Supplementary Materials for AE TextSpotter

Wenhai Wang[1], Xuebo Liu[2], Xiaozhong Ji[1], Enze Xie[3], Ding Liang[2]
ZhiBo Yang[4], Tong Lu[1,✉], Chunhua Shen[5], and Ping Luo[3]

[1]National Key Lab for Novel Software Technology, Nanjing University
[2]SenseTime Research [3]The University of Hong Kong
[4]Alibaba-Group [5]The University of Adelaide
wangwenhai362@smail.nju.edu.cn, {liuxuebo, liangding}@sensetime.com,
shawn_ji@163.com, xieenze@hku.hk, zhibo.yzb@alibaba-inc.com,
lutong@nju.edu.cn, chunhua.shen@adelaide.edu.au, pluo@cs.hku.hk

A Appendix

A.1 Ambiguous Images Selection

To keep the objectivity of TDA-ReCTS validation set, we designed two rules to select ambiguous images from the training set of IC19-ReCTS [3], and then randomly sample 1,000 images among them as the validation set.

For an image, we consider its text lines as L, and the internal characters of the ith text line as C^i_{in}.

An image is regarded as a sample with large character spacing, if at least one text line in the image has large character spacing. The character spacing of a text line $\ell_i \in L$ is large, if it satisfies Eqn. 1.

$$\frac{\sum_{c^i_j \in C^i_{in}} \min_{k \neq j, c^i_k \in C^i_{in}} \mathcal{D}_c(c^i_j, c^i_k)}{\sum_{c^i_j \in C^i_{in}} \text{Scale}(c^i_j)} > 2, \tag{1}$$

where c^i_j means the jth character in the internal characters C^i_{in}. $\mathcal{D}_c(\cdot)$ denotes the Euclidean distance between the center points of two character boxes. Scale$(\cdot)$ signifies the scale of a character box, which is calculated by the square root of the box area.

Moreover, an image is considered as a sample with juxtaposed text lines, if the image has a pair of text lines aligned to the top, bottom, left or right direction, and characters in them have similar scales. Two text lines (*i.e.*, ℓ_i and ℓ_j) are aligned, if they satisfy Eqn. 2. The characters in two text lines (*i.e.*, ℓ_i and ℓ_j) have similar scales, if they satisfy Eqn. 3.

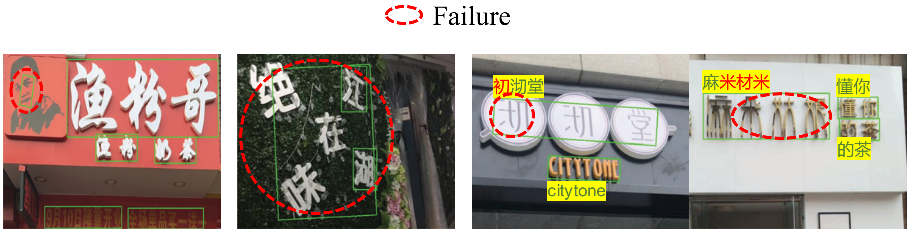
$$\frac{|C^i_{in}| \min_{d \in \{t,b,l,r\}} \mathcal{D}_d(\ell_i, \ell_j)}{\sum_{c^i_k \in C^i_{in}} \text{Scale}(c^i_k)} < \frac{1}{10}. \tag{2}$$

$$\frac{9}{10} \leq \frac{|C^j_{in}| \sum_{c^i_k \in C^i_{in}} \text{Scale}(c^i_k)}{|C^i_{in}| \sum_{c^j_k \in C^j_{in}} \text{Scale}(c^j_k)} \leq \frac{10}{9}. \tag{3}$$

Here, $|C^i_{in}|$ is the number of internal characters of ith text line. $\mathcal{D}_{d=\{t,b,l,r\}}(\cdot)$ represents the absolute difference of the **t**op/**b**ottom/**l**eft/**r**ight of two text line boxes.

