Scene Text Detection and Recognition

1st Tran Kien Quoc - 19127535
Ho Chi Minh University of Science

2nd Pham Duc Duy - 19127379
Ho Chi Minh University of Science

***Abstract*—Scene text recognition has many applications in many fields. This paper explores some solutions for the problem.**

***Index Terms*—scene text, detection, recognition**

I. Introduction

A. Motivation

Text is a system of symbols used to record, communicate, of inherit culture, and is the most important carrier of information in the human world. As one of the most influential inventions of humanity, text has played an important role in human life.

The rich and precise semantic information carried by text is important in a wide range of vision-based application:

- Image search
- Traffic automation
- Industrial automation
- Instant translation

In the past few years, scene text understanding has received rapid growing due to a large amount of everyday scene images that contain texts. Therefore, scene text detection is attracting increasing attention in the computer vision community. Many work has been made and the performance is increase every year with the rapid improvement of some popular method like deep learning, segmentation, classification.

B. Challenges

Recognizing text in natural scenes, also known as scene text recognition (STR), is usually considered as a special form of optical character recognition (OCR). In contrast to OCR in documents, STR remains challenging because of many factors, such as:

- Background: text in natural scenes can appear on anything. Therefore, the text can appear on some complex background making it hard to recognize due to the texture of the background (backgrounds with patterns or objects with a shape that is extremely similar to any text).
- Form: text appears in multiple colors with irregular fonts, different sizes, and diverse orientations. The diversity of text makes STR more difficult and challenging than OCR in scanned documents which has regular font, consistent size, and uniform arrangement.
- Noise: non-uniform illumination, low resolution, and motion blurring of the input image may cause failure to STR.
- Accessibility: scene text is captured randomly, various shape of text cause by perspective, languages... increase the difficulty of recognizing characters and predicting text strings.

C. Problem statements

Given an image with natural background, we have to detect the text deppicted in the image and convert into digital formats.

- Input: An image, grayscale or RGB.
- Output:
  - Detection: Detected text regions, with each characters enclosed in bouding boxes.
  - Recognition: Detected characters are regconized and converted into digital formats, easier for computers to process.

II. Related works

Traditional methods. In this context, "traditional methods" refer to methods that don't rely on deep learning. Traditional methods often regard text detection and recognition as two separate tasks.

- [1] first proposed a Maximally Stable Extremal Regions [2] based method. This involves no training from real world data, and doesn't separate the task of text recognition from text localizaiton, or an end-to-end method.
- [3] proposed a MSER-based method, a method for blob analysis. Character candidates are extracted using MSER, close characters are grouped together with a metric learning algorithm. Then a classifer is used to eliminate blobs with high probability of non-text. Finally text candidates are converted into true text by a text classifier.
- [4] utilizes Higher order correlation clustering [5] after a MSER graph is built.
- [6] proposed edge-preserving MSER (eMSER), and 3 novel cues: Stroke Width, Percetual Divergence, Histogram of Gradients of Edges. Proposed a Bayesian method to integrate these 3 cues. Developed a random Markov field to exploit the inherent dependencies between characters.

Deep learning-based methods.

- EAST [7] is a pipeline consisting of feature extraction using PVANet [8] to extract the features of the image, image segmentation using a Fully Convolutional Network [9] for text segmentation. The post processing steps involve NMS to merge the results.

References

[1] L. Neumann and J. Matas, *A Method for Text Localization and Recognition in Real-World Images*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 6494, p. 770–783.

[2] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” p. 10.

[3] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, “Robust text detection in natural scene images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, p. 970–983, May 2014, arXiv: 1301.2628.

[4] L. Kang, Y. Li, and D. Doermann, “Orientation robust text line detection in natural images,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2014, p. 4034–4041. [Online]. Available: https://ieeexplore.ieee.org/document/6909910

[5] S. Kim, S. Nowozin, P. Kohli, and C. D. Yoo, “Higher-order correlation clustering for image segmentation,” p. 9.

[6] Y. Li, W. Jia, C. Shen, and A. v. d. Hengel, “Characterness: An indicator of text in the wild,” *arXiv:1309.6691 [cs]*, Sep 2013, arXiv: 1309.6691. [Online]. Available: http://arxiv.org/abs/1309.6691

[7] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, “East: An efficient and accurate scene text detector,” *arXiv:1704.03155 [cs]*, Jul 2017, arXiv: 1704.03155. [Online]. Available: http://arxiv.org/abs/1704.03155

[8] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, “Pvanet: Deep but lightweight neural networks for real-time object detection,” *arXiv:1608.08021 [cs]*, Sep 2016, arXiv: 1608.08021. [Online]. Available: http://arxiv.org/abs/1608.08021

[9] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *arXiv:1411.4038 [cs]*, Mar 2015, arXiv: 1411.4038. [Online]. Available: http://arxiv.org/abs/1411.4038