

LAND COVER CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK WITH REMOTE SENSING DATA AND DIGITAL SURFACE MODEL

Bo Liu, Shihong Du*, Xiuyuan Zhang

Institute of Remote Sensing and GIS, Peking University, Beijing 100871, China - dshgis@hotmail.com

Commission III, WG III/1

KEY WORDS: Land cover classification, CNN, VHR remote sensing images, DSM, CRF, Deep features

ABSTRACT:

Land cover map is widely used in urban planning, environmental monitoring and monitoring of the changing world. This paper proposes a framework with convolutional neural network (CNN), object-based voting and conditional random field (CRF) for land cover classification. Both very-high-resolution (VHR) remote sensing images and digital surface model (DSM) are inputs of this CNN model. To solve the “salt and pepper” effect caused by pixel-based classification, an object-based voting classification is performed. And to capture accurate boundary of ground objects, a CRF optimization using spectral information, DSM and deep features extracted through CNN is applied. Area one of Vaihingen datasets is used for experiment. The experimental results show that method proposed in this paper achieve an overall accuracy of 95.57%, which demonstrate the effectiveness of proposed method.

1. INTRODUCTION

Land cover classification using remote sensing imagery was widely used in the past decades (Friedl et al., 2002). With the development of remote sensing technology and computer vision technology, land cover mapping using remote sensing imagery has also developed significantly.

In the beginning, pixel-based methods with handcrafted features were used for land cover classification, such as maximum likelihood method, decision tree method, random forest, and support vector machine (Otakei and Blaschke, 2010; Rodriguez-Galiano et al., 2012). For these methods, the definition of handcrafted features is critical. The quality of classification results largely depends on the quality of handcrafted features definition. However, handcrafted features are not enough to effectively distinguish different categories in most cases. With the advent of neural networks and deep learning, artificial neural network (ANN) based multilayer perceptron (Kavzoglu and Mather, 2003) and CNN based convolution operation (Scott et al., 2017) were adopted. Both ANN and CNN can extract features automatically and get classification results through training samples. However, ANN loses spatial information of pixels because of the inputs of ANN are spectral features or other features of pixels, which may reduce the classification accuracy of land cover. Fortunately, the inputs of CNN are image patches and the spatial information of pixels are retained. Although fully convolutional network (FCN) (Shelhamer et al., 2017) can get an end-to-end classification result and has been used in aerial scenes (Mao et al., 2019), fully labelled samples which are hardly obtained in many cases are needed for FCN training. Therefore, pixel-based CNN is more effective for land cover classification with few samples.

Nevertheless, those pixel-based methods will cause “salt and pepper” effect and cannot capture the precise outline of ground objects. However, Object-based image analysis (OBIA) methods (Blaschke, 2010; Blaschke et al., 2014) can greatly

reduce “salt and pepper noise” and get a rough boundary of ground objects. Unfortunately, image objects are irregular polygon and the inputs of CNN are regular patches. So far, there is not an excellent way to use CNN to classify image objects directly. Therefore, combining pixel-based CNN method and object-based voting method is an appropriate way for land cover classification (Zhao et al., 2017a). Additionally, contextual information is critical for land cover classification, especially for capturing precise outline of ground objects. Many researchers have utilized CRF to take advantage of contextual information (Albert et al., 2017).

In addition to optical remote sensing data, DSM acquired by LiDAR or some other sensors is also very helpful for land cover classification (Wai et al., 2015; Szostak et al., 2020.). From height information contained in DSM, extra information can be obtained, and this information can be obtained from spectral features. For example, spectral features of the same roof may be different, but the height will not be much different. Another example, spectral features of building and road may be extremely similar, but the heights of these two ground objects are totally different. Therefore, DSM is a very important data source for land cover classification.

