The Building Area Recognition in Image Based on Faster-RCNN

Xuguang Wang
North China Electric Power University (Baoding)
Baoding, China
wang_xuguang@163.com

Qin Zhang
North China Electric Power University (Baoding)
Baoding, China
prototypef3@163.com

Abstract—**Building area recognition technology is a rising up research topic in the field of pattern recognition and it has a wide application prospect in the fields of military, city planning and other fields. This paper combines building area recognition algorithm based on Faster-RCNN and some of the problems in it are discussed respectively and the corresponding solutions are given. Through grabbing a streetscape at Baidu API and real shot in street with SLR camera, building my own image dataset and doing some corresponding training and tests based on Faster-RCNN after images are annotated. The experimental results show that the algorithm above-mentioned can achieve better recognition results. The main work of this article is collecting a large number of image samples and building own image dataset; annotating the images; building and training building recognition model based on Faster-RCNN and combining ZF convolutional neural network to do building area recognition.**

Keywords—Building recognition; Image dataset; Faster-RCNN; Convolutional neural network

I. Introduction

In recent years, with the continuous improvement computing speed of computer, identifying buildings through computer has been more and more widely used. In addition, the recognition of urban buildings is a major application field of pattern recognition, and it plays a very important role and has important scientific significance and practical value in the field of city planning and management, city disaster monitoring, city environmental pollution monitoring, geographic information database updates and construction of digital city, monitoring of military target and force deployment and so on. In the above applications, buildings are important features of the urban or suburban surface, and buildings are the target of obtaining information, Therefore, it has become an important task in the field of application to study how to automatically identify the area of the building in the image, and it has also become one of the most important content in the image understanding.

One of the basic tasks of image understanding is to identify and extract artificial targets from satellite images or street view images. In particular, buildings are an important feature of urban areas. The technology of extracting buildings is widely used in the fields of urban mapping, urban planning, geographic information engineering and military reconnaissance. In these applications, there are many ways to identify buildings automatically.

Automatic recognition of buildings refers to the recognition of the building area and obtain the location and the height of a building from a single (multiple) aerial satellite or a streetscape image with a computer as an auxiliary mean. Generally speaking, problems can be described as giving one or more images of a certain area, automatically detecting the buildings in the area and obtaining relevant information. Specific applications will be different because of different requirements.

II. Related Work

The research on building detection technology can be traced back to the mid and late 80s of last century. The research team, led by the Nevatia [1] in University of Southern California, was the first to carry out the relevant work under the support of United States Defense advanced research project management agency, its research results represent a direction of research in this field. Since then, the significance of building recognition technology has been recognized by more scholars. O. Henricsson [2], F.Lang [3] and H.Wiman [4] studied how to extract 3D information of buildings from images to reconstruct buildings.

Since convolutional neural network CNNs [5] has been applied to various fields, the target recognition [6], voice recognition [7] and other fields have developed rapidly. In the field of target recognition, after the RCNN [8] model and Fast-RCNN [9] model proposed by Girshick R and others before, in order to further reduce the running time of the detection network, Microsoft Shaoqing Ren proposed the latest target detection method Faster-RCNN [10]. At present, there are not many researches on building recognition and extraction based on Faster-RCNN model. But in real life, a lot of data are obtained from the camera on the street or satellite images. And the analysis of buildings is an indispensable part of the life with science and technology. In order to obtain information from these images and save manpower, we use deep convolutional neural network to effectively identify the buildings in the images, which depletes the time consumed by human-eye recognition.

