Store sign text recognition for wearable navigation assistance system

Kaite Xiang, Kaiwei Wang*, Lei Fei and Kailun Yang

State Key Laboratory of Modern Optical Instrumentation, Zhejiang University

wangkaiwei@zju.edu.cn +86-571-8795-3154

Abstract. With the development of computer vision, wearable computing technologies not only have changed our lifestyle, but also have provided much convenience for vulnerable road users, especially the Visually Impaired (VI) pedestrians. VI people have difficulties in locating and socializing due to the limitations of traditional assistive tools, e.g., the inability to recognize text. Text plays a significant role in various aspects, which can convey abundant semantic information of the scene. In recent years, text detection and recognition has made huge progress which makes it possible for VI people to understand the surroundings by using scene text information. In this paper, a text recognition system is proposed to help VI people to perceive store sign text. Firstly, we locate the text on the sign with the aim of leading VI pedestrian to reach the destination store. Towards this end, an objection detection network is integrated into the system to extract Regions of Interest (ROI) in complex real-world scenarios. In order to fulfil real-time assistance, an efficient detection network named Single Shot MultiBox Detector (SSD) has been made light-weight and embedded in the wearable system. Secondly, we leverage an open-source optical character recognition (OCR) instrument to recognize the detected text. Afterwards, we introduce the collected dataset and critical training tips for the task. Finally, a comprehensive set of experiments on our dataset demonstrates that our approach significantly improves the precision and make the recognition robust even in real-world settings. Based on our approach, the wearable system can feedback the recognized text in real time and assist the VI people during their every independent navigation.

1. Introduction

According to the World Health Organization, there are 253 million people estimated to be visually impaired and 36 million people totally blind in the world [1]. VI people lack the capability to sense the surrounding environments so that they have difficulties in locating themselves and socializing with other individuals. Current assistive tool such as white cane cannot provide reliable assistance for ambient smart living of the visually impaired. In this sense, an intelligent navigation assistance system is clearly desirable.

In the research line of assisted navigation, we have developed generations of prototypes for the visually impaired [2]. On the customized wearable system Intoer [3], we have implemented traversability detection [4], obstacle detection, water hazards detection and so on, but have not integrated the function of text information recognition. Text information acts as the high-level semantic information of scenes, which is the key element to understand the content of the images. Traditional OCR is mainly applied to document text, while document text is of the similar fonts, sizes and high-resolution. However, scene text in the real world has the problems of low resolution,

complex background, irregular typesetting, etc. With the urge of International Conference on Document Analysis and Recognition (ICDAR) [5], scene text recognition has made rapid progress. On the other hand, text detection also benefits from the advancement of deep learning algorithms.

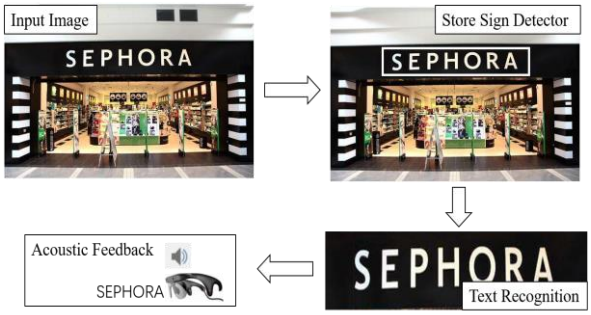


Figure 1. Overview of the store sign text recognition framework.

Figure 2. The wearable navigation assistance system.

In the context of assisted navigation, there is no need for the system to detect all the text information of a scene, because much text information is of none sense for VI people and recognizing all the text of an image is time-consuming.

Following the rationale, we consider to locate the text into specific object like number of bus lines, doorplates, and store sign. In other words, we transform the text detection task into an objection detection task. In this system, we focus on the localization of text information at store sign and the text recognition. With the development of computer science especially AI technology, text recognition has gotten on a fair treat like Convolutional Recurrent Neural Network (CRNN) [6], and so on. In this system, we choose an open-source OCR instrument produced by BIDU [7].