This paper proposes a simple framework for land cover classification using VHR image and DSM. First of all, CNN is trained using samples obtained from VHR image and DSM. Then, initial pixel-based classification result and deep features are acquired. Subsequently, VHR image is segmented to image objects and object-based voting is performed to get intermediate result. Finally, a CRF optimization using spectral features, DSM and deep features is utilized to refine intermediate result, and final result is obtained.

The rest of this paper is organized as follows. Section 2 will introduce the details of proposed framework. Experimental results will be shown in Section 3. And the last section is the conclusion.

* Corresponding author

2. METHODOLOGY

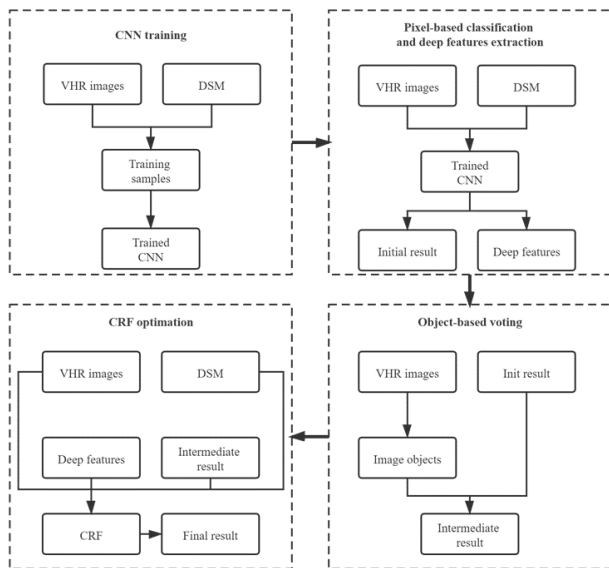


Figure 1. The framework of proposed method.

The proposed method consists of 4 steps, as is illustrated in Figure 1.

- (1) CNN training: Train CNN with training samples obtained from VHR images and DSM.
- (2) Pixel-based classification and deep features extraction: Input VHR images and DSM into the trained CNN to get the initial classification result and pixel-based deep features.
- (3) Object-based voting: Get image objects by multiresolution segmentation, and intermediate result is obtained by voting based image objects.
- (4) CRF optimization: Optimize intermediate result using CRF with spectral features, DSM and deep features.

2.1 CNN training

2.1.1 The structure of proposed CNN: The most basic CNN structure consists of convolutional layers, pooling layers and fully connected layers (Krizhevsky et al., 2017). Convolutional layers are used to extract multilayer features from inputs, and one convolutional layer is represented as follow formula:

$$X_l = f(X_{l-1}W_l + B_l) \quad (1)$$

where the X_l denotes output of l^{th} convolutional layer, the W_l and the B_l denotes convolutional kernels and bias of l^{th} convolutional layer respectively, the f denotes activation function which can be sigmoid, tanh and rectified linear units (ReLU). Pooling layers are used to compress features extracted by convolutional layers to reduce computational cost. And fully connected layers are used to get classification results using features extracted by convolutional layers and pooling layers. Because of loss of information caused by pooling layers, the pooling layers in proposed CNN structure are replaced by convolutional layers with stride two. Additionally, batch normalization layers are utilized after each convolutional layer and fully connected layer (except the last fully connected layer) in proposed CNN structure. Batch normalization layers guarantee that the distribution of features extracted by each

layer is the same, which is helpful for classification. And the activation functions of all layers used in this CNN structure are ReLU (except the last fully connected layer utilizes softmax).

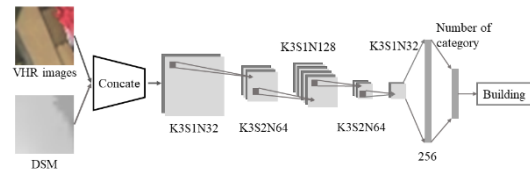


Figure 2. The structure of proposed CNN.

