An Improved System For Real-Time Scene Text Recognition

Haojin Yang, Cheng Wang, Xiaoyin Che, Sheng Luo, Christoph Meinel
Hasso Plattner Institute (HPI), University of Potsdam, Germany
P.O. Box 900460,
D-14440 Potsdam
{haojin.yang, cheng.wang, xiaoyin.che, sheng.luo, meinel}@hpi.de

ABSTRACT

In this paper we showcase a system for real-time text detection and recognition. We apply deep features created by Convolutional Neural Networks (*CNNs*) for both *text detection* and *word recognition* task. For text detection we follow the common localization-verification scheme which already shown its excellent ability in numerous previous work. In text localization stage, textual regions are roughly detected by using a MSERs (*Maximally Stable Extremal Regions*) detector with high recall rate. False alarms are then eliminated by using a CNNs classifier, and remaining text regions are further grouped into words. In the word recognition stage, we developed an skeleton-based text binarization method for segmenting text from its background. A CNNs based recognizer is then applied for recognizing character. The initial experiments show the powerful ability of deep features for text classification comparing with commonly used visual features. Our current implementation demonstrates real-time performance for recognizing scene text by using a standard PC with webcam.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*

General Terms

Algorithms, Demonstration, Experimentation

Keywords

Video OCR, Scene Text Recognition, Multimedia Indexing, Real-Time System

1. INTRODUCTION

In the last decade digital libraries and web video portals have become more and more popular. The amount of video data available on the World Wide Web (*WWW*) is growing rapidly. According to the official statistic-report of the popular video portal YouTube[1], 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 on the web. Generally, it lacks metadata for indexing such kind of videos, since the only searchable textual content is often the title given by the video 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 features, which directly depicts the video content. In general text displayed in a video can be categorized into scene text and overlay text [9]. 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 OCR (*Optical Character Recognition*) result.

In this work, we address both text detection and recognition issues for video images. In the text detection, we follow the commonly used localization-verification scheme, in which a MSERs (*Maximally Stable Extremal Regions*) [11] detector is applied intended to identify text candidate regions with high recall rate. Then candidate regions are verified by a text/non-text classifier which is trained based on Convolutional Neural Networks (*CNNs*). Finally, remaining text regions are grouped into words relying on their position and size information. For text recognition we have developed a novel skeleton-based binarization algorithm in order to separate text from complex background to make it processible for OCR engines. We further developed a CNN-based word recognizer for processing scene text content. Our initial experimental results demonstrate the superiors ability of CNN features for text classification comparing to other commonly used visual features such as Scale-Invariant Feature Transform (*SIFT*) [10] and Histogram of Oriented Gra-

WOODSTOCK '97 El Paso, Texas USA

[1] https://www.youtube.com/yt/press/statistics.html

dients (*HOG*) [12] etc. The demonstrated system achieved real-time performance[2] on a standard PC platform (3.2 GHz CPU×4, 4G RAM) with webcam.

The rest of the paper is organized as follows: section 2 reviews previous work, whereas the section 3 describes the system architecture and involved techniques. Section 4 provides initial experimental results and demonstrates exemplary real-time recognition results. Section 5 concludes the paper with an outlook on future work.

2. RELATED WORK

Most of proposed scene-text detection 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*) [5], MSERs [13], Oriented Stroke [14] etc. One of the significant benefits of CCs based method is its computational efficiency since the detection is often an one pass process across image pixels. The sliding window based methods as e.g., [17, 3, 16, 7] usually apply representative visual features to train a machine learning classifier for text detection. Here hand-crafted features [16, 10, 2] as well as deep features [17, 3, 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 [17, 3, 7], sliding window based methods with deep features achieved promising accuracy for end-to-end text recognition. However, their proposed approaches may hard to achieve sufficient performance for real-world 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.

3. SYSTEM DEMONSTRATION

In [17, 7], the authors were intended to achieve the best end-to-end text recognition accuracy. Therefore in the 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. At this point, we argue that this kind of design may lower the overall system efficiency. From our experiences we learned that the OCR engine is normally less efficient than the text detector for processing same amount of image pixels concerning running time. Since we want to design a system with real-time capability, we thus keep the text detection process as accurate as possible, and only pass the text candidates with high confidence to the recognition stage.

