Research on Urban Building Extraction Method Based on Deep Learning Convolutional Neural Network

Beiru Lv\textsuperscript{1,2,3}, Ling Peng\textsuperscript{1,3}, Tong Wu\textsuperscript{1,2,3} and Ruonan Chen\textsuperscript{1,2,3}

1. Aerospace Information Research Institute, Chinese Academy of Sciences
2. University of Chinese Academy of Sciences
3. Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences

plqiqi@126.com

Abstract. Buildings’ location information is an important basic data for the construction of refined smart city. Using remote sensing images to obtain building information is both high-quality and efficiency. However, urban village building is a very special category in remote sensing image of urban area. The distribution patterns of buildings inside the urban village is quite unique, such as high building density, narrow streets and lanes, concentrated illegal buildings, which brings various high-risk security risks. Therefore, how to use remote sensing technique to detect the structural characteristics and explore the spatial distribution patterns of urban village buildings through massive urban image data is of great significance to both planning and governing urban villages. Traditional remote sensing image building extraction methods have the problems of large workload, low efficiency and slow update, and the buildings in the urban village has complex building types and serious interference of deposits. To tackle the above issues, this paper adopts the deep learning convolutional neural network algorithm which is the advanced method in the field of computer vision, and proposes a building extraction method based on high-resolution remote sensing images. In this paper, Mask R-CNN, an instance segmentation method, is used to extract urban village buildings. After trained on 678 samples, this method reached 66\% mAP on test dataset. In addition, spatial distribution patterns of the research area are analysed based on the detection result. The spatial analysis result shows that buildings in the research area are with small average building area and large building density and do not comply with regulations of People's Republic of China on the Planning and Design Standards for Urban Residential Areas.

1. Introduction

With the increasing amount of remote sensing data, remote sensing plays an increasingly important role in real-time updating of digital city databases and the researching of city internal structure. Buildings are important parts of urban areas. However, since urban village building is a very special category of urban architectures, how to use remote sensing to detect the structural characteristics and explore the spatial distribution patterns of urban village buildings through massive urban image data is of great significance to planning and governing urban villages.

At present, many scholars have conducted researches on urban villages using remote sensing data. Wei et al [1]comprehensively used multi-spectral high-resolution-1 (GF-1) images and high-resolution TerraSAR-X radar (TSX) images to extract urban village information. Zhao et al [2]used Tencent user density data, building profile data and POI data, combined with building height and density. Dare et al
used high-resolution remote sensing images to monitor and map informal settlements. Mayunga [4] used high-resolution satellite imagery to extract dense informal settlements. In general, most researches on urban villages focused on the overall identification of urban villages, and there is a lack of researches on the building detection and spatial pattern analysis inside the urban villages.

Due to the complex building types in the urban village and the serious interference of deposits, it is difficult to extract the urban village buildings based on the edge and texture features in the traditional way. Therefore, in this paper, Mask R-CNN is used for remote sensing image building extraction. And multiple analysis methods are applied on the detection results to discover the spatial distribution pattern of buildings inside urban village. Experiments show that Mask R-CNN is a high efficiency and effective method for urban village buildings extraction. And the spatial distribution of the urban village buildings in research area shows that there are some problems and deficiency of the distribution of buildings in this area, and it do not compliance with relevant regulations[5].

2. Algorithm Description

2.1. Network Structure

Mask R-CNN [6] is a multi-task deep neural network based on Faster R-CNN [7]. It combines the ideas of Fully Convolutional Networks (FCN) and Feature Pyramid Networks (FPN) [8]. Mask R-CNN has three phases: In the first phase, the feature map of input image is extracted by the Mask R-CNN backbone network (i.e. ResNet101 [9] and FPN). In the second phase, Region Proposal Network (RPN) is used to generate region proposal boxes for the target, and then the region proposal boxes are filtered to get the Regions of Interest (ROIs). In the third phase, category, location and the binary mask of corresponding building target is predicted for each ROI. The overall structure of the network is shown in Figure 1.

