https://doi.org/10.5392/IJoC.2018.14.1.001

Deep-Learning Approach for Text Detection Using Fully Convolutional Networks

Trieu Son Tung and Gueesang Lee
Dept. of ECE,
Chonnam National University, Kwangju, 500-070, Korea

ABSTRACT

Text, as one of the most influential inventions of humanity, has played an important role in human life since ancient times. The rich and precise information embodied in text is very useful in a wide range of vision-based applications such as the text data extracted from images that can provide information for automatic annotation, indexing, language translation, and the assistance systems for impaired persons. Therefore, natural-scene text detection with active research topics regarding computer vision and document analysis is very important. Previous methods have poor performances due to numerous false-positive and true-negative regions. In this paper, a fully-convolutional-network (FCN)-based method that uses supervised architecture is used to localize textual regions. The model was trained directly using images wherein pixel values were used as inputs and binary ground truth was used as label. The method was evaluated using ICDAR-2013 dataset and proved to be comparable to other feature-based methods. It could expedite research on text detection using deep-learning based approach in the future.

***Key Words**: Text Detection, FCN, Deep Learning, Nature Scene Image.*

1. INTRODUCTION

As a product of human abstraction and manipulation, the text in natural scenes directly carries high-level semantics. This property makes the text that is present in natural images and videos a special, important source of information, as can be seen in Fig. 1. The rich and precise information that is embodied in text can be very beneficial to a variety of vision-based applications such as image search, target geolocation, human-computer interactions, robot navigation, and industrial automation. Consequently, automatic text detection that offers a means to access and utilize textual information in images and videos has become an active topic in the computer-vision and document-analysis research fields.

However, natural-scene textual location is an extremely difficult task. The major challenges in the scene text detection can be generally categorized into the following three types [1], [2]:

1) Diversity of the scene text: In contrast to the characters in document images, which are usually of regular fonts, single colors, consistent sizes, and of uniform arrangements, natural-scene texts may bear entirely different fonts, colors, scales, and orientations, even in the same scene.
2) Complexity of the background: The backgrounds in natural-scene images and videos can be very complex. Elements like signs, fences, bricks, and grasses are virtually undistinguishable from true text, and therefore, they can easily cause confusion and errors.
3) Interference factors: Various interference factors- for instance, tonoise, blurring, distortion, low resolutions, nonuniform illumination, and partial occlusion-may give rise to failures in the scene text detection.

Fig. 1. Examples of the Text in Natural-scene Images [3].

To tackle these challenges, a rich body of approaches has been proposed, and substantial progress has been achieved. Over the past two decades, researchers have proposed numerous methods for the text detection in natural images or videos. The current approaches for text detection mostly employ a bottom-up pipeline, and commonly, their starting points are aspects like low-level characters or techniques such

* *Corresponding author, Email: gslee@jnu.ac.kr*
Manuscript received Aug. 21, 2017; revised Jan. 30, 2018; accepted Mar. 15, 2018

as stroke detection. In spite of extensive research on feature based methods, the performance has been saturated and a revolutionary approach has be employed to record a breakthrough.

Recently, the architectures that are based on the deep *Convolutional Neural Networks* (CNNs) have advanced the general-object detection process substantially [25]. Especially, the *Fully Convolutional Network* (FCN) achieved a great semantic-segmentation success using a pixelwise prediction [26]. Fig. 2 shows how FCN works. Instead of binarized results or classifications in the upper part of Fig. 2, FCN generates a heat map as the final result as shown in the lower part of Fig. 2. This idea is applied to text detection which inputs a natural scene image and the detection results appear as a heat map which is the exact response required for text detection.

However, due to the lack of a deep supervision, the multiscale responses that are produced at the hidden layers are less meaningful, leading to less satisfactory results. In this paper, a *Supervised FCN* for the text detection in natural-scene images is presented. The present paper is structured as follows: Section 2 explains the existing feature based methods and Section 3 provides the details of the method, Section 4 presents the experiment results, and Section 5 addresses the conclusion.