3.1 Text Localization

In this step we apply a MESRs based detector to roughly detect text regions from the input video frame with high recall rate. MSERs define an Extremal Region (*ER*) as a connected component of image pixels having intensity contrast against its boundary pixels [11]. Those ERs can be controlled by tuning the intensity thresholds. A low threshold would result in a large number of low-level ERs which are separated by small contrast differences amount pixels. While increasing the threshold the low-level ERs will be merged to construct a higher level one. In our current implementation we adapted the detector to ensure that 95% of text regions will be remained as candidates from the input images. All candidate regions will be further verified by using a grouping method and CNN classifier.

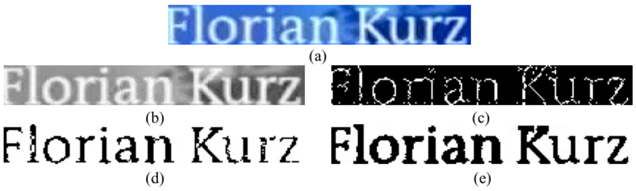

Figure 1: Seed selection and region growing results: (a) original text line image, (b) grayscale image, (c) skeleton image, (d) seed-selection result, (e) region growing result

3.2 Text Verification

In the verification stage, we first roughly generate the text-region-pairs from ERs according to their positions. Then each region-pair will be verified by using several heuristics including normalized color distance, object-center distance, minimal height ratio, maximal intensity distance. This procedure can help to filter out many simple false alarms. The remaining difficult non-text patterns as e.g., windows, blocks, trees, garden fens will be handled by CNN classifier.

To verify the text regions, we trained a text/non-text classifier based on CNNs. Our networks have two convolutional layers with 20, 50 filters respectively. We apply *Maxout* pooling layer [6] after each convolution with group size 2. The input image size is 24×24 and the convolutional kernel size is 5×5. The sequence continues by a inner-product layer (500 filters), ReLU layer and another inner-product layer (2 filters). Inspired by [6] and [7] we also utilized dropout (0.5) combined with maxout layers. The output of the last layer is fed into a SVM [4] classifier to obtain a binary decision. We gathered about 180k positive and 360k negative character samples from various datasets to build the training and test dataset. The verification process follows the sliding-window metrics, and we use majority voting to distinguish text objects and false alarms.

3.3 Text Segmentation

Text pixels from the detected text lines need to be separated from background by applying appropriate segmentation/binarization techniques for further OCR processing. We developed a novel skeleton-based approach for video text binarization, which consists of three steps: First, we determine the text gradient direction for each text line object 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 orientations. The region grows iteratively until it reaches the character boundary. Figure 1 shows the workflow of the algorithm. Our skeleton based binarization methods achieved the first and second place in ICDAR 2011 text segmentation challenge for born digital images. More details of this method can be

[2] Similar to [13], we consider the real-time ability of a video text recognition system if its response time is comparable to a human.

found in [18].

3.4 Word Recognition

In this step verified text objects are first separated into words. Then we also apply sliding-window metrics to recognize characters within each word. We use a similar setup for training the character recognizer as the text/non-text classifier. The only difference is that the output of the second inner-product layer is 62, and the final output is fed into softmax-classifier for creating 62-way classification result. Unlike the recognition procedure in [17, 7], before character recognition we try to find the character positions and boundaries by using contour detection method. Subsequently, for each character candidate we produce a set of recognition responses by adjusting the sliding-window positions, e.g. shift the boundary position to left or to right with certain percentages. The word response is created based on character candidates with best response-scores. Here several post-processing techniques could be applied to further improve the recognition result. As e.g., by using non-maximal suppression we could remove the duplicated words from the result. We could also use spell-checking tool to create suggested word lexicon for each word response, and apply beam search algorithm to find the final result. The post-processing procedures will be studied and implemented as the next step of our work.

4. INITIAL EXPERIMENTAL RESULTS

Most of machine learning based text detection and recognition methods take use of representative features to discriminate text from other objects. The ability of applied visual features directly affect the system accuracy and efficiency. Deep learning techniques have been applied to numerous challenging research problems, and created break-record improvements in recent several years. In this work, we conducted experiments to investigate the accuracy of CNNs deep features comparing to several commonly used visual features for text classification. The selected hand-crafted visual features include SIFT [10], SURF [2], HOG [12] and eLBP (*Edge-based Local Binary Pattern*) [1]. The parameters of each hand-crafted feature have been optimized by using exhaustive-search metrics. Our CNNs classifier is trained based on *Caffe* framework [8].

