Journal of Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Land use and land cover classification for rural residential areas in China using soft-probability cascading of multifeatures

Bin Zhang
Yueyan Liu
Zuyu Zhang
Yonglin Shen

SPIE.

Bin Zhang, Yueyan Liu, Zuyu Zhang, Yonglin Shen, "Land use and land cover classification for rural residential areas in China using soft-probability cascading of multifeatures," *J. Appl. Remote Sens.* **11**(4), 045010 (2017), doi: 10.1117/1.JRS.11.045010.

Land use and land cover classification for rural residential areas in China using soft-probability cascading of multifeatures

Bin Zhang,[a,*] Yueyan Liu,[a] Zuyu Zhang,[a] and Yonglin Shen[b]
[a]China University of Geosciences, Land Resource Management Department, Wuhan, China
[b]China University of Geosciences, School of Information Engineering, Wuhan, China

Abstract. A multifeature soft-probability cascading scheme to solve the problem of land use and land cover (LULC) classification using high-spatial-resolution images to map rural residential areas in China is proposed. The proposed method is used to build midlevel LULC features. Local features are frequently considered as low-level feature descriptors in a midlevel feature learning method. However, spectral and textural features, which are very effective low-level features, are neglected. The acquisition of the dictionary of sparse coding is unsupervised, and this phenomenon reduces the discriminative power of the midlevel feature. Thus, we propose to learn supervised features based on sparse coding, a support vector machine (SVM) classifier, and a conditional random field (CRF) model to utilize the different effective low-level features and improve the discriminability of midlevel feature descriptors. First, three kinds of typical low-level features, namely, dense scale-invariant feature transform, gray-level co-occurrence matrix, and spectral features, are extracted separately. Second, combined with sparse coding and the SVM classifier, the probabilities of the different LULC classes are inferred to build supervised feature descriptors. Finally, the CRF model, which consists of two parts: unary potential and pairwise potential, is employed to construct an LULC classification map. Experimental results show that the proposed classification scheme can achieve impressive performance when the total accuracy reached about 87%. [DOI: 10.1117/1.JRS.11.045010]

Keywords: high-spatial-resolution remote sensing; image classification; midlevel feature learning; conditional random field.

Paper 170296 received Apr. 17, 2017; accepted for publication Nov. 10, 2017; published online Dec. 1, 2017.

1 Introduction

Image classification is crucial in the interpretation of remote sensing images with high spatial resolution (HSR).[1] The availability of HSR remote sensing imagery obtained from satellites (e.g., WorldView-2, IKONOS, QuickBird, ZY-3C, GF-1, and GF-2) increases the possibility of accurate Earth observations. Such HSR imagery provides highly valuable geometric and detailed information, which is important for various applications, such as precision agriculture, security applications, and damage assessment for environmental disasters and land use.[2] In these applications, mapping a high-resolution image for land use and land cover (LULC) is particularly relevant.

In terms of LULC classification using remote sensing images, Landsat series satellite imagery with medium resolution is important in regional LULC and land use/cover change studies.[3–7] In processing high-resolution remote sensing images, numerous classification algorithms, such as the object-oriented approach,[8–10] based on the classification of a support vector machine (SVM)[11–13] and Markov random fields (MRF)[14–18] are being developed.

Local features[19–23] have been successfully applied to image retrieval, semantic segmentation, and scene understanding. These features gained popularity in the remote sensing community

*Address all correspondence to: Bin Zhang, E-mail: zhangbin@cug.edu.cn

because of their robustness in rotation, scale changes, and occlusion. Sparse coding is one of the most effective approaches to group local features and performs well in object categorization, scene-level land use classification, etc.[24–36] The sparse coding method combined with max-pooling and spatial pyramid matching (SPM) can be used to learn midlevel features. In this approach, a class type is represented by the distribution of a set of visual words, which are usually obtained by unsupervised K-means clustering of a set of low-level feature descriptors. However, visual words are learned in an unsupervised manner, resulting in less discriminative midlevel features. This characteristic reduces the accuracy of classification. Several conventional low-level features, such as spectral features, are neglected in the building of midlevel features. Some studies have resolved this drawback and effectively incorporated spectral and local features.[33,34] Hu et al.[37] developed a method that combines convolutional neural networks (CNN) and sparse coding to learn discriminative features for scene-level land use classification, and impressive results were obtained when the total accuracy reached about 96%. However, this method is limited by the lack of information of the LULC class type, because the parameters of a CNN model are estimated by the ImageNet dataset.[38]

In addition to feature learning, the selection of a classifier is particularly important for LULC classification based on high-resolution remote sensing images. Many classification methods, such as maximum likelihood, MRF, and SVM models, have been developed. The SVM classifier is widely used for various computer vision tasks and LULC classification, because this model has shown advantages on high-dimension feature space. MRF[39] and conditional random field (CRF)[40] are structured output models that consider interactions of random variables. These approaches have been successfully developed in remote sensing[14–17,41] and computer vision communities.[42–49] Moser et al.[14] proposed an LULC classification for high-resolution remote sensing images based on the MRF model. However, the results of this model always exhibit an oversmoothed appearance.[9,48] Another drawback of the MRF is its difficulty in processing high-dimension feature space. The CRF model overcomes these drawbacks and shows advantages on image classification and semantic segmentation.

Thus, we establish an LULC classification framework for HSR remote sensing images by exploiting labeled data based on midlevel feature learning and the SVM classifier to achieve multifeature soft-probability feature descriptors, and we employ a CRF classification method to jointly model the unary and pairwise costs.

In this paper, a multifeature soft-probability cascading and CRF (MFSC-CRF) classification model is designed to learn discriminative midlevel features in a supervised manner. First, we extracted the spectral, gray-level co-occurrence matrix (GLCM), and dense scale-invariant feature transform (DSIFT) features as low-level feature descriptors. Three types of midlevel feature descriptors are achieved by adopting sparse coding, superpixel segmentation, and max-pooling methods. Then, the probability that some labeled samples belong to LULC classes can be calculated. The three probability values are cascaded to construct the feature descriptors for each superpixel. Finally, the CRF model is introduced to generate the LULC classification.

The supervised learned feature descriptors can be obtained using the SVM classifier with training samples. This classifier has been demonstrated to effectively incorporate low-level features. Using the CRF classifier, the local spatial relationship between the neighboring superpixels is considered by combining the learned feature descriptor. Thus, the proposed method achieves better classification results than traditional methods.

The rest of this paper is structured as follows. In Sec. 2, the proposed method for midlevel feature learning and soft-probability cascading and CRF classification is presented. In Sec. 3, the experiments on the rural residential area dataset of Wuhan are discussed. Conclusions are drawn in Sec. 4.

2 MFSC-CRF Classification Framework

An HSR remote image classification framework for LULC classification is proposed. This method is based on midlevel feature learning by integrating sparse coding and the CRF method to utilize spectral, structural, and spatial contextual information. Three kinds of typical features, namely, GLCM, DSIFT, and spectral features, are selected to construct the low-level features.