![Figure 1. Mask R-CNN network structure](image)

2.2. Loss Function

As shown in equation (1), the loss function of Mask R-CNN consists of three parts, which includes classification loss $L_{cls}$, boundary box detection loss $L_{bbox}$ and mask segmentation loss $L_{mask}$.

$$L = L_{cls} + L_{bbox} + L_{mask}$$  \hspace{1cm} (1)

The classification loss $L_{cls}$ is shown as equation (2).

$$L_{cls} = \frac{1}{N_{cls}} \sum_{i} \log[p_{i} p_{i} + (1 - p_{i})(1 - p_{i})]$$  \hspace{1cm} (2)
Where \( i \) represents the subscript of the ROIs in the feature map; \( N_{cls} \) represents the number of categories; \( p_i \) represents the probability that the \( i \)-th ROIs are predicted to be positive samples; \( p_i^* = 1 \) when the ROI is a positive sample, while \( p_i^* = 0 \) when the ROI is a negative sample.

The boundary box loss \( L_{bbox} \) is shown in equation (3).

\[
L_{bbox} = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^*\text{smooth}_L(t_i, t_i^*)
\]

(3)

Where \( p_i^* = \begin{cases} 1, & \text{foreground} \\ 0, & \text{background} \end{cases} \), \( \lambda, N_{cls}, N_{reg} \) should be set according to actual training needs to maintain the balance of the RPN loss value; \( t_i^* \) represents the four translation scaling parameter from positive sample ROIs to its corresponding real label.

The mask loss \( L_{mask} \) is shown in equation (4).

\[
L_{mask}(Clsc) = \text{Sigmoid}(Clsc)
\]

(4)

where \( Clsc \) represents the category of the target.

3. Data Set and Pre-processing

We collected 88 urban village building samples with the spatial resolution of 0.11m from Google Earth imagery. Regarding data augmentation, these samples were 90 degrees, 180 degrees and 270 degrees counterclockwise rotated and horizontal and vertical flipped. Finally, we obtained 678 samples of urban village buildings. Each pixel of the samples was classified as building and non-building. In this paper, the ratio of training set and validation set is 7:3.

![Image of data augmentation](image)

Figure 2. Data augmentation

4. Experiment and Algorithm Evaluation

4.1. Algorithm Implementation

This paper adopts Mask R-CNN network based on deep learning framework Keras and TensorFlow. The algorithm of using Mask R-CNN for urban village building extraction is shown in Figure 3.
Figure 3. Mask RCNN for urban village building extraction

The model was trained with batch size 32, learning rate 0.001, and the loss function no longer decreased at the 85th epoch, staying around 0.8.

4.2. Urban Village Building Detection and Accuracy Evaluation

In order to quantitatively evaluate the performance of the algorithm, this paper adopts mAP (mean Average Precision) as the evaluation criterion. mAP is an indicator for measuring the detection accuracy. A P-R curve can be drawn for each category of multiple categories of object detection based on Precision and Recall. AP is the area under the curve, mAP is the average of multiple categories of APs. Classification results confusion matrix is shown in Table 1.

Table 1. Mixed matrix of categorized results.

| Actual category | Positive | Negative |
|-----------------|----------|----------|
| Detection category | TruePositive(TP) | TrueNegative(TN) |
| False | FalsePositive(FP) | FalseNegative(FN) |

TP is the number that predicts a true positive category as a positive category; TN is the number that predicts a true negative category as a negative category; FP is the number that predicts a true negative category as a positive category; FN is the number that predicts a true positive category as a negative category.

The equation of Recall, Precision and mAP is shown below:

Recall = \frac{TP}{TP + FP} \tag{5}

Precision = \frac{TP}{TP + FN} \tag{6}
\[
\text{mAP} = \frac{\sum \text{AveragePrecision}_C}{\text{Num(Classes)}}
\]

In this paper, IOU (Intersection Over Union) is used to judge whether the detected urban village building is correct or not. When \( \text{IOU} > 0.5 \), the building is predicted as a positive sample.

\[
\text{IOU} = \frac{\text{detected building area} \cap \text{label building area}}{\text{detected building area} \cup \text{label building area}}
\]

By experimenting, we obtain 0.66 mAP on test set.

