International Journal of Control, Automation and Systems 17(6) (2019) 1599-1610
http://dx.doi.org/10.1007/s12555-018-0578-8

ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Text Detection with Deep Neural Network System Based on Overlapped Labels and a Hierarchical Segmentation of Feature Maps

Hong-Hyun Kim, Jea-Ho Jo, Zhu Teng, and Dong-Joong Kang*

Abstract: This paper proposes a three-level framework to detect texts in a single image. First, a salient feature map of text is extracted using a Fully Convolutional Network (FCN) that achieves good performance in semantic segmentation. Label combination using both boxes of word and characters level is proposed to improve the detection of uneven boundaries of text regions. Second, in the feature map of FCN, the text region has a higher probability value than the background region, and the coordinates in the character area are very close to each other. We segment the text area and the background area by using the characteristics of text feature map with Hierarchical Cluster Analysis (HCA). Finally, we applied a Convolutional Neural Networks (CNN) to classify the candidate text area into text and non-text. In this paper, we used CNN which can classify 4 classes in total by separating the background area and three text classes (one character, two characters, three characters or more). The text detection framework proposed in this paper have shown good performance with ICDAR 2015, and high performance especially in Recall criterion, finding more texts than other algorithms.

Keywords: Deep neural netwrok, detection framework, text detection, text localization.

1. INTRODUCTION

With rapid advances in information technology, the computer vision technologies based on machine learning have been increasingly in demand. The development of computational hardware has made it possible to use techniques that were previously difficult to use because of their high computational complexity. In addition, the large amount of data that can be collected via the internet has provided growth for intelligent technologies with high recognition ability. Booming advance in machine learning for visual analysis is now able to distinguish objects with complex and dynamic shapes, such as people, cars, and airplanes, from images containing complex backgrounds. Such technologies are leading the development of new intelligent industries based on products including unmanned vehicles and intelligent robots. Image data contains real world information and directly provide information for tasks such as determining the position of an object, and text recognition.

In this work, we address text recognition in images. Among the various types of information included in images, the text is one of the most specific forms of information that can be used to express the properties of an object.

Text can provide precise local information that can be used for various purposes, such as identifying a surrounding location or building, as well as distance analysis and website searches. People easily recognize text under complicated backgrounds. Even if they cannot determine what it means, they can discriminate whether it is text or not. On the other hand, machines do not have this innate capability. Automated text recognition presents many difficulties.

The text includes a large number of fonts, sizes, colors, and languages. It has inconsistent features. Therefore, it is not easy for a machine to distinguish between text and background in images. In a complex street view of a city, for example, it is first necessary to detect the position corresponding to the text, before recognizing the text.

A Fully Convolutional Network (FCN) is robust against noise and exhibits good performance at pixel level classification, such as object segmentation and edge detection [1]. FCN replaces the fully connected layer at the top of an existing Convolutional Neural Network (CNN) [2] used for classification with a convolution layer. This makes it possible to use the position information of objects in images.

Manuscript received August 13, 2018; revised December 26, 2018; accepted January 30, 2019. Recommended by Associate Editor Kang-Hyun Jo under the direction of Editor Euntai Kim. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (2016R1A2B4007608), and National IT Industry Promotion Agency(NIPA) grant funded by the Korea government(MSIT) (S0249-19-1019).

Hong-Hyun Kim, Jae-Ho Jo, and Dong-Joong Kang are with the School of Mechanical Engineering, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea (e-mails: {hong12, tjdakqjqtk, djkang}@pusan.ac.kr). Zhu Teng is with the School of Computer and Information Technology, Beijing Jiaotong University, No. 3 Shangyuancun Haidian District Beijing 100044, P. R. China (e-mail: zteng@bjtu.edu.cn).
* Corresponding author.

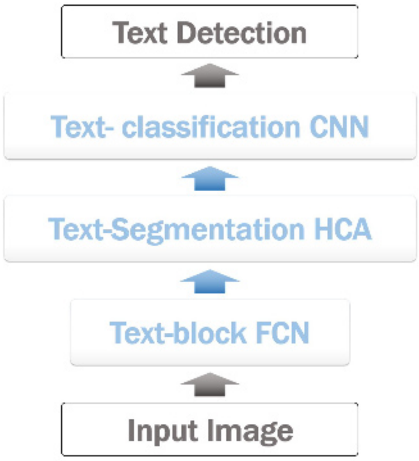


Fig. 1. System flowchart for text detection.

In addition, the size of the input image is not limited.

In this work, we use FCN for text detection in images with complex backgrounds. Fig. 1 shows the flowchart of the algorithm proposed in this paper. Through FCN, the input image provides a salient feature map, which takes a block in the text area. Next, pixels corresponding to the text in the feature map are clustered using Hierarchical Cluster Analysis (HCA) [3] to distinguish the text area from the background area. The divided regions are classified into text and non-text using CNN.

Our contributions can be summarized as follows:

First, we propose a new labeling method for FCN network learning. This method uses a character-box label that represents individual characters as well as a word-box label that indicates the location of text group. In this way, the network can consider the character spacing, color, size, etc., which can vary between individual characters, while representing the overall text characteristics of combinations of individual characters.

Second, we apply HCA algorithm to the extracted feature map to precisely segment the text area and the background area. We show that this method performs better than the binary classification method using a certain threshold.

Our third contribution is to subdivide the category of the text into more detailed classes when learning CNN for classifying text/non-text. The proposed method achieved high performance in text detection and false positive removal.

The proposed method consists of three steps using a hierarchical approach. Each step works independently and this method has the advantage in terms of expandability.

The paper is organized as follows: Section 2 analyzes the existing text detection method in detail. Section 3 describes the method used for text detection in this study. Experiments and results are presented in Sections 4 and 5, respectively.

2. RELATED WORKS

2.1. Text localization

The conventional approach for text recognition is mainly based on a sequential procedure [4, 5]. Fig. 2 shows the process of recognizing text in an image, which involves text detection and text recognition. The text detection step includes localization to know where the text is located and verification to discriminate whether the detected area is text. Text recognition consists of segmenting each character and identifying what the segmented characters are.