Based on above observations, this paper aimed to achieve store sign text detection and recognition of street scenes. Combining the objection detection technology with text recognition, our system can feedback the name of address of the shop which is on the board of store sign before the VI person like a certain hospital, convenience store, and etc. The system can not only help the VI pedestrians reach the specific destination, but also can be combined with a localization system to improve the precision in determining positions during navigation.

Our contributions can be summarized as follows: (1) A wearable navigation assistance system (as shown in Figure 1) incorporating store sign detection and text recognition is proposed. (2) A dataset of store sign is created by ourselves, and will be offered to community at *http://wangkaiwei.org/downloadeg.html*. (3) The detection precision has been significantly improved by augmenting the dataset and fine-tuning a network using various kinds of hyper-parameters. (4) The main insight gained from our experiments is that in essence, the robustness of text detection is not only a matter of CNN architecture or learning capacity, but also a matter of data diversity and accuracy. (5) We prove that the system can detect and recognition the text of store signs within a certain distance with high precision and speed, which makes it suitable for real-world assisted navigation that require both robust and real-time operation.

2. Related Work

Scene text detection and recognition has been developing since decades ago, and our system is of high-relationship with them besides objection detection.

2.1. Scene text detection

Scene text detection is addressing the task of locating the area of text in an image, which has been progressed from traditional methods based on hand-crafted features into neural networks. For classical

methods, they utilize the features of text like colour, stroke width, and etc. Epshtein et al. [8] came up with a method named Stroke Width Transform (SWT) based on the stroke width of texts. With the similar purpose, Shi et al. [9] utilized Maximally Stable Extremal Regions (MSER) to produce text region proposal. In spite of the fact that they can handle some cases, they can't be applied to complex circumstances. Over the past years, deep learning methods standout over traditional approaches, which dramatically augment the performance, even in cluttered application conditions. An epoch-making method named Connectionist Text Proposal Network (CTPN) [10] was proposed by Tian et al. Benefited from the Region Convolutional Neural Network (R-CNN) [11], it combines Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), and improves the text detection precision to a great extent. In addition, Zhou et al. [12] combined segmentation with bounding-box regression and optimized PVANet [13] to obtain favourable results. Zhang et al. [14] regarded text region as the objection needed to be segmented, and combined with other features of text to detect text. Along with instance-level segmentation, He et al. [15] utilized FCN [16] to detect arbitrarily oriented texts.

2.2. Scene text recognition

Scene text recognition is to recognize the text on the picture that only contains text, which can achieve decent performance for document text, but not as good as when it comes to real-world scenes. Nowadays, with the proliferation of hardware resources and computer science. Many methods based on deep learning have been proposed and various open-source tools can be used for research or commercial production. For illustration, CRNN is an end-to-end text recognition network that directly outputs the text of the image. It has the characteristic of high speed and can be used to recognize diverse languages. On the other hand, IFLYTEK [17], Tencent [18] et al. have developed their own OCR tools which can be utilized by the research community.

2.3. Objection detection

Objection detection is one of the most promising applications propelled by deep learning. Objection detection network is roughly divided into two classes, one-stage model and two-stage model. For two-stage model, models like R-CNN series have high precision but low speed because of a large number of floating-point calculations. Comparatively, for one-stage model like YOLO [19] and SSD [20] also has a high precision but higher speed than two-stage models. Additionally, many light-weight networks like SqueezeNet [21] and MobileNet [22] have been proposed which makes it possible for Convolutional Neural Networks (CNNs) to be deployed in portable devices like mobile phones and wearable systems.

3. Methodology

3.1. System overview

For the hardware parts, as is shown in Figure1, our store sign text recognition framework needs to be integrated into the wearable system Intoer (Figure 2) consisting of a pair of wearable smart glasses and a mobile processor. The pair of the glasses is comprised with a RGB-Depth sensor and a bone-conduction headphone. The RGB image of the sensor serves as the input of our system. Meanwhile, the depth image of the sensor can be used to feedback users the distance between the store and themselves, and the result is transmitted by the bone-conduction headphone in acoustic ways.