4.3. Analysis of the spatial distribution pattern of buildings in urban villages

In this paper, 189,000 square meters of area in Pingfang Village, Haidian District, Beijing was selected as the research area, the building inside this area was detected by this model and the distribution of the buildings was analyzed. Then we input the results to ArcGIS for spatial analysis including: kernel density estimation, direction analysis based on standard deviation ellipse, nearest neighbour analysis and area statistic. The results are shown in Figure 6 and Table 2.

| Average nearest neighbor spacing (m) | Average building area (m²) | Building density (%) |
|--------------------------------------|----------------------------|----------------------|
| 0.90                                 | 75.08                      | 43.75                |

Figure 4. Detection results  
Figure 5. PR curve  
Figure 6. Spatial analysis result  

Table 2. Nearest neighbour analysis and area statistic results.
The result of Kernel density estimation shows that buildings in this area are evenly distributed in each small area divided according to the road network. The ellipse direction analysis based on the standard deviation shows that the overall building orientation of the area is north-south orientation, which is in compliance with the building orientation specifications of the Beijing. The average building area of this area is 75.08 m², and the average nearest neighbor distance of the building is 0.90 m. After eliminated land under construction and urban roads based on nuclear density estimation results, the building density of the area is 43.75%, and the green space rate is 5.12%. It do not compliance with the regulations of People's Republic of China on the Planning and Design Standards for Urban Residential Areas [5].

5. Conclusion
This study proposes an urban village building extraction method based on Mask R-CNN. We train the model on 678 urban village building samples to extract the urban village buildings and analyze the spatial patterns of them. This method reached 0.66 mAP on test set. Based on the detection results, this paper analyses the spatial distribution pattern of buildings in a district of Pingfang Village, Haidian District, Beijing. The results of spatial analysis show that the buildings in this area are mainly north-south oriented, densely distributed and relatively uniform, with small average building area and large building density. However, there are still some shortcomings in this paper. The extraction accuracy of extraction of urban village buildings does not reach the accuracy of the extraction of ordinary buildings. Through the analysis of the results, it is considered that the low accuracy is mainly caused by the inherent characteristics of the urban village buildings, such as large building density, complex building types, and many disturbances such as building materials. Also the insufficient of image resolution and the number of samples is also effect the extraction accuracy. The above problems will be the focus of follow-up research.

6. Acknowledge
This work was supported by the Beijing Municipal Science and technology Project (grant No. Z191100001419002).

References
[1] C. Wei, T. Blaschke, and H. Taubenböck, "Monitoring of “urban villages” in Shenzhen, China from high-resolution GF-1 and TerraSAR-X data," SPIE Remote Sensing, 2015.
[2] Y. Zhao, G. Chen, G. Chen, X. Liu, and N. Niu, "Coupling multi-source big data extraction of village buildings in the city -- taking tianhe district of guangzhou city as an example, "Geography and GeographicInformation Science, vol. 34, no. 05, pp. 7-13+1, 2018.
[3] P. M. Dare and C. S. Fraser, "Cover: Mapping informal settlements using high resolution satellite imagery," International Journal of Remote Sensing, vol. 22, no. 8, pp. 1399-1401, 2001.
[4] S. D. Mayunga and S. D. Mayunga, "Semi-automatic building extraction in informal settlements from high-resolution satellite imagery," 2006.
[5] Urban residential planning and design standards. (2018).
[6] G. G. Kaiming He, Piotr Dollár, Ross Girshick, "Mask R-CNN," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
[7] S. R. K. H. R. G. J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," presented at the IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.
[8] T. Y. Lin, P. Dollár, R. Girshick, K. He, and S. Belongie, "Feature Pyramid Networks for Object Detection," 2016.
[9] T. Hirata, T. Kuremoto, M. Obayashi, S. Mabu, and K. Kobayashi, "Deep Belief Network Using Reinforcement Learning and Its Applications to Time Series Forecasting," Cham, 2016, pp. 30-37: Springer International Publishing.