In this work, we studied text detection during text analysis. Several methods for detecting text have been used, such as the sliding window based method [6–10] and the connected component analysis (CCA) method [4, 11–15]. The sliding window based method searches for text while a specified window slides within an image, recognizing text in the window using machine learning. CCA method extracts a candidate region of letters from an image, analyzing the rules between letters and clustering it into text. The Maximally Stable Extremal Region (MSER) algorithm is mainly used in CCA method [11, 14, 15]. MSER is an algorithm that detects all areas where the brightness value can be discriminated from the surrounding area. It detects not only the text areas but also the non-text areas that are distinct from the background areas. It is important to minimize the number of non-text areas when detecting text. Recently, CNN [2] based on Deep Neural Networks (DNN) have been used to extract features for character recognition [6, 16–18].

2.2. DNN (Deep neural network) – based detection

Recently, CNN based on DNN has been utilized in various detection problems [2, 19–22]. It is also used in text detection [16]. CNN relies on training data to generate feature maps for classifying objects. This is useful for detecting text with inconsistent characteristics that are difficult to model in hand-crafted features. On the other hand, FCN with all layers replaced with convolutional layer achieved good performance in the text detection field [1]. FCN is not restricted by the size of the input image.

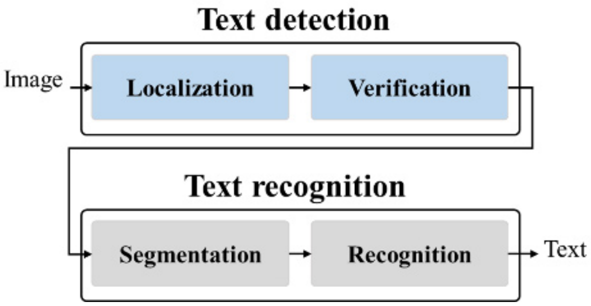


Fig. 2. Process of text recognition.

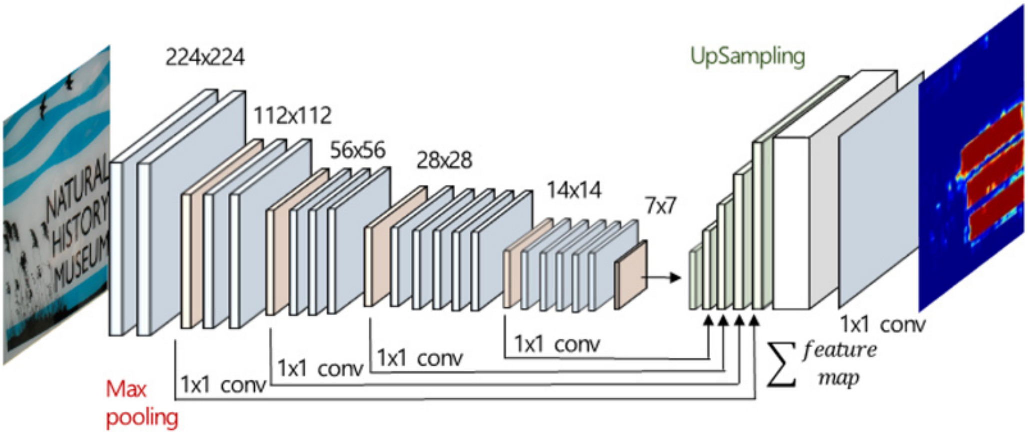


Fig. 3. FCN (Fully Convolutional Networks) for detecting text box.

As a result, the area including text can be detected pixel-wise. Also, since FCN is based on CNN, it can detect various types of texts and is robust against noise.

Usually, the CNN is not used alone for text detection but used in combination with CCA method. CCA is a grouping of adjacent pixels and MSER is mainly used in CCA. After CNN extracts features that represent characters, MSER is used to distinguish between the text area and the background area. This is useful for removing noises from complex natural images, including windows and leaves. In addition, the text detection method combining MSER and CNN can detect text regardless of the scale of images [16]. This makes it possible to solve the problem of the sliding window technique by repeatedly inspecting images at various scales. As a result, the processing time for detection can be reduced.

In this paper, we propose a framework consisting of three steps for text detection. In the first step, FCN extracts a feature map representing the location of the text. In the second step, HCA is applied to the feature map extracted from the FCN to extract text candidate areas. In the third step, another CNN is used to classify text candidate areas in four categories.

3. METHODS

Zhang *et al.* [23] used FCN for text detection. For training CNN networks, they used word boxes label information alone. The learned FCN extracted text blocks representing the coarse locations of the text in the feature map. The CCA algorithm is applied to the text block as a method for detecting multi-oriented text.

In word labeling, networks extract the characteristics of the text by comparing the text with the surrounding background in a box containing words, as shown in Fig 4. However, word-based labeling does not take into account variations such as the height, size, and location of the letters in the text.

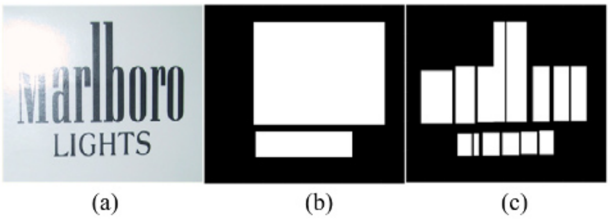


Fig. 4. Text Region labeling method. (a) Original image, (b) word-box labeling, (c) character-box labeling.

In this work, we propose two labeling methods that allow the network to learn a variety of text. The first method is to include text in block form in the same way as the existing labeling method [23]. This means a Minimum Boundary Rectangle (MBR) box that can contain all text areas. As shown in Fig. 4(b), the text block contains all the characteristics of text such as character spacing, color similarity, character height, and size similarity. The text block is the most useful labeling method for CNN. However, when the size or length of individual characters varies in the text block, it does not accurately represent the unique characteristics of the text.

The second method is a labeling method that considers only the characteristics of text. In Fig. 4(a), individual letters are not constant in size and height. However, with some rules, it takes the form of the text. These attributes of text are an important consideration when learning networks for text detection. Therefore, the characters are individually labeled as shown in Fig. 4(c). This can include the characteristics of each character. However, if only character box labels are used for learning networks, they will not know the overall shape of the text, such as the shape, color, and spacing similarity between characters.

In this paper, we propose a method to detect the text area by superimposing two labeling methods that take into account the unique characteristics of text. Fig. 5 shows the

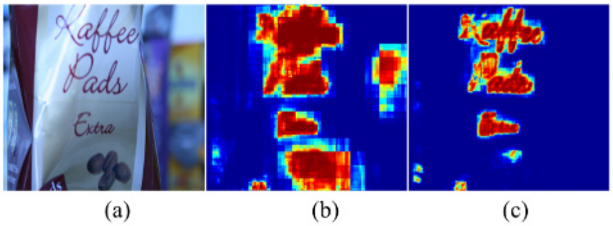


Fig. 5. The results for feature mapping in word-box labeling and our method. (a) Original image, (b) word-box labeling, (c) our method (word and character labeling).

