

CNN BASED DETECTION OF BUILDING ROOFS FROM HIGH RESOLUTION SATELLITE IMAGES

Liu Hang, Guoyin Cai

(Department of Remote Sensing Science and Technology, School of Geomatics and Urban Informatics, Beijing University of Civil Engineering and Architecture, 100044)

Email: cgyin@bucea.edu.cn

KEY WORDS: Building roofs; CNN; VHR images

ABSTRACT:

The detection and reconstruction of building have attracted more attention in the community of remote sensing and computer vision. Light detection and ranging (LiDAR) has been proved to be a good way to extract building roofs, while we have to face the problem of data shortage for most of the time. In this paper, we tried to extract the building roofs from very high resolution (VHR) images of Chinese satellite Gaofen-2 by employing convolutional neural network (CNN). It has been proved that the CNN is of a higher capability of recognizing detailed features which may not be classified out by object-based classification approach. Several major steps are concerned in this study, such as generation of training dataset, model training, image segmentation and building roofs recognition. First, urban objects such as trees, roads, squares and buildings were classified based on random forest algorithm by an object-oriented classification approach, the building regions were separated from other classes at the aid of visually interpretation and correction; Next, different types of building roofs mainly categorized by color and size information were trained using the trained CNN. Finally, the industrial and residential building roofs have been recognized individually and the results have been validated individually. The assessment results prove effectiveness of the proposed method with approximately 91% and 88% of quality rates in detection industrial and residential building roofs, respectively. Which means that the CNN approach is prospecting in detecting buildings with a very higher accuracy.

1. INTRODUCTION

High-resolution remote sensing images can provide massive surface feature information with rich texture and spectral characteristics, so they have been widely used in map mapping. As one of the most important components of urban regional features, accurate and timely acquisition of building information plays a very important role in mapping, urban planning, land use survey, digital city and other fields. Fast, accurate and intelligent extraction of building information has always been one of the most important contents in the field of remote sensing image processing [1].

In recent years, researchers have tried to use a variety of methods to achieve automatic extraction of buildings from high-resolution remote sensing images. Anwen et al. has integrated image information such as geometric structure and spectral characteristics of buildings to achieve the extraction of buildings [2]. Zhao Lingjun et al. used the idea of layer-by-layer processing to detect the geometric information of buildings and realize accurate extraction of buildings [3]. Houlei et al. use Hough transform to extract the outline of buildings by combining geometric and grey features [4]. Gong Danchao et al. proposed a method of building detection based on boundary line, which realized automatic building extraction. However, these traditional algorithms only pay attention to the geometric features of buildings [5]. Although they achieve the purpose of automatically extracting buildings, they are not accurate, are

greatly affected by the environment, and rely on manual operation. On the whole, there are many difficulties in automatically extracting buildings from remote sensing images at present, which are mainly manifested in the following three aspects:

(1) The size difference of urban buildings is relatively large, and buildings that are automatically extracted according to geometric or structural characteristics are usually larger buildings, and smaller buildings will be missed or wrongly detected;

(2) The general appearance of buildings is regular (such as square residential buildings), but the appearance of urban buildings is complex and varied with different shapes (such as oval gymnasiums). If only manually designed shape features have no context information, it is difficult to extract all types of buildings.

(3) Urban buildings are often shaded by dense trees, and low-rise buildings are often shaded by high-rise buildings, which makes it extremely difficult to extract complete and complete buildings.

Among many different types of ground objects such as buildings, water bodies, vegetation, etc., the shapes of the ground objects such as water bodies and vegetation are mostly irregular shapes, accounting for a small proportion, and the importance of these ground object types in urban planning and construction of smart cities is far less than that of buildings and

roads, so extracting buildings and roads for change monitoring of urban construction is the core idea of this paper.

2. RESEARCH BACKGROUND

Extracting buildings from high resolution remote sensing images. In a broad sense, high-resolution images can be expressed as high spatial resolution, high temporal resolution and reduction of spectral band number. In a narrow sense, high-resolution images refer to high resolution. The high-resolution remote sensing images discussed in this paper are not specifically defined as narrow definition. As the definition of high resolution in the field is not consistent, some are defined by 2 m or sub-meter level. This paper adopts the standards used in most recent literatures and considers remote sensing images with spatial resolution up to meter level as high resolution remote sensing images.