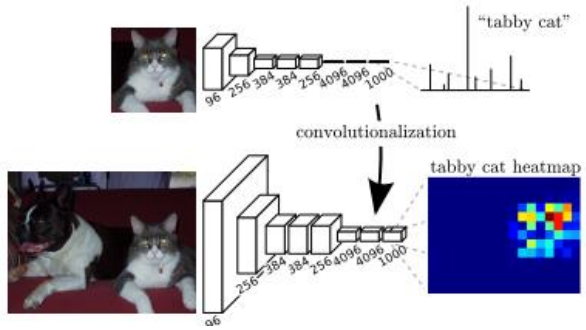


Fig. 2. Transforming fully connected layers into convolution layers enables a classification net to output a heat map. Adding layers and a spatial loss produces an efficient machine for end-to-end dense learning [26].

2. FEATURE BASED RELATED WORKS

The following three main method types exist: texture-based methods, component-based methods, and hybrid methods.

The texture-based methods [4]-[7] treat texts as a special type of texture and make use of their textual properties such as the local intensities, filter responses, and wavelet coefficients to distinguish between the textual and nontextual areas in the images. These methods mostly handle horizontal texts and are sensitive to rotations and scale changes.

To handle multilingual texts (mainly Chinese and English), Lyu et al. [8] proposed a coarse-to-fine multiscale search scheme. The scheme uses properties such as the strong edges and high contrast of texts to distinguish between the textual and nontextual regions. Moreover, this algorithm provides a local adaptive binary strategy to segment the detected textual areas. Similar to many other approaches, this method involves numerous rules and parameters, so it is very difficult to deal with different qualities and texts of different types.

Different from the conventional methods, Zhong et al. [9] proposed an interesting algorithm that can directly detect the text in the discrete-cosine-transform (DCT) domain. The advantage of this algorithm lies in its high efficiency, as a predetection image decoding is unnecessary; however, the detection accuracy of this method is limited.

To speed up the text-detection procedure, Chen et al. [4] proposed a fast text detector. The detector is a cascade Adaboost [10] classifier in which each weak classifier is trained from a feature set. The feature pool includes the mean strength, intensity variance, horizontal difference, vertical difference, and gradient histogram. The detection efficiency of this method is significantly higher than those of the other algorithms [11]-[13], but the detection accuracy is limited regarding real-world images.

Recently, Wang et al. [14] proposed a method for the locating of specific words in natural scenes. First, the single characters are detected using a sliding window. Then, the possible combinations are scored according to the intercharacter structural relationships. Finally, the most similar combinations are selected from the given list as the output results. Unlike the traditional text-detection methods, this algorithm can only detect the given-list words and is incapable of handling words beyond this list. In reality, however, a word list that contains all of the possible cases is not always available for each image, narrowing the method-applicability range compared with the other text-detection methods.

The component-based methods [15]-[19] first extract the candidate components using a variety of ways (e.g., color clustering or extreme region extraction), and then they filter out the nontextual components using manually designed rules or automatically trained classifiers. Generally, the efficiency of these methods is much higher because the number of to-be processed components is relatively small. In addition, these methods are insensitive to rotations, scale changes, and font variations. In recent years, the component-based methods have become the mainstream option in the field of scene text detection.

Making use of the nearly constant character-stroke width property, Epshtein et al. [15] proposed the following new image operator: *Stroke Width Transform* (SWT). This operator provides an easy way to recover the character strokes from edge maps and it can efficiently extract the textual components of different scales and directions from complex scenes. However, this method also comes with a series of human-defined rules and parameters, and only horizontal texts are considered.

Based on SWT [15], Yao et al [16] proposed an algorithm that can detect the text of the arbitrary orientations in natural images. This algorithm is equipped with a two-level classification scheme and two sets of rotational and rotation-invariant features that are specially designed for the capturing of the intrinsic characteristics of the characters in natural scenes.

Huang et al. [17] presented a new SWT-based operator called *Stroke Feature Transform* (SFT). To solve the mismatch problem of the edge points in the original SWTs, the SFT introduced a color consistency and constrains the relations of

the local edge points, producing improved component-extraction results. The SFT detection performance on the standard datasets is significantly higher than those of the other methods, but only for horizontal text.

Neumann et al. [20] proposed a text-detection algorithm based on the maximally stable extremal regions (MSER). This algorithm extracts the MSER regions from the original images as candidates, and eliminates the invalid candidates using a trained classifier. At a later stage, the remaining candidates are grouped into textual lines using a series of connection rules. Such connection rules, however, can only adapt to horizontal or nearly horizontal text, so this algorithm is unable to handle texts with larger inclination angles.