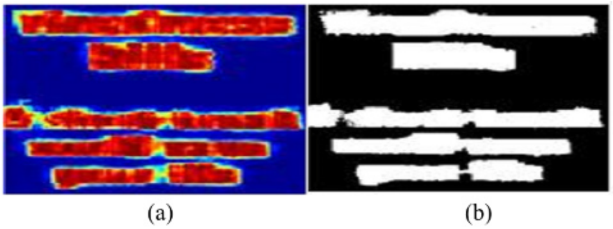


Fig. 6. The results for feature mapping binarized by the threshold. (a) The feature map; (b) the binarized feature map.

effect of separating text when learning FCN according to each labeling method. Fig. 5(b) is the result of using only word-box labeling. Fig. 5(c) is the result of using word-box labeling and character labeling together. The red region has a probability close to 1, which means the pixel corresponding to the text. Intuitively, when FCN learned using our method, which extracts feature maps, it clearly distinguished text better than a conventional method using only word box labeling. And it is robust to noise.

3.1. Text localization using FCN

MSER analyzes similarities on the pixel level to detect a region of interest, while FCN detects text by learning the difference between the text and the background regions in the entire image.

The convolutional layer of FCN at the top finds features of the character inside the text area. The convolutional layer of FCN at the bottom compresses the features of the shape of the text form, and finds global features that represent the text and the background areas. It is suitable for detecting text of various sizes and shapes.

A feature map showing the position of the learned object can be extracted from FCN. Since there is no fully connected layer, it is possible to label effectively on pixel-level regardless of image size.

In this paper, as shown in Fig. 3, we used a convolution structure of five levels excluding the fully-connected layer in the VGG 16 network [24]. In the input image, a lower level layer extracts more local features, and a higher level layer extracts more global features. Through the pooling layer and the convolution layer, text features of various scales can be extracted. Using the convolution layer and the pooling layer, networks can learn weights that extract features ranging from small-sized text to large-sized text. In networks, each convolutional step is followed by a 1×1 convolutional layer and an up-sampling to generate feature maps of the same size. The same feature maps are concatenated with a 1×1 convolutional layer to generate the feature maps of the two channels.

3.2. Region segmentation on feature maps using HCA

FCN is used as a filter for extracting the text candidate regions. However, it can't classify the text areas. The feature maps obtained from FCN are given a probability value in the range $[0,1]$. Using the probability values the text candidates can be extracted as regions. In the method using conventional FCN the threshold value has been used to divide the pixels corresponding to the background regions and the pixels corresponding to the text regions. Pixels representing the text areas have high probability values. In the feature maps, however, some pixels have high probability values even though they are in non-text areas. Therefore, it is difficult to distinguish the text using only the probability value of the pixel in the feature map.

Fig. 6(a) shows the feature map, and Fig. 6(b) shows the binarized feature map.

On the other hand, pixels in the text areas are very close to each other. To use these spatial characteristics, we applied HCA [3]. HCA performs clustering using not only the value of pixels but also a measure of distance between pairs of data. It has been mainly used for low noise classification problems.

In this paper, HCA was applied to feature vectors, which are composed of 3 dimensions. (x coordinate, y coordinate, and value of the activated pixel in the feature map resulting from FCN).

The following is an expression that describes HCA.

$$d(r,s) = \|\overline{x_r} - \overline{x_s}\|_2\,,$$
$$\bar{x}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} x_{ri}, \tag{1}$$

where $d(r,s)$ is the Euclidean distance between two separate clusters r and s. The distance of two clusters is distance of two centroids. $\bar{x}_r$ and $\bar{x}_s$ are centroid of the two clusters r and s. n_r is the number of data in cluster r. x_{ri} is the ith data in cluster r.

The Euclidean distance of each data is calculated according to (1). Then, the data adjacent to each other are clustered. The centroid of the clustered data is considered

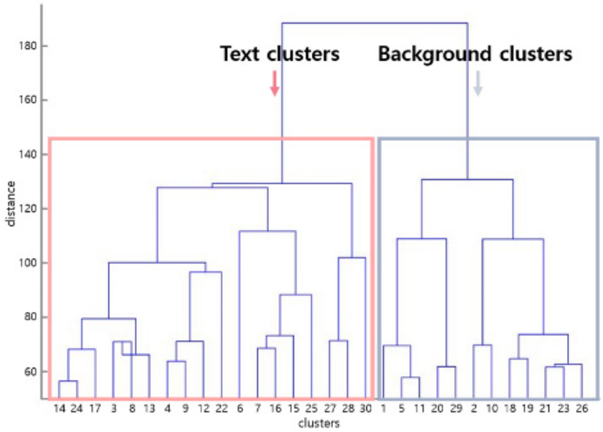


Fig. 7. The results of the clustering given by the dendrogram.

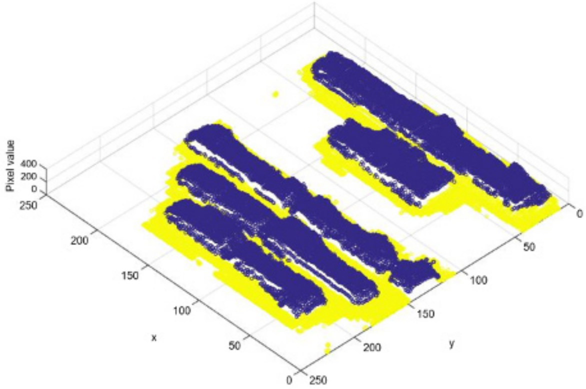


Fig. 8. Feaure map divided into text and background areas using HCA.

new data. The new data again cluster with the nearest data. The process repeats until it becomes one cluster.

Fig. 7 shows the results of the clustering given by the dendrogram. The top of the dendrogram indicates that all the pixels have become a cluster. Fig. 8 is a feature map that divides the text areas and the background areas using the dendrogram. Pixels corresponding to the background regions have a low probability value, so they are clustered at the bottom. Pixels corresponding to the text areas have a high probability value and the property of being adjacent to each other, and they are clustered at the top. This method is more effective for text segmentation than the binary segmentation technique using a certain threshold. Fig. 10 shows the result of binary image comparison. When HCA was applied, the text areas were clearly distinguished.