For the software parts, our system is composed of two components. One is an object detection network aimed at text detection, the other one is an open-source text recognition tool used for recognizing the text information on the store sign. Finally, we utilize some open-source tools to feedback the acoustic information. In view of the trade-off between latency and accuracy, we choose to use the SSD as the store sign detector rather than other networks. Because SSD possesses a higher speed and precision than YOLO, and it is much faster than R-CNN series networks. For text detection tools, we choose

one produced by BIDU which is easy to use and integrated into our system. The detection and recognition pipeline is visualized in Figure 3.

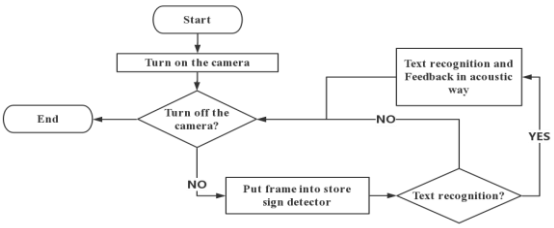


Figure 3. The processing pipeline of the system.

Figure 4. Example of the BIDU's OCR.

3.2. Text Recognition tools

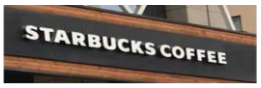
Based on the research, there are various kinds of off-the-shelf text recognition tools in the research community. By comparison we choose the BIDU's text recognition tool, because of its high-precision, multilingual applicability and convenience of development for different programming language users.

Its usage is quite easy that we only need to register a BIDU's account and use it according to correlation tutorial. As is shown in Figure 4, if we use Python as the programming language, its output is a Python dictionary which consists of *log-id*, *words_result_num* and *words_result*. And we only need to concentrate on the latter two outputs.

Someone may raise the doubts that now that the tool can detect and locate all the texts of an image, is there any necessity for us to develop our own text detection module? It can be explained by our initial experiment. As is displayed in Figure 5a, if we input the whole picture, the speed is low and the outputs of the picture consists of many erroneous and meaningless recognition results. For the purpose of assisting VI people to obtain the text information of the store, only the name, address or telephone number of the store are meaningful. For this reason, it's necessary for us to develop an efficient store sign detector. As is shown in Figure 5b, if we input the store sign region of the image, the output is precise and fast and it can improve the targeting of critical text information.

Number	Recognition Results
1	STARBUCXSCO
2	F

(a)



Number	Recognition Results
1	STARBUCKS COFFEE

(b)

Figure 5. An initial experiment with the Text Recognition tool.

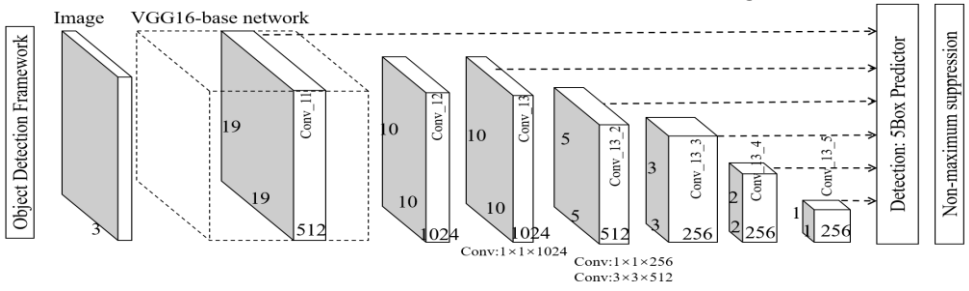


Figure 6. Architecture of SSD.

