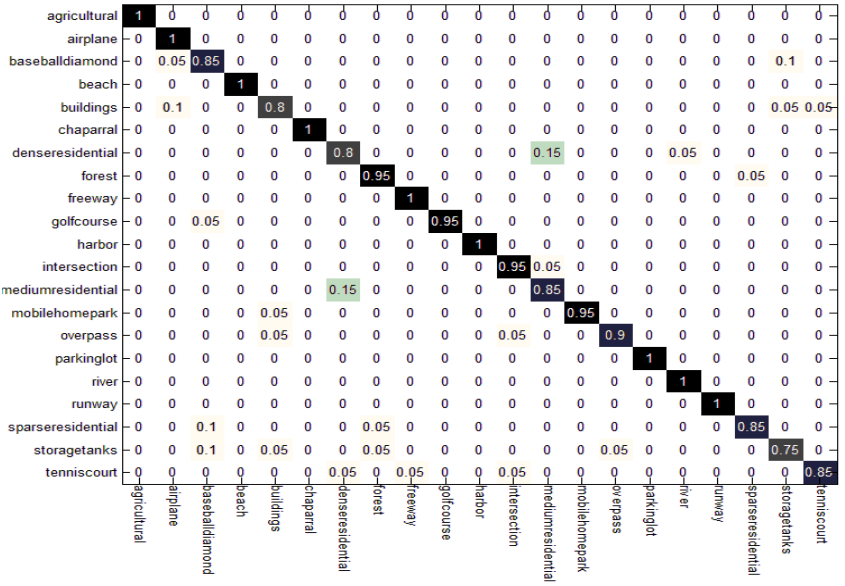

Fig. 3. Confusion matrix of the 21-class land-use dataset using our classification method

4 Conclusions and Future Work

This paper has presented a hierarchical multiple feature fusion approach for high-resolution satellite scene categorization. More specifically, we made three main contributions. First, we selectively extracted four complementary image descriptors. Further, we used HIK-SVM and L1R-LRC in different classification stages. Above all, the posteriori information we applied in Level-II classification was the product of four probabilities from Level-I outputs. Based on the presented experimental results and analysis, we can conclude that the proposed two-level classification model can yield state-of-the-art results. Furthermore, our classification model is also flexible to other extensions, such as new features, classifiers, and fusion methods. In future work, we will also extend our work to large-scale satellite scene categorization.

Acknowledgments. The research was supported in part by the Chinese National Natural Sciences Foundation grants (NSFC) 61271401 and China Postdoctoral Science Foundation (CPSF) 20110491187.

References

1. Dai, D.-X., Yang, W.: Satellite Image Classification via Two-layer Sparse Coding with Biased Image. IEEE Geoscience and Remote Sensing Letters 8, 173–176 (2011)
2. Amarsaikhan, D., Douglas, T.: Data Fusion and Multisource Image Classification. International Journal of Remote Sensing 25, 3529–3539 (2004)
3. Li, C.-S., Castelli, V.: Deriving Texture Feature Set for Content-Based Retrieval of Satellite Image Database. In: Proceedings of International Conference on Image Processing, vol. 1, pp. 576–579 (1997)
4. Xia, G.-S., Yang, W., Delon, J., Gousseau, Y., Sun, H., Maitre, H.: Structural High-resolution Satellite Image Indexing. In: Proceedings of ISPRS, pp. 298–303 (2010)
5. Gehler, P., Nowozin, S.: On Feature Combination Methods for Multiclass Object Classification. In: IEEE International Conference on Computer Vision, pp. 221–228 (2009)
6. Sheng, G.-F., Yang, W., Xu, T., Sun, H.: High-resolution Satellite Scene Classification Using Sparse Coding Based Multiple Features Combination. International Journal of Remote Sensing 33, 2395–2412 (2012)
7. Fernando, B., Fromont, E., Muselet, D., Sebban, M.: Discriminative Feature Fusion for Image Classification. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2012) (2012)
8. Yang, Y., Newsam, S.: Bag-of-visual-words and Spatial Extensions for Land-use Classification. In: ACM SIGSPATIAL GIS, California, America (2010)
9. Guo, Y.-M., Zhao, G.-Y., Pietikäinen, M.: Discriminative Features for Texture Description. Pattern Recognition 45(10), 3834–3843 (2012)
10. Wengert, C., Douze, M., Jégou, H.: Bag-of-colors for Improved Image Search. In: Association for Computing Machinery, Multimedia (2011)
11. Maji, S., Berg, A.C., Malik, J.: Classification Using Intersection Kernel Support Vector Machines is Efficient. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2008)
12. Lee, S.I., Lee, H., Abbeel, P., Andrew, Y.N.: Efficient L1 Regularized Logistic Regression. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006), Boston, MA, USA (2006)