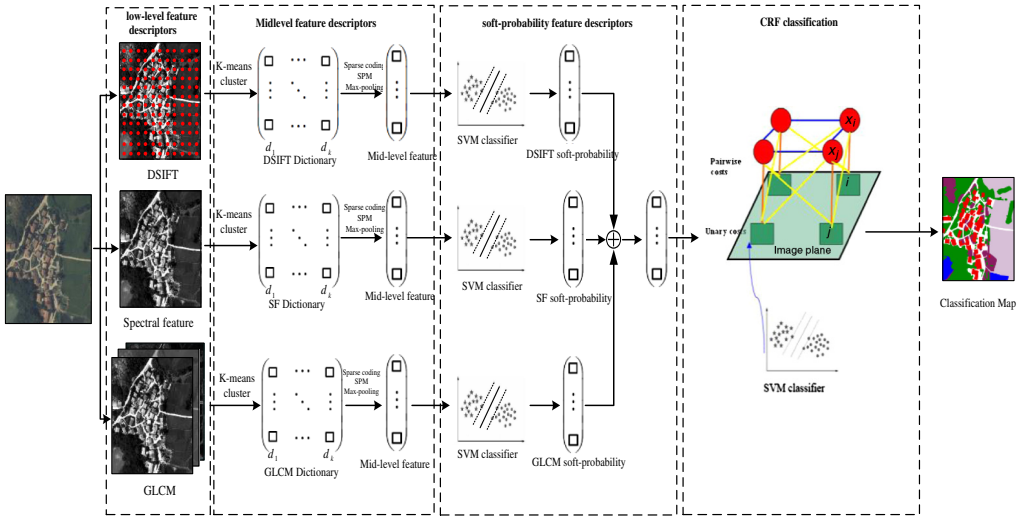


Fig. 1 Flowchart of the MFSC-CRF classification framework.

The whole pipeline of the MFSC-CRF classification framework consists of two main steps, namely, feature learning and CRF classification (Fig. 1).

Midlevel feature descriptors are achieved during the feature learning step using the three features by combining sparse coding, SPM, and max-pooling method. The probability can be calculated by the SVM classifier using the training samples. The resulting probability values form the new discriminative feature descriptors.

During CRF classification, the CRF model is introduced to classify the superpixels according to the land cover class types. The probability feature descriptor from the first step is considered in this step, and an SVM classifier is adopted to construct the unary potentials. The pairwise potentials can be acquired by calculating the distance between neighboring superpixels. The graph-cut-based α-expansion algorithm is executed to obtain the classification result of the CRF models.

2.1 *Midlevel Feature Descriptors*

As discussed above, three typical features are adopted for the low-level feature descriptors, and the details are described as follows.

1. Spectral features: Features on the Earth reflect, absorb, transmit, and emit electromagnetic energy from the sun. A measurement of energy commonly used in remote sensing of the Earth is reflected energy (e.g., visible light, near-infrared, etc.) coming from land and water surfaces. The amount of energy reflected from these surfaces is usually expressed as a percentage of the amount of energy striking the objects. The band values of remote sensing images are used as the spectral features in this article.
2. GLCM: GLCM is a texture measurement to many image analyses. In this article, GLCM is extracted by ENVI software. Eight features are achieved, which are called as mean, variance, homogeneity, contrast, dissimilarity, entropy, etc. They are normalized to form feature vectors.
3. DSIFT: DSIFT descriptors are computed at points on a regular grid. At each grid point, the descriptors are computed over four circular support patches with different radii, and, consequently, each point is represented by four SIFT descriptors. Multiple descriptors are computed to allow for scale variation between images.[50]

The low-level feature descriptors are extracted from images, and each feature descriptor has size T. The visual dictionary $\mathbf{D}$ of K visual words obtained by unsupervised K-means clustering algorithm can be defined as follows:

$$\mathbf{D} = [d_1, d_2, \cdots, d_k] \in R^{T \times K}, \tag{1}$$

where each d_k is represented as a linear classifier with bias and calculated as follows:

$$d_k = [\mathbf{D}_{k,1}, \mathbf{D}_{k,2}, \cdots, \mathbf{D}_{k,T}]^T \in R^T. \tag{2}$$

An encoding scheme based on the classification score obtained by each dictionary word is used, instead of sparse coding to encode each descriptor. This step is suggested in Ref. 23. If v is a descriptor vector, its coding vector $\mathbf{f}_{d_k}(\alpha^{i_l})$ corresponding to dictionary $\mathbf{D}$ is given as follows:

$$\mathbf{f}_{d_k}(\alpha^{i_l}) = [\langle \alpha_k^1 . d_k \rangle, \cdots, \langle \alpha_k^N . d_k \rangle] \in R^N. \tag{3}$$

Intuitively, the descriptor α should be similar only to a few words in the dictionary if the visual words of dictionary $\mathbf{D}$ are sufficiently discriminative. Therefore, the vector $\mathbf{f}_{d_k}(\alpha^{i_l})$ is expected to have only a few values that are greater than zero.

Given a dictionary $\mathbf{D}$ and a set of segmented superpixel regions L over an image, we represent the image by spatial max-pooling. For each superpixel region, $l \in [1, \cdots, N_S]$ of image i, where N_S represents the number of superpixels extracted from the image, let α_j^l be a descriptor vector extracted from region l, where $j \in [1, \cdots, N_l]$ indexes the N_l image pixels extracted from region l. Thus, given a dictionary $\mathbf{D}$, region l can be encoded using max spatial pooling, as follows:

$$\begin{aligned} \mathbf{x}_{i_l,\mathbf{D}} &= [\max_{j \in N_l} \langle \alpha_j^l, d_1 \rangle, \cdots, \max_{j \in N_l} \langle \alpha_j^l, d_K \rangle] \in R^K, \\ \mathbf{x}_{\mathbf{D}}(i) &= [\mathbf{x}_{\mathbf{D}}(l_1), \cdots, \mathbf{x}_{\mathbf{D}}(l_{N_S})] \in R^{K \times N_S}, \end{aligned} \tag{4}$$

where $\mathbf{x}_{i_l,\mathbf{D}}$ represents the midlevel feature descriptor of superpixel l. $\mathbf{x}_{\mathbf{D}}(i)$ represents the midlevel feature descriptor of image i. If the midlevel features of the pixels in the segmentation region are more similar to some of the visual words, these features can be used to represent the characteristics of the region, and the similarity is measured for the whole region.

2.2 *Probability Feature Descriptors*

Let $\mathbf{x}_{\mathbf{D}}$ be the midlevel feature vector of an image. This feature represents a vector in a K-dimensional space with a dictionary $\mathbf{D}$. If three different types of features (DSIFT, spectral band, and GLCM) are used in the sparse coding phase, then an image can be represented by three different corresponding vectors. That is, each image i can be represented by the following vectors:

$$\begin{aligned} \mathbf{x}_{\mathbf{D}_1}(i) &= [\mathbf{x}_{\mathbf{D}_1}(l_1), \cdots, \mathbf{x}_{\mathbf{D}_1}(l_{N_S})] \in \mathbf{R}^{K1 \times N_S}, \\ \mathbf{x}_{\mathbf{D}_2}(i) &= [\mathbf{x}_{\mathbf{D}_2}(l_1), \cdots, \mathbf{x}_{\mathbf{D}_2}(l_{N_S})] \in \mathbf{R}^{K2 \times N_S}, \\ \mathbf{x}_{\mathbf{D}_3}(i) &= [\mathbf{x}_{\mathbf{D}_3}(l_1), \cdots, \mathbf{x}_{\mathbf{D}_3}(l_{N_S})] \in \mathbf{R}^{K3 \times N_S}, \end{aligned} \tag{5}$$

where $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}_3$ are the dictionaries extracted from the DSIFT and spectral features, l represents the superpixels, and $K1$, $K2$, and $K3$ are the dictionary sizes. These two kinds of midlevel features combined with training samples are used to estimate the SVM classifier parameters and calculate the probability of vectors belonging to each LULC class, respectively.