A.2 Failure Analysis

As demonstrated in previous experiments, the proposed AE TextSpotter works well in most cases. It still fails in some intricate images, such as text-like regions (see Fig. 1(a)), strange character arrangement (see Fig. 1(b)), and texts in fancy styles (see Fig. 1(c)). Most text-like regions can be removed with a high classification score threshold, but there are still some inevitable error detections. In Fig. 1(b), characters are arranged in “x” shape, which does not follow common writing habit. Therefore, it is difficult to detect text lines with extremely strange character arrangements. The detection of text lines in fancy styles is feasible. However, due to fancy styles, it is hard to correctly recognize text contents in these text lines.



(a) Text-Like Region (b) Strange Arrangement (c) Text Lines in Fancy Styles

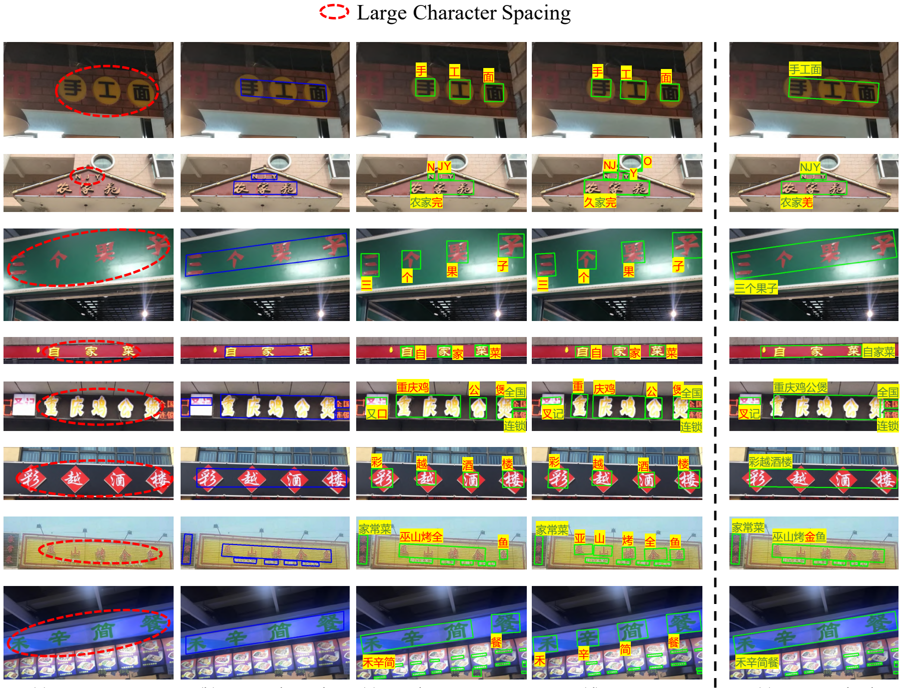
Fig. 1. Failed examples.

A.3 Visual Comparisons between AE TextSpotter and State-of-The-Art Methods.

In this section, we present the effectiveness of our AE TextSpotter by comparing visual results of different methods. Specifically, we visualize and analyse the results predicted by Mask TextSpotter [1], FOTS [2], and our AE TextSpotter on TDA-ReCTS. For a fair comparison, both methods are trained on the 19,000 training images in IC19-ReCTS [3] (except 1,000 images in TDA-ReCTS). In the testing phase, we scale the short side of test images to 800. Fig. 2 and Fig. 3 show the examples of large character spacing and juxtaposed text lines, respectively. In these examples, our method is clearly better than Mask TextSpotter and FOTS. From these results, we can find that the proposed AE TextSpotter has the following abilities.

