The Trend and Achievement on Scene Text Reading

Anna Zhu, Seiichi Uchida
（AIT-ISEE, Kyushu University)

1 Introduction

Text reading in natural scene images is an open and challenging problem due to the significant variations of the appearance of the text itself and its interaction with the context. In the past decades, the relevant researchers used features like color, edge, local texture and geometry for this task. Nowadays, they prefer to extract character or text line features by deep learning (DL) methods, such as, convolutional neural network (CNN), fully convolutional network (FCN) and recurrent neural network (RNN). With the tool of DL, the evaluation of both text detection and text recognition reached new heights. This paper surveyed the new trends and achievements of this field. We decompose text reading problems to two parts, text detection and text recognition. The traditional methods are firstly reviewed in a short amount of space. Then, we focus on DL-based methods introduction. The comparison and analysis of these two kinds of methods are drawn at last.

2 Scene text detection

Two years ago, the objective of text detection is defined as detecting text components precisely as well as grouping them into candidate text regions with as little background as possible. Text detection methods are roughly classified to two categories, connected component analysis (CCA) -based and sliding window-based. In past decades, CCA based methods, such as widely-used SWT or MSER [1], are mostly built on bottom-up strategy that starts from stroke or character candidate extraction to text line construction and region verification. The Sliding window-based methods use multi-scale windows and machine learning tools for character/non-character classification. However, both kinds of methods consider text as fragments and do character-level classification, which is neither robust nor discriminative.

To overcome the problem, more principled methods that jointly identify a group of text strings on one scale images are proposed. In [2], an R-CNN inspired method generates, and then regresses the word bounding box proposals to text regions. In [3], a FCN model is exploited to predict the saliency of text region in a holistic manner. In [4], a cascaded convolutional text network jointing two customized convolutional networks detect text from coarse to fine. These kinds of methods use DL tools for text representation which are very discriminative. Meanwhile, those models detect text on one-scaled image that avoids computational complexity for exhaustive searching on multi-scale, multi-aspect ratio sliding windows. The overall evaluation is improved to 86% (F-score) on ICDAR 2013 dataset. Thus, the trend can be drawn as designing fast and robust end to end text detection system.

3 Scene text recognition

Text recognition converts image regions into strings. The traditional methods mostly adopt bottom-up approaches, where individual characters are firstly detected using sliding window, connected components, or Hough voting. Then, the detected results are integrated to words by means of dynamic programming, connected random field, lexicon search, etc. Nowadays, the end-to-end text recognition system is preferred and high performances are achieved based on the DL framework. Typically, the CNN is used to extract character-level features or classify characters directly. In [2], a whole-word method recognizes entire text regions by a deep CNN with 90K dictionary. In [5], a robust text recognizer with automatic rectification is proposed which concatenate CNN and long short-term memory (LSTM) to recognize several types of irregular text, including perspective text and curved text. These end-to-end systems require only images and associated text labels for training and are convenient to be deployed in practical systems. The DL-based methods improved the accuracy to 98.6% with full lexicon on ICDAR 2003 datasets.

4 Conclusion

This short paper surveyed the recent trends of applying the deep learning into text reading. The performance is significant. So, is the deep learning tools the terminator for this task? We will see.

References

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

[2] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading text in the wild with convolutional neural networks[J]. International Journal of Computer Vision, 2016, 116(1): 1-20.

[3] Zhang Z, Zhang C, Shen W, et al. Multi-oriented text detection with fully convolutional networks[J]. arXiv preprint arXiv:1604.04018, 2016.

[4] He T, Huang W, Qiao Y, et al. Accurate Text Localization in Natural Image with Cascaded Convolutional Text Network[J]. arXiv preprint arXiv:1603.09423, 2016.

[5] Shi B, Wang X, Lv P, et al. Robust Scene Text Recognition with Automatic Rectification[J]. arXiv preprint arXiv:1603.03915, 2016.