The probability vectors of the different midlevel feature descriptors can be represented as follows:

$$\begin{aligned} \mathbf{P}1 &= [p_1(l_1), \cdots, p_1(l_{N_S})] \in R^{KL \times N_S}, \\ \mathbf{P}2 &= [p_2(l_1), \cdots, p_2(l_{N_S})] \in R^{KL \times N_S}, \\ \mathbf{P}3 &= [p_3(l_1), \cdots, p_3(l_{N_S})] \in R^{KL \times N_S}, \end{aligned} \tag{6}$$

where KL represents the number of land cover classes. The MFSC feature descriptors for the final classification are given as follows:

$$\mathbf{P} = [\mathbf{P}1, \mathbf{P}2, \mathbf{P}3] \in \mathbf{R}^{KL3 \times N_S}, \tag{7}$$

where $KL3$ represents the size of the feature descriptors, and this value is thrice the number of LULC classes. The size of MFSC feature descriptors is much smaller than the size of midlevel feature descriptors as in Eq. (5).

2.3 *CRF Classification Model*

The CRF model for the final classification of high-resolution remote sensing images is proposed. The CRF is defined over a set of superpixels ν extracted from the image I. Each superpixel $i \in \nu$ is associated with a class label $x \in \mathcal{L} = \{1, \cdots, L\}$. The labeling of the image is denoted by the vector $x \in \mathcal{L}^{|\nu|}$. The interaction among various superpixels of the CRF is captured by the set of edges $\varepsilon \in \nu \times \nu$, where each edge $e_{ij} \in \varepsilon$ corresponds to a pair of superpixels $i, j \in \nu$ that share a boundary.

The CRF energy, which consists of unary and pairwise costs, can be formulated as follows:

$$E(x, I) = \lambda_U \sum_{i \in \upsilon} \psi_i^U(x_i, I) + \lambda_P \sum_{e_{i,j} \in \varepsilon} \psi_{ij}^P(x_i, x_j, I), \tag{8}$$

where $\lambda_U \geq 0$ and $\lambda_p \geq 0$ are the relative weights of the unary and pairwise potentials, respectively.

The unary potential, which is expressed as $\psi_i^U(x_i, I)$ in Eq. (8), models the cost of assigning a class label $x_i \in \mathcal{L}$ to superpixel i in image I. This potential is defined as the score of a kernel SVM classifier for class x_i applied to an MFSC feature vector of superpixel i described in Eq. (7). The classifier for class l is trained using the MFSC feature vector extracted from the superpixels in the training set. This vector is labeled as l. The radial basis function (RBF)-$\chi 2$ kernel is adopted for SVM classification.

The pairwise potential, $\psi_{ij}^P(x_i, x_j, I)$, models the cost of assigning labels x_i and x_j to the neighboring superpixels i and j, respectively. When a CRF formulation is used for classification, the pairwise potentials are usually used to ensure the smoothness of the label assignments. A contrast sensitive cost is used as follows:

$$\psi_{ij}^P(x_i, x_j, I) = \frac{L_{ij}\delta(x_i \neq x_j)}{1 + \|\bar{I}_i - \bar{I}_j\|}, \tag{9}$$

where L_{ij} is the length of the shared boundary between superpixels i and j, and $\bar{I}_i$ and $\bar{I}_j$ are the gray mean values of superpixels i and j, respectively. The parameters in Eq. (8), λ_U and λ_P, are estimated by the cutting plane method, the details of which are described in Ref. 49. The classification result of the CRF models could be achieved by solving Eq. (8).

3 Experimental Results

We conduct experiments using the high-resolution aerial images to evaluate the effectiveness of the proposed MFSC-CRF framework for LULC classification. Based on the study of Jain et al.'s[49] work, comparative experiments are conducted by combining feature descriptors and classification methods. We compared the different methods using single-object class accuracy and total accuracy. The low-level feature, midlevel feature, and classifier associated with SF-SVM, U-SVM, GLCM-SVM, MFSC-SVM, SF-CRF, U-CRF, GLCM-CRF, and MFSC-CRF are reported in Table 1. The details are described as follows.

1. SF-SVM: This method uses only the unary segmentation cost. Spectral features are considered low-level features in this technique. After midlevel feature learning, the SVM method is adapted to achieve classification results. This method is very similar to the simultaneous orthogonal matching pursuit method proposed by Chen et al.[51]
2. U-SVM: This method is similar to SF-SVM, but they differ in the selection of low-level features. As described in Ref. 26, the DSIFT feature is considered as the low-level feature, and the SVM classifier is used for superpixel level classification.
3. GLCM-SVM: The GLCM feature is considered as the low-level feature in this method, and the SVM classifier is used for superpixel level classification.

Table 1 Information of different classification methods.

Method	Low-level feature	Midlevel feature	Classifier
SF-SVM	Spectral features	Sparse coding and max-pooling [Eq. (4)]	SVM
U-SVM	DSIFT	Sparse coding and max-pooling [Eq. (4)]	SVM
GLCM-SVM	GLCM	Sparse coding and max-pooling [Eq. (4)]	SVM
MFSC-SVM	Spectral features, DSIFT, and GLCM	MFSC [Eq. (7)]	SVM
SF-CRF	Spectral features	Sparse coding and max-pooling [Eq. (4)]	CRF
U-CRF	DSIFT	Sparse coding and max-pooling [Eq. (4)]	CRF
GLCM-CRF	GLCM	Sparse coding and max-pooling [Eq. (4)]	CRF
MFSC-CRF	Spectral features, DSIFT, and GLCM	MFSC [Eq. (7)]	CRF

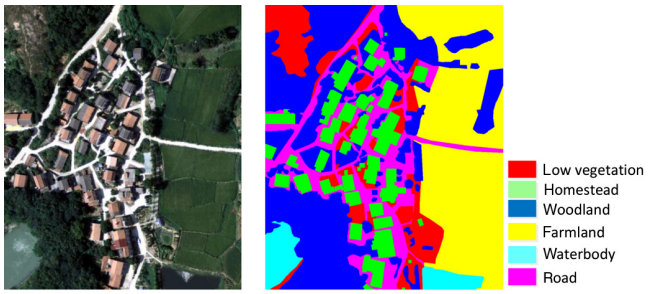
4. MFSC-SVM: Multifeature soft-probability is used for the feature vector in this method, and SVM is adopted for LULC classification.
5. SF-CRF: Spectral feature is considered as the low-level feature in this method, which is combined with sparse coding and CRF to achieve the classification results.
6. U-CRF: Sparse coding and the CRF model are used in this technique, and DSIFT is considered as the low-level feature, as described in Ref. 48.
7. GLCM-CRF: GLCM is considered as the low-level feature descriptor in this model, in which CRF is adopted for classification.
8. MFSC-CRF: Probabilities are considered as feature descriptors in this proposed method, in which CRF is adopted for supervised classification.

The experimental results are evaluated using three kinds of accuracies, namely, the accuracy of each class, overall accuracy (OA), and kappa coefficient (Kappa). OA is the fraction of correctly classified pixels, based on all pixels of that ground-truth class. For a fair comparison, the classification results with the highest OA are selected for all classification algorithms. The effect of the number of training samples is further investigated in relation to the MFSC-CRF model.

