Text Detection in Street View Images by Cascaded Convolutional Neural Networks

Po-Wei Chang
Dept. of Computer Science
and Information Engineering
National Central University
Taoyuan, Taiwan
Email: zedest@g.ncu.edu.tw

Guan-Xin Zeng
Dept. of Computer Science
and Information Engineering
National Central University
Taoyuan, Taiwan
Email: zadays9057@gmail.com

Po-Chyi Su
Dept. of Computer Science
and Information Engineering
National Central University
Taoyuan, Taiwan
Email: pochyisu@csie.ncu.edu.tw

***Abstract*—Considering traffic/shop signs in street view images convey a large amount of information such as locations of pictures taken or effects of advertisement etc., a text detection mechanism for street view images is proposed in this research. To deal with relatively complicated content of street views in urban areas, the proposed scheme consists of two major parts. First, since various interference caused by pedestrians, buildings, vehicles appearing in images will significantly affect the detection performance, a Fully Convolutional Network is employed to locate street signs. Next, another neural network, i.e., Region Proposal Network, will help to extract text lines in the identified traffic/shop signs. Both horizontal and vertical text-lines will be extracted. The experimental results show that the proposed scheme is feasible, especially in processing complex streetscape.**

Keywords—text detection, sign detection, street view, fully convolutional network, region proposal network

I. INTRODUCTION

Texts, signs or artificial graphics in street view images often convey plentiful information. These identifiable markers including traffic/shop signs, posters and slogans etc. usually draw a lot of attention and can thus be viewed as regions of interest in such images. Locating the corresponding areas in streetscape may help to extract the image-related information, such as the locations of pictures taken, or to evaluate the effects of advertising signs, etc. This research aims at detecting texts in street view images. It should be noted that street-view object recognition is challenging because of relatively complex content. An example is shown in Fig. 1, in which cluttered backgrounds such as contours of buildings, roads or trees exist. Store or road signs may also overlap each other or be hidden by other objects and further complicate the related detection or recognition.

Fig. 1. Cluttered backgrounds of street-view images

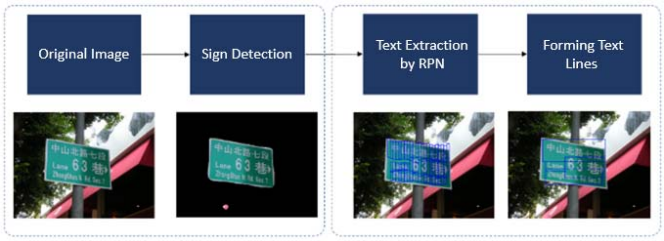


Fig. 2. The flow chart of the proposed scheme

It should be noted that analyzing various patterns or shapes of targeted objects/texts in such cases by traditional methods is difficult. Due to the surging power of computing facilities and availability of large volume of data, convolutional neural network [1] is considered a potential solution to achieving more effective detection. Besides, as also indicated in Fig. 1, street signs are often superimposed on flat areas to clearly convey the necessary information. This hints that locating entire signs, instead of searching texts themselves, would be a more practical way, in our opinions. Therefore, the proposed mechanism is composed of two parts. The first part is to locate the store or traffic signs in the imagery data based on Fully Convolutional Network (FCN). Next, the related texts are extracted from the regions with detected signs using the Region Proposal Network (RPN) [3]. The proposed scheme can thus be viewed as a cascaded structure to achieve more reliable results in locating texts in street views. The rest of this paper is organized as follows. Sec. II details the proposed mechanism and Sec. III shows the experimental results to demonstrate the feasibility of the proposed scheme. Conclusions and future work will be provided in Sec. IV.

II. PROPOSED METHOD

Fig. 2 shows the flowchart of the proposed scheme, including the sign detection with FCN, region of interest or text detection with RPN and text line grouping. We will explain each step in the following subsections.

2.1. Sign detection based on FCN

The proposed scheme is designed by referring the original structure of FCN in [2]. In the network architecture of so-called FCN-32s described in [2], the first half contains five sets of

convolutions and pooling layers to extract the image features, resulting in a feature map with 1/32 the original image size. The 1x1 convolution layer is then used to classify each pixel on the feature map. By applying deconvolution with interpolation, the prediction map has the same dimension as the input image. It should be noted that the first-half structure is eventually VGGNet [4], a commonly used neural network. The other two schemes, FCN-16s and FCN-8s, were further proposed in [2] to reduce miss detections of small objects caused by sampling. The network architecture of FCN-8s is further shown in Fig. 3 (a) After 2x-upsampling, the output of the pooling layer 4 is combined and 2x-upsampled again. Then it is combined with the output of the pooling layer 3. By doing so, it is possible to utilize the image features that have not been pooled so that smaller objects will have better chances to be retained.

