Scene text detection and recognition system for visually impaired people in real world

Lei Fei, Kaiwei Wang*, Shufei Lin, Kailun Yang, Ruiqi Cheng, and Hao Chen
State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, 38# Zheda Road, Hangzhou 310027, China

ABSTRACT

Visually Impaired (VI) people around the world have difficulties in socializing and traveling due to the limitation of traditional assistive tools. In recent years, practical assistance systems for scene text detection and recognition allow VI people to obtain text information from surrounding scenes. However, real-world scene text features complex background, low resolution, variable fonts as well as irregular arrangement which make it difficult to achieve robust scene text detection and recognition. In this paper, a scene text recognition system to help VI people is proposed. Firstly, we propose a high-performance neural network to detect and track objects, which is applied to specific scenes to obtain Regions of Interest (ROI). In order to achieve real-time detection, a light-weight deep neural network has been built using depth-wise separable convolutions that enables the system to be integrated into mobile devices with limited computational resources. Secondly, we train the neural network using the textural features to improve the precision of text detection. Our algorithm suppresses the effects of spatial transformation (including translation, scaling, rotation as well as other geometric transformations) based on the spatial transformer networks. Open-source optical character recognition (OCR) is used to train scene texts individually to improve the accuracy of text recognition. The interactive system eventually transfers the number and distance information of inbound buses to visually impaired people. Finally, a comprehensive set of experiments on several benchmark datasets demonstrates that our algorithm has achieved an extraordinary trade-off between precision and resource usage.

Keywords: Assistive technology, scene text detection, scene text recognition, object classification network

1. INTRODUCTION

Limited by traditional assistive tools, almost 253 million visually impaired people around the world have difficulties in socializing and traveling.[1] Text is one of the most expressive means of communication. It can be embedded into documents or scenes as a way of information exchange.[2] In recent years, practical scene text detection and recognition systems such as screen-reading software allow visually impaired people to obtain text information from outside world which could profoundly change their life. However, the text information in real life has complex background, low resolution, variable fonts and irregular arrangement, making it difficult to implement scene text detection and recognition. Most of the current assistive tools used by the blind are embedded screen readers in intelligent devices, which are far from being able to read the surrounding scene text.

The problems of text detection and recognition in images and videos have received increasing attention in recent years. As in many fields of computer vision, the scene text analysis benefits from deep learning algorithms, and the accuracy of the method has been significantly improved.[3] Some methods of text detection and recognition have been come up within recent years, and the following methods have aroused the attention from the research community of assistive technology. For text detection, Connectionist Text Proposal Network (CTPN)[4] is an end-to-end trainable text detector that has achieved state-of-the-art results both in detection accuracy and efficiency. Convolutional Recurrent Neural Network (CRNN)[5] takes input images of varying dimensions and produces predictions with different lengths, which can predict scene text flexibly. However, these deep neural networks have large parameters and computational complexity with large size of input image.

Real-time object detection system is used to quickly locate useful scene text areas. Compared with the approach to detect the area of text over the whole image, extracting the area of ROI (bus) firstly not only locates the area of text but also reduce the computational cost of the next location system. Meanwhile, the information of object recognition is able to enrich the context information of the surrounding.

* Corresponding author. Email: wangkaiwei@zju.edu.cn.

Based on the analysis, this paper aims to achieve text localization and interpretation in complex scenes. Through the integration of object detection technology and character recognition, the surrounding environment is richly described. In order to provide the assistance in real-world navigation and socialization, this paper focus on detecting and recognizing some common scene texts which VI people come across usually in their daily life, such as license number of buses, traffic signs and store signs. Finally, the recognition results will be transfer to VI people by speech interaction.