3.3. Network Architecture

As is explained as before, we choose SSD module as the system's store sign detector serving as the text location. The objection detector is composed of two components as is depicted in figure 6. The first one is a base net which is used to extract meaningful features of object which is normally pretrained on ImageNet [23]. In order to verify the effectiveness of our method, we choose a relatively highly precise network, VGG [24] as our base net. Another one is an auxiliary structure like image pyramid, which utilizes multi-scale feature maps so that the network can detect objects of different sizes and length-width ratios with the assistance of default boxes. As is shown in Figure 5, the final detection layer extracts different feature maps (*conv_11, conv_13, conv_13_2, conv_13_3, conv_13_4 and conv_13_5*). After the inference stage, we apply a confidence threshold filter and Non-Maximun Suppression (NMS) algorithm to obtain the final results. In the sense, we choose the detected regions of store sign and put them into the text recognition module.

4. Experiments

4.1. Dataset

In order to train the text region detector, we collect and annotate a dataset consisting of 3000 street pictures which has the regions of store sign. Eighty percent of the dataset serves as the train set, while the others serves as the test set. We gather the images by taking pictures with our navigation assistance system and portable cameras, as well as searching via web crawlers. For camera parts, we especially consider to capture the scene which has store signs blocked out by tree, street lamp or transmission pole, because such diversity of the dataset can improve the robustness of text region detector. In this regard, it is robust and effective for real-world scenes and complex conditions observed by wearable navigation assistance systems for visually impaired people. For the web crawler parts, we search the samples via the key words like convenience stores, street scene and so on. Undoubtedly, the search results still have much noise, we need to filter the invalid results by removing those samples manually. Then, we label the pictures as the style of VOC2007 [25], the annotation file is as depicted in Figure 7, which consists of picture information like filename, size and objection information like objection name (in this task, the objection name is sign) and bounding box locations.

Figure 7. Annotation example.

(a)

(b)

Figure 8. Data Augmentation**.**

In order to further enhance the generalization ability of the detector, we apply an array of data augmentation methods like colour jittering and random cropping as exhibited in Figure 8. In Figure 8a, it shows the effect of colour jittering which can help the system attain the robustness to different illumination conditions and colour deviations. In Figure 8b, it shows the effect of random cropping allowing the detector to learn the essential features of the store sign.

4.2. Results

We have used the Adam optimization algorithm [26] to train the SSD model, which allows the model to converge quite quickly. The initial learning rate is 0.001 with learning rate decay factor of 0.97 by exponential method, and weight decay is 0.0005. Our experiments are conducted on Tensorflow.

During the training stage, we train different models by different kinds of methods mainly focused on the adjustment of the dataset and other hyper-parameters like iteration numbers or default box ratios detailed in [20].

(a)

(b)

Figure 9. Training Step1: the effect of the model in step 1.

(a) (b)

Figure 10. Training Step2: (a) is the low confidence of the model in step1, (b) is the higher confidence of the model in step 2.

Step1. In order to validate our idea, we simply train the model with the iteration number of around 7000 by only using the dataset of the camera part, we find the total loss converge quickly. We test the model by some pictures, we can find that the model can detect the store sign of the pictures quite precisely, but detect some incorrect region which normally has the text region like billboard, traffic sign and etc. as shown in Figure 9. Therefore, we draw a conclusion that our model can detect the region that has text, but lack the ability to detect the store sign.

Step2. As is known to all, iteration number is of great significance for training a reliable model in deep learning. After the verification stage, we begin to train the model for real-world assistance. We increase the iteration number from 7000 to around 20000, and train it by only using the dataset of the camera part as step 1. We can find that our model's detection of the store sign region is more precise and the confidence is higher as shown in Figure 10. However, the model still faces the problem of detecting some fallible regions like Figure 9 (b). Therefore, only increasing the iteration number is of no use, it's the extension of dataset that counts. Only if we extend the dataset can our model learn to extract more abstract store sign features, because data diversity plays an essential role in reinforcing the robustness of the text recognition.