To be able to input VHR images and DSM simultaneously, proposed CNN structure has two inputs. The structure of proposed CNN is shown in Figure 2 (Batch normalization layers are not shown). This CNN consists of a concatenation layer before convolution layers, which is used to combine VHR images and DSM. Additionally, there are five convolutional layers, two fully connected layers and six batch normalization layers. The KaSbNc denotes that the kernel size, the stride and the number of filters are a, b and c respectively. And the number of neurons of the first fully connected layer and the last fully connected layer are 256 and number of category respectively.

2.1.2 Parameters of CNN training: Samples for CNN training are image patches with patch size of 32, including VHR image patches and DSM patches. The training optimizer, the training batch size and the learning rate are Adam, 32 and 0.0001, respectively. Furthermore, early stopping method is used to prevent overfitting (Yao et al., 2007).

2.2 Pixel-based classification and deep features extraction

The output of the last fully connected layer is probability that a pixel belongs to each class, and the output of the first fully connected layer is a 256-dimensional features. Take each pixel as the center to capture an image patch with patch size of 32 and input it to the trained CNN, the category and a 256-dimensional features of each pixel can be obtained. For image boundary, pixels in image patches without values are set to zero. Therefore, initial classification result by pixel-based classification and pixel-based deep features are obtained, and they will be used in next step.

2.3 Object-based voting

2.3.1 Image segmentation: To reduce the “salt and pepper” effect caused by pixel-based classification, an object-based voting is performed. VHR image first is segmented into image objects. The segmentation algorithm used in this paper is multiresolution segmentation (MRS) (Baatz and Schape, 2000), which uses spectral heterogeneity and shape heterogeneity to segment images. There are three main parameters in this algorithm, including scale, shape weight and compactness weight. In this paper, these three parameters are 10, 0.1 and 0.5 respectively.

2.3.2 Voting based image objects: Through initial result and image objects, object-based voting can be performed. Consider the number of categories is M , then the category is $C = \{C_k | k = 1, 2, \dots, M\}$. An image object O contains N pixels, and initial labels of these pixels are $\{l_i | i = 1, 2, \dots, N \& \& l_i \in C\}$. Then label of the image object can be obtained by:

$$p_c = \sum_{i=1}^N \text{int}(l_i == c) / N, c \in C \quad (2)$$

$$l_o = \arg \max(p_c), c \in C \quad (3)$$

where the p_c is the probability that the image object belongs to category c , and l_o is the category of image object O . Thus, categories of all image objects can be obtained.

2.4 CRF optimization

To capture precise outline of ground objects, CRF optimization based pixels with spectral features, DSM and deep features is performed (Zhao et al., 2017b). The energy function is defined as follow:

$$E(l) = \sum_{p \in P} D_p(l_p) + \sum_{\{p,q\} \in N} w_{(p,q)} V_{(p,q)}(l_p, l_q) \quad (4)$$

Where the $\sum_{p \in P} D_p(l_p)$ and $\sum_{\{p,q\} \in N} w_{(p,q)} V_{(p,q)}(l_p, l_q)$ denote unary and pairwise terms, respectively. The l denotes all land cover categories. The goal is to assign each node p (a pixel) a label l_p . The unary term $D_p(l_p)$ can be formulated as:

$$D_p(l_p) = -\log P(l_p) \quad (5)$$

where the $P(l_p)$ denotes the probability of pixel p belonging to l_p which is the output of last fully connected layer of CNN. $w_{(p,q)}$ encourages discontinuities to coincide with intensity edges, which has a significant impact on optimization. In this framework, spectral features, height information obtained from DSM and deep features extracted through CNN are combined to construct it. $w_{(p,q)}$ can be formulated as:

$$w_{(p,q)} = \exp\left(-\frac{s_{p,q}^2}{D}\right) + \exp\left(-\frac{h_{p,q}^2}{D}\right) + \exp\left(-\frac{d_{p,q}^2}{D}\right) \quad (6)$$

$$D = \text{dist}(p, q) \times 2\sigma^2 \quad (7)$$

where the $s_{p,q}$, the $h_{p,q}$ and the $d_{p,q}$ denote heterogeneities of spectral features, height features and deep features between pixel p and pixel q , respectively. The $\text{dist}(p,q)$ denotes the Euclidean distance between pixel p and pixel q . And σ is used to adjust value range. $V_{(p,q)}(l_p, l_q)$ is defined as:

$$V_{(p,q)}(l_p, l_q) = \begin{cases} 0, & l_p = l_q \\ 1, & l_p \neq l_q \end{cases} \quad (8)$$