III. Faster Rcnn

Faster-RCNN designs a region proposal network (RPN) to generate the recommended area (region proposals). The emergence of RPN takes the place of previous methods such as Selective Search [11] and Edge Boxes [12]. It shares the convolution feature of the whole graph with the detection

978-1-5386-6057-7/18/$31.00 ©2018 IEEE
DOI 10.1109/SDPC.2018.00130

network, making the area recommendation detection almost no time consuming, greatly reducing the detection time and improving the detection efficiency. Both the region proposal network (RPN) and the previously proposed Fast-RCNN require an original feature extraction network, that is, the following gray box:

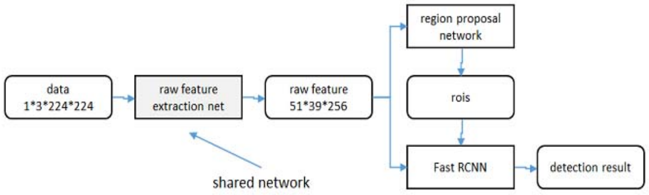


Fig.1. The flow process diagram of target recognition [17]

The previous target detection network needs to use the regional recommendation algorithm to speculate the target location first, such as SPPnet and Fast R-CNN, which have reduced the time of detection network operation, but the calculation of area proposal is still time-consuming. So, under such a bottleneck, RBG and Kaiming He and others give Region Proposal to CNN, which begins with the RPN (Region Proposal Network) area recommendation network to extract detection areas. The test results on the PASCAL VOC2007 dataset showed that compared to Selective Search (SS), the recall rate of the RPN method decreased little when the candidate regions generated by each graph fell from 2000 to 1000 and then dropped to 300, indicating that the purpose of using the RPN method was more explicit.

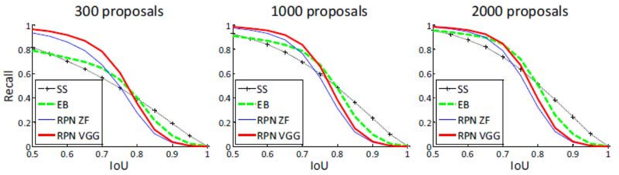
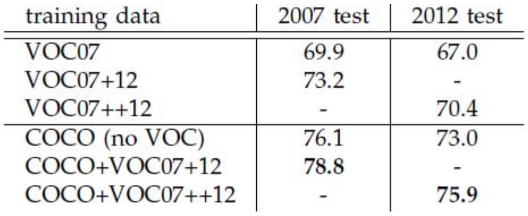


Fig.2. Recall rate curves of different methods tested on PASCAL VOC2007 dataset [10]

Using a larger Microsoft COCO database for training, the test results are about 6% higher than the accuracy rate directly on the PASCAL VOC2007, which indicates that the mobility of Faster-RCNN is good and there is no fitting phenomenon.



training data	2007 test	2012 test
VOC07	69.9	67.0
VOC07+12	73.2	-
VOC07++12	-	70.4
COCO (no VOC)	76.1	73.0
COCO+VOC07+12	**78.8**	-
COCO+VOC07++12	-	**75.9**

Fig.3. Comparison of test accuracy of different datasets [10]

A. Region Proposal Networks

RPN (Region Proposal Networks) is a network [13] based on total convolution. It can simultaneously predict the target area frame of each location of the input picture and the score of target, that is, the probability value of the real target. RPN is trained by the way of end-to-end in order to generate high quality region suggestion box for Faster-RCNN classification detection. Because our ultimate goal is to make the original Fast-RCNN object detection network [14] share the calculation of convolutional layer, then using the full convolution network to simulate the process [15]. The RPN structure is shown in the following figure:

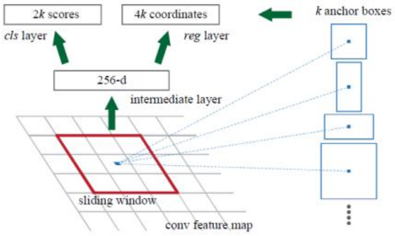


Fig.4. Region Proposal Network (RPN) [10]

B. Loss Function

Following the definition of multi-task loss, the objective function is minimized. The loss function of an image in Faster-RCNN is defined as Equation (1).