Since the rise of remote sensing technology in the last century, with the increasing number of remote sensing satellites launched into space by various countries in the world and the maturity and progress of shooting and other related technologies, the number of remote sensing images available to people has increased rapidly, and the resolution of the images themselves is also continuously improving. All these make the application prospect and scope of remote sensing technology and the remote sensing data obtained continuously expand. Nowadays, remote sensing image data has become the main carrier for people to study the ground space information. People can acquire various high-resolution earth observation data quickly and in real time through remote sensing satellites and other ways. The development and progress of remote sensing technology have opened a new window for human to study the earth, and the application of remote sensing images has been paid more and more attention by various countries in the world.

Table 1 shows some major remote sensing satellites in various countries in recent decades. In the table, PAN is a panchromatic remote sensing image and MS is a multispectral remote sensing image.

Satellite	Launch time	Country	Spatial resolution (PAN)	Spatial resolution (MS)
QuickBird	2001	U.S.A	0.61m	2.44m
Orb View3	2003	U.S.A	1m	4m
Arrow 2	2006	The Republic of Korea	1m	4m
ALOS	2006	Japan	2.5m	10m
Remote Sensing Satellite II	2007	China	2m	-
WorldView-1	2007	U.S.A	0.5m	-
GeoEye-1	2007	U.S.A	0.41m	1.65m
WorldView-2	2009	U.S.A	0.5m	1.8m
Helios 2B	2009	France	0.25m	-
ZY-3	2012	China	2.1m	6m
Gaofen-2	2014	China	1m	4m
Beijing-2	2015	China	1m	4m

Table 1.High resolution remote sensing satellite launched successfully

Table 1 show that the resolution of remote sensing images that can be taken by remote sensing satellites launched into space by various countries in the world is increasing year by year. The cameras are civilian and some commercial satellites and the shooting resolution of some military reconnaissance satellites can reach centimetre level. With the increasing number of satellites and their real-time collection, the remote sensing data available to people are multiplying, but only 5% of such a large

amount of remote sensing data has been properly utilized. Therefore, once pointed out in the construction of the digital earth that while we are faced with thirst for knowledge, a large amount of data is not utilized.

Therefore, it is increasingly important and urgent to develop new analysis and processing methods for remote sensing images, especially the extraction technology for certain specific target information. The traditional interpretation method of manual interpretation not only requires the staff to have professional knowledge in relevant aspects, but also with the increase of remote sensing data and the improvement of image resolution, the interpretation work will become more time-consuming and thus affect the accuracy of the results. Nowadays, all countries in the world have increased their investment in the development of new remote sensing image processing methods. Therefore, it is very meaningful to study this topic.

Using high-resolution remote sensing images to extract artificial ground object information is not only more efficient than manual on-the-spot investigation, but also people can obtain desired remote sensing data at any time when shooting from space. Therefore, the use of such data to extract information from artificial targets has gradually been adopted in many fields such as urban construction, environmental monitoring, etc. Artificial ground objects are the main content of high-resolution remote sensing images in urban areas, most of which are roads and buildings. Buildings on remote sensing images have obvious positioning characteristics. The detection of buildings is not only related to the automation level of ground object mapping, but also provides the premise for image matching and understanding. Secondly, as the main part of artificial ground objects in urban areas, buildings can be identified and extracted from remote sensing images, which can provide help for cartography, data acquisition and automatic updating of geographic information systems. So, how to identify and extract building roofs is one of the important research topics in remote sensing image information extraction

3. DEVELOPMENT STATUS OF ROOF EXTRACTION OF BUILDINGS

Buildings are an important feature of urban areas. The technology of extracting buildings has been widely used in the fields of urban mapping, urban planning, and geographic information engineering and military reconnaissance. In these applications, many automatic building extraction methods have been developed.

In the modern research on image understanding, the extraction of artificial ground objects is one of its main research contents, among which the automatic extraction of buildings, as a hot research topic with wide application prospects, has attracted worldwide attention in recent decades. Summarize the methods of building extraction from remote sensing images that have been proposed so far. Generally speaking, they can be divided into the following two categories:

The first one is the method of extracting DSM (Digital Surface Model) data. This kind of method studies the height difference between buildings and surrounding ground objects to realize the detection and extraction of buildings.

The second type is to use the knowledge and processing methods of image processing, image analysis, pattern recognition and other disciplines to complete the detection and

extraction of buildings in images. Because this kind of method does not need other auxiliary data, it is adopted in many research and practical applications. Research on automatic building extraction can be traced back to the 1980s. After entering the 1990s, as the research value in this field has been gradually widely recognized, its research has also gradually begun.

The research on automatic building extraction technology starts from the feature analysis of the ground floor, and then gradually combines with the new methods in the current pattern recognition and image processing fields to move towards a fully automatic and more accurate direction. The following will introduce the method of extracting building roofs in this article.