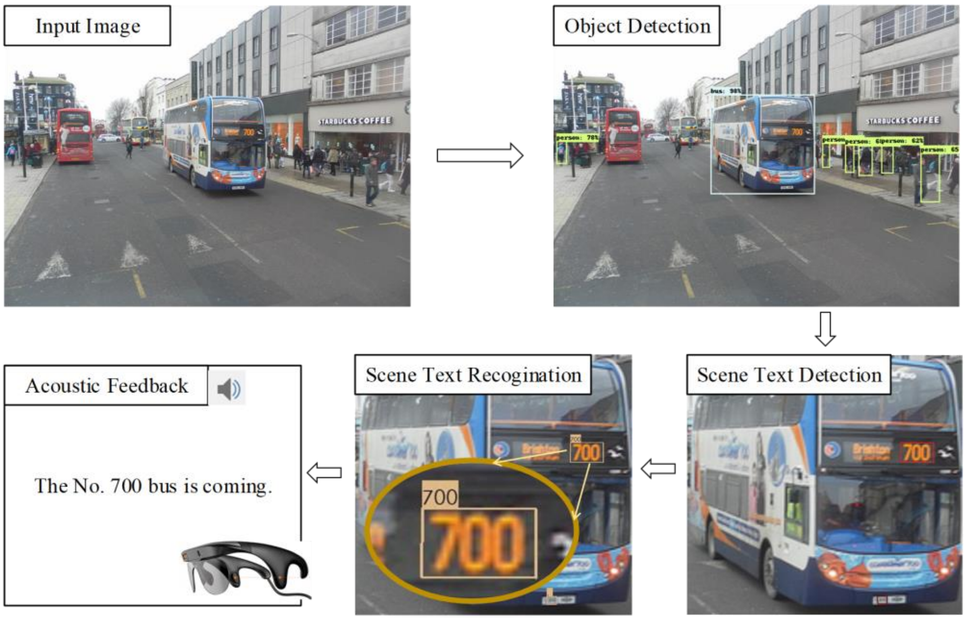


Figure 1. Overview of the scene text detection and recognition system

Our contributions are as follows: (1) A system (as shown in Figure 1) combining object recognition and character recognition is proposed. (2) The network of object detection and character recognition is designed to improve the accuracy of recognition. (3) We prove that the system achieves the desired detection speed and accuracy.

2. RELATED WORK

For VI people, traditional way of text acquisition is based on Braille which is written with embossed paper. However, Braille has the disadvantages of slow scanning speed, long learning time and low penetration rates. Therefore, more and more researches on text detection and recognition in the computer vision community. In this research line, several new systems that try to solve at least one of the two sub-tasks (text detection and text recognition) have been proposed.

Scene text Detector locates the text over the entire input image. Text detection tasks, much easier applied in documents than in scene images, have attracted wide attention in recent years. In some traditional methods, features are manually designed to capture the attributes of scene text, while methods based on deep learning features are learned directly from training data[6]. By detecting edge or extracting extremal region, Stroke Width Transform (SWT)[7] and Maximally Stable Extremal Regions (MSER)[8,9] based methods show weakness dealing with challenging scenarios, such as geometric distortion and low resolution.

As many areas of computer vision, scene text field, which has achieved great progress on accuracy, has benefited greatly from the deep learning technology[10]. Huang et al.[11] used MSER and then incorporated it with a deep convolutional neural network (CNN) which is a strong classifier to prune false positives. The method proposed by Jaderberg et al.[12] scanned the image using sliding-window and produced a text saliency map for each scale with a CNN model. Later, CNN and aggregated channel features (ACF) were used by Jaderberg et al.[13] to hunt for candidate words and to refine them with regression method. Without sharing computational resources between sliding windows or region proposals, training and testing processes are slow in early approaches. Directly predicting text boxes using the output of object detection and segmentation, such as Faster R-CNN[14], SSD[15] and fully convolutional network (FCN)[16], have significantly improved the accuracy of text detection. Zhang et al.[17] utilized FCN for text regions generation and use component projection for orientation estimation. Tian et al. [7] come up with a vertical anchor mechanism that jointly predicts location and text/non-text score of each fixed-width proposal, which considerably improved the localization accuracy. TextBoxes++[18, 19]detects arbitrary-oriented scene text with both high accuracy and efficiency in a single forward passing network.

Text Recognition is to identify machine readable alphabet sequences from images that only contain text. Jaderberg et al. [13] took a cropped region of a single word, achieved text recognition with a convolutional neural network as one of the words in a 90k-class dictionary. CRNN is an open-source text recognition module that directly outputs the character sequence of the given input image, which is also end-to-end trainable. Connectionist Temporal Classification (CTC), designed by Graves et al. [5] to eliminate pre-segmenting data in speech recognition, is used in a bidirectional Long-Short Term Memory (LSTM)[20] to construct CRNN. Moreover, spatial transformer network (STN) and a sequence recognition network (SRN)[21] form a recognition model to recognize several types of irregular text.

object detection framework. Light-weight CNNs, such as SqueezeNet[22] and MobileNet[23], are more practicable to implemented on FPGAs and other memory-limited hardware. SqueezeNet has fewer parameters and computational cost by replacing 3×3 filters with 1×1 filters while maintaining competitive accuracy. MobileNet uses depth-wise separable convolutions and attains comparable results among light-weight models. Meanwhile, MobileNet can be deployed as an effective base network of both Faster-RCNN[14] and SSD[15] framework.