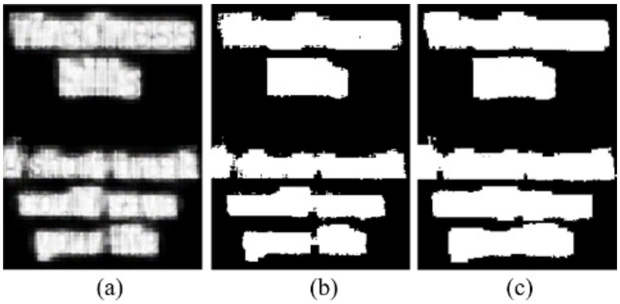


Fig. 10. Binarization image comparison: (a) the result image of FCN, (b) binarization based on constant threshold(the value is 0.5), (c) binary image based on HCA.

3.3. Text verification from classification using CNN

Fig. 11 shows the results of performing blob detection on the result images of HCA. Most of the detected images include texts, however, some images are not. The another CNN categorizes images resulting from HCA into text and non-text.

In order to classify an image into specific categories, a fully-connected layer is required. However, when the fully-connected layer is added in CNN, the size of the input image must be fixed. But the detected blob images contain characters of various lengths and background noises as shown in Fig. 11. There is a difference in detec-

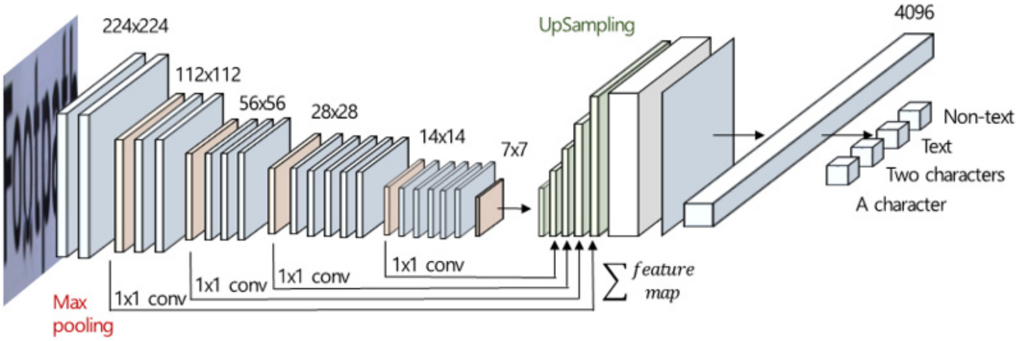


Fig. 9. CNN architecture for text classification.

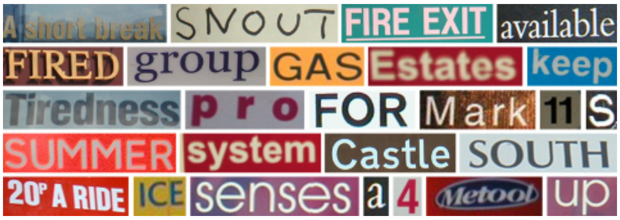


Fig. 11. The croped images obtained as results of HCA.

tion size according to the number of characters between text images.

Therefore, we subdivided the text class for improving the performance of CNN for various sizes and text lengths. Four categories for one character, two characters, three or more characters, and non-text were assigned to the classification classes, where one character, two characters, three or more characters are regarded as the text. The background area is classified as the non-text.

Fig. 9 shows CNN structure used in this paper. At the end of FCN structure in Fig. 3, a fully connected layer is added to classify the text and the non-text. The blob images resized to 224x224 are entered into CNN. And CNN decides what the input image belongs to which category.

4. EXPERIMENTS

In this work, we used Python and the Keras library to implement the proposed method; Keras is a high-level neural networks API, written in Python, and runs seamlessly on the CPU and GPU. We ran a workstation (i7-4770, 32GB RAM, GTX TitanX and Windows 10 64bit-OS) for all the experiment. The GPU was run under the CUDA Toolkit, cuDNN environment.

4.1. FCN learning

ICDAR (International Conference on Document Analysis and Recognition) is a conference on research topics including character, symbol recognition, printing, handwriting recognition, document analysis, document understanding, camera, and video-based scene text analysis. ICDAR also provides a certified public text dataset for the field of text detection. In order to learn FCN, we generated about 30,000 augmentation images using 229 images provided by the Focused Scene Text of ICDAR 2015 using the data augmentation method [25].

Fig. 12 presents the examples of SynthText Dataset [26]. SynthText is a dataset created by artificially synthesizing text on a complex background for learning. Approximately 850,000 images are provided for learning and the dataset contains 8 million words.

We have trained FCN using about 900,000 images from both ICDAR and Synthtext. The size of the images is 224 × 224, and the labeling value of each image is superimposed on word and letter labeling. The objective function is binary cross-entropy [27]. For FCN learning, batch size 32, momentum 0.9, learning rate 0.0001, and epoch 25 were set as parameters.

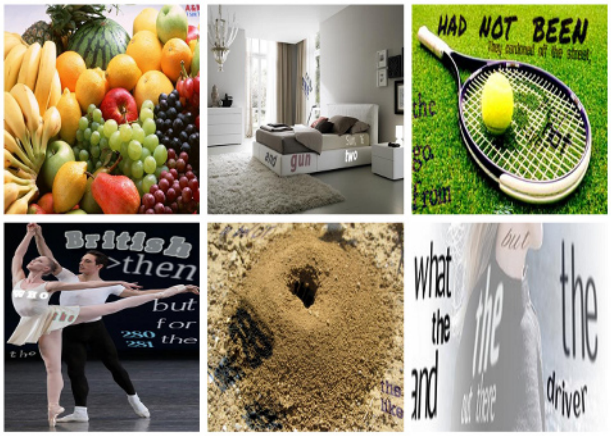


Fig. 12. The examples of SynthText dataset.

4.2. Text classification from another CNN

In this paper, CNN is used for the binary classification of text and non-text. In order to improve the classification performance, networks were learned for all four classes including non-text.

The network learning process is as follows. The front part of the network to extract features used the VGG16 network [24] pre-trained on ImageNet [2]. In order to train an upsampling layer and a fully connected layer, the data generation process is required. Images used for data generation are provided by ICDAR.

