SceneTextReg: A Real-Time Video OCR System

Haojin Yang, Cheng Wang, Christian Bartz, Christoph Meinel
Hasso Plattner Institute (HPI), University of Potsdam, Germany
P.O. Box 900460,
D-14440 Potsdam
{haojin.yang, cheng.wang, meinel}@hpi.de
{christian.bartz}@student.hpi.uni-potsdam.de

ABSTRACT

We showcase a system for real-time video text recognition. The system is based on the standard workflow of text spotting system, which includes text detection and word recognition procedures. We apply deep neural networks in both procedures. In text localization stage, textual candidates are roughly captured by using a Maximally Stable Extremal Regions (*MSERs*) detector with high recall rate, false alarms are then eliminated by using Convolutional Neural Network (*CNN*) verifier. For word recognition, we developed a skeleton based method for segmenting text region from its background, then a CNN based word recognizer is utilized for recognizing texts. Our current implementation demonstrates a real time performance for recognizing scene text by using a standard laptop with webcam. The word recognizer achieves competitive result to state-of-the-art methods by only using synthetical training data.

CCS Concepts

•Computing methodologies → Computer vision; Visual content-based indexing and retrieval; •Computer systems organization → *Real-time systems;*

Keywords

Video OCR; Multimedia Indexing; Deep Neural Networks

1. INTRODUCTION

The amount of video data available on the World Wide Web (*WWW*) is growing rapidly. According to the official statistic-report of YouTube, 100 hours of video are uploaded every minute. Therefore, how to efficiently retrieve video data on the WWW or within large video archives has become a very important and challenging task.

On the other hand, due to the rapid popularization of smart mobile and wearable devices, large amounts of self-recorded “lifelogging” videos are created. Generally, it lacks metadata for indexing such video data, since the only searchable textual content is often the title given by the uploader, which is typically brief and subjective. A more general solution is highly desired for gathering video metadata automatically.

Text in video is one of the most important high-level semantic feature, which directly depicts the video content. In general text displayed in a video can be categorized into scene text and overlay text (or artificial text). In contrast to overlay text, to detect and recognize scene text is often more challenging. There are numerous problems affecting the recognition results, as e.g., texts appeared in a nature scene image can be in a very small size with high variety of contrast; motion changes of the camera may affect the size, shape and brightness of text content, and may lead to geometrical distortion. All of those factors have to be considered in order to obtain a correct recognition result.

Most of proposed scene-text recognition methods can be briefly divided into two categories, either based on connected components (*CCs*) or sliding windows. The CCs based approaches include Stroke Width Transform (*SWT*) [4], MSERs [11], Oriented Stroke [12] etc. One of the significant benefits of CCs based method is its computational efficiency since the detection is often a one pass process across image pixels. The sliding window based methods as e.g., [14, 3, 13, 7] usually apply representative visual features to train a machine learning classifier for text detection. Here hand-crafted features [13, 9, 1] as well as deep features [14, 7] can be applied, and text regions will be detected by scanning the whole image with a sub-window in multiple scales with a potential overlapping. In [14, 3, 7], sliding window based methods with deep features achieved promising accuracy for end-to-end text recognition. [15] propose to consider scene text detection as a semantic segmentation problem, by which their Fully Convolutional Network (*FCN*) performs per-pixel prediction for classifying text and background. However, their proposed approaches may hard to achieve sufficient performance for real-time application due to the expensive computation time.

In our approach, we intended to take advantages from both categories, i.e. the computation benefit of CCs based algorithm and the powerful text-classification ability of deep features. The demonstrated system achieves real-time performance[1] on a standard laptop (3.2 GHz CPU×4, 8G RAM, NVIDIA GeForce 860M) with webcam.

MM ’16 October 15-19, 2016, Amsterdam, Netherlands

ACM ISBN 978-1-4503-3603-1/16/10.
DOI: http://dx.doi.org/10.1145/2964284.2973811

[1] Similar to [11], we consider the real-time ability of a video text recognition system if its response time is comparable to a human.

2. SYSTEM DESIGN

In this section, we will briefly describe the main work flows of the system, and report evaluation results on ICDAR 2015 Robust Reading Competition Challenge 2 - Task 3 “Focused Scene Word Recognition”.

2.1 Text Detection

In [14, 8, 7], the authors were intended to achieve the best end-to-end text recognition accuracy. Therefore in text detection step, their systems have been tuned to produce text candidates with high recall, and the subsequent recognition engines will further eliminate the false alarms. Since recognition procedures are often time consuming, we thus keep the text detection result as accurate as possible, and only pass the text candidates with high confidence to the recognition stage. We apply MESRs [10] based detector to roughly detect character candidates from the input video frame with high recall rate. All candidate regions are further verified by using a grouping method and CNN classifier. The applied binary CNN classifier has been trained by using deep network similar to [5].