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*)$$

$$+\lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (1)$$

In Equation (1), i is the index number of anchor in mini-batch; p_i is the possibility of anchor predicted as a target; *GT* tags: $p_i^* = \begin{cases} 0\ negative\ label \\ 1\ positive\ label \end{cases}$;$t_i = \{t_x, t_y, t_w, t_h\}$ is a vector that representing the 4 parameterized coordinates of the predicted bounding box.t_i^* is the coordinate vector of ground truth corresponding to positive anchor, shown in Equation (2)-(5).

$$t_x = \frac{x - x_a}{w_a},\ t_y = \frac{y - y_a}{h_a}, \quad (2)$$

$$t_w = log\left(\frac{w}{w_a}\right),\ t_h = log\left(\frac{h}{h_a}\right), \quad (3)$$

$$t_x^* = \frac{x^* - x_a}{w_a},\ t_y^* = \frac{y^* - y_a}{h_a}, \quad (4)$$

$$t_w^* = log\left(\frac{w^*}{w_a}\right),\ t_h^* = log\left(\frac{h^*}{h_a}\right), \quad (5)$$

$L_{cls}(p_i, p_i^*)$are two categories, that is, target and non target logarithmic loss, shown in Equation (6).

$$L_{cls}(p_i, p_i^*) = -log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \quad (6)$$

$L_{reg}(t_i, t_i^*)$ is regression loss, calculation with $L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$, R is smooth L1 function.

The improvement on generating proposed area of Faster RCNN makes the detection efficiency improved effectively and provides the possibility of building the Faster-RCNN model for building identification.

IV. Establishment of Image Database

At present, the main databases are PASCAL VOC2007 and PASCAL VOC2012. If you need to identify domestic buildings, you need to make data sets yourself. Make data sets in accordance with the format of the PASCAL VOC2007 dataset and most of the pictures come from the streetscape of Taiyuan, Shanxi. The road distribution map of the main urban area of Taiyuan is shown below, including spatial location, road length, road level and so on. The coordinate system is WGS1984. It can be divided into six grades: national highway, provincial road, county road, Township Village Road, nine grade road and other roads.

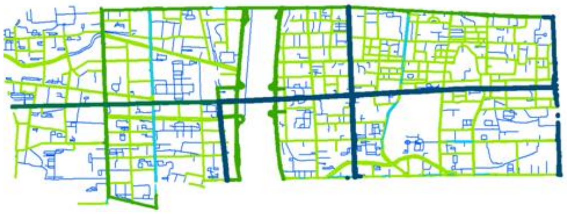


Fig.5. Road distribution map of the main urban area of Taiyuan

The building distribution map of the main urban area of Taiyuan is shown in Figure 6, including the following information: spatial location information (latitude and longitude, contour range, coordinate system is WGS1984); house property information (number, existence state, address, number of floors, structure type, building age, area, perimeter, etc.).

Fig.6. The building distribution map of the main urban area of Taiyuan

A. Data Sampling

According to their coordinate information and street information, we can grab the street view on the map. The pictures after grabbing are renamed as VOC2007 data format.

Among them, two groups of street view pictures of Taiyuan were taken for comparison training, one group of 500, and one group of 2000. There are also 1000 camera scenes for Baoding street view for more tests.

B. Annotation

Before training, annotate the data first, select the building frame in the picture and save the coordinates. As shown in the following figure:

Because of human tagging, it cannot avoid the situation of annotation error, which may lead to the length of the coordinates beyond the limit of the coordinate frame, which will lead to the loss of the model training in the subsequent model is infinitely large, which will cause the model not convergence, and cannot detect the target. In order to avoid this situation, this paper puts forward a test link after the data annotation, which will output the name of the picture corresponding to the coordinates or the difference of the coordinates that exceed the picture size and make a new annotation.

After the end of the data tag, transform the txt documents that generated to store coordinates to the files in XML format, and proportionately generate the train, trainval, val, and test files based on the generated XML files. Since this dataset is not large enough compared to the dataset of Girshick R, set the proportion of files in train and trainval to 70%, that is, 70% of the data is used for training. And set different learning rate to train the ZF-net [16] model and test the effect.