Finally, category of each pixel can be inferred by minimizing the energy function (4) using α -expansion algorithm (Boykov et al., 2001).

3. EXPERIMENTS

3.1 Datasets

The experimental datasets used in this paper is area one of Vaihingen datasets, including a VHR image, a DSM image and a ground truth image, as is shown in Figure 3. The image size of these three images are 1919×2569 and the VHR image contains near infrared, red and blue three bands. There are five categories in experimental area, including impervious, building, low vegetation, tree and car. Five percent of pixels of each

category are selected as samples for CNN training, and the rest 95 percent of pixels of each category are used for accuracy assessment.

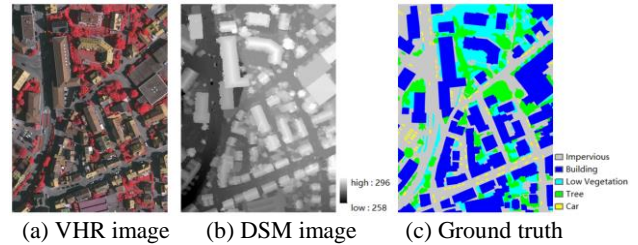


Figure 3. Datasets for experiments.

3.2 Experimental results

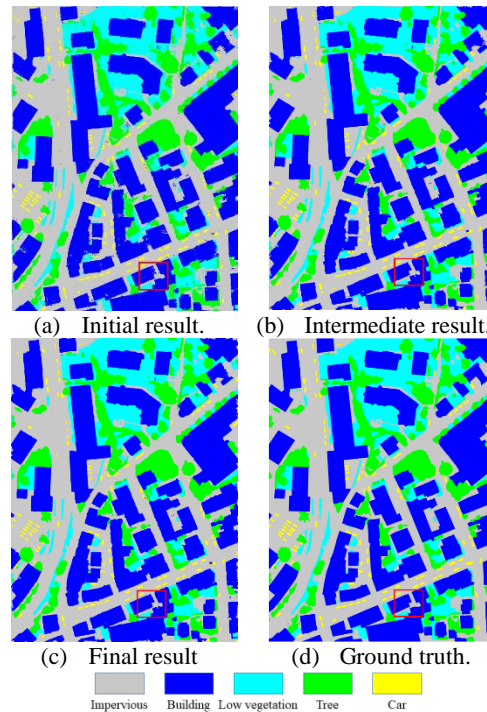


Figure 4. Classification results.

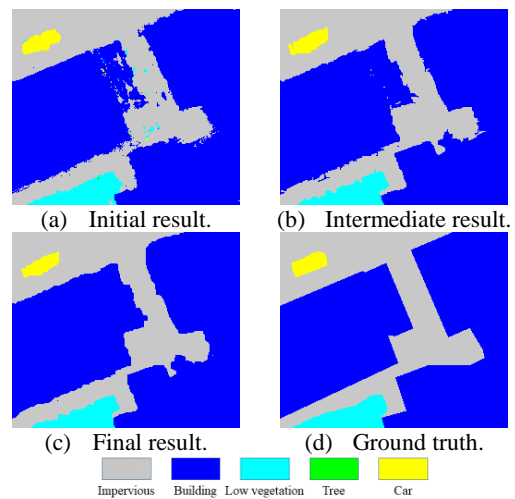


Figure 5. Classification results of red box in Figure 4.