To achieve assisted navigational for VI people, we have embedded semantic traversable area detection,[24,25] zebra crossing detection[26], traffic lights detection[27], water hazards detection[28] and road barriers recognition[29] algorithm under challenging scenarios on the customized wearable system Intoer.[37] By deploying deep neural networks on scene text recognition, we propose a scene text recognition system to help visually impaired people take public transports. Our system is able to detect route number of the coming bus and notify visually impaired people by voice. With the scene text detection system, VI people can obtain more practical help.

3. METHODOLOGY

3.1 System overview

Our proposed system includes two main components, including object detection and scene text recognition as shown in Figure 1. As a visual aid for VI people, Intoer[30] (As shown in Figure 2) consists of a pair of wearable smart glasses and an intelligent microcomputers. The pair of smart glasses integrates an RGB-Depth sensor (RealSense R200) and a bone-conduction headphone. The color images captured by RGB-Depth camera serve as input of the object recognition network. The interactive part of the recognition results is achieved through speech feedback from bone-conduction headphones.

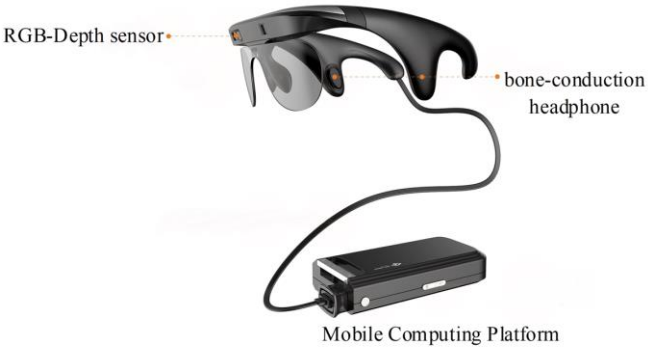


Figure 2. Visual Aid for VI people: Intoer

Our system extracts the text features of a scene (e.g., bus station) from a segmented region of interest (e.g., coming bus), and trains a convolutional neural network to detect scene texts (e.g., route number). Firstly, we propose a high-performance neural network for object classification to obtain ROI. In order to satisfy the requirement of real-time detection, a light-weight deep neural network (e.g., MobileNet) was built. Secondly, we train the neural network using the textural features to improve text detection precision. Open source OCR is used to train characters individually to improve the accuracy of text recognition. The interactive system eventually transfers the information of the number of inbound buses to visually impaired people.

3.2 Object Detection Framework

We serve VGG16 as an effective base network of Single Shot Detector (SDD), which is used to obtain the results of classification and bounding box prediction. The object recognition network is depicted Figure 3. Based on the SSD network, the neural convolutional network structure is designed as follows. We use VGG-16 as the base network, and add extra 6 convolutional layers for feature extraction. Since the size of extra convolutional layers are decreasing, the predictions can be implemented at multiple scales. In addition to target detection on the final feature map, the prediction is carried out on 5 special features selected previously. Detection process is carried out not only with feature maps (conv13_2, conv13_3, conv_13_4, conv_13_4), but also on the feature graph (conv_11, conv_13) of base network to ensure that the network has robust detection performance on small targets. The network obtains rich high-level semantic information by designing hierarchical structure. The classification results of object detection will be used in the subsequent scenario analysis and description.