3.1 *Experimental Data Description*

3.1.1 *Experimental datasets (testing site 1)*

The first test image is captured over the rural residential area in Wuhan city, Hubei Province, China, through unmanned aerial vehicle aerial photography, including red, green, and blue three spectral bands. The image is of 1024×1200 pixels, with spatial resolution of 0.2 m and three multispectral channels. An overview of this dataset is shown in Fig. 2(a). The corresponding ground truth is shown in Fig. 2(b). The testing image was segmented to 52,654 superpixels



(a) (b)

Fig. 2 Wuhan rural residential area dataset (testing site 1): (a) RGB and (b) ground-truth images (low vegetation, homestead, woodland, farmland, waterbody, and road).

Table 2 Class information of Wuhan rural residential area dataset of testing site 1.

Class name	Training samples	Testing samples
Low vegetation	100	6055
Homestead	100	6518
Woodland	100	17,710
Farmland	100	13,022
Waterbody	100	2294
Road	100	7055

using the simple linear iterative clustering method. Six classes of interest, namely, low vegetation, homestead, farmland, waterbody, road, and woodland, are considered and listed in Table 2. Rural homestead is the main type of rural residential land and is more scattered. This class contains various houses, walls, and other facilities with spatial correlation and semantic structure characteristics. The other five class types are mainly land cover types. A total of 100 training samples for each LULC class type is used from the reference ground-truth data, and the remaining samples are used to evaluate the accuracy. The results are shown in Table 2.

3.1.2 *Experimental datasets (testing site 2)*

This testing image is also captured over the rural residential area in Wuhan city, Hubei Province, China. The image is of 1113×1777 pixels, with spatial resolution of 0.2 m and three multispectral channels. Compared with testing site 1, testing site 2 is larger and has a more complex scene. More trees are around the homesteads in this rural residential area, and the shadow effect is more obvious. This image is a challenging task for LULC classification. The ground-truth image corresponding to the high resolution image (HRI) has been classified manually into the six most common LULC classes. The classification data (label images) are shown in Fig. 3(b). The testing image was segmented to 92,441 superpixels. Similar to testing site 1, six classes of interest are considered and described in Table 3, which also shows the number of the training and testing samples for each class. The training samples are randomly chosen from the reference ground-truth data and are shown in Table 3. The dictionary size is set to 500, and 20,000 pixels are randomly selected for the training dictionary via the K-means clustering method. A total of 500 training samples per LULC class is randomly selected for classifier parameters (Table 3).

3.2 *Experimental Results and Analysis for Testing Site 1*

The experimental results for testing site 1 are reported to validate the effectiveness of the proposed MFSC-CRF for LULC classification. The classification accuracies of the various midlevel

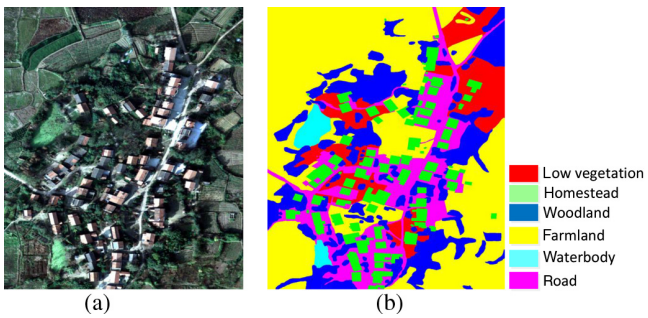


Fig. 3 Wuhan rural residential area dataset (testing site 2): (a) RGB and (b) ground-truth images. (low vegetation, homestead, woodland, farmland, waterbody, and road).

Table 3 Class information of Wuhan rural residential area dataset of testing site 2.

Class name	Training samples	Testing samples
Low vegetation	500	8964
Homestead	500	9215
Woodland	500	20,304
Farmland	500	41,528
Waterbody	500	1781
Road	500	10,649

feature learning methods, namely, SF-SVM, GLCM-SVM, U-SVM, MFSC-SVM, GLCM-CRF, and U-CRF, which are different combinations of low-level feature descriptors and classifier, are compared. The SVM classifier with RBF kernel has been proven to be successful in supervised classification of high-dimensional HRI data. Among the SVM-based methods, MFSC-SVM achieves better classification results than the other three methods [Figs. 4(c)–4(f)]. However, the SVM algorithm, in which any neighborhood spatial contextual information is not considered, results in high isolated salt-and-pepper classification noise, because neighborhood interactions are not considered in the algorithms.

For the MFSC-CRF algorithm, which is proposed to combine different effective features, the oversmoothing is less serious in Fig. 4(e), as is shown in the red boxes of Figs. 4(e) and 4(h). Moreover, the boundaries of homestead are better preserved. By contrast, SF-SVM is more focused on the spectral information. Thus, the classification remarkably depends less on the structural information, which probably explains the misclassification of U-CRF.

The quantitative performances with the highest classification accuracies obtained by SF-SVM, U-SVM, GLCM-SVM, MFSC-SVM, SF-CRF, U-CRF, GLCM-CRF, and MFSC-CRF are reported in Table 4. The best result of each column are in bold. The results show that

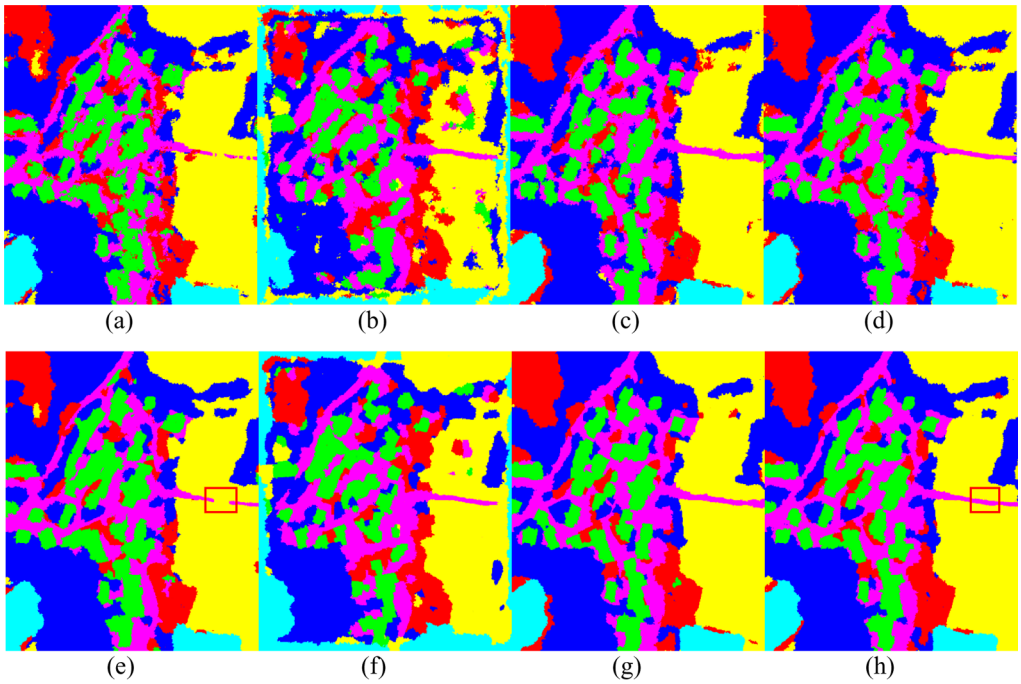


Fig. 4 Classification of Wuhan rural residential area datasets (testing site 1): (a) SF-SVM, (b) U-SVM, (c) GLCM-SVM, (d) MFSC-SVM, (e) SF-CRF, (f) U-CRF, (g) GLCM-CRF, and (h) MFSC-CRF. [The red rectangles in (e) and (h) are used to indicate the difference in the classification results].