2.2 Text Segmentation

We developed a novel skeleton-based approach for text segmentation, which will simplify the further OCR process. In short, we determine the text gradient direction for each text candidate by analyzing the content distribution of their skeleton maps. We then calculate the threshold value for seed-selection by using the skeleton map which has been created with the correct gradient direction. Subsequently, a seed-region growing procedure starts from each seed pixel and extends the seed-region in its north, south, east, and west directions. The region grows iteratively until it reaches the character boundary. This method achieved the first place in ICDAR 2011 text segmentation challenge for born digital images.

2.3 Word Recognition

In this step verified text candidates are first separated into words. Word recognition is accomplished by performing joint-training of CNN and LSTM [6] based Recurrent Neural Network, followed by a standard spell checker. In order to train the deep network we developed a data engine, which created several millions of synthetical training data by considering different scene factors such as font style, background, lighting effect, contrast ratio etc.

2.4 Evaluation Result

We evaluated our word recognizer by using ICDAR 2015 Robust Reading Competition Challenge 2 - Task 3 “Focused Scene Word Recognition” dataset (refer to IC15) in the unconstrained manner[2]. Since the output of our word recognizer is case-insensitive, we thus only consider the evaluation results of the category “Word Recognition Rate (Uppercase)” (WRR-U) from the ICDAR 2015 online evaluation system. We also created evaluation result ignored capitalization and punctuation differences using this dataset (refer to I.C.P), and compared with the best known methods [2, 7].

Table 1 shows the comparison results to previous methods on IC15 dataset. In *I.C.P* evaluation, our current result outperforms JOINT-model from [7], but comes behind Google’s photoOCR by 1.9%. We didn’t considered the DICT-model from [7], since its result is constrained to lexicons. According to the *IC15 ranking* results, our approach is currently not able to outperform the results created by commercial organizations such as SRC-B, Baidu-IDL, Megvii and Google, but improves on the next best one (NESP) by almost 18% of WRR-U. Our result is still competitive to commercial organizations, regarding that we have only used synthetical training data, and applied more succinct network architecture by taking into account of the execution speed. In contrast, [2] applied several millions of manually labelled samples, and the processing time of their system is around 1.4 secs per image. To process a 640 × 480 image, the system from [15] needs about 20 secs on CPU or 1 sec on GPU only for text localization. Therefore, our system is obviously superior regarding running time. This can be proved by our demo video captured by using a laptop and a video camera[3]. The OCR analysis has been performed on every input frame from the camera.

	Description	WRR-U
I.C.P	PhotoOCR [2]	0.876
	Our result	**0.857**
	Jaderberg’s JOINT-model [7]	0.818
IC15	SRC-B-TextProcessingLab*	0.8895
	Baidu-IDL*	0.872
	Megvii-Image++ [15]	0.8603
	PhotoOCR [2]	0.853
	Our result	**0.826**
	NESP	0.6484
	PicRead	0.6192
	PLT	0.6311
	MAPS	0.6329
	Feild’s Method	0.5233
	PIONEER	0.5571
	Baseline	0.4658
	TextSpotter	0.2813

Table 1: Evaluation result on IC15. The baseline method is from a commercially available OCR system. We are intended to include results that are not constrained to a pre-defined lexicon. However the methods marked with * are not published, therefore they are not distinguishable. (Last check: 20/05/2016)

3. DEMO SETUP

We will show case the proposed video ocr system in an interactive manner. The input video stream will be captured by using a live camera, and the ocr result will be directly displayed on the computer screen. The hardware needed to be prepared by the authors are a laptop and a video camera.

4. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[2] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven. Photoocr: Reading text in uncontrolled conditions. In

[2] The OCR results are not constrained to a given lexicon.

[3] https://vimeo.com/140059065
http://www.tudou.com/programs/view/hMkGCsWEe2Y/

The IEEE International Conference on Computer Vision (ICCV), December 2013.

[3] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng. Text detection and character recognition in scene images with unsupervised feature learning. In *Proc. of International Conference on Document Analysis and Recognition*, ICDAR '11, pages 440–445, Washington, DC, USA, 2011. IEEE Computer Society.

[4] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Proc. of International Conference on Computer Vision and Pattern Recognition*, pages 2963–2970, 2010.

[5] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082v4*, 2014.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[7] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Deep structured output learning for unconstrained text recognition. In *International Conference on Learning Representations*, 2015.

[8] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *Proc. of European Conference on Computer Vision (ECCV)*. Springer, 2014.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761 – 767, 2004. British Machine Vision Computing 2002.

[11] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3538–3545, June 2012.

[12] L. Neumann and J. Matas. Scene text localization and recognition with oriented stroke detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 97–104, Dec 2013.

[13] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1457–1464, Nov 2011.

[14] T. Wang, D. Wu, A. Coates, and A. Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308, Nov 2012.

[15] C. Yao, J. Wu, X. Zhou, C. Zhang, S. Zhou, Z. Cao, and Q. Yin. Incidental scene text understanding: Recent progresses on icdar 2015 robust reading competition challenge. *arXiv preprint arXiv:1511.09207v2*, 2016.