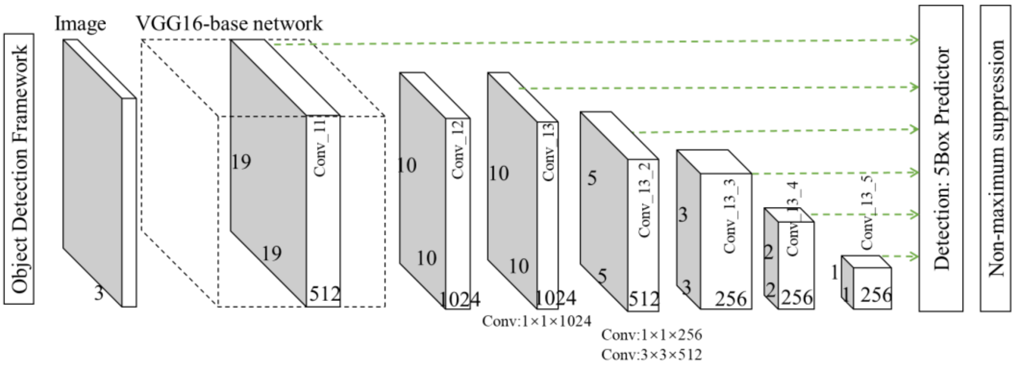


Figure 3. Proposed architecture of the object detection network

3.3 Scene text detection and recognition

The text detection network is based on Connectionist Text Proposal Network (CTPN)[4] which extends architecture of the Region Proposal Network (RPN)[14]. The architecture of CTPN is shown as Figure 4. First, VGG16 is used as the base net to extract features. We get the output feature maps from conv5_3, whose size is W×H×C. And slide 3×3 spatial window on the feature maps thus each window can get a 3×3×C feature vector. A bidirectional LSTM takes the features from the previous step as input and outputs a feature vector of W×256 dimension. In detail, the bidirectional LSTM combines a forward (left to right) and a backward (right to left) LSTMs. Then a 512D fully connected layer links output layer which contains 2k vertical coordinates, 2k text/non-text scores, side-refinement offsets of k anchors. It results in a densely predicted text proposal, such that a standard non-maximum suppression algorithm is used to filter out the redundant boxes. Finally, a graph based text row construction algorithm is used to merge the text segments into a text line that will be input text recognition network for text decoding.

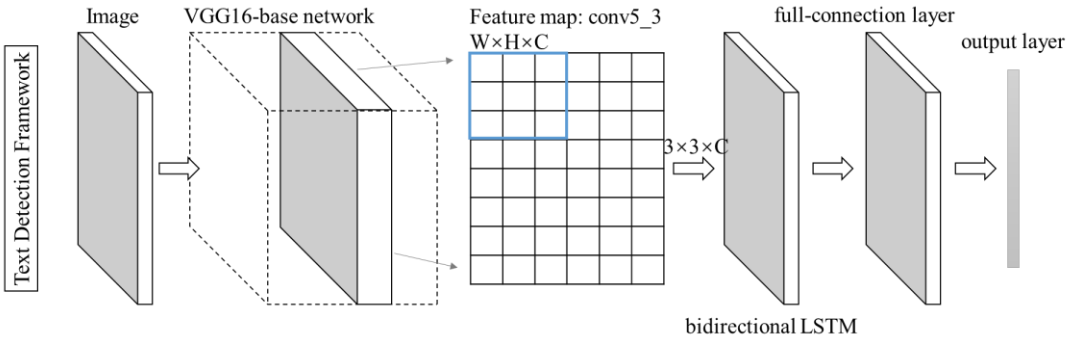


Figure 4. The architecture of scene text detection: Connectionist Text Proposal Network.

Text recognition network is able to take input images of varying dimensions and produce predictions with different lengths using CRNN.[5] It integrates feature extraction, sequence analysis and sequence decoding algorithm, which are implemented in different layers. Pictures and the corresponding words need to be annotated, character level tagging is not required. The network (shown in Figure 5) extracts the features from the input images through the multi-layer CNN, and obtains features maps which contain rich semantic information. After that, each feature map is converted into a feature sequence in sequence analysis, and each column of the feature graph is extracted from left to right to construct a feature sequence. In this manner, the length of the sequence equals to the width of the feature graph, while the dimension of each vector in the sequence equals to the height of the feature map to be multiplied by depth. In order to increase the description capacity of these regions, LSTM is used to analyze the feature sequences. It has the characteristics of memorizing past context, such that the long distance correlation can be captured in a single direction in the sequence. Additionally, bidirectional LSTM[5] serves as the recurrent neural network (RNN) for combining output on opposite directions to get a two-way long short memory network. This network can simultaneously analyze the long distance correlation both from left to right and from right to left, so that feature sequence of its output contains rich context information, which is important for robust scene context recognition.

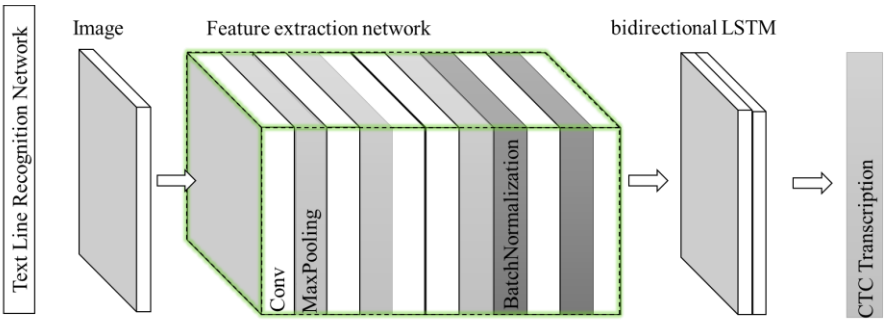


Figure 5. The architecture of scene text recognition network. Different grayscales are used to represent Convolution layer, Max Pooling layer, Batch Normalization layer respectively.