Table 4 Classification accuracy for Wuhan rural residential area using dataset of testing site 1 with different classifiers.

Methods	Accuracy (%)						OA (%)	Kappa
	Low vegetation	Homestead	Woodland	Farmland	Waterbody	Road		
U-SVM	55.7 ± 5.6	65.4 ± 7.8	47.8 ± 1.8	68.9 ± 6.2	94.4 ± 2.5	57.9 ± 4.1	64.9 ± 0.7	0.596 ± 0.000059
SF-SVM	74.3 ± 15.2	83.8 ± 1.2	79.5 ± 7.6	91.4 ± 3.1	97.4 ± 0.8	72.9 ± 5.1	83.4 ± 0.3	0.795 ± 0.000042
GLCM-SVM	73.9 ± 6.0	87.5 ± 8.2	77.8 ± 0.9	92.4 ± 3.2	97.9 ± 1.9	71.4 ± 16.7	83.3 ± 0.7	0.795 ± 0.000096
MFSC-SVM	77.6 ± 1.2	87.7 ± 1.2	80.1 ± 3.0	94.7 ± 0.9	98.1 ± 0.3	75.9 ± 3.2	85.7 ± 0.1	0.823 ± 0.000050
U-CRF	59.0 ± 8.4	67.9 ± 12.7	52.9 ± 4.0	79.7 ± 24.7	97.1 ± 2.3	58.1 ± 8.1	69.1 ± 0.8	0.639 ± 0.000087
SF-CRF	74.6 ± 17.5	85.6 ± 5.1	81.3 ± 8.8	92.5 ± 3.8	97.8 ± 0.5	73.4 ± 8.4	84.4 ± 0.6	0.807 ± 0.000092
GLCM-CRF	73.2 ± 5.0	89.0 ± 8.5	79.7 ± 1.4	93.3 ± 3.4	97.8 ± 0.2	72.3 ± 16.5	84.2 ± 0.7	0.805 ± 0.000097
MFSC-CRF	**78.2 ± 2.2**	**88.7 ± 0.9**	**81.4 ± 3.5**	**95.1 ± 0.9**	**98.4 ± 0.1**	**76.6 ± 2.8**	**86.3 ± 0.1**	**0.830 ± 0.000044**

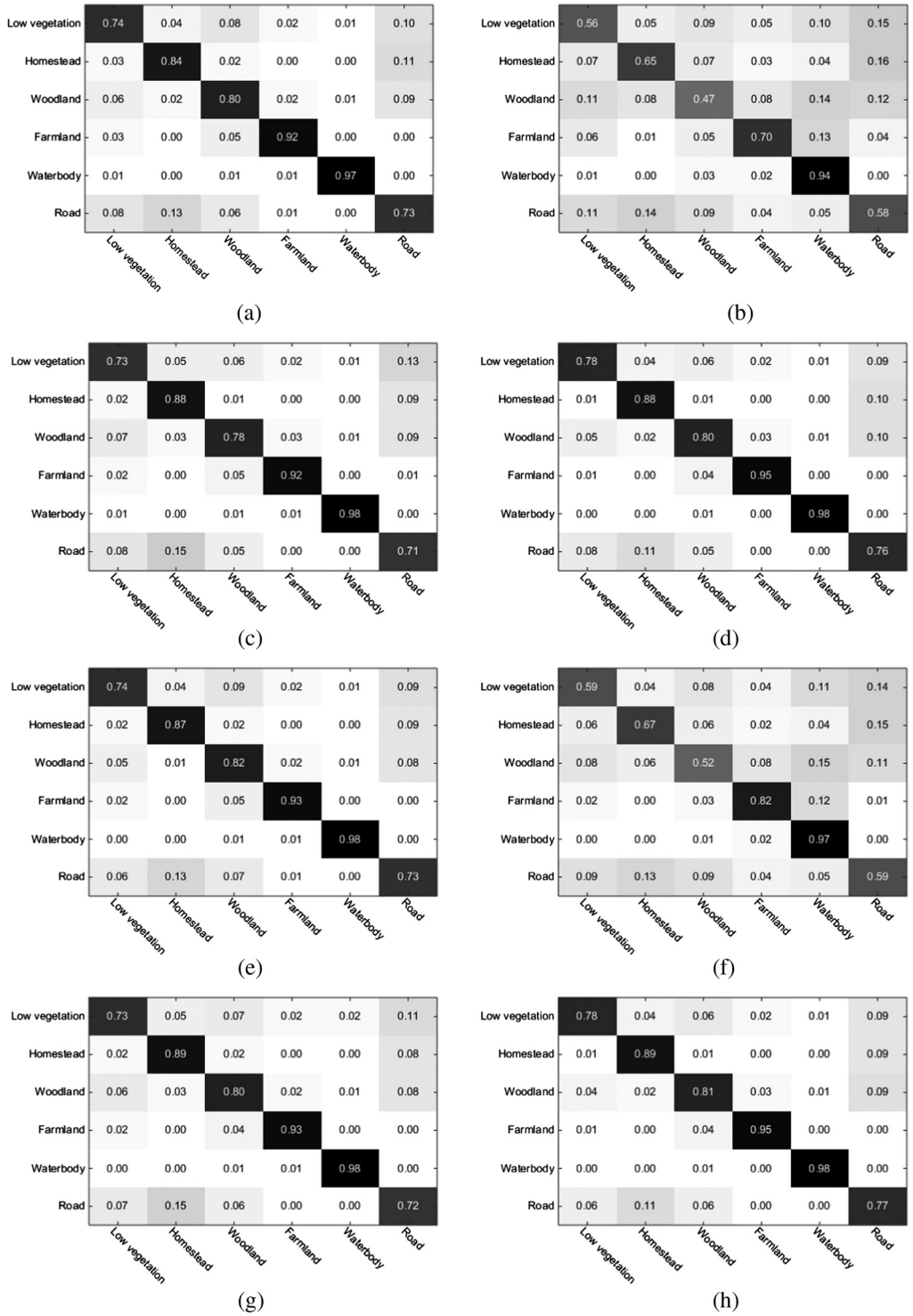


Fig. 5 Confusion matrices on Wuhan rural residential area datasets (testing site 1): (a) SF-SVM, (b) U-SVM, (c) GLCM-SVM, (d) MFSC-SVM, (e) SF-CRF, (f) U-CRF, (g) GLCM-CRF, and (h) MFSC-CRF.

the algorithms in which spatial contextual information are considered significantly outperformed the SVM classification in classification accuracy. Moreover, the accuracy of MFSC-CRF is higher than the three other CRF-based classification methods (i.e., SF-CRF, U-CRF, and GLCM-CRF), indicating that the MFSC-CRF can adaptively incorporate different low-level feature descriptors. With GLCM as the low-level feature descriptor, the GLCM-CRF method

achieves much higher accuracy than the SF-SVM, SF-CRF, U-SVM, and U-CRF. This result shows that GLCM can be very effective for LULC classification. In the dataset of the testing site 1 of Wuhan rural residential area (Table 4), the reported quantitative performance of MFSC-CRF exhibits the improvement in OA. Additionally, the 21% higher accuracy (from 64.9% to 86.3%) of MFSC-CRF compared with U-SVM shows that MFSC-CRF focuses more on spatial contextual information. Thus, spatial contextual information and other effective feature descriptors should be considered. Finally, the MFSC-CRF obtains the highest accuracy.