The training data includes some background around the text. Each training data was collected by increasing the word width to the left and right by 30 percent and the word height up and down by 40 percent. The image input to CNN is 224 × 224 × 3 (height × width × channel). Fig. 13 shows the text extracted from the images provided by ICDAR. The first row is one character, the second row is two characters, the third row is three or more characters, and the fourth row is a non-text image, which means noisy background images. For the CNN learning, batch size 32, momentum 0.9, learning rate 0.0001, and epoch 25 were used as parameters. And the objective function was the categorical cross-entropy method.

5. RESULTS

5.1. FCN-based multiple labeling

In this section, we qualitatively compare the results obtained when the network was learned by multiple labeling methods, and the word-box labeling method, respectively. We also analyzed how the text candidate regions were detected according to the labeling method. In addition, the quantitative results are compared by analyzing the

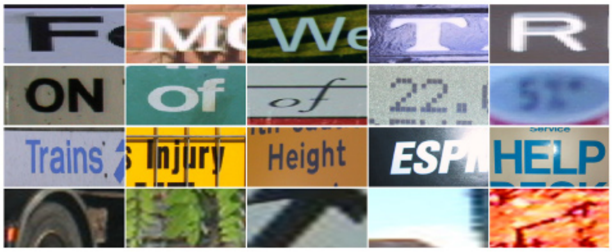


Fig. 13. The examples of training dataset for CNN.

detected text region and the removed background region based on pixel-wise ground truth. Two FCN networks were used in the experiment. The only differences between the two methods were in network structure depending on the labeling method. That is, the hyper-parameters used for learning were the same. In addition, the same data was used for network learning.

Fig. 14 presents qualitative results showing how much of the background noise is removed. The proposed method removed much more background noises than the existing labeling method. The lane located at the lower right of the picture has characteristics that are distinct from the background. The FCN network uses labeling information given in pixel units by a rectangular box. Because the word-box labeling method contains a lot of background pixels as well as text in the boxes set to ground truth, it may be difficult to distinguish between background and text in a network learned with word-box labeling. Previous methods show that the lane is represented by text.

On the other hand, our proposed multiple labeling methods use character box labeling as well as word-box labeling. The weights of the FCN network are learned to distinguish between background and text, and at the same time to distinguish fine differences in characters existing in a word-box. As shown in the middle column of Fig. 14, our proposed method distinguishes the background area and the text better than the existing method.

Fig. 15 shows how precisely the text is represented in the feature map. It can be seen that the proposed method represents the text or the characters more precisely. The first row in Fig. 15 shows the main problems with word-box labeling. There is a word '.com' under the word 'jungle' in the image. Clearly, there is a problem when the word box including 'jungle' and the word box including 'com' overlap each other. On the other hand, the proposed multiple labeling methods include character-box labeling, which means it can distinguish another text in the word-box. As a result, the FCN network learned with the multiple labeling method extracts feature maps that can detect text more precisely. Also, FCN learned with word-box labeling does not precisely represent the height and shape of the text. On the other hand, the proposed method clearly represents the line and the form of the text and space between the texts.

Fig. 16 shows a case where the proposed method fails to remove false positives. In the image, the piano keyboard has a distinct color from the background, a similar size, and a uniform spacing between the keys. This attribute is very similar to the text. This made text detection difficult.

Fig. 17 shows the results of a pixel-wise quantitative evaluation for each method on the ICDAR verification dataset. The pixel distribution represents the text regions and the background regions, for feature maps extracted by each method. Here, GT means ground truth corresponding to a positive region and a negative region in the image. If the pixel value corresponding to each region on the feature

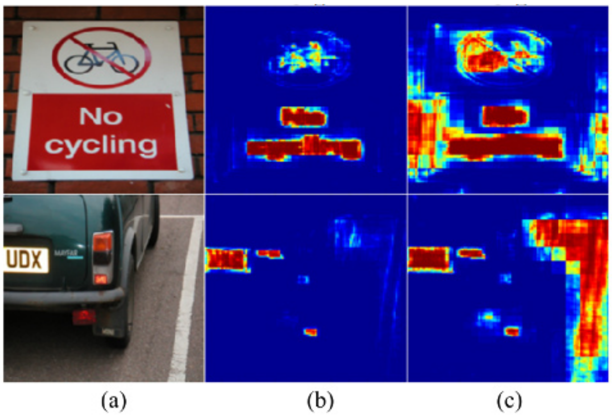


Fig. 14. The results of the background noise removal according to labeling method. (a) Original image; (b) the proposed method(multiple labeling); (c) the exting labeling method (word-box labeling) [23].

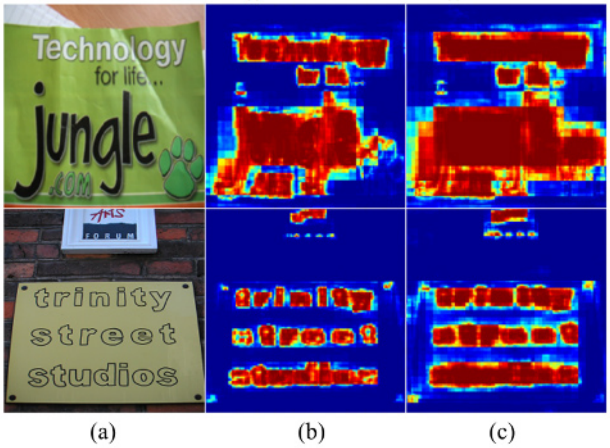


Fig. 15. The results of FCN feature map according to labeling method. (a) Original image; (b) the proposed method(multiple labeling); (c) the exting labeling method (word-box labeling) [23].