Step3. In order to avoid erroneous detected results, we adopt the method extending the dataset via web crawlers which haven't been filtered by our rules as detailed in section 4.1. The total loss decreases with a huge fluctuation as shown in Figure11.a. We consider the unfiltered dataset results in the fluctuation, because the original dataset only contains the images captured by us, whose characteristic is unitary. But the extended dataset is full of store signs with different orientations, sizes, styles and structures, which contributes to improving the store sign feature extraction ability of the model so that it will reduce the chance of detecting the wrong store sign regions. Undoubtedly, the confidence of the model will decrease. In the end, we find that although the confidence decreases for some samples, but it solves the condition detecting the wrong regions as showed in Figure 12. Still, decreasing confidence may trigger new problems like omitting the real store sign regions.

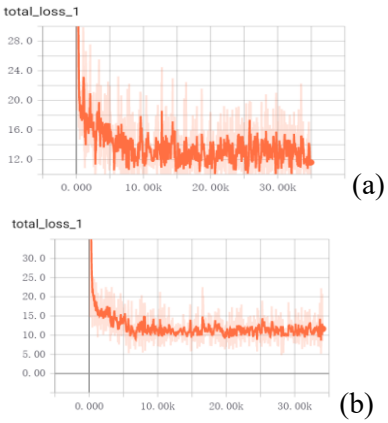


Figure 11. Curve of Loss function: (a) is curve of the model in step 3, (b) is curve of the model in step 4.

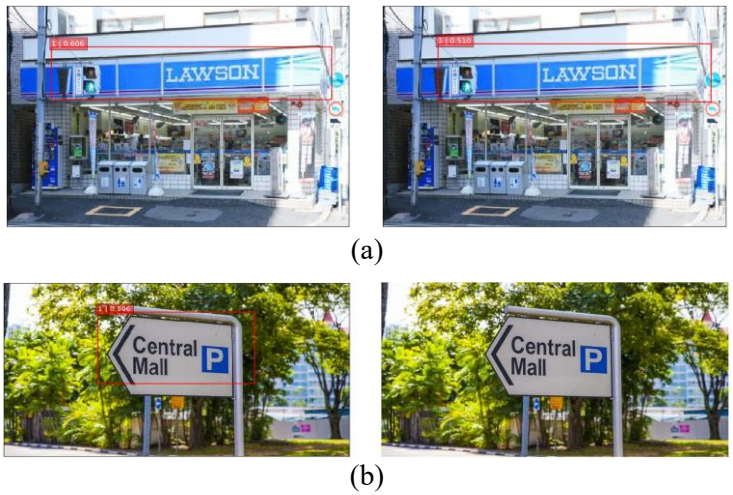


Figure 12. Training step 3: (a) is the side effect of decreasing confidence, (b) is the positive effect of eliminating wrong regions.

Figure 13. Filtered samples (dense store sign picture).

Figure 14. Training Step 4: (a) is the bad effect of former model, (b) is the improvement effect of the model in step 4.

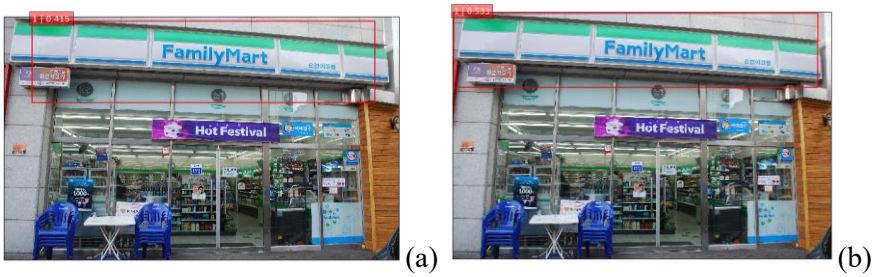


Figure 15. Training Step 5: (a) is the effect of former model, (b) is the improvement effect of the model in step 5 (better marginality).