We basically refined FCN-8s to design a more suitable network architecture for our street view sign detection model. The original FCN in [2] was trained using the Pascal VOC dataset [5] containing 59 classes. Although a "sign" category is included, the performance on sign detection in street view images is not satisfactory. Our model contains only two classes, the shop/road signs and background. An example is shown in Fig. 4, in which a shop/road sign is manually labelled as red and the background is left as black. For the signs that are vague or too far away, it's difficult to classify their classes as their boundaries are not easy to be clearly defined. These ambiguous objects are thus labeled as "ignored" (marked with green) in training, which means that their existence does not affect the network weight adjustment when training the network.

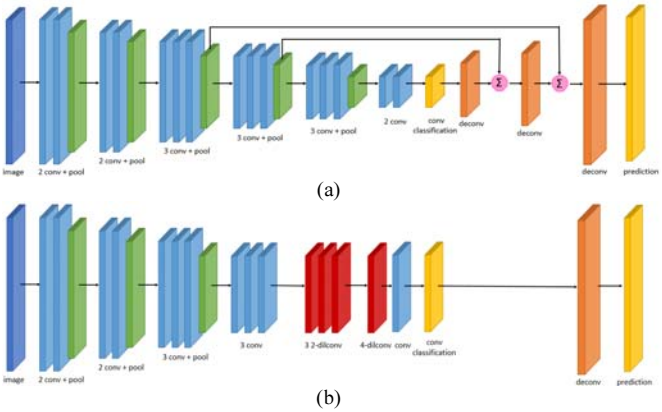


Fig. 3. (a) is the FCN-8s network architecture and (b) is the shop/road sign detection network architecture in the proposed scheme

Fig. 4. The labelling of signs

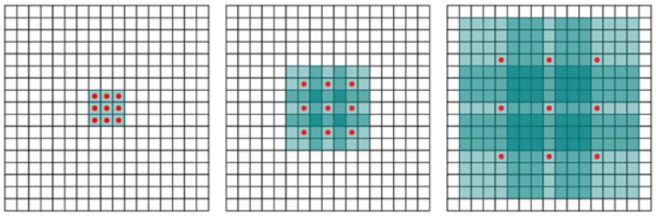


Fig. 5. The receptive field of (a) 3x3, (b), 5x5 and (c) 7x7 dilated convolutions

The architecture of the proposed scheme is shown in Fig. 3 (b), from which we can see that the first half of the proposed structure and FCN-8s are the same so the pretrained model from VGGNet can be used to save training time. Some convolution layers in the second half are replaced with dilated convolution layers [6] as illustrated in Fig. 5 to alleviate the information loss problem during the pooling process and obtain the same eceptive field (RF) as two pooling layers. . Fig. 5 (a) is the general 3x3 convolution with RF being 3x3. We use a 2-dilated convolution connected behind the 3x3 convolution to achieve 7x7 RF, equivalent to doing a pooling and convolution on RF, as shown in Fig. 5 (b). Fig. 5 (c) shows a 4-dilated convolution connected behind 2-dilated convolution to acquire 15x15 RF, equivalent to the RF obtained by secondary pooling and convolution. To be more specific, we replace the pooling layer 4 and convolution layer 5 with the 2-dilated convolution. The pooling layer 5 and the subsequent convolutional layer are also replaced by one 4-dilated convolution. We can thus reduce the pooling process twice but it still has the same RF as the original architecture of FCN-8s, which outperforms FCN-16s and FCN-32s. In addition, we only need to distinguish between street signs and backgrounds. A convolution kernel of 1x1 is adopted and a convolutional layer with two output channels for classification. An example is shown in Fig. 6, in which the regions containing signs are kept and other parts are viewed as the background painted with black. Both shop and road signs are displayed with reasonably good precision, which will facilitate the subsequent text detection if necessary.