We utilized the commonly used evaluation methodology: classification accuracy and F-measure scores in the experiment. The evaluation dataset was created by Wang et al.[17], which consists of 15000 (text: 5000, non-text: 10000) cropped character image samples for training and another 7500 (text: 2500, non-text: 5000) samples for testing. All samples were extracted from ICDAR 2003 dataset. SVM classifier with RBF-kernel has been used for all test runs. The parameter optimization of SVM has been executed by cross-validation.

Table 1 shows the classification results, from which it is easy to realize that the CNNs deep features significantly outperformed the other visual features for scene text classification. It once again proved the strong ability of deep features for solving computer vision problem. CNNs feature improves on the second best one (eLBP) by **14%** of classification accuracy.

Figure 2 demonstrates an exemplary end-to-end text recognition result of our system by using a webcam in real-time.

Feature	Accuracy	F_1 Measure	Recall	Precision
HOG	0.58	0.27	0.32	0.23
SURF	0.78	0.58	0.46	0.81
SIFT	0.78	0.65	0.69	0.61
eLBP	0.83	0.73	0.76	0.69
CNNs	**0.97**	**0.95**	**0.93**	**0.97**

Table 1: Text classification results by using hand crafted visual features and CNNs deep features

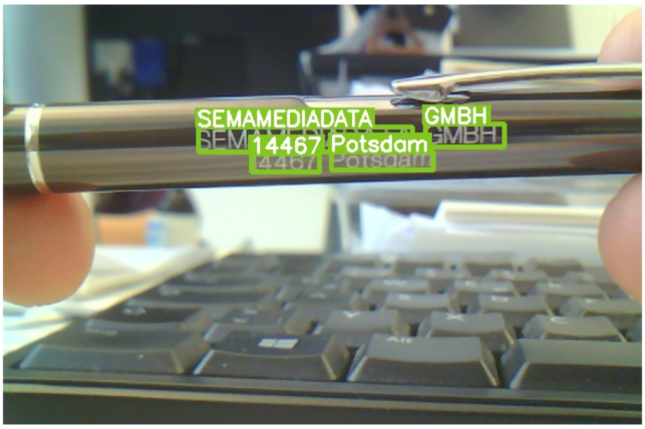

Figure 2: Exemplary recognition result of our system by using a webcam in real-time

The detailed evaluation of our skeleton-based text segmentation method can be found in [18].

5. CONCLUSION AND FUTURE WORK

In this paper, we showcased a system for text detection and recognition in nature scene videos. This system achieved real-time recognition performance by using a standard computer with webcam. The proposed system consist of text detection, text verification, text segmentation and word recognition processes. We use a MESRs detector to rapidly create text candidate regions, and verify them by applying a CNNs based text classifier. The text contents are extracted by using a novel skeleton-based binarization method, which is followed by a CNNs based word recognizer for obtaining end-to-end recognition result.

In the future work, as inspired by [15], we plan to investigate various filter-operations in CNNs, intended to further improve system efficiency. In this paper, so far, we only reported our initial experimental results. The detailed evaluation for text detection, word recognition and end-to-end text recognition will be performed by using appropriate ICDAR benchmarks.

6. REFERENCES

[1] M. Anthimopoulos, B. Gatos, and I. Pratikakis. A two-stage scheme for text detection in video images. *Journal of Image and Vision Computing*, 28:1413–1426, 2010.

[2] 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.

[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] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[5] 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.

[6] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1319–1327, 2013.

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

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[9] K. Jung, K. I. Kim, and A. K. Jain. Text information extraction in images and video: a survey. *Pattern Recognition*, 37(5):977–997, 2004.

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

[11] 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.

[12] B. T. N. Dala. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[13] 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.

[14] 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.

[15] A. Sironi, B. Tekin, R. Rigamonti, V. Lepetit, and P. Fua. Learning separable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints):1, 2014.

[16] 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.

[17] 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.

[18] H. Yang, B. Quehl, and H. Sack. A framework for improved video text detection and recognition. *Multimedia Tools and Applications*, pages 1–29, 2012. 10.1007/s11042-012-1250-6.