4. CNN INTRODUCTIONS AND COMPARISON

4.1 CNN Brief

Convolutional neural network (CNN) is a deep neural network model with convolution layer, which has become a research hotspot in the field of speech analysis and image recognition [6]. Its weight sharing network structure makes it more similar to biological neural network, reducing the complexity of the network model and the number of weights. Structurally, CNN is a multi-layer neural network, mainly composed of convolution layer, sub-sampling layer and full connection layer. Each layer is composed of a plurality of two-dimensional planes, and each plane is composed of a plurality of independent neurons. Figure 1 shows the classic structure of LeNet-5 [7].

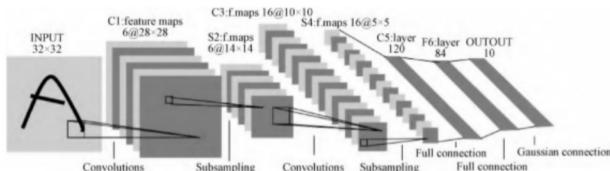


Figure 1. LeNet-5 structure

In the convolution layer of the convolution neural network, one neuron is only connected with some neighbouring neurons. In a convolution layer of CNN, there are usually several feature maps. Each feature map consists of some neurons arranged in a rectangle. The neurons of the same feature map share weights, and the weights shared here are convolution kernels. Convolution kernel, which is usually initialized in the form of random decimal matrix, will be learned in the training process of the network to obtain reasonable weights. The direct benefit of shared weight (convolution kernel) is to reduce the connection between the layers of the network, while reducing the risk of fitting.

Subsampling is also called pooling, and usually has two forms: mean subsampling and max subsampling. Subsampling can be regarded as a special convolution process. Convolution and subsampling greatly simplify the complexity of the model and reduce the parameters of the model.

4.2 CNN and Traditional Pattern Recognition Neural Networks

Traditional pattern recognition neural network (NN) algorithm is based on gradient descent and learning from a large number of input sample feature data, and has the ability to identify and classify different target samples. These traditional pattern

recognition methods include KNN, SVM, NN and other methods. They have an unavoidable problem, that is, they must manually design algorithms to extract features from input images. In the process of feature extraction, various invariance problems, the most common ones, need to be considered, such as rotation invariance, illumination invariance, and scale invariance, rotation invariance through calculation of image gradient and angle, and illumination influence avoidance through normalization. The scale pyramid is constructed to realize scale invariance, in which SIFT and SURF are typical representatives of such features. In addition, based on contour HOG features, LBP features, etc., then the feature data is used as input to select suitable machine learning methods, such as KNN, SVM, etc. to realize classification or recognition. One of the biggest drawbacks of these methods is that the design process of feature extraction is completely dependent on human beings. There are too many human factors, which do not give full play to the ability of machine to actively learn and extract features. The advantage is that people can completely control every detail of feature extraction and every feature data.

The deep learning method represented by CNN realizes the recognition and classification of objects. The feature extraction is completely handed over to the machine. The whole feature extraction process does not need manual design and is completed automatically by the machine. Feature extraction is realized through convolution of different filter, thus distortion and illumination can be kept unchanged to a certain extent, and scale invariance can be realized through maximum pooling layer sampling. While keeping the three invariance of traditional feature data, manual design details can be minimized in feature extraction method, and computer computing capability can be brought into play through supervision and learning to actively search for appropriate feature data.

Therefore, convolutional neural network has the following advantages over traditional feature extraction and pattern recognition methods:

- (1) Training is relatively easy, without complicated feature extraction process, which can be said to reduce the learning threshold of image recognition, so that more people who know the data find a shortcut to learn image processing and computer vision.
- (2) Convolution layer reduces the number of parameters and memory requirements by sharing weight parameters compared with traditional neural networks.
- (3) The distortion, distortion and pixel migration of the image remain stable and have certain invariance characteristics.

5. ROOF DETECTION OF HIGH RESOLUTION SATELLITE IMAGES BASED ON CNN

5.1 Research Methods and Processes

Compared with the traditional shallow machine learning classification model SVM and Boosting algorithm [8-9], the deep convolution neural network in the deep learning algorithm has deeper network structure and stronger nonlinear fitting ability, and can extract information layer by layer from pixel-level original data to abstract semantic concepts. From the point of view of feature extraction, compared with the traditional feature extraction algorithm SIFT (Scale-Invariant Feature Transform) [10], the deep convolution neural network can automatically learn features from a large number of data, instead of using manually designed features. Deep convolution neural

network has strong learning ability and efficient feature expression ability, which makes it have outstanding advantages in extracting global and local features of images, and brings new ideas for image segmentation.