The CTC (connectionist temporal classification) model is connected to the last layer of RNN network for sequence learning. Each point t of a sequence with length T outputs a vector representing the prediction probability at the last layer of the RNN network. The label of each sample point t may be one of L different labels or 'blank' label ($L' = L \cup$), so the label of sequence T has L'^{T} types. Each possibility of L'^{T} is called a "PATH" π, and its conditional probability is shown in equation (1).

$$p(\pi|\mathrm{X}) = \prod_{t=1}^{T} y_{\pi_t}^{t}, \forall \pi \in L'^{T} \tag{1}$$

Remove the repeated places and blanks in each road. for example, prediction results of 'a-bc', 'a--bbc' are all 'abc'. Therefore, as shown in equation (2), the total conditional probability p is the sum of the conditional probability of the path with the same prediction result.

$$p(l|\mathrm{X}) = \sum_{\pi \in \mathrm{B}^{\ 1}(l)} p(\pi|\mathrm{X}) \tag{2}$$

The loss function is set in equation (3), so the gradient descent method can be used to optimize the whole system.

$$O^{ML}(S, N_W) = \sum_{(x,z)\in S} \ln(p(z|\mathrm{X})) \tag{3}$$

In the sequence decoding step, the repeated letters will be merged into a single letter and blank symbol will be removed. For example, '-sw-e-eett-'will get the word output 'sweet' after decoding.

3.4 Combination of networks and Text to Speech (T2S) conversion

Small object detection network describes the location and attribute of the object in the scene, determines whether the object has a text attribute and determines whether the object area should be taken into the text detection and recognition network. For example, we think that the text information of the bus or book area detected is worth reading, so we segment the identified objects and take them into the character recognition network. The results enable bus route number and book titles to be successfully acquired. Based on the description of objects and the extraction of text details, the assistive prototype has a richer understanding of scenes. Bone-conduction headphones are used to feedback voice and transfer the test results to the blind.

4. EXPERIMENTAL RESULTS

We measure text recognition performance of our system in street view text (SVT) [31], MSRA-TD 500 [32] and Tian Chi ICPR [33]datasets. The word detection subsystem is evaluated on MSRA-TD 500 datasets, whose languages of text in the images are mainly Chinese and English. Tian Chi ICPR datasets contain manifold images with a large number of Chinese text databases, and it is the main database for training our text recognition network. Tian Chi ICPR datasets contain about 10 thousand pictures, divided into training and validation set in the ratio of 9:1. Synth Text datasets[10] were used as standard datasets to form the pre-trained models.

Test set of MSRA-TD 500 datasets was used to evaluate the accuracy of text detection and recognition respectively. 77.65% of recall, 72.52% of precision and 75.00% of F1-score are obtained on text detection which are comparable to the state-of-art results[34]. In text recognition, the system achieves 78.22% accuracy which obtains gratifying results in reading Chinese and English hybrid text.

Representative images are selected from the datasets to measure the entire system as shown in Figure 1. The results of bus detection and text recognition are represented in the Figure 6 (a) and (b), respectively. Our system not only locates precisely the bus and pedestrian but also recognizes robustly the number of the coming bus that is ‘1R’.As shown in Figure 6 (c), (d), there are books and a mouse on the table and the cover of the book is ‘OpenCV3’, ‘编程入门 (Programming Introduction)’ and so on. Our system successfully identifies English characters as well as Chinese characters with high accuracy. Object detection system shows different classes in different colors in the graph and marks the confidence levels simultaneously. White, yellow, blue and green represent the categories of objects: bus, person, book and mouse respectively. Text detection and recognition results are represented by yellow box and black text.