Step4. In order to avoid the fluctuation of loss function in step3, increase the confidence and improve the detection precision and robustness of the model, we choose to adjust the dataset like removing pictures with tiny store sign, low resolution, and etc. As is explained in step 3, the extended dataset is too abundant that makes it hard for the model to extract general features of a store sign leading to a low confidence score in step 3. Therefore, we need to modify and adjust the dataset. For example, some pictures like Figure 13 have many small store signs (dense store signs), when we train the model,

we need to resize the pictures into 300 × 300. The store signs of the pictures are so small that we can't recognize the store sign regions by ourselves not to mention to let the model learn to detect the regions. In addition to the small store signs, some pictures like uncommon store signs which will decrease the confidence score of the model and lead to the fluctuation of loss function. Moreover, some images full of watermarks may consume some parameters to represent the feature of watermark rather than the pure store sign. After adjusting the dataset and training, we find that our loss function converges quickly and gently (Figure11.b) and its performance is much better than step 3 (Figure14). The main insight gained from this experiment is that in essence, the robustness of text detection is not only a matter of CNN architecture or learning capacity, but also a matter of data diversity and accuracy.

Table 1. Comparison of Different Steps' Model Performance.

Step	mAP
1	0.729
2	0.774
3	0.748
4	0.777
5	0.808

Step5. In view of the fact that normal store signs' length-width ratio is larger than 1.0, therefore, we choose to wipe off the default boxes whose length-width ratio is less than 1.0 and increase the default boxes whose length-width ratio is larger than 1.0. Because the default boxes correspond to the parameters that we need to train. If we don't wipe off the default boxes whose length-width ratio is less than 1.0, many parameters don't make sense and are negative for training. After training, we find that our detection is slightly more precise than before as shown in Figure 15.

After the training stage, we utilize the mean Average Precision (mAP) as the evaluation index to test our models. The results are shown in Table 1. We can draw the conclusion that our training steps indeed improve the performance of the detector. When testing on a cost-effective GPU GTX 1080Ti, the system runs at 48 FPS when inputting a 544 × 960 video file, which is sufficient to provide real-time feedback and assistance.

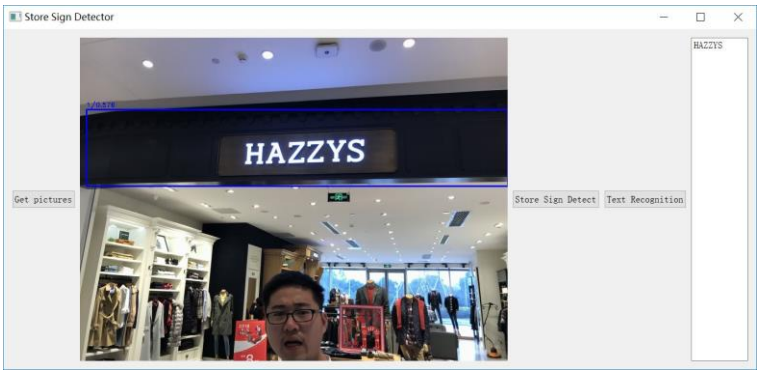


Figure 16. Software interface of the system.

4.3. Software interface of the system

Finally, we integrate the store sign detector and text recognition instrument, and our system can feedback in acoustic ways. Apart from algorithmic experiment, we produce a demo software by PyQt

as shown in Figure16. The software can detect and recognize the store sign text of the pictures selected by us, which has been provided to the community for benchmarking store text recognition.

5. Conclusion and future work

According to the demands of the VI people, we propose a novel approach to integrate SSD and text recognition instrument to achieve the target of recognizing text on the store sign. According to the experiments, our wearable system can fulfil the real-time store sign text recognition. Future work will involve improving the detection speed like replacing the base net VGG to MobileNet, and developing our own text detection tool so that we don't have to rely on the Internet to obtain recognition results. In addition, we are determined to research on the importance-aware semantic segmentation to provide higher-level assistance for visually impaired pedestrians in complex intersections and roundabouts.