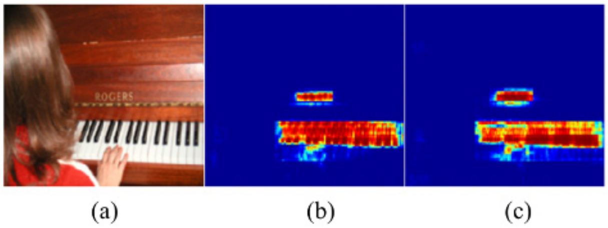
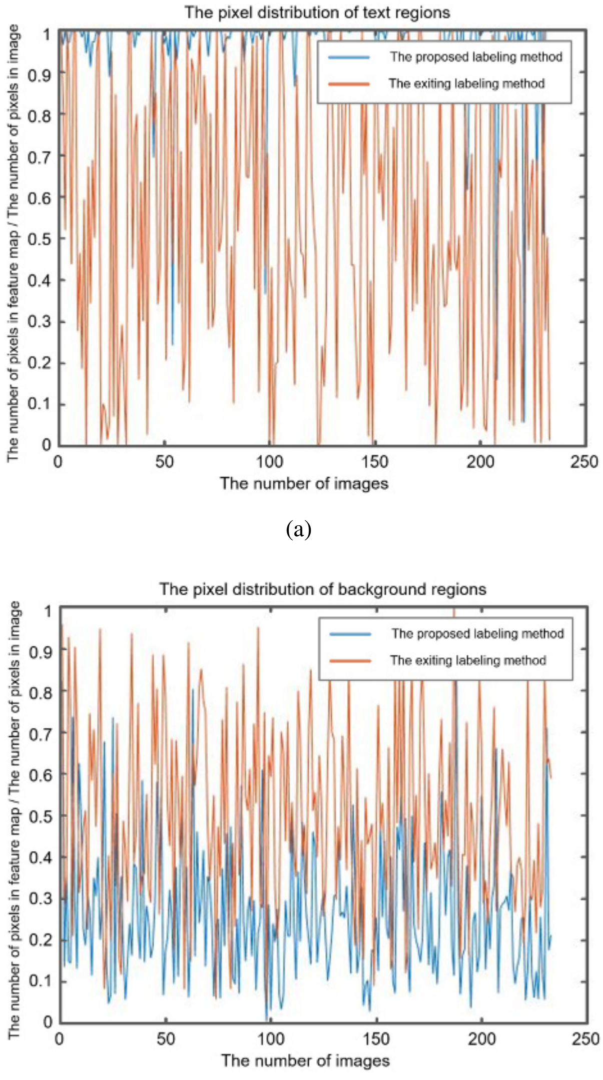


Fig. 16. The example of failure using the proposed method. (a) Original image; (b) the proposed method (multiple labeling); (c) the exting labeling method (word-box labeling) [23].



(a)

(b)

Fig. 17. The results of pixel-wise quantitative evaluation for each method on ICDAR 2015. (a) The pixel distribution of text regions; (b) the pixel ditribution of background regions.

Table 1. Performance comparison of HCA and the binarization method on ICDAR 2015.

	Recall	Precision
Binarization method	76.0 %	80.2 %
HCA	83.4 %	81.0 %

Table 2. The results of thext classfication for CNN.

	Recall	Precision
Two classes	94.3 %	95.3 %
Four classes	98.1 %	95.1 %

map is larger than 0, it is regarded as a response. In other words, if the pixel value in the text area of the ground truth is greater than 0 and the pixel value in the background area is 0, it means that the networks were learned well.

Fig. 17(a) is a graph that evaluates pixel values on a feature map corresponding to the text regions. While the conventional method achieved an average of 0.55, the proposed method achieved an average of 0.97 on images used for validation. Fig. 17(b) is another graph that evaluates pixel values on a feature map corresponding to the background regions. The proposed method achieved an average error of 0.26. On the other hand, the existing method shows an average error of 0.5. Based on these qualitative and quantitative evaluation results, we have confirmed that the FCN learned with multiple labeling is superior to the FCN learned only with word-box labeling.

5.2. Segmentation based on HCA

Feature maps extracted through FCN have values in the range $[0,1]$ in which the value is a probability value. The positive region is close to 1 and the negative region is close to 0. A positive area means a text area. Also, the pixels in the text area are very closely clustered. Based on this characteristic, we applied HCA to segment the text area and the background area. For a comparative experiment, we replaced HCA with simple binarization method in a framework consisting of 3 steps. Table 1 shows the results of the comparison using the verification data of ICDAR 2015. HCA showed better performance than the binarization method, which had a recall of 83.4% and precision of 81.0%.

5.3. The results of text classification

We determined the text candidate regions using FCN and HCA. Then, blob detection was applied to the text candidate regions. The detected regions were identified as text and non-text using a classifier. In this paper, we segmented the text categories considering the diversity of the text. Table 2 shows the results for Precision and Recall, comparing the existing binary classification method and our proposed method.