Figure 5 shows the confusion matrices of different classification methods with various feature descriptors and classifiers. The methods, which used only spectral features as low-level feature descriptors (SF-SVM and SF-CRF), misclassified homestead to road with 14%. The reason is that the two LULC types have similar spectral characteristics, and all belong to the impermeable surface. The GLCM- (GLCM-SVM and GLCM-CRF) and MFSC-based methods (MFSC-SVM and MFSC-CRF) are less serious than the SF-based methods. The MFSC-CRF method incorporates different low-level feature descriptors and results in 89% accuracy for homestead.

3.3 *Experimental Results and Analysis for Testing Site 2*

The resulting maps for the visual classification for this testing image are shown in Figs. 6(a)–6(h). The quantitative classification results of the different classification methods are shown in Table 5 (The best result of each column is in bold) and Figs. 7(a)–7(h). The proposed MFSC-CRF method achieves the highest OA and Kappa than SF-SVM, U-SVM, GLCM-SVM, MFSC-SVM, SF-CRF, U-CRF, and GLCM-CRF. Compared with SF-SVM and U-SVM, the MFSC-SVM method achieves remarkably enhanced OA and homestead accuracy. Compared with GLCM-SVM, the classification accuracy of the MFSC-SVM method shows ~3% improvement for each LULC class. Considering neighborhood spatial contextual information, the quantitative performance of MFSC-CRF shows 0.1% accuracy improvement (from 87.4% to 87.5%) compared with MFSC-SVM method.

3.4 *Parameter Sensitivity Analysis*

The performance of the proposed MFSC-CRF method is further evaluated using different numbers of training samples. Testing image 1 is selected for parameter sensitivity analysis, and

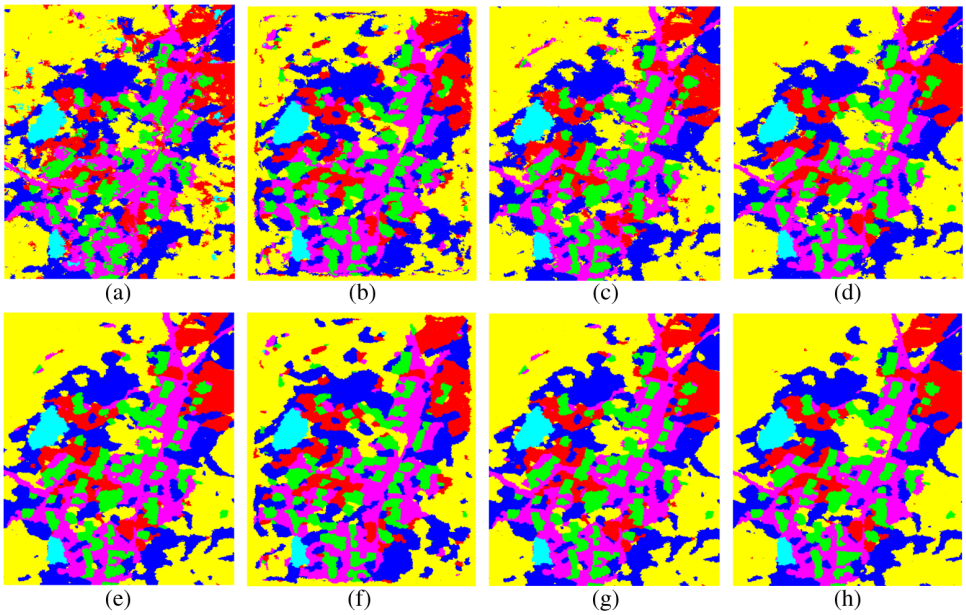


Fig. 6 Classification of Wuhan rural residential area dataset (testing site 2): (a) SF-SVM, (b) U-SVM, (c) GLCM-SVM, (d) MFSC-SVM, (e) SF-CRF, (f) U-CRF, (g) GLCM-CRF, and (h) MFSC-CRF.

Table 5 Classification accuracy for Wuhan rural residential area dataset of testing site 2 with different classifiers.

Methods	Accuracy (%)						OA (%)	Kappa
	Low vegetation	Homestead	Woodland	Farmland	Waterbody	Road		
U-SVM	86.6 ± 1.2	80.2 ± 0.5	76.4 ± 2.9	85.4 ± 0.5	98.3 ± 0.6	74.0 ± 0.9	83.5 ± 0.1	0.601 ± 0.000036
SF-SVM	87.7 ± 0.2	82.7 ± 3.0	76.1 ± 0.2	85.4 ± 0.4	98.2 ± 0.3	74.4 ± 1.8	84.1 ± 0.1	0.616 ± 0.000022
GLCM-SVM	86.4 ± 1.4	82.7 ± 4.4	76.6 ± 0.8	83.1 ± 3.5	97.9 ± 0.1	74.5 ± 1.8	83.6 ± 0.2	0.612 ± 0.000013
MFSC-SVM	91.2 ± 1.9	85.7 ± 1.9	**81.5 ± 0.5**	**89.5 ± 0.1**	96.6 ± 0.3	79.6 ± 1.0	87.4 ± 0.1	0.671 ± 0.000040
U-CRF	87.4 ± 0.3	80.9 ± 0.2	76.9 ± 3.0	85.7 ± 0.7	**98.4 ± 0.4**	75.5 ± 2.2	84.1 ± 0.3	0.614 ± 0.000160
SF-CRF	88.3 ± 0.1	83.8 ± 1.8	76.3 ± 1.0	86.0 ± 0.3	98.2 ± 0.2	75.4 ± 1.9	84.6 ± 0.1	0.625 ± 0.000059
GLCM-CRF	87.0 ± 1.9	82.0 ± 3.9	77.1 ± 0.6	83.4 ± 2.6	97.9 ± 0.1	76.2 ± 1.3	84.0 ± 0.1	0.619 ± 0.000007
MFSC-CRF	**91.2 ± 0.6**	**85.9 ± 2.7**	81.4 ± 0.5	89.4 ± 0.2	96.7 ± 0.3	**80.4 ± 0.7**	**87.5 ± 0.1**	**0.674 ± 0.000084**