SWT [15] and MSER [20] are two representative methods in the field of scene text detection that constitute the basis of many of the subsequent works [16].

The great success of the sparse representation in facial recognition and image denoising has inspired numerous research studies [21], [22]. For example, Zhao et al. [23] constructed a sparse dictionary from training samples and used it to judge whether a particular area in an image contains text. However, the restricted generalization ability of the learned sparse dictionary means that this method is unable to handle issues like rotations and scale changes.

Shivakumara et al. [24] also proposed a method for multioriented text detection. This method extracts the candidate regions by implementing a clustering in the Fourier–Laplace space, and the regions are then divided into distinct components using *skeletonization*. However, these components generally only correspond to text blocks instead of strokes or characters as well. This method cannot be directly compared with the other methods quantitatively, since it is unable to detect the characters or words directly.

3. PROPOSED METHOD

In the past few years, most of the leading methods in scene text detection are based on character detection. In the early practices, a large number of manually designed features were used to identify the characters with strong classifiers. Recently, a number of works have achieved great performances [27], [28], where the CNN has been adopted as a character detector. However, the performance of the character detector is limited due to the following three aspects: First, the characters are susceptible to several conditions such as blurring, nonuniform illuminations, low resolutions, and disconnected strokes; second, a great quantity of the background elements are similar in appearance to the characters, making them extremely difficult to distinguish; and third, the variations of the actual characters such as the fonts, colors, and languages increases the classifier-learning difficulty. By comparison, text blocks possess more distinguishable and stable properties. Both local and global text-block appearances are useful cues to distinguish between the textual and nontextual regions.

The FCN, the proposed deep CNN, has achieved a great performance regarding the pixel-level recognition tasks such as the object segmentation [26] and the edge detection [29]. This kind of network is very suitable for the detection of text blocks, owing to several advantages, as follows:

- It simultaneously considers both the local and global contextual information.
- It is trained in an end-to-end manner.
- As it benefits from the removal of the fully connected layers, the FCN is efficient in pixel labeling.

With such great successes, however, the lack of a deep supervision at the intermediate stages means that the responses produced at the hidden layers are less meaningful, leading to less satisfactory results. The proposed network is used to label the salient text-block regions in a holistic manner.

The Visual Geometry Group (VGG) 16-layer net [30] was converted into the proposed text-block detection model that is illustrated in Fig. 2. The first five convolutional stages are derived from the VGG-16 net. The receptive field sizes of the convolutional stages are variable, contributing to the capturing of the contexual information of different sizes at different stages. Each convolutional stage is followed by a deconvolutional layer to generate feature maps of the same size. Finally, the fully connected layers are replaced at each stage with a 1x1 convolutional layer and a sigmoid layer to efficiently derive the pixel-level prediction. The network is shown in Fig. 3 and an example in shown in Fig. 4.

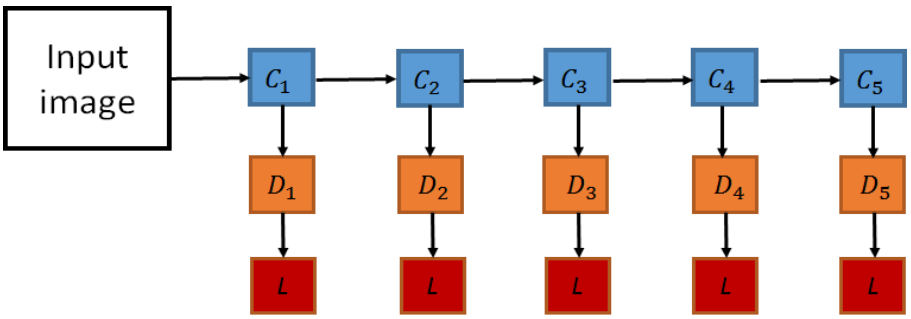
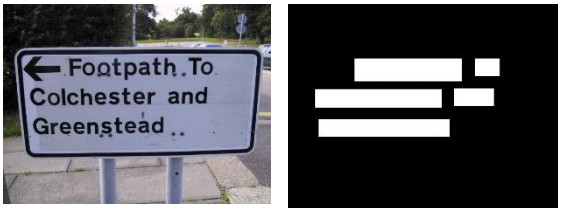