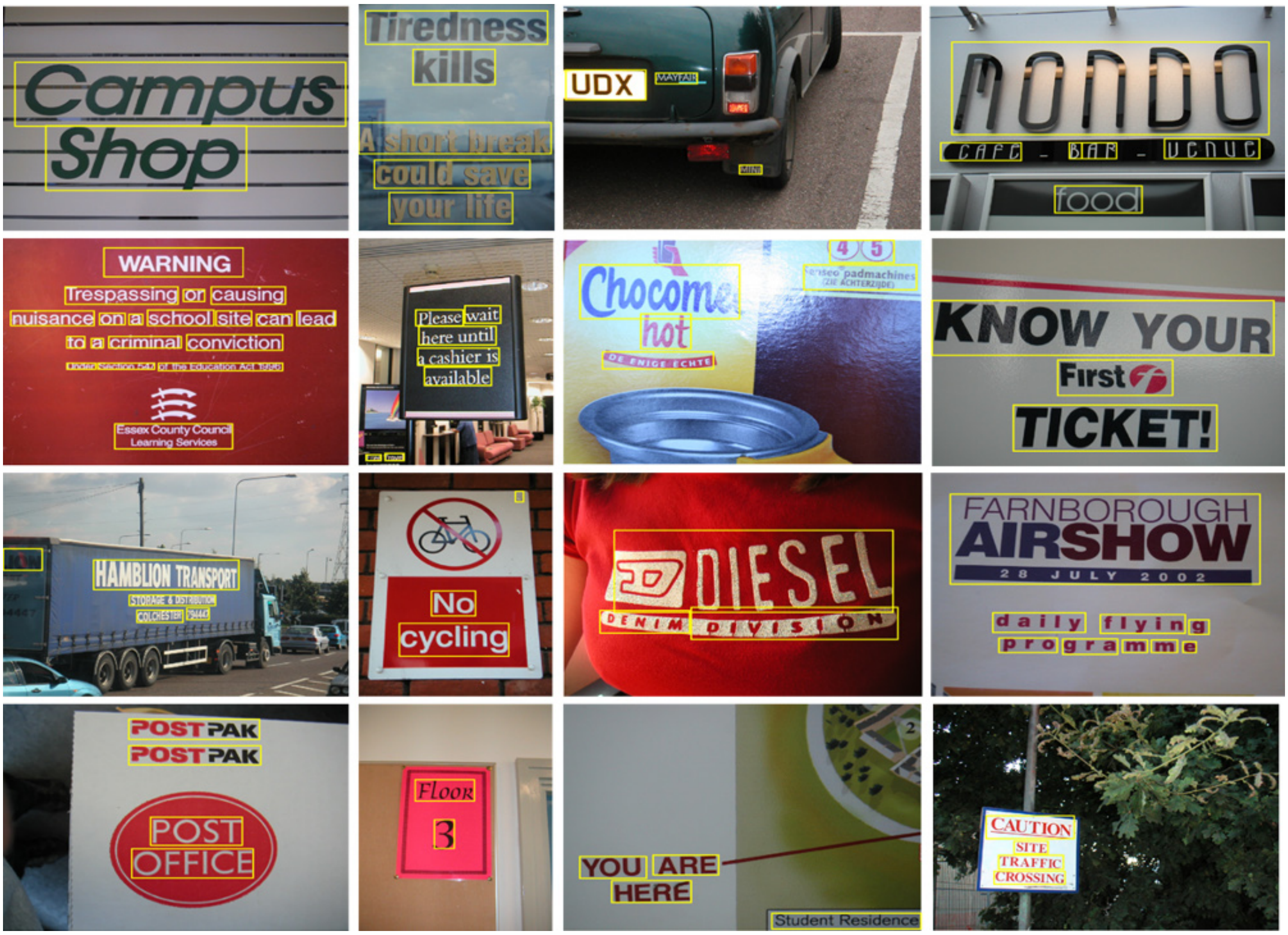


Fig. 18. Examples of text detection.

From the validation dataset of ICDAR 2015, we used 1,090 images with text and 2,240 images without text for evaluation. Both methods showed almost the same performance on Precision.

On the other hand, our method outperformed the binary method on Recall. High recall value means that the background is not categorized as text. Thus, our method is more effective in removing false positives.

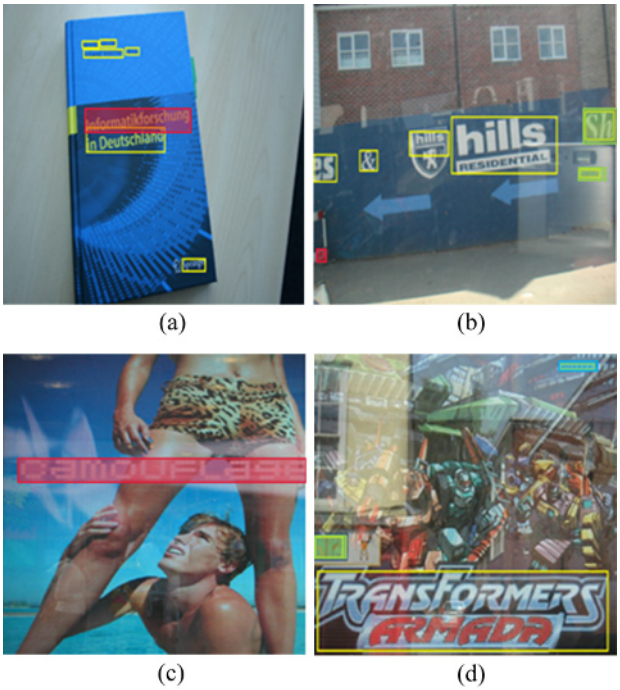
5.4. Text decision

Fig. 18 shows detection examples using the proposed method on the dataset for validation. The text was robustly detected in a complex scene. It was well detected even when the font size was large.

Fig. 19 shows incorrect detection/miss-detection or false detection examples using the proposed method. In this case, the text is not included in the ground truth, so it is classified as a false detection. The red box shows when text is not detected. The not detected region was classified as text in HCA but classified as non-text by CNN. Increasing the width and height of many characters seems to have influenced CNN learning.

In Fig. 19(b), the green box is the result of detecting text reflected in the glass that is not included in the ground truth. It is not included in GT so it is classified as a false detection. Examples like this indicate that the method proposed in this paper can even robustly detect texts reflected on glass windows. The green box in Fig. 19(c) is also detected text not included in the ground truth. The blue box has a shape similar to text and is a false detected area. In Fig. 19(d) the red box had been detected as text in the early stage of FCN learning, but it was classified as background as learning progressed. The text is in braille. Braille text is distinct from the background, but the number of braille texts is much lower than texts consisting of consecutive characters. For this reason, it seems that the braille text cannot be distinguished in the feature map.

Table 3 shows the result when our proposed method was applied to the ICDAR 2015 dataset for validation. The proposed method achieved 0.8340, 0.8099, 0.8218 in Recall, Precision, and Hmean, respectively. The method presented in this paper ranked 10th in the published papers. In addition, it was ranked second for Recall, by effectively eliminating false positives or background noise.



(a) (b)

(c) (d)

Fig. 19. Examples of false and non-detected images.

Table 3. Text detection performance evaluated using IC-DAR 2015 [25].

	Recall (%)	Precision (%)	Hmean (%)	Time (s)
CTPN	82.98	92.88	87.69	0.140
SCUT-HCII	84.22	95.10	87.32	-
CASIA_USTB	82.59	89.50	85.91	-
MSER_Binary_CNN	82.17	89.12	85.61	-
StradVision	80.15	90.93	85.20	–
VGGMaxNet_cmb	77.32	92.18	84.10	-
MCLAB_FCN	79.65	88.40	83.80	-
Text_CNN	76.29	92.69	83.69	4.600
Ours	**83.40**	**80.99**	**82.18**	**0.362**
IWRR2014	78.65	85.89	82.11	-
HUST_MCLAB	76.05	87.96	81.58	-
BUCT_YST	73.88	84.64	78.90	-
USTB_TexStar	69.28	88.80	77.83	-
SWT	73.24	81.53	77.16	-
BayesText	67.05	84.58	74.80	-
TextSpotter	64.97	87.49	74.56	-
I2R_NUS_FAR	70.92	75.71	73.24	-

5.5. Computing time

The proposed method consists of three steps. Table 4 shows the processing time for each step. The time means the average value. Step 1 and Step 3 were based on convolutional neural network. Therefore, its processing time depends on the performance of GPU.

Table 4. The computing time for each step.

Step	Processing Time (s)
Step 1: Text-block FCN	0.010
Step 2: Text-segmentation HCA	0.337
Step 3: Text-classification CNN	0.015
Overall system	**0.362 (2.76 fps)**

Table 5. The computing time compared on different methods. Note that each method has a different dataset and various image scale.