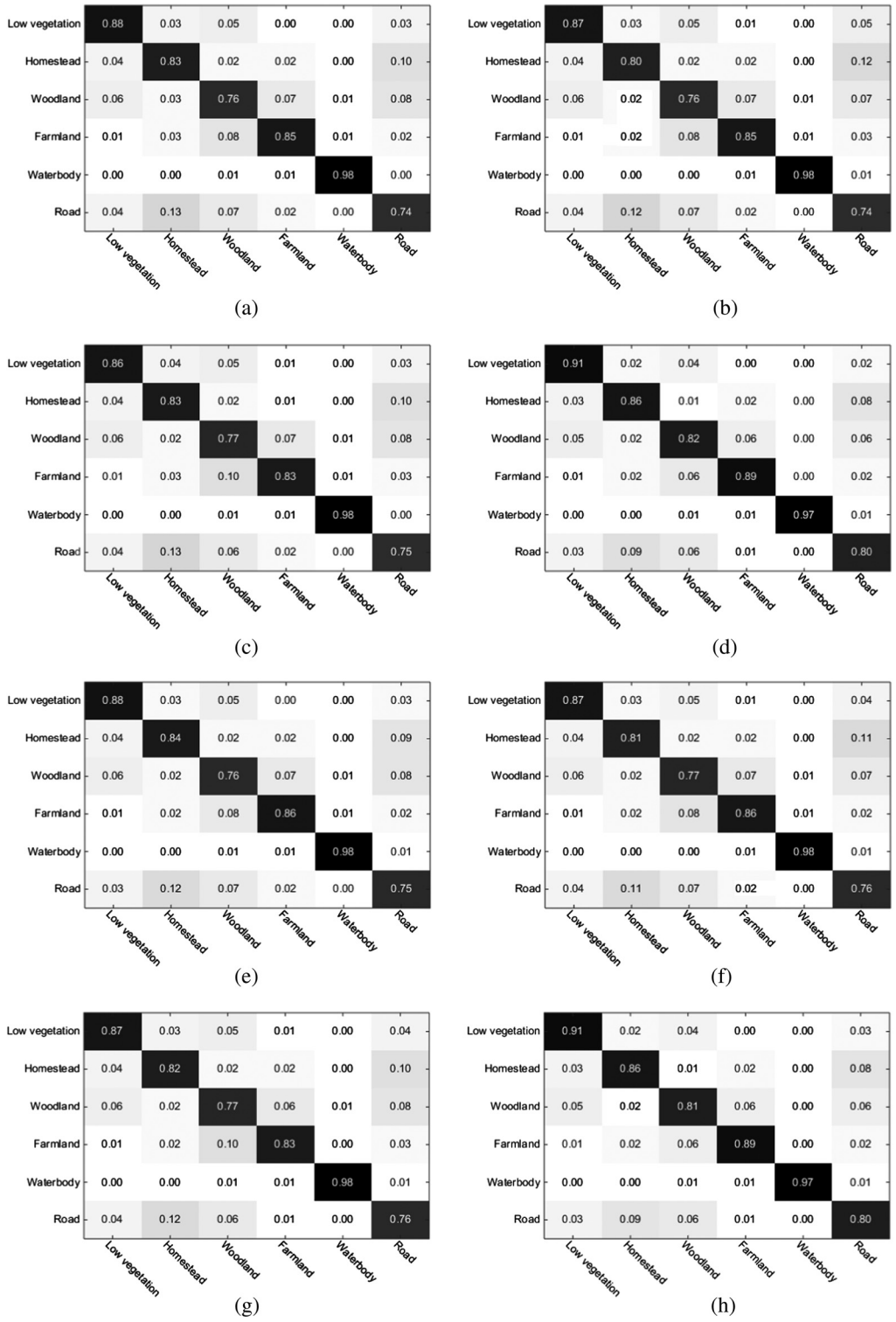


Fig. 7 Confusion matrices on Wuhan rural residential area datasets (testing site 2): (a) SF-SVM, (b) U-SVM, (c) GLCM-SVM, (d) MFSC-SVM, (e) SF-CRF, (f) U-CRF, (g) GLCM-CRF, and (h) MFSC-CRF.

the effects of training sample numbers on the MFSC-CRF algorithms are examined. Different sizes ranging from 100 to 1000 are tested with an interval of 100 for each LULC class.

As shown in Fig. 8, the classification accuracy of MFSC-CRF initially increases for the datasets with gradual increase in the number of training samples per class (from 85.6% to 93.2%).

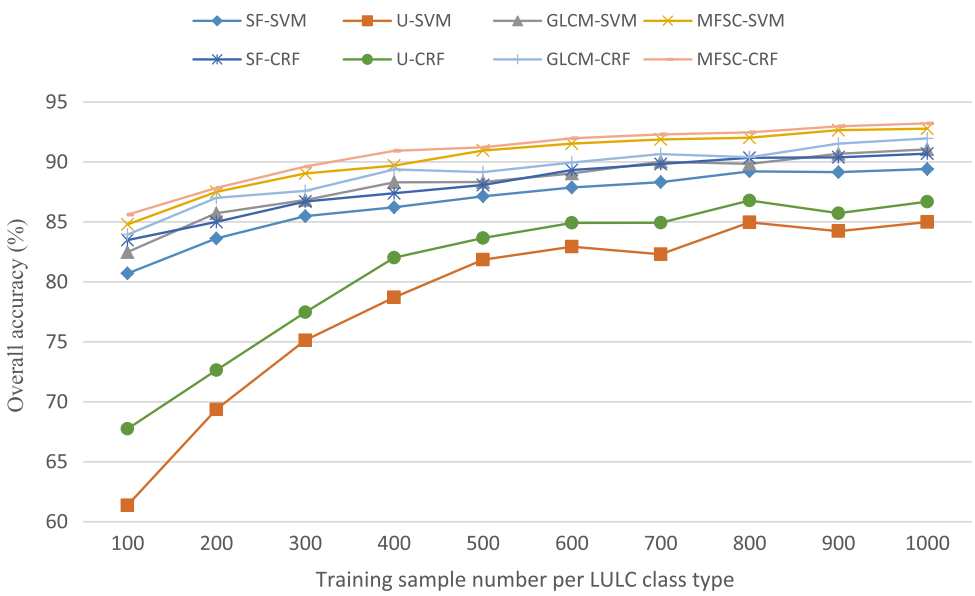


Fig. 8 Effect of training data size on the classification results for Wuhan rural residential area dataset of testing site 1.

The classification accuracy of MFSC-CRF is slightly higher than GLCM-CRF (from 84.0% to 92.0%) and MFSC-SVM (from 85.0% to 92.8%) classification approaches with Wuhan rural residential area dataset of testing site 1. The accuracy then remains roughly constant when the training sample number is set to 900 but slightly decreases. Moreover, the classification accuracy of the proposed method remains higher than the other seven methods at each training number. The training samples are randomly selected from the overall ground truth, and the remaining samples are used to evaluate the classification accuracies. The experiments show that the classification accuracies of the methods incorporating spatial contextual information (i.e., SF-CRF, U-CRF, GLCM-CRF, and the proposed MFSC-CRF) are all better than SVM-based classification methods. Moreover, the MFSC-CRF method is more robust than the other classification methods with different training samples.

4 Conclusion

A classification method for HSR remote sensing images based on MFSC and CRF models is proposed. The proposed MFSC-CRF method can effectively incorporate spectral, structural, and textural features, as well as spatial contextual information. Midlevel feature learning based on sparse coding is very important in image classification, and the proposed feature combination method can significantly improve the classification accuracy by effectively combining three complementary features, namely, DSIFT, spectral bands, and GLCM. Experiments on the Wuhan residential area datasets also show that the GLCM features can achieve more promising results than the original spectral features. This method is an open model, very convenient to cascade different features to improve the accuracy of image classification. Recently, the convolution neural network is widely used in image classification and achieved good results. However, the convolution neural network model requires a large number of training samples to train the parameters. Therefore, our next step is to use a small amount of training samples to fine-tune the convolution neural network model so that it can be effectively applied to remote sensing image classification applications.

References

1. D. Li, L. Zhang, and G. S. Xia, "Automatic analysis and mining of remote sensing big data," *Acta Geod. Cartogr. Sin.* **43**, 1211–1216 (2014).
2. G. S. Xia et al., "Structural high-resolution satellite image indexing," in *ISPRS TC VII Symp. 100 Years ISPRS*, Vol. 38, pp. 298–303 (2010).