Initial result, intermediate result and final result are shown in Figure 4. And classification results of red box in Figure 4 are

illustrated in Figure 5. It can be seen from Figure 5(a) that severe “salt and pepper” effect is occurred in classification result obtained by pixel-based CNN, and the outline of ground objects is not clear. After object-based voting, intermediate result is obtained and is shown in Figure 5(b). It can be seen that “salt and pepper” effect is almost not exist. To capture precise outline of ground objects, CRF optimization is performed and the result is shown in Figure 5(c). Comparing with initial result and intermediate result, the outline of ground objects is indeed more accurate.

To further verify the effectiveness of object-based voting and CRF optimization. F1 scores for each class and overall accuracies of initial result, intermediate result and final result are shown in Table 1. It is illustrated in Table 1 that object-based voting and CRF optimization are critical in proposed method. Especially classification accuracies have been greatly improved after object-based voting. As for the decline in F1 score of Car, it is mainly due to the car is too small, which leads to inaccurate segmentation boundaries. Smaller segmentation scale can get higher classification accuracies. CRF optimization is mainly used to capture accurate boundary, that is why there is no major change in classification accuracies after CRF optimization. However, precise outline of ground objects can be captured after CRF optimization, which can be seen in Figure 5. the decline in F1 score of Car after CRF optimization is also caused by the car is too small. Overall, object-based voting and CRF optimization are still essential steps of the proposed method.

	Initial	Intermediate	Final
Impervious	95.69	95.75	95.73
Building	97.56	97.67	97.63
Low vegetation	91.09	92.36	92.43
Tree	93.01	93.34	93.64
Car	82.93	82.79	81.57
Overall	95.27	95.54	95.57

Table 1. F1 scores for each class and overall accuracies of classification results (%).

Additionally, to validate the importance of DSM, pixel-based CNN classification and object-based voting without DSM data was performed. F1 scores for each class and overall accuracies are illustrated in Table 2. It is obvious that F1 scores of all categories except Car and overall accuracy of approach with DSM are higher than these accuracies of approach without DSM. As for Car, DSM has basically no effect on its classification accuracy. From this, DSM is very helpful for land cover classification.

	Without DSM	With DSM
Impervious	95.06	95.75
Building	97.45	97.67
Low vegetation	90.57	92.36
Tree	91.98	93.34
Car	82.80	82.79
Overall	94.82	95.54

Table 2. F1 scores for each class and overall accuracies comparison between approach without DSM and approach with DSM (%).

3.3 Comparison with the state-of-the-art methods

There are many researchers evaluating their land cover classification methods with Vaihingen datasets. To verify the effectiveness of the proposed method, a comparison of

classification accuracy is performed. Two approaches are selected to compare with our method. One approach was proposed by Audebert et al. (Audebert et al., 2018), the other approach was proposed by Sun et al. (Sun et al., 2020). Audebert proposed two architectures to fusion RGB image and DSM. These two architectures are V-FuseNet (Early fusion) and SegNet-RC (Later fusion), respectively. Sun et al. utilized an effective deep FCN ensemble and fully connected CRF for 2D Semantic Labeling Contest of Vaihingen dataset. F1 scores for each class and overall accuracies of different methods are shown in Table 3. It can be seen from Table 3 that our method can get highest F1 scores for all categories except Car and overall accuracies, which illustrates the effectiveness of the proposed method. In addition to higher accuracies, the proposed method also used the fewest samples. Approaches proposed by Audebert et al. and Sun et al. were based FCN, therefore fully labelled samples were needed. For proposed method, only 5% samples are used for training. Overall, whether it is from the perspective of classification accuracy or practical application, proposed method is better.

	SegNet-RC	V-FuseNet	Sun et al.	Ours
Impervious	91.0	91.0	93.0	95.73
Building	94.5	94.4	95.6	97.63
Low vegetation	84.4	84.5	85.6	92.43
Tree	89.9	89.9	90.3	93.64
Car	77.8	86.3	84.5	81.57
Overall	89.8	90.0	91.2	95.57

Table 3. F1 scores for each class and overall accuracies of different methods (%).