Fig. 3. FCN architecture used in the proposed method

(a) (b) (c) (d) (e) (f)

Fig. 4. Example of Feature-map Generation from the Usage of the Supervised Fully Convolutional Network (FCN). (a) Input image, (b)-(e) feature maps from stages 1-4, and (f) feature maps for stage 5 and the final salient map

In the training phase, the pixels within the bounding box of each text line or word are considered as the positive region for the following reasons: First, the regions between the adjacent characters are distinct in contrast to the other nontextual regions; second, the global textual structure can be incorporated into the model; and third, the bounding boxes of the text lines or words can be easily annotated and obtained. An example of the ground-truth map is shown in Fig. 5. The sigmoid-loss function and the stochastic gradient descent are used to train this model.



(a) An input image (b) The ground-truth map

Fig. 5. Illustration of the ground-truth map

In the testing phase, the salient map of the textual regions is first computed using the trained Supervised FCN model. Then, the pixels whose values are larger than 200 are labeled as the foreground, whereas those with values that are less than 200 are labeled as the background. Finally, the connected pixels are grouped together to form the text blocks. An example of the text-block detection is shown in Fig. 6.

Fig. 6. Detection Examples of the Proposed Method

4. EXPERIMENT RESULTS

The proposed method was evaluated using the International Conference on Document Analysis and Recognition (ICDAR)-2013 dataset. The ICDAR-2013 dataset is a horizontal-text database that was used in the 2011–2013 ICDAR competitions. This dataset consists of 229 training images and 233 testing images. A number of the result examples are shown in Fig. 5. Our proposed method performs comparable to other state of the art methods. Especially FCN based approach for multi-oriented text has recorded the top precision, however our method is better in recall. Even though overall scale in F-measure does not outperform all of the existing method, we believe that our method can be improved with proper pre-processing or post-processing. Recall is the rate of true positives over false negatives. If recall is low, only a part of true positive is guaranteed and there are more false negatives, which means the target object is not fully extracted. In text detection, if the bounded rectangle includes higher rate of recall, the result can be further improved by proper post-processing. Pre-processing for the text detection includes binarization or feature selection for the detection algorithm, which could result in different outcome. Post-processing is any kind of improvement to the generated result, which includes application of another approach to the first generated result.

In the training phase, 10,000 450-x-450 patches were randomly cropped from the training dataset as training examples. To compute the salient testing-phase map, each image was resized to three scales due to the memory limitation, so the respective maximum values between the height and the width are 300, 400, and 500 px. The proposed network was fine-tuned using the pretrained VGG-16 network. The learning rate is 10^{-8} and the number of iterations is 80000.

The comparison is supported by the precision and the recall, which are shown in Table 1 and are represented by the following equations.

$$\text{F-score} = 2 * \frac{\text{Precision*Recall}}{(\text{Precision+Recall})} \quad (1)$$

Table 1. Performance of Different Algorithms Evaluated on the International Conference on Document Analysis and Recognition (ICDAR)- 2013 Dataset.

Method	Precision	Recall	*F*-measure
Proposed	0.75	**0.84**	**0.76**
CASIA_NLPR [3]	0.79	0.68	0.73
I2R_NUS_FAR [3]	0.75	0.69	0.72
Oriented FCN [28]	0.88	0.78	0.83
Yin et al. [31]	0.84	0.65	0.73
Text Spotter [32]	0.88	0.65	0.74
Symmetry [33]	0.88	0.74	0.80

5. CONCLUSION

In this paper, a novel text-detection framework for natural-scene images is presented. The main idea is the use of the Supervised FCN for semantic labeling. Based on the superior performance of the proposed method compared with those of the other competing methods, it has been verified that the combining of the local and global properties for the purpose of text localization is research-worthy.

ACKNOWLEDGMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by MEST (NRF-2015R1D1A1A01060172 and NRF-2017R1A4A1015559) and by Chonnam National University(Grant no. 2016-2884). The corresponding author is Gueesang Lee.

REFERENCES

[1] C. Yao, X. Zhang, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Rotation-invariant features for multi-oriented text detection in natural images," PLoS One, vol. 8, no. 8, 2013.