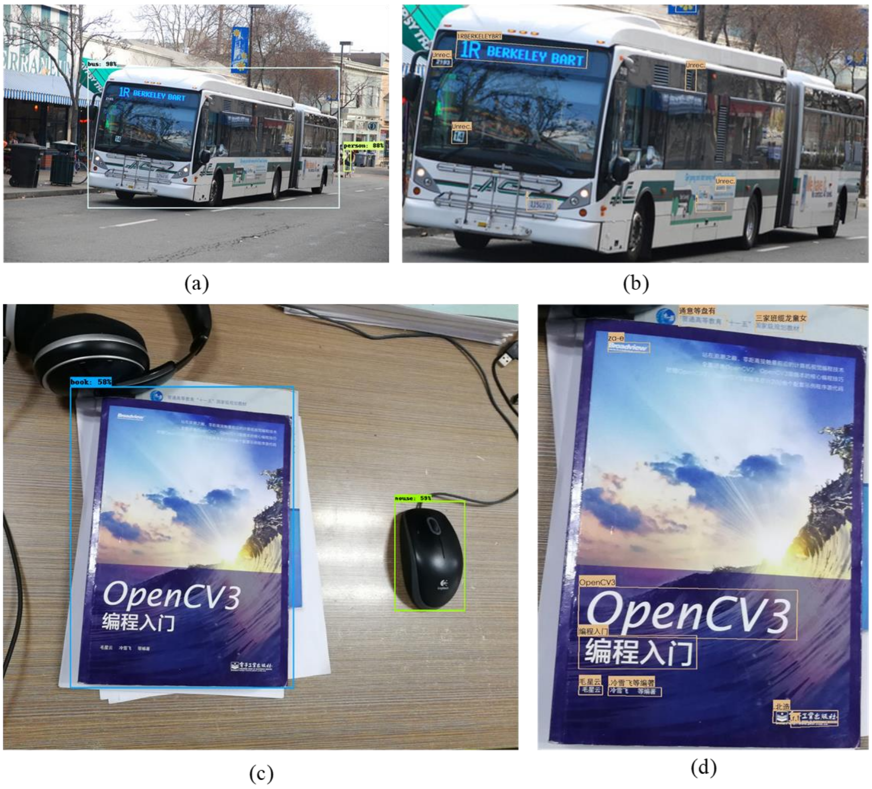


Figure 6. Some examples of the results. A bus and a pedestrian are identified in (a) and text information on the bus is recognized in (b). A book and a mouse are identified in (c) and text information on the bus is recognized in (d). (a) is obtained from COCO2017 datasets. (d) is captured by a mobile phone camera.

Most of the experiments on scene text recognition have been aimed at English characters. In order to apply the system to real-world scenarios, we also add Chinese character library to expand the existing lexicon, and re-train the character recognition system (shown in Figure 5). Tian Chi ICPR datasets are used to fine-tune the pre-trained model obtained from the Synth Text dataset. We trained text recognition network using stochastic gradient descent (SGD) with text line images which are cropped from Tian Chi ICPR datasets and normalized into same height. Learning rate is set at 0.01 firstly, and decayed to 0.005 after 20K training iterations. The training results are shown in Figure 7. We found that increasing the batch size benefits the model to converge faster, and the training loss ends up at 0.23. This design is capable of describing scene texts for this region by training weakly labeled data in deep neural network.

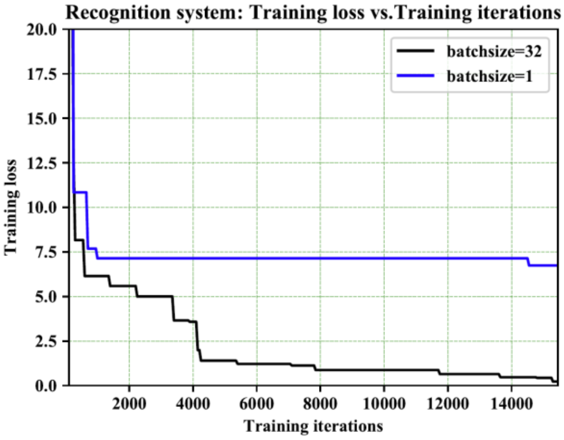
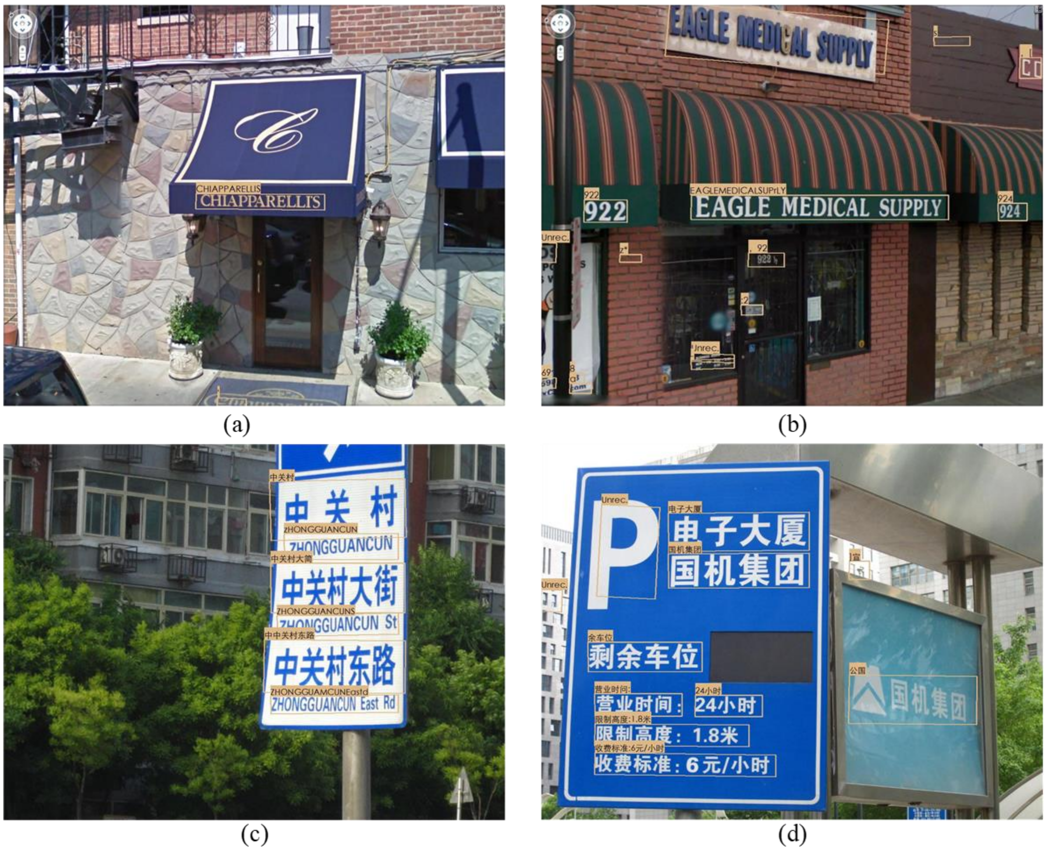