4. CONCLUSION

This paper proposes a land cover classification approach using CNN with remote sensing data and DSM. Experimental results show that object-based voting can greatly reduce the “salt and pepper” effect and CRF optimization can capture more accurate outline of ground objects than pixel-based classification and image objects. Additionally, DSM is very helpful for land cover classification. The final overall accuracy of proposed approach is 95.57%. However, there are still some shortcomings. Compared with ground truth, the outline of ground objects in final result is still not accurate enough, which should be solved in future research. Additionally, experiments were performed in urban areas only because of lack of experimental data in a natural landscape.

REFERENCES

- Albert, L., Rottensteiner, F., Heipke, C., 2017. A higher order conditional random field model for simultaneous classification of land cover and land use. *ISPRS-J. Photogramm. Remote Sens.*, 130, 63-80.
- Audebert, N., Le Saux, B., Lefevre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS-J. Photogramm. Remote Sens.*, 140, 20-32.
- Baatz, M., Schape, A., 2000. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. Paper presented at the Angewandte Geographische Informations- Verarbeitung XII.

- Blaschke, T., 2010. Object based image analysis for remote sensing. *ISPRS-J. Photogramm. Remote Sens.*, 65, 2-16.
- Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., van der Meer, F., van der Werff, H., van Coillie, F., 2014. Geographic object-based image analysis – Towards a new paradigm. *ISPRS-J. Photogramm. Remote Sens.*, 87, 180-191.
- Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23, 1222-1239.
- Friedl, M.A., Mciver, D.K., Hodges, J.C.F., Zhang, X.Y., Muchoney, D., Strahler, A.H., Woodcock, C.E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., Schaaf, C., 2002. Global land cover mapping from MODIS: algorithms and early results. *Remote Sens. Environ.*, 83, 287-302.
- Kavzoglu, T., Mather, P.M., 2003. The use of backpropagating artificial neural networks in land cover classification. *Int. J. Remote Sens.*, 24(23), 4907-4938.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM.*, 60(6), 84-90.
- Mao, L., Hua, Y., Zhu, X.X., 2019. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp, 12416-12425),
- Otukei, J.R., Blaschke, T., 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *Int. J. Appl. Earth Obs. Geoinf.*, 12S, S27-S31.
- Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS-J. Photogramm. Remote Sens.*, 67, 93-104.
- Scott, G.J., England, R., Starns, W.A., Marcum, R.A., Davis, C.H., 2017. Training deep convolutional neural networks for land-cover classification of high-resolution imagery. *IEEE Geosci. Remote Sens. Lett.*, 14(4), 549-553.
- Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39, 640-651.
- Sun, X., Shen, S., Hu, Z., 2020. Vaihingen 2D Semantic Labeling. International Society for Photogrammetry and Remote Sensing. <http://www2.isprs.org/commissions/comm2/wg4/vaihingen-2d-semantic-labeling-contest.html> (NLPR3)(20 April 2020).
- Szostak, M., Pierezykowski, M., Likus-Cieslik, J., 2020. Reclaimed area land cover mapping using sentinel-2 imagery and LiDAR point clouds. *Remote Sens.*, 12(2), 261.
- Wai, T. Y., Ahmed, S., Nagwa, E. A., 2015. Urban land cover classification using airborne LiDAR data: A review. *Remote Sens. Environ.*, 158, 295-310.
- Yao, Y., Rosasco, L., Caponnetto, A., 2007. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2), 289-315.
- Zhao, W.Z., Du, S.H., Emery, W.J., 2017a. Object-based convolutional neural network for high-resolution imagery classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, 10(7), 3386-3396.
- Zhao, W.Z., Du, S.H., Wang, Q., Emery, W.J., 2017b. Contextually guided very-high-resolution imagery classification with semantic segments. *ISPRS-J. Photogramm. Remote Sens.*, 132, 48-60.