6. References

[1] Bourne, R.R., Flaxman, S.R., Braithwaite, T., Cicinelli, M.V., Das, A., Jonas, J.B., Keeffe, J., Kempen, J.H., Leasher, J., Limburg, H. and Naidoo, K., *Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis.* The Lancet Global Health, 5(9), 2017, pp.e888-e897.

[2] Yang, K., Wang, K., Bergasa, L.M., Romera, E., Hu, W., Sun, D., Sun, J., Cheng, R., Chen, T. and López, E., *Unifying Terrain Awareness for the Visually Impaired through Real-Time Semantic Segmentation. Sensors*, 18(5), 2018, p.1506.

[3] KrVision: Intoer: auxiliary glasses for people with visual impairments (in Chinese). *http://www.krvision.cn/cpjs/*

[4] Yang, K., Wang, K., Hu, W. and Bai, J., *Expanding the detection of traversable area with RealSense for the visually impaired. Sensors*, 16(11), 2016, p.1954.

[5] Ye, Q. and Doermann, D., *Text detection and recognition in imagery: A survey*. IEEE transactions on pattern analysis and machine intelligence, 37(7), 2015, pp.1480-1500.

[6] Shi, B., Bai, X. and Yao, C., *An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition*. IEEE transactions on pattern analysis and machine intelligence, 39(11), 2017, pp.2298-2304.

[7] BIDI AI: *http://ai.baidu.com/*.

[8] Epshtein, B., Ofek, E. and Wexler, Y., *Detecting text in natural scenes with stroke width transform.* In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 2963-2970.

[9] Shi, C., Wang, C., Xiao, B., Zhang, Y. and Gao, S., *Scene text detection using graph model built upon maximally stable extremal regions.* Pattern recognition letters, 34(2), 2013, pp.107-116.

[10] Tian, Z., Huang, W., He, T., He, P. and Qiao, Y., October. *Detecting text in natural image with connectionist text proposal network*. In European conference on computer vision, Springer, Cham. 2016, pp. 56-72.

[11] Girshick, R., Donahue, J., Darrell, T. and Malik, J., *Rich feature hierarchies for accurate object detection and semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[12] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W. and Liang, J., July. *EAST: an efficient and accurate scene text detector.* In Proc. CVPR, 2017, pp. 2642-2651.

[13] Kim, K.H., Hong, S., Roh, B., Cheon, Y. and Park, M., *PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection*. arXiv preprint arXiv:1608.08021, 2016.

[14] Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W. and Bai, X., *Multi-oriented text detection with fully convolutional networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4159-4167.

[15] He, D., Yang, X., Liang, C., Zhou, Z., Alexander, G., Ororbia, I.I., Kifer, D. and Giles, C.L., 2017, July. *Multi-scale FCN with Cascaded Instance Aware Segmentation for Arbitrary Oriented Word Spotting in the Wild*. In CVPR, 2017, pp. 474-483.

[16] Long, J., Shelhamer, E. and Darrell, T., *Fully convolutional networks for semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431-3440.
[17] IFLYTEK CO., LTD. AI: *https://www.xfyun.cn/services/wordRecg*.
[18] Tencent AI: *https://cloud.tencent.com/product/ocr*.
[19] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., *You only look once: Unified, real-time object detection*. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.
[20] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C., October. *SSD: Single shot multibox detector*. In European conference on computer vision, Springer, Cham. 2016, pp. 21-37.
[21] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K., *Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size*. arXiv preprint arXiv:1602.07360, 2016.
[22] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., *Mobilenets: Efficient convolutional neural networks for mobile vision applications*. arXiv preprint arXiv:1704.04861, 2017.
[23] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., *Imagenet: A large-scale hierarchical image database*. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. Ieee. 2009, pp. 248-255.
[24] Simonyan, K. and Zisserman, A., *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556, 2014.
[25] Everingham, M., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A., 2008. *The pascal visual object classes challenge 2007 (voc 2007) results (2007)*.
[26] Kingma, D.P. and Ba, J., *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980, 2014.