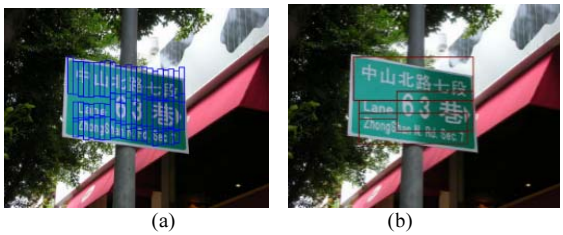
(a) (b)

Fig. 6. An example of "sign" detection. (a) is the source or original street view image and (b) is the result with the detected areas containing signs

2.2. Text detection

Faster R-CNN [3] is a powerful tool for object detection. To improve the speed of its predecessors, one network layer, RPN, is developed to search regions of interest in images using multi-scale anchors on the feature map and can help to calculate the loss from training data. RPN can locate objects that meet the size of some pre-defined anchors on the feature map quite accurately, and then apply the regression to make the bounding

anchors closer to the targeted objects. It is intuitive to think of a text line as a series of fixed height text proposals. We choose to employ the anchor design in Connectionist Text Proposal Network (CTPN) [7] in which each anchor has a fixed width and varying heights. Each anchor may thus include certain parts of characters in text-lines. We train the neural network using ResNet architecture [8] conjunct with RPN. The RPN layer has 14 anchors with different heights ranging from 6 to 300, all of which have a fixed width of 8 pixels. Again, we used a 1x1 convolution layer to classify the anchors on the RPN feature map. The regions are classified into text or background, and given a "text" score indicating the probability that it may be a text area. The text score can be used to find the most suitable anchor if some of them overlap each other. Another 1x1 convolution layer is used to calculate the regression parameter for every anchor generated by RPN. The usage of varying heights comes from the observation that predicting the height of these small anchors is considered easier and could be more accurate than predicting the location of entire text line. After the model prediction, quite a few small anchors with different heights are available. We then exclude the anchors that do not reach the threshold of text score. Non Maximum Suppression (NMS) is adopted to further exclude overlapping anchors (with lower scores) if IoU (Intersection over Union) is larger than 0.7. One example is shown in Fig. 7 (a), in which many blue short anchors are detected and they cover the road sign pretty well.



(a) (b)

Fig. 7. (a) The result using the RPN and (b) the merged text lines

(a) (b) (c) (d)

Fig. 8. (a) is the test image and (b) is the result of sign detection. (c) is the text lines extracted by RPN and (d) is the result of excluding non-target areas.

2.3. Connecting text-lines

After finding the anchors containing texts, these same-width anchors have to be grouped into a text-line, which may increase the accuracy of text identification as these texts are related. We determine if two anchors belong to the same text line by checking their intersection, size and distance. To accommodate possible gaps between texts, letters and characters, the distance between two anchors being less than or equal to the anchor's width will be allowed. Their vertical coordinates have to be similar as well. We then connect the related anchors to form a text-line with a score calculated by averaging the anchor scores in the text-line. Fig. 7 (b) shows the formed text lines of Fig. 7 (a).

After connecting the anchors to form text-lines, some of them could still be partially overlapped. If the IoU of overlapping part is too large, NMS is used to decide which text-line to keep. False alarms are further reduced by excluding the anchors or text lines outside the signs detected by FCN. Some smaller text lines are also removed. A more complex example is illustrated in Fig. 8. Fig. 8 (a) is the investigated image and Fig. 8 (b) shows the areas of road signs extracted by our sign detection model. Fig. 8 (c) shows the resultant text lines following the previously mentioned procedures. Fig. 8 (d) is the filtered result by taking detected sign regions into account.

2.4. Locating texts

It should be noted that English writing is mostly horizontal while Chinese characters can be written horizontally and vertically. Since our test images contain a lot of Chinese road and shop signs, both horizontal and vertical text-lines have to be correctly aligned. In fact, the network trained by a large number of horizontal texts will not detect vertical texts well. A simple strategy here is to detect horizontal and vertical texts separately and then both results are checked to decide which direction is more appropriate. Fig. 9 (a) shows one example, in which vertical and horizontal text-line detections are marked with different colors. Although most text-lines can be found, it is possible that the same texts are detected repeatedly. This ambiguity is resolved by designing another NMS for deciding which text-line to keep depending on the aspect ratio, the area, the score, and the intersection of two text-lines. Since horizontal and vertical text-line detections are processed separately so their scores may not be used directly. The new score is adjusted according to the aspect ratio (the long side of a text-line divided by the short side) and the occupied area. To be more specific, the larger the aspect ratio is, the more probable that the text-line is along the correct direction. If the aspect ratio is lower than a threshold, which is set as 3, an area weighting factor will also be obtained through dividing the investigated text-line area by the largest one in the image. The scores are multiplied by the aspect ratio and the possible area weighting factor and then compared to determine the correct text-line direction. As shown in Fig. 9 (b), those text-lines along the wrong directions are excluded.