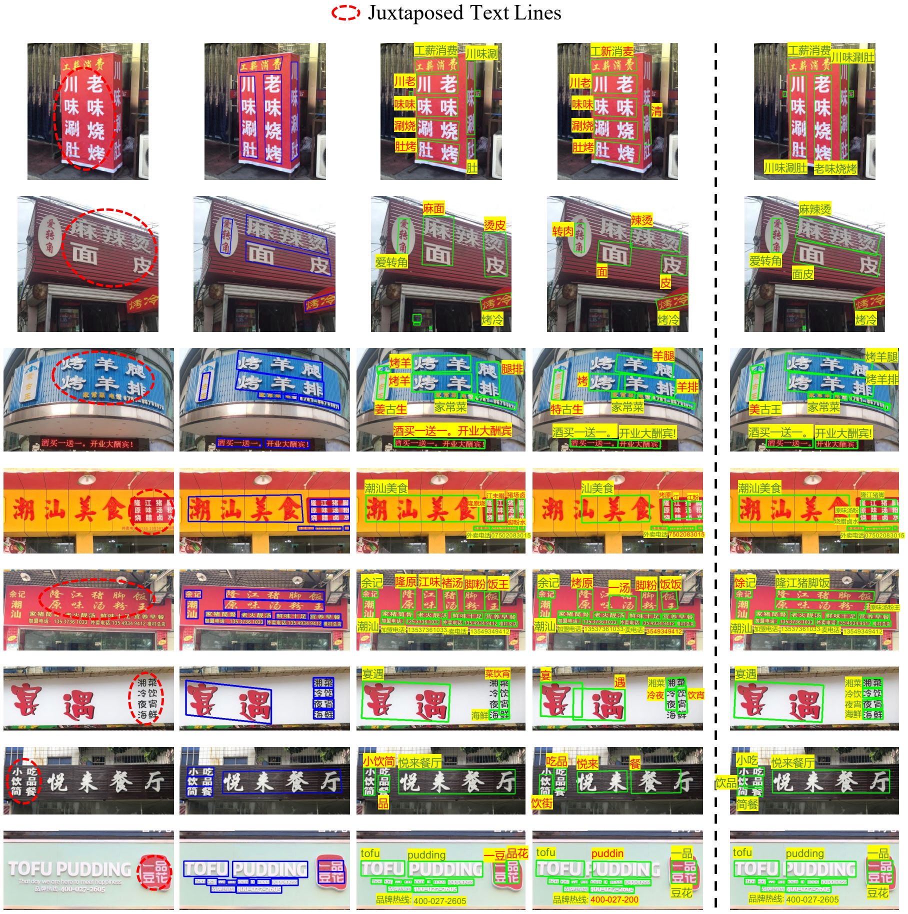
- Detecting and recognizing text lines with large character spacing;
- Detecting and recognizing juxtaposed text lines;
- Detecting and recognizing the text lines with various orientations;
- Detecting and recognizing the multi-language (*e.g.* English, Chinese and Arabic numerals) text lines;

- Thanks to the strong feature representation, our method is also robust to complex and unstable illumination, different colors and variable scales.



(a) Input Image (b) Ground-Truth (c) Mask TextSpotter (d) FOTS (e) Our Method

Fig. 2. Examples of text lines with large character spacing. (a) are original images. (b) are ground-truths. (c) are results of Mask TextSpotter [1]. (d) are results of FOTS [2]. (e) are results of our AE TextSpotter.



(a) Input Image (b) Ground-Truth (c) Mask TextSpotter (d) FOTS (e) Our Method

Fig. 3. Examples of juxtaposed text lines. (a) are original images. (b) are ground-truths. (c) are results of Mask TextSpotter [1]. (d) are results of FOTS [2]. (e) are results of our AE TextSpotter.

References

1. Liao, M., Lyu, P., He, M., Yao, C., Wu, W., Bai, X.: Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE transactions on pattern analysis and machine intelligence (2019)
2. Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., Yan, J.: Fots: Fast oriented text spotting with a unified network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5676–5685 (2018)
3. Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., Wang, L., Wang, D., Liao, M., Yang, M., et al.: Icdar 2019 robust reading challenge on reading chinese text on signboard. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1577–1581. IEEE (2019)