Figure 7. Training loss of Character recognition system with different batch size



(a)

(b)

(c)

(d)

Figure 8. Some examples of text detection and recognition results are represented by yellow box and black text. (a) (b) are obtained from SVT datasets. (c) (d) are obtained from MSRA-TD500 datasets. ‘Unrec.’ is abbreviation of ‘Unrecognized’.

In Figure 8, we show the results of word detection in both Chinese and English, which demonstrate the robustness of our system in most situations. Admittedly, there are still some cases that make it difficult to correctly recognize the character. For example, the display screen of bus is array LEDs, which results in the wrong segmentation of text area. The text region of objects with high speed is so blurred that our system fails to identify them. The expansion of datasets with Chinese characters has challenged the accuracy of the original recognition results. In the future research, we will enhance the adaptability and accuracy of the scene text detection and recognition system.

5. CONCLUSION

In this paper, the design and implementation of a prototype for text detection and recognition has been discussed. The combination of the object detection network and the scene text recognition network shows robust adaptability and provides rich scene information for VI people in different scenarios. Our scene text detection and recognition system can be applied not only to the number detection of buses on roads, but also to bibliographic identification in library, as well as commodity and price identification in supermarkets. In this sense, it can be applied to assistive devices for visually impaired people widely and efficiently. Meanwhile, the proposed algorithm provides scene text descriptions with higher accuracy and better real-time performance. When combined with other navigational algorithms in this research project, this system can provide comprehensive, precise and efficient travel aids for visual impaired people both indoors and outdoors.

REFERENCES

[1] Antonarakis, E. S., Lu, C., Wang, H., Luber, B., Nakazawa, M., Roeser, J. C., Chen, Y., Mohammad, T. A., Chen, Y., Fedor, H. L., Lotan, T. L., Zheng, Q., De Marzo, A. M., Isaacs, J. T., Isaacs, W. B., Nadal, R., Paller, C. J., Denmeade, S. R., Carducci, M. A., et al., “AR-V7 and Resistance to Enzalutamide and Abiraterone in Prostate Cancer,” N. Engl. J. Med. **371**(11), 1028–1038 (2014).

[2] Ye, Q. and Doermann, D., “Text Detection and Recognition in Imagery: A Survey,” IEEE Trans. Pattern Anal. Mach. Intell. **37**(7), 1480–1500 (2015).

[3] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V. R., Lu, S., Shafait, F., Uchida, S. and Valveny, E., “ICDAR 2015 competition on Robust Reading,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR **2015–Novem**, 1156–1160, IEEE (2015).

[4] Tian, Z., Huang, W., He, T., He, P. and Qiao, Y., “Detecting text in natural image with connectionist text proposal network,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **9912 LNCS**, 56–72 (2016).

[5] 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 Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017).

[6] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W. and Liang, J., “EAST: An efficient and accurate scene text detector,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 **2017–Janua**, 2642–2651 (2017).

[7] Patel, H. A., Shekokar, K. S. and Student, M. E., “Detecting Text or character in Natural Scenes with Stroke Width Transform,” Int. J. Innov. Res. Comput. Commun. Eng. (An ISO Certif. Organ. **3297**(5), 1–8 (2007).