[2] C. Yao, X. Bai, B. Shi, and W. Liu. Strokelets, "A learned multi-scale representation for scene text recognition," CVPR, 2014.

[3] ICDAR 2013 robust reading competition, http://dag.cvc.uab.es/icdar2013competition, 2014.

[4] X. Chen and A. Yuille, "Detecting and reading text in natural scenes," CVPR, 2004.

[5] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images," Pattern Recognition, vol. 28, no. 10, 1995, pp. 1523-1535.

[6] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," IEEE Trans. PAMI, vol. 25, no. 12, 2003, pp.1631-1639.

[7] J. Gllavata, R. Ewerth, and B. Freisleben, "Text detection in images based on unsupervised classification of high-frequency wavelet coefficients," ICPR, 2004.

[8] B. Leibe and B. Schiele, "Scale-invariant object categorization using a scale-adaptive mean-shift search," Pattern Recognition, 2004, pp. 145-153.

[9] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," IEEE Trans. CSVT, vol. 15, no. 2, 2005, pp. 243-255.

[10] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," IEEE Trans. PAMI, vol. 22, no. 4, 2000, pp. 385-392.

[11] P. Viola and M. Jones, "Fast and robust classification using asymmetric adaboost and a detector cascade," Proc. of NIPS, 2001.

[12] V. Wu, R. Manmatha, and E. M. Riseman, "Finding text in images," ACM Int. Conf. Digital Libraries, 1997.

[13] C. Wolf and J. M. Jolion, "Extraction and recognition of artificial text in multimedia documents," Formal Pattern Analysis and Applications, vol. 6, no. 4, 2004, pp. 309-326.

[14] K. Wang and S. Belongie, "Word spotting in the wild," ECCV, 2010.

[15] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," *CVPR, 2010.*

[16] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," CVPR, 2012.

[17] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," IEEE Trans. Image Processing, vol. 20, no. 9, 2011, pp. 2594-2605.

[18] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," ICCV, 2013.

[19] A. Jain and B. Yu, "Automatic text location in images and video frames," Pattern Recognition, vol. 31, no. 12, 1998, pp. 2055-2076.

[20] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," ACCV, 2010.

[21] J. Wright, A. Y. Yang, and A. Ganesh, "Robust face recognition via sparse representation," IEEE Trans. PAMI, vol. 31, no. 2, 2009, pp. 210-227.

[22] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," IEEE Trans. Image Processing, vol. 15, no. 12, 2006, pp. 3736-3745.

[23] M. Zhao, S. Li, and J. Kwok, "Text detection in images using sparse representation with discriminative dictionaries," Image and Vision Computing, vol. 28, no. 12, 2010, pp. 1590-1599.

[24] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A laplacian approach to multi-oriented text detection in video," IEEE Trans. PAMI, vol. 33, no. 2, 2011, pp. 412- 419.

[25] Y. Pan, X. Hou, and C. Liu, "A hybrid approach to detect and localize texts in natural scene images,"

IEEE Trans. Image Processing, vol. 20, no. 3, 2011, pp. 800-813.
[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," CVPR, 2015.
[27] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," ECCV, 2014.
[28] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," CVPR, 2016.
[29] S. Xie and Z. Tu, "Holistically-Nested Edge Detection," ICCV, 2015.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," ICLR, 2015.
[31] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao, "Multi-orientation scene text detection with adaptive clustering," IEEE Trans. on PAMI, vol. 37, no. 9, Jan. 2015, pp. 1930-1937.
[32] A. Zamberletti, L. Noce, and I. Gallo, "Text localization based on fast feature pyramids and multi-resolution maximally stable extremal regions," ACCV workshop, 2014.
[33] Z. Zhang, W, Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," CVPR, 2015.

Trieu, Son Tung
He received the B.S. degree in Microelectronics from Hanoi University of Science and Technology (HUST) in 2014. He received M.S. degree from the Electronics and Computer Science department of Chonnam National University, Republic of Korea in 2017. His research interests are multimedia and image processing, vision tracking, and pattern recognition.

GueeSang Lee
He received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, Republic of Korea, in 1980 and 1982, respectively. He received the Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering at Chonnam National University, Republic of Korea. His primary research interests are image processing, computer vision, and video technology.