3. P. Gong et al., "Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data," *Int. J. Remote Sens.* **34**(7), 2607–2654 (2013).
4. Z. Zhu et al., "Assessment of spectral, polarimetric, temporal, and spatial dimensions for urban and peri-urban land cover classification using Landsat and SAR data," *Remote Sens. Environ.* **117**, 72–82 (2012).
5. A. Paul, M. A. Peter, and J. C. Paul, "Fine spatial resolution simulated satellite sensor imagery for land cover mapping in the United Kingdom," *Remote Sens. Environ.* **68**, 206–216 (1999).
6. O. Debeir et al., "Textural and contextual land-cover classification using single and multiple classifier system," *Photogramm. Eng. Remote Sens.* **68**, 597–605 (2002).
7. K. Jia et al., "Land cover classification of finer resolution remote sensing data integrating temporal features from time series coarser resolution data," *ISPRS J. Photogramm. Remote Sens.* **93**, 49–55 (2014).
8. T. Blaschke et al., "Geographic object-based image analysis towards a new paradigm," *ISPRS J. Photogramm. Remote Sens.* **87**, 180–191 (2014).
9. Y. Zhong, J. Zhao, and L. Zhang, "A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.* **52**(11), 7023–7037 (2014).
10. K. Biro et al., "Exploitation of TerraSAR-X data for land use/land cover analysis using object-oriented classification approach in the African Sahel Area, Sudan," *J. Indian Soc. Remote Sens.* **41**(3), 539–553 (2013).
11. M. Ustuner, F. Balik Sanli, and B. Dixon, "Application of support vector machines for land use classification using high-resolution rapid eye images: a sensitivity analysis," *Eur. J. Remote Sens.* **48**, 403–422 (2015).
12. S. D. Jawak et al., "Advancement in land cover classification using very high resolution remotely sensed 8-band WorldView-2 satellite data," *Int. J. Earth Sci. Eng.* **6**(2), 1742–1749 (2013).
13. X. Huang and L. Zhang, "An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.* **51**(1), 257–272 (2013).
14. G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images," *Proc. IEEE* **101**(3), 631–651 (2013).
15. L. Wang and Q. Wang, "Subpixel mapping using Markov random field with multiple spectral constraints from subpixel shifted remote sensing images," *IEEE Geosci. Remote Sens. Lett.* **10**(3), 598–602 (2013).
16. A. Voisin et al., "Classification of very high resolution SAR images of urban areas using copulas and texture in a hierarchical Markov random field model," *IEEE Geosci. Remote Sens. Lett.* **10**(1), 96–100 (2013).
17. W. Yang et al., "SAR-based terrain classification using weakly supervised hierarchical Markov aspect models," *IEEE Trans. Image Process.* **21**(9), 4232–4243 (2012).
18. X. L. Li et al., "A survey on scene image classification," *Sci. Sin. Inf.* **45**, 827–848 (2015) (in Chinese with English abstract).
19. H. Bay et al., "Speeded-up robust features (SURF)," *Comput. Vis. Image Understanding* **110**(3), 346–359 (2008).
20. G. S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *Int. J. Comput. Vision* **88**(3), 382–403 (2010).
21. G. S. Xia, J. Delon, and Y. Gousseau, "Accurate junction detection and characterization in natural images," *Int. J. Comput. Vision* **106**(1), 31–56 (2014).
22. D. G. Lowe, "Distinctive image features from scale-invariant key points," *Int. J. Comput. Vision* **60**, 91–110 (2004).
23. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 886–893 (2005).
24. H. Goncalves, L. Corte-Real, and J. Goncalves, "Automatic image registration through image segmentation and SIFT," *IEEE Trans. Geosci. Remote Sens.* **49**(7), 2589–2600 (2011).

25. A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.* **52**(1), 439–451 (2014).
26. F. F. Li and P. Perona, "Bayesian hierarchy model for learning natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 524–531 (2005).
27. Y. Boureau et al., "Ask the locals: multi-way local pooling for image recognition," in *IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2651–2658 (2011).
28. Y. Cao et al., "Spatial-bag-of-features," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3352–3359 (2011).
29. Y. Huang et al., "Salient coding for image classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1753–1760 (2011).
30. Z. L. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(11), 2651–2664 (2013).
31. W. Yang, X. Yin, and G. S. Xia, "Learning high-level features for satellite image classification with limited labelled samples," *IEEE Trans. Geosci. Remote Sens.* **53**(8), 4472–4482 (2015).
32. H. Lobel, R. Vidal, and A. Soto, "Learning shared, discriminative, and compact representations for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(11), 2218–2231 (2015).
33. G. F. Sheng et al., "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.* **33**(8), 2395–2412 (2012).
34. K. Qi et al., "Land-use scene classification in high-resolution remote sensing images using improved correlatons," *IEEE Geosci. Remote Sens. Lett.* **12**(12), 2403–2407 (2015).
35. S. S. Chen and Y. L. Tian, "Pyramid of spatial relations for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.* **53**(4), 1947–1957 (2015).
36. F. Hu et al., "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **8**(5), 2015–2030 (2015).
37. F. Hu et al., "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.* **7**(11), 14680–14707 (2015).
38. K. Nogueira, O. A. B. Penatti, and J. A. D. Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recogn.* **61**, 539–556 (2017).
39. J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. R. Stat. Soc. Ser. B* **36**(2), 192–236 (1974).
40. J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Int. Conf. on Machine Learning*, Williamstown (2001).
41. G. S. Xia, C. He, and H. Sun, "Integration of synthetic aperture radar image segmentation method using Markov random field on region adjacency graph," *IET Radar Sonar Navig.* **1**(5), 348–353 (2007).
42. B. Y. Liu and X. M. He, "Multiclass semantic video segmentation with object-level active inference," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, Massachusetts (2015).
43. X. M. He and G. Stephen, "An exemplar-based CRF for multi-instance object segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, Ohio (2014).
44. Y. S. Ming, H. D. Li, and X. M. He, "Connected contours: a contour completion model that respects closure-effect," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, Rhode Island (2012).
45. X. M. He, R. Zemel, and M. Carreira-Perpinan, "Multiscale conditional random fields for image labelling," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, DC (2004).
46. V. Michele and F. Vittorio, "Semantic segmentation of urban scenes by learning local class interactions," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Boston, Massachusetts (2015).

47. V. Michele and F. Vittorio, "Structured prediction for urban scene semantic segmentation with geographic context," in *Joint Urban Remote Sensing Event*, Lausanne, Switzerland (2015).
48. P. Zhong and R. Wang, "Learning conditional random fields for classification of hyperspectral images," *IEEE Trans. Image Process.* **19**(7), 1890–1907 (2010).
49. A. Jain et al., "Visual dictionary learning for joint object categorization and segmentation," *Lect. Notes Comput. Sci.* **7576**(5), 718–731 (2012).
50. A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *IEEE Int. Conf. on Computer Vision*, pp. 1–8 (2007).
51. Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.* **49**(10), 3973–3985 (2011).

Bin Zhang received his BS, MS, and PhD degrees from the School of Electronic Information, Wuhan University, in 2007, 2009, and 2013, respectively. He is currently working at China University of Geosciences. His research interests include image classification, scene-level land use classification, and deep learning.

Biographies for the other authors are not available.