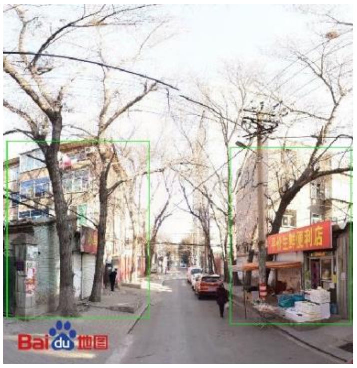
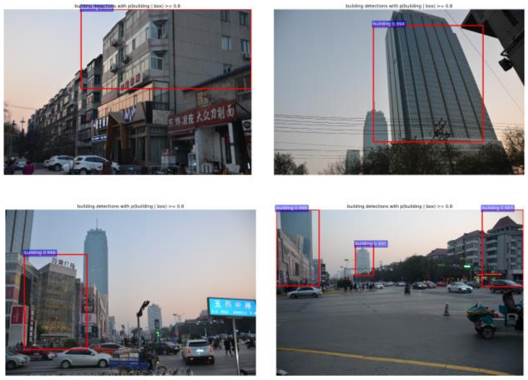


Fig.7. Picture data annotation

V. Experimental Results and Analysis

After training, a model file was generated. First, test the model effect of the 500 labeled images dataset.



building detections with p(building | box) >= 0.8
building detections with p(building | box) >= 0.8
building detections with p(building | box) >= 0.8
building detections with p(building | box) >= 0.8

Fig.8. Model effect of the 500 labeled images

The following are the results of training model with 2000 images.

From the above test results, we can see that the larger number of data sets is, the better recognition effect is, the more accurate the identification of buildings is, and more target buildings are identified by one picture. In addition, the above test pictures are all taken by camera, and it is found that the detection network takes a long time to identify the test picture of high resolution, about 0.4s, while the resolution of images that grabbed from the internet are lower, the detection time is shorter, about 0.04s.

The following are the test results of some web grabbed streetscape pictures.