(a)

(b)

Fig. 9. Dealing with horizontal and vertical text lines. (a) The detection results of vertical text lines (green) and horizontal text lines (red) and (b) is the combined detection result (marked with red)

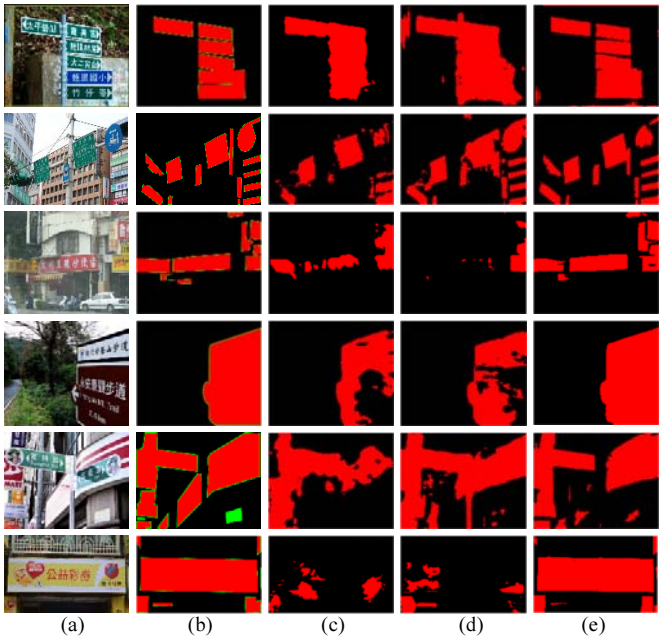
III. Experimental Results

3.1. Network training

It is rather difficult to find in the current literature the similar cases with the image data that we would like to test as most of existing datasets are collected for general text detection only, or specifically used for detecting road signs. In our tests, many images were collected from the Internet or the photos taken by ourselves to obtain complex street view images in urban areas. We collected 569 images, including a number of local street scenes, 504 of which were used as the training images, 65 as the test images and 1/10 in the training images as the validation set. In order to collect enough training data set, data augmentation is adopted to expand the training image set to about 30,000 images. We observed the iteration times and loss values in the training stage to ensure that this sign detection model can be trained well. We then used the dataset "Robust Reading Challenge on Multi-lingual Scene Text Detection and Script Identification" [9] to train our text detection model since these street view images may contain many complicated patterns and different languages. The number of images in this training dataset is around 5,000.

Table 1: The comparison of the original FCN and the proposed method

Method	Accuracy	Precision	Recall	F-Measure
FCN-32s[2]	0.83	0.73	0.58	0.58
FCN-8s[2]	0.85	0.81	0.58	0.61
Our model	0.96	0.92	0.9	0.91



(a) (b) (c) (d) (e)

Fig. 10. The comparison of sign detection results: (a) street view image, (b) ground truth of image, and sign detection result of (c) FCN-8s, (d) PSPNet [10] and (e) the proposed model

3.2. Comparison of detection results

Fig. 10 shows the comparison between the trained model and the original FCN. We can see that the model specifically trained for the sign detection outperforms the original FCN-8s and PSPNet [10] as the boundaries of signs match the actual contours better. According to the detection criteria adopted by FCN, the detection results of our test image set mentioned in Sec. 3.1 are listed in Table 1, which demonstrates that the trained model has very good performance in detecting signs.

The performance of the proposed text detection model is then evaluated using two datasets. The first set is the same street view dataset collected by ourselves and mentioned in Sec. 3.1; the second set is ICDAR 2013 [11]. The ground truth of the test set is in the form of text-lines. The results are compared by the precision, recall, and comprehensive evaluation index, F-Measure. For the ICDAR 2013 dataset, Table 2 shows the comparison, in which the bold numbers indicate the best results while the italic numbers are the runners-up. We can see that CTPN and the proposed scheme are ranked the top two methods. It is worth noting that the recall rate of the proposed model is similar to that of CTPN, but the precision rate is slightly lower. The reason is that our methodology tends to find more potential text-lines and then filter out errors through the sign detection. However, this comparison using ICDAR 2013 doesn't adopt the sign detection because many texts in this dataset do not appear on signboards but on books, clothes, etc. Table 3 shows the comparison for the street view dataset. The proposed method has a much higher recall rate and F-measure than CTPN, with similar precision rates. Higher recall rates come from the fact that the proposed method is specifically designed for street

views containing many vertical texts, which are quite common in certain streetscapes. The experiments also verify the functions of sign detection as the recall rate is raised from 0.61 to 0.73 in the proposed scheme.