Approach	Device	FPS
Yao *et al.* [28]	K40m	1.61
Tian *et al.* [29]	-	7.14
Zhou *et al.* [30]	Titan X	6.52
Zhang *et al.* [23]	Titan X	0.476
Ours	**Titan X**	**2.74**

On the other hand, HCA algorithm used in Step 2 is based on CPU operation. The steps consisted of clustering, linkage, and blob detection, which took a relatively long time. The overall system took about 0.362 s (2.76 fps). Most of the previous methods listed in Table 3 did not report the computing time.

Table 5 shows the computation time compared to the state of the art methods. Note that each method was not tested in completely the same environment and the purpose of the experiment may be slightly different. Some approaches are not restricted to the text localization task.

Our method is not as fast as the state of the art methods. However, the method has three steps, so it has advantages in expandability.

6. CONCLUSIONS

In this paper, we proposed a framework consisting of three steps for text detection. The first step uses FCN to detect text areas in images in pixel-wise. In the second step, HCA is applied to divide and segment the text areas and the background areas. In the third step, CNN is used to classify the divided areas in step 2 into the text areas and the non-text areas.

Multiple labeling method combining word-box labeling and character-box labeling extracted more the text areas with distinct boundaries than the conventional only word-box labeling method. And, it was robust to noises. Using HCA showed better performance than using the binarization method in text and background region segmentation. Finally, classification into four categories when classifying text (one letter, two letters, three letters or more) and non-text showed higher performance at removing false positives than the binary classification (text/non-text).

Each step is independent of each other. It is advantageous in terms of expandability and viewpoint of modularity. Therefore, it is possible that performance will be improved by using the latest algorithm for each step. Another advantage is that it can be used as a kind of detection framework.

In future work, we will construct a framework of three steps as a single network and detect text areas and study OCR (Optical Character Recognition) by combining text detection and recognition algorithms.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[3] L. Rokach and O. Maimon, "Clustering methods," *Data Mining and Knowledge Discovery Handbook*, pp. 321-352, Springer, Boston, MA. 2005.

[4] Q. Ye and D. Doermann, "Text detection and recognition in imagery: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1480-1500, 2015.

[5] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2963-2970, June 2010.

[6] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," *European Conference on Computer Vision*, pp. 512-528, Springer, Cham, September 2015.

[7] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 97-104, 2013.

[8] J. J. Lee, P. H. Lee, S. W. Lee, A. Yuille, and C. Koch, "Adaboost for text detection in natural scene," *Proc. of International Conference on Document Analysis and Recognition (ICDAR)*, pp. 429-434, September 2011.

[9] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004.

[10] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 4, pp. 256-268, 2002.

[11] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761-767, 2004.

[12] L. Neumann and J. Matas, "Real-time scene text localization and recognition," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3538-3545, June 2012.

[13] Y. F. Pan, X. Hou, and C. L. Liu, "Text localization in natural scene images based on conditional random field," *Proc. of 10th International Conference on Document Analysis and Recognition (ICDAR'09)*, pp. 6-10, July 2009.

[14] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," *Proc. of Asian Conference on Computer Vision*, pp. 770-783, Springer, Berlin, Heidelberg, November 2010.

[15] L. Neumann and J. Matas, "Text localization in real-world images using efficiently pruned exhaustive search," *Proc. of International Conference on Document Analysis and Recognition (ICDAR 2011)*, pp. 687-691, September 2011.

[16] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," *Proc. of European Conference on Computer Vision*, pp. 497-511, September 2014.

[17] Y. Zheng, Q. Li, J. Liu, H. Liu, G. Li, and S. Zhang, "A cascaded method for text detection in natural scene images," *Neurocomputing*, vol. 238, pp. 307-315, May 2017.

[18] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111-3122, Nov. 2018.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, arXiv:1409.1556, 2014.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[21] H. H. Kim, J. K. Park, J. H. Oh, and D. J. Kang, "Multi-task convolutional neural network system for license plate recognition," *International Journal of Control, Automation and Systems*, vol. 15, no. 6, pp. 2942-2949, 2017.

[22] J. K. Park and D. J. Kang, "Unified convolutional neural network for direct facial keypoints detection," *The Visual Computer*, pp. 1-12, 2018. DOI: 10.1007/s00371-018-1561-3

[23] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4159-4167, 2016.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, arXiv:1409.1556, 2014.

[25] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, and F. Shafait, "ICDAR 2015 competition on robust reading," *Proc. of 13th International Conference on on Document Analysis and Recognition (ICDAR)*, IEEE, pp. 1156-1160, August 2015.

[26] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2315-2324, 2016.

[27] P. D. Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of Operations Research*, vol. 134, no. 1, pp. 19-67, 2005.

[28] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," arXiv preprint arXiv:1606.09002., 2016.

[29] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," *Proc. of European Conference on Computer Vision*, pp. 56-72. Springer, Cham. October 2016.

[30] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: an efficient and accurate scene text detector," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2642-2651, July 2017.

Hong-Hyun Kim received his B.S. degree in School of Mechanical Engineering from Pusan National University, Busan, Korea in 2012. He is currently in Unified Master and Doctor's course at the same graduate school. His research interests include deep learning, machine learning, and pattern recognition.

Jae-Ho Jo received his B.S. degree in School of Computer Science and Engineering from Pusan National University, Pusan, Korea in 2014 and an M.S. degree in Mechanical Engineering from Pusan National University, Busan, Korea in 2016. Now, he is a research engineer in Hanhwa-Techwin R&D Center. His research interests include machine learning, and computer vision.

Zhu Teng received her B.S. and Ph.D. degrees in Automation from Central South University, China, 2006 and in Mechanical Engineering of Pusan National University, Korea, 2013, respectively. She is now an associate professor in the School of Computer and Information Technology, Beijing Jiatong University. Her current research interests are visual tracking, deep learning, and computer vision.

Dong-Joong Kang received his B.S. degree in Precision Engineering from Pusan National University, Busan, Korea, in 1988 and his M.S. and Ph.D. degrees in Mechanical, and Automation & Design Engineering from KAIST, Korea, in 1990 and 1999, respectively. Now, he is a professor at the School of Mechanical Engineering in Pusan National University. He is also an associate editor of the International Journal of Control, Automation, and Systems since 2007. His current research interests are machine vision, machine learning, and visual inspection in factory.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.