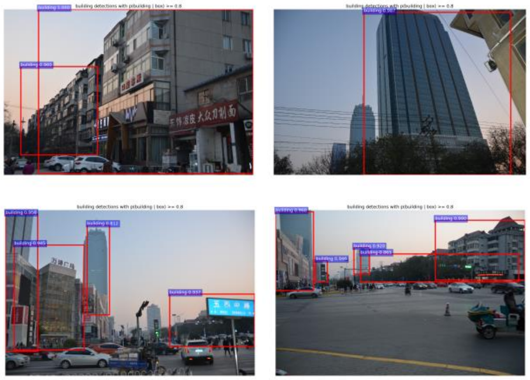


Fig.9. Model effect of the 2000 labeled images

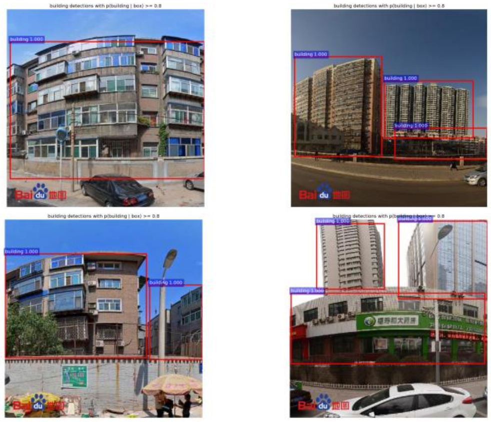


Fig.10. Test result of the model trained by 500 images dataset

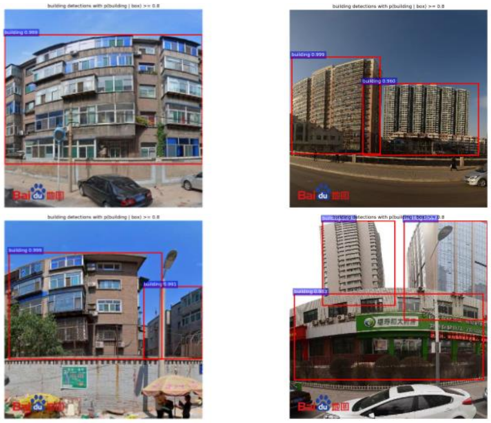


Fig.11. Test result of the model trained by 2000 images dataset

It can be seen from the test results that the test results of the two models are not significantly different. It shows that the detection effect is related to the resolution of the test picture. Because the training dataset is mostly composed of streetscape pictures grabbed by the internet, the effect of testing streetscape pictures is better. The number of datasets has little effect on the test results, but when testing images of different resolutions, large number of training data has an obvious advantage.

VI. Conclusions

Under the circumstance of different number of datasets, study to use Faster RCNN network to detect the building area in the outdoor scene of the SLR camera and internet grabbed pictures respectively. It is found that the larger the number of training dataset is, the better the recognition effect is. In addition, the higher the resolution of the test picture is, the longer the recognition time will be. And when model learning is carried out under different learning rates, it is found that the high rate of learning may lead to the problem of gradient explosion. The loss value is infinite, resulting in the final model cannot identify buildings. The learning rate is too small, which will lead to prolonged training time, and it is likely to achieve local optimum and identify problems. The final model was trained at a learning rate of 0.0001.

References

[1] A.Huertas and R.Nevatia, Detecting Buildings in Aerial Images, *Computer Vision, Graphics and Image Processing,* 41(2), Feb 1988: 131-152.

[2] O.Henricsson, The role of color attributes and similarity grouping in 3-D building reconstruction, *Computer Vision and Image Understanding* 72(2). 1998.163-184.

[3] F.Lang and W.Forstner, Surface reconstruction of man-made objects using polymorphic mid-level features and generic scene knowledge, *Z.plrotogramm, Fernerkundung* 6/96, 1996, 193-201.

[4] H.Wiman, Least squares matching for three dimensional building reconstruction, in *proceedings, Automatic Extraction of Man-Made Objects from Aerial and Space Imagers(II)* Birkhauser. Basel, Switzerland, 1997,pp.223-232.

[5] Lecun Y, Boser B E, Denker J S, et al. Handwritten digit recognition with deep neural networks[C]//2016 IEEE Intelligent Vehicles Symposium, IV 2016 IEEE, Gotenburg, Sweden, June 19-22. Sweden: IEEE, 2016: 1115-1120.

[6] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//13th European Conference on Computer Vision, ECCV 2014, Zurich, Switzerland, September 6-12, 2014. Springer: Cham, 2014:346-361.

[7] Hinton G, Deng L, Yu D., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[8] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, United states, June, 23-28, 2014. USA: IEEE, 2014: 580-587.

[9] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision, Santiago, Chile, December 11-18, 2015. USA: IEEE, 2015: 1440-1448.

[10] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region-proposal networks[C]//29th Annual Conference on Neural Information Processing Systems, NIPS 2015, Montreal, QC, Canada, December 7-12, 2015. Canada: NIPS Conference, 2015: 91-99.

[11] Uijlings J R R, Van De Sande K E A, Gevers T, er al. Selective search for object recognition[J]. International journal of computer vision, 2013,104(2): 154-171.

[12] Zitnick C L. Dollar P. Edge Boxes: Locating object proposals from edges[C]//13th European Conference on Computer Vision, ECCV 2014,

Zurich, Switzerland, September, 6-12, 2014. Zurich: Springer, Cham, 2014: 391-405.

[13] Long J. Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, United states, June 7-12, 2015. USA: IEEE, 2015: 3431-3440.

[14] R. Girshick, “Fast R-CNN”, in IEEE International Conference on Computer Vision(ICCV), 2015.

[15] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation”, in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR),*2015.

[16] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision, Zurich, Switzerland, September, 6-12, 2014. Zurich: Springer, Cham, 2014: 818-833.

[17] https://img-blog.csdn.net/20160415134137682