Table 2: The comparison of text detection on ICDAR 2013

Method	Precision	Recall	F-Measure
Faster-RCNN	0.79	0.71	0.75
Yin [12]	0.88	0.66	076
Neumann [13]	0.82	0.71	0.76
FASText [14]	0.84	0.69	0.77
Zhang [15]	0.88	0.74	0.80
TextFlow [16]	0.85	**0.76**	0.80
CTPN [7]	**0.93**	*0.75*	**0.83**
The proposed text-detection model	*0.89*	*0.75*	*0.81*

Table 3: The comparison of text detection in street view images

Method	Precision	Recall	F-Measure
CTPN [7]	**0.67**	0.40	0.50
The proposed method without sign detection	0.61	0.61	0.62
The proposed method with sign detection	0.66	**0.73**	**0.69**

IV. Conclusions and Future Work

In this research, we propose a method for detecting texts in street-view images by cascaded convolutional neural networks, i.e., FCN and RPN. The experimental results demonstrate that the proposed method can effectively detect the regions with clear shop and traffic signs. For the scenes containing more complex and changing texts, we can also locate such areas if they can be identified by the human's eyes. Because the proposed method is based on the detection of complete shop or traffic signs, miss detections can be decreased by determining the relevant positions first. For the design of fully convolution network architecture, we use the dilated convolution to increase the receptive field and prevent the feature map from reducing too much by the multiple pooling processes, so that the smaller signs can be kept. In the near future, we will try to employ FPN [17] to pursue more performance enhancement in the detection of object profile, and improve the detection of smaller targets. The accuracy of sign detection has to be further increased so that the subsequent detection/classification of content can be benefited. The purpose of this design is to automatically search the locations of images containing some man-made information. The scheme should continue to analyze the actual meaning by text and/or logo recognition. Automatically detecting the geographical location based on the information extracted from street-view images is another research objective.

Acknowledgement

This research is supported by the Ministry of Science and Technology in Taiwan, R.O.C., under Grants 106-2221-E-008 -003-MY3 and MOST 107-2634-F-008-002.

References

[1] T. Wang, D. J. Wu, A. Coates, A. Y. Ng, "End-to-end text recognition with convolutional neural network." IEEE International Conference on Pattern Recognition (ICPR), 2012.

[2] N. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.

[3] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards realtime object detection with region proposal networks,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," 2007.

[6] F. Yu, V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," International Conference on Learning Representations (ICLR), 2016

[7] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 56–72

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of CVPR, pages 770–778, 2016.

[9] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, W. Khlif, M. L. Muzzamil, J.-C. Burie, C.-l. Liu, and J.-M. Ogier, "ICDAR2017 Robust Reading Challenge on multi-lingual scene text detection and script identification – RRC-MLT," in Document Analysis and Recognition (ICDAR), 2017 14th International Conference on. IEEE, 2017

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. arXiv:1612.01105, 2016.

[11] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. de las Heras, et al. ICDAR 2013 robust reading competition. In ICDAR 2013, pages 1484–1493. IEEE, 2013

[12] X.C. Yin, X. Yin, K. Huang, H.W. Hao: Robust text detection in natural scene images. IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI) 36, 970– 983 (2014)

[13] L. Neumann, J. Matas: Real-time lexicon-free scene text localization and recognition. In IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI) (2015)

[14] M. Busta, L. Neumann, J. Matas: Fastext: Efficient unconstrained scene text detector (2015), in IEEE International Conference on Computer Vision (ICCV)

[15] Z. Zhang, W. Shen, C. Yao, X. Bai: Symmetry-based text line detection in natural scenes (2015), in IEEE Computer Vision and Pattern Recognition (CVPR)

[16] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, C.L. Tan: Text flow: A unified text detection system in natural scene images (2015), in IEEE International Conference on Computer Vision (ICCV)

[17] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and ´ S. Belongie. Feature pyramid networks for object detection. In CVPR, 2017. 2, 4