This paper proposes a cascade full convolution neural network based on depth learning algorithm to extract buildings from remote sensing images. In this paper, the original remote sensing image is defined as R:

$$R = \{R^{(1)}, R^{(2)}, R^{(3)}, \dots, R^{(n)}\} \quad (1)$$

Where: R (n) is a single multi-band remote sensing image. Defining the label image of multi-channel cell level as G:

$$G = \{G^{(1)}, G^{(2)}, G^{(3)}, \dots, G^{(n)}\} \quad (2)$$

Where: G(n) is a single label image, and the label image is also called the true value. The algorithm proposed in this paper needs to predict the pixel-level building segmentation image \hat{G} from the remote sensing image S. The prediction result is a binary image, that is, the pixel point with a value of 0 is non-building and the pixel point with a value of 1 is building.

For roof inspection, CNN's processing flow is as shown in Figure. 2. Firstly, the training samples are simply regularized and adjusted to a uniform size. in order to prevent poor quality data from appearing in the first few samples of the training samples from adversely affecting the training process, the training process adopts batch processing, i.e. a fixed number of training samples are randomly selected as a small sample input each time, and the BP algorithm is used to update the weight value of each small sample once, so as to stop the training when a certain number of iterations are reached or the error reaches a given threshold. Input the test data into the trained CNN model, and finally get the classification results through forward propagation.

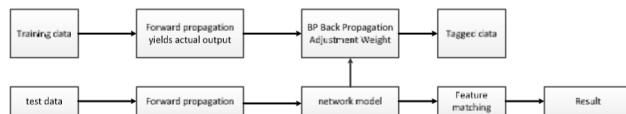
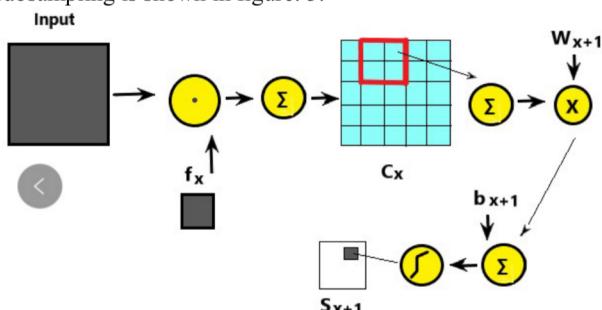


Figure 2. CNN processing flow

The forward propagation phase mainly includes convolution and sampling. The detailed process of convolution and subsampling is shown in figure. 3.



network. This paper introduces a very important "blocking" mechanism to ensure that the computer will not be down due to insufficient display memory capacity, and also to ensure the normal training of the network.

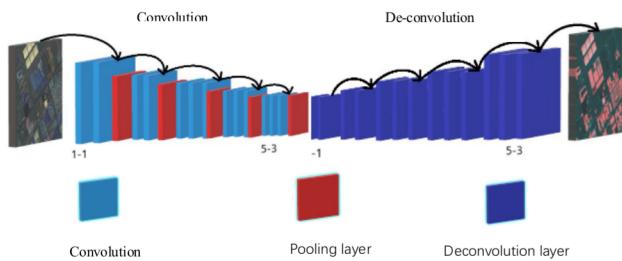


Figure 4. Cascade full convolution neural network structure

6. EXPERIMENTS AND RESULT ANALYSIS

The data used in this experiment is Gaofen-2, which is the image data of Changchun in 2016 and 2019 respectively. And find out the areas with large roof changes from 2016 to 2019, and select three areas with obvious changes. The extraction results are shown in Figure 5.

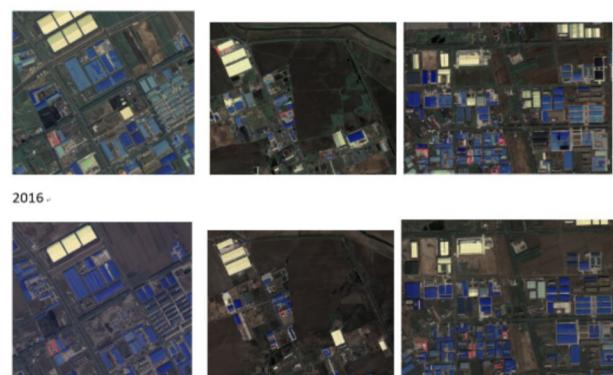


Figure 5.Comparison of original images

The extraction results are shown in Figure 6.