[8] Neumann, L. and Matas, J., “A method for text localization and recognition in real-world images,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **6494 LNCS**(PART 3), 770–783 (2011).

[9] Neumann, L. and Matas, J., “Real-Time Lexicon-Free Scene Text Localization and Recognition,” IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1872–1885 (2016).

[10] Gupta, A., Vedaldi, A. and Zisserman, A., “Synthetic Data for Text Localisation in Natural Images,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2315–2324 (2016).

[11] Huang, W., Qiao, Y. and Tang, X., “Robust scene text detection with convolution neural network induced MSER trees,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **8692 LNCS**(PART 4), 497–511 (2014).

[12] Jaderberg, M., Vedaldi, A. and Zisserman, A. B. T.-E. C. on C. V., “Deep Features for Text Spotting,” Eur. Conf. Comput. Vis., 512–528 (2014).

[13] Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A., “Reading Text in the Wild with Convolutional Neural Networks,” Int. J. Comput. Vis. **116**(1), 1–20 (2014).

[14] Ren, S., He, K., Girshick, R. and Sun, J., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017).

[15] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y. and Berg, A. C., “SSD: Single shot multibox detector,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **9905 LNCS**(1), 21–37 (2016).

[16] Shelhamer, E., Long, J. and Darrell, T., “Fully Convolutional Networks for Semantic Segmentation,” IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 2005/09/30, 640–651 (2017).

[17] Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W. and Bai, X., “Multi-Oriented Text Detection with Fully Convolutional Networks,” Proc. IEEE Conf. Comput. Vis. Pattern Recognition., 4159–4167 (2016).

[18] Liao, M., Shi, B., Bai, X., Wang, X. and Liu, W., “TextBoxes: A Fast Text Detector with a Single Deep Neural Network,” AAAI, 4161–4167 (2016).

[19] Liao, M., Shi, B. and Bai, X., “TextBoxes++: A Single-Shot Oriented Scene Text Detector,” IEEE Trans. Image Process. **27**(8), 3676–3690 (2018).

[20] Hochreiter, S. and Schmidhuber, J., “Long Short-Term Memory,” Neural Comput. **9**(8), 1735–1780 (1997).

[21] Shi, B., Wang, X., Lyu, P., Yao, C. and Bai, X., “Robust Scene Text Recognition with Automatic Rectification,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 4168–4176 (2016).

[22] 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 <0.5MB model size,” Med Image Comput Comput Assist Interv **15**(Pt 1), 2013/01/05, 348–356 (2016).

[23] 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 Prepr. arXiv1704.04861 (2017).

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

[25] Yang, K., Wang, K., Bergasa, L. M., Romera, E., Hu, W., Sun, D., Sun, J., Cheng, R., Chen, T. and L ópez, E., “Unifying terrain awareness for the visually impaired through real-time semantic segmentation,” Sensors (Switzerland) **18**(5) (2018).

[26] Cheng, R., Wang, K., Yang, K., Long, N. and Hu, W., “Crosswalk navigation for people with visual impairments on a wearable device,” J. Electron. Imaging **26**(05), 1 (2017).

[27] Cheng, R., Wang, K., Yang, K., Long, N., Bai, J. and Liu, D., “Real-time pedestrian crossing lights detection algorithm for the visually impaired,” Multimed. Tools Appl., 1–21 (2017).

[28] Yang, K., Wang, K., Cheng, R., Hu, W., Huang, X. and Bai, J., “Detecting traversable area and water hazards for the visually impaired with a pRGB-D sensor,” Sensors (Switzerland) **17**(8), 1890 (2017).

[29] Lin, S., Wang, K., Yang, K. and Cheng, R., “KrNet: A kinetic real-time convolutional neural network for navigational assistance,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **10897 LNCS**, K. Miesenberger and G. Kouroupetroglou, Eds., 55–62, Springer International Publishing, Cham (2018).

[30] KrVision., “Intoer,” <http://www.krvision.cn/cpjs/>.

[31] Wang, K. and Belongie, S. J., “Word spotting in the wild,” Eur. Conf. Comput. Vis., 591–604 (2010).

[32] Yao, C., Bai, X., Liu, W., Ma, Y. and Tu, Z., “Detecting texts of arbitrary orientations in natural images,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 1083–1090 (2012).

[33] Aliyun., “icpr_mtwi_2018_challenge,” <https://tianchi.aliyun.com/markets/tianchi/icpr_mtwi_2018_challenge> (11 August 2018).

[34] Zhu, Z., Liao, M., Shi, B. and Bai, X., “Feature Fusion for Scene Text Detection,” 2018 13th IAPR Int. Work. Doc. Anal. Syst., 193–198 (2018).