Figure 6. Extraction Results of Cascaded Full Convolutional Neural Networks

The comparative results are shown in Figure 7: Red areas show where no building exist in either period. Green areas show where no building exist in 2016, while buildings exist in 2019. Maroon areas show where buildings exist in 2016, while no building exist in 2019.

Sea green areas show where buildings exist in both 2016 and 2019.

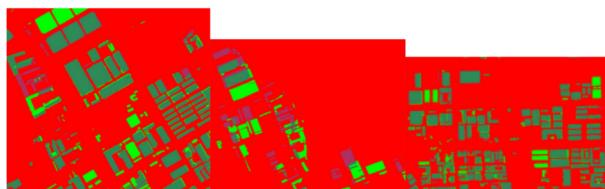


Figure 7. The changes between 2016 and 2019

In this paper, a cascade full convolution neural network structure is proposed. The traditional full convolution neural network is improved, cascade structure is introduced into the deep convolution neural network, and the flow of information in the network is improved by using methods such as feature reuse and feature enhancement. In addition, the phenomenon of gradient disappearance in the network reverse propagation process is avoided, and the depth of the network layer can be greatly increased. The addition of empty convolution can ensure that the convolution neural network increases the receptive field without pooling processing, so that the output characteristic image of each convolution contains more information, which is helpful to fuse the global information and local information of the image and improve the segmentation accuracy. The idea of image blocking is introduced to ensure that the network can be trained normally, and finally a network model with higher accuracy is obtained. The advantage of the network structure proposed in this paper is that only one original remote sensing image needs to be input, and no other image pre-processing methods such as cropping, scaling, whitening, principal component analysis, normalization, etc. are required, so that a high-precision building extraction result map with the same size can be obtained. Finally, the building extraction accuracy of this paper is better than other mainstream methods.

In the future work, we should try to adopt deeper network structures, such as ResNet with more than 100 network layers, Inception with smaller convolution kernel and wider network structure, etc. These frontier works in the field of visual recognition can be expanded in the field of remote sensing image processing. Super-parameters of deep convolution neural network gradually abandon empirical values, and Bayesian optimization method is needed to find the optimal value. Derivation of more adaptive activation functions and loss functions with better performance will improve the accuracy, which are the contents to be explored in future work.

7. OUTLOOKS

Extracting buildings from high-resolution remote sensing images to avoid other targets has always been a very complex problem, requiring a lot of time and effort. Although this article has achieved some phased results, limited by the lack of time and ability, there is still room for improvement. In the future, it will try to do further study and research from the following aspects:

- (1) To further improve the image segmentation algorithm, the existing ground object segmentation and recognition methods have more false segmentation. In future experiments, a better segmentation method can be selected, and the segmented ground object can be used as input data and added into the training set to achieve higher recognition accuracy.
- (2) Increase the amount of training data. Compared with various other data sets on the network, the current amount of training data is still less, which will lead to inaccurate training models and easy over-fitting.

(3) On the premise of increasing data volume, further improve the depth of the model, extract deeper feature information, improve the network, increase various strategies, and avoid the interference of artificial ground objects except buildings.

REFERENCES

- [1] Yan Y. Research review on the extraction of high spatial resolution remote sensing images. *Digital Technology and Application*, 2012(7):75-78.
- [2] An W, Yang J F, Zhao Xi, et al. Building automatic extraction of high resolution remote sensing images. *Science of Surveying and Mapping*, 2014, 39(11):80-84.
- [3] Zhao L J. Research on extraction method of high resolution SAR image buildings. Changsha: National Defense Science and Technology University, 2009.
- [4] Hou L, Yin D, You X J. The automatic extraction method of buildings in a remote sensing image. *Computer simulation*, 2006, 23(4):184-187.
- [5] Gong D C, Zhang Y S. Research on the semi-automatic extraction of buildings based on aerial images. *Surveying and mapping*, 2002 (10):15-18.
- [6] Bottou L, Bengio Y, Lecun Y. Global training of document processing systems using graph transformer networks. *Computer Vision and Pattern Recognition*. 1997, 6:564-579.
- [7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012:1097-1105.
- [8] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, 20(3):273-297.
- [9] Yoav Freund, Robert Schapire, E. A Decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997, 55(1):119-139.
- [10] Lowe, David G. Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer*, 1999:1150-1157.

© 2020. This work is published under
<https://creativecommons.org/licenses/by/4.0/>(the “License”). Notwithstanding
the ProQuest Terms and Conditions, you may use this content in accordance
with the terms of the License.