Text Detection based on Edge Enhanced Contrast Extremal Region and Tensor Voting in Natural Scene Images

Van Khien Pham*, Soo-Hyung Kim*, Hyung-Jeong Yang*, Guee-Sang Lee*

Abstract

In this paper, a robust text detection method based on edge enhanced contrasting extremal region (CER) is proposed using stroke width transform (SWT) and tensor voting. First, the edge enhanced CER extracts a number of covariant regions, which is a stable connected component from input images. Next, SWT is created by the distance map, which is used to eliminate non-text regions. Then, these candidate text regions are verified based on tensor voting, which uses the input center point in the previous step to compute curve salience values. Finally, the connected component grouping is applied to a cluster closed to characters. The proposed method is evaluated with the ICDAR2003 and ICDAR2013 text detection competition datasets and the experiment results show high accuracy compared to previous methods.

Keywords : CER | MSER | SWT | Tensor voting | Text line information | Text detection

I. INTRODUCTION

Text detection in natural scene images is an initial and essential step for text understanding of image processing fields. Natural scene images can be taken from any environment, which introduces many challenges to the precise identification of text regions. Although a lot of methods have been published for text detection, it is still challenging to identify text in natural scene images. These problems are due to several issues, such as the variation of illumination changes, different font styles, sizes, colors, orientations, and low contrast as well as complex backgrounds.

Text detection methods can be approximately classified into five main groups: texture based, edge based, connected component based, stroke based, learning based and others. Texture based approaches rely on the observation that characteristics of text areas, such as their texture or coefficient values in transformed domains, are different from those of non-text areas. These approaches [1, 2] can detect and localize texts exactly even when images are noisy but the computational time is very high and the accuracy depends on the direction of text arrangement. Edges are stable features for text detection in natural scenes [3, 4, 5]. An edge detector is used first, followed by morphological operations to obtain text regions from the background and false positives are removed. The major problem of edge based methods is that it is sensitive to illumination changes. Connected component (CC) based methods [6, 7] group small connected components into larger ones until all texts are detected in the input image. The non-text regions are eliminated based on heuristic rules or by classifiers. CC based approaches cannot segment text regions exactly without prior knowledge of the text position and scale. These methods have high processing time and non-text regions are hard to distinguish from text regions. In stroke based methods [8, 9], SWT uses robust features for text detection by analyzing strokes, which remain almost the same for a single character in natural scene images. Significant changes in stroke within non-text regions result in irregularity. Character stroke candidates are verified by feature extraction, extracted by segmentation, and collected

* Dept. of ECE, Chonnam National University , Korea

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by MEST (NRF-2015R1D1A1A01060172) and by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-00383).

Manuscript : 2017 09 04
Revised : 2017. 11. 24

Confirmation of Publication: 2017.11.27
Corresponding Author : Guee-Sang Lee
e-mail : gslee@jnu.ac.kr

together by clustering. These approaches are easy to do in these specific applications but complex backgrounds make text strokes difficult to detect.

(a) (b)

Fig. 1. The results from Chen [15] and Toan [16]

Learning based methods provide some features to train a classifier (SVM [10], Neural Network classifier or deep Convolutional Neural Network classifier [11, 12]). The results of these methods show high accuracy, however the disadvantages of learning based methods are with training huge images. These methods have very high processing time and it is hard to choose the best features for text areas.

In MSER methods [13, 14], the MSER algorithm detects the best area detectors because of its robustness against scale, view point, and illumination changes. There are two types of MSERs, which are light regions on dark backgrounds and the dark regions on light backgrounds. They function by distinctly removing the mutual influence, but many background regions are also identified. MSERs can extract not only text regions but also non-text regions, which include uncertain regions. It is hard to clearly decide between text and non-text. Chen et al. [15] combined edge enhanced MSERs and SWT to detect small letters in original images with limited resolution. Candidate characters are clustered into lines and additional checks are implemented to remove false positives. However, this method did not have a robust way to remove non-text regions in the final step, which leaves too many false positives. Additionally, this method cannot handle complex backgrounds such as highlighting, illumination changes, which are shown in Fig. 1(a).

Toan et al. [16] is the first paper that uses tensor voting for detecting text in natural images. Tensor voting is a very good approach for removing non-text regions. First, this method extracts the center point with vertical edge detection. Then, tensor voting extracts text line information by using the curve saliency value and the curve normal vector at each character, which is useful to eliminate non text regions. However, this method has some false positive and false negative text areas in many scene images. Moreover, as the only input to the tensor voting process, the centroids obtained from these connected components of edge maps in the initial step cannot cover all text regions in the image such as highlighting, shadow and blur. We can see this in Fig. 1(b), where text detection is illustrated by red rectangles. Since the publication of the original paper, there has been limited progress in this direction.

We have several open challenges to solve for text detection. First, our method can handle several complex backgrounds such as highlighting, low contrast, blur, various text sizes and font texts. Second, edge enhanced CER is good for extracting candidate text areas in the initial step. Finally, tensor voting is very effective at eliminating non-text areas in the final process.

In summary, we suggest a novel and robust method based on edge enhanced CER, SWT, and tensor voting for text detection. The contributions of our proposed method can be described as:

- Firstly, edge enhanced CER is extracted from original images based on color information and edge information in natural scenes, where the foreground connected components are considered as text candidates.
- In the stroke width filtering process, distance mapping is used to estimate values for the stroke width, which eliminates non-text areas.
- We next take the center point of each candidate text region as a token that corresponds to a point or curve segment. This is an important step in the process. The curve saliency value in a text area is a larger value than that in non-text area. Therefore, we know the text location and can eliminate non-text regions exactly.
- Finally, grouped connected components of the candidate text areas are combined according to the text line information. The outline for our method is shown in Fig. 2.

This paper is organized as follows: section 2 describes the proposed method. Section 3 details the experimental method. Finally, our paper is concluded in section 4.

II. PROPOSED METHOD

In this paper, we will present more details of our

approach in the following steps. Each step will be described clearly to present the key ideas behind our solutions for complex background images. Given an input image, edge enhanced CER are efficiently extracted from the image, finding regions of similar intensities. Characters in most languages have similar stroke widths and the stroke widths of characters have low variation. The center points of these candidate text regions are typically close together and are mostly aligned on a line or a smooth curve. Thus, the curve saliency of an input token corresponding to a center point for a non-text area has lower curve saliency than that of a token in a text area. Thanks to edge enhanced CER, SWT, and tensor voting, the proposed method can effectively identify text areas in complex background images.

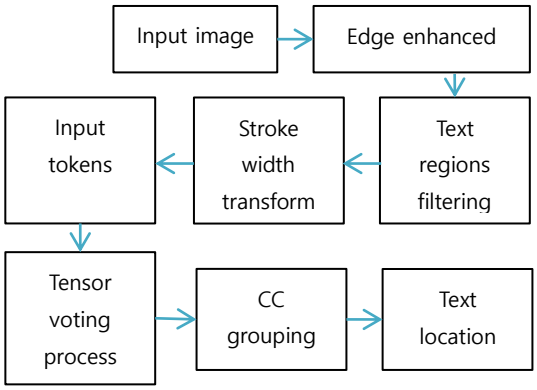


Fig. 2. The overview of our method

2.1 Edge enhanced CER detection

CER is used as the connected component generation method in our system to extract better candidate text regions than MSERs. MSER is a connected component for a suitable threshold image [17]. The word "extremal" refers to the property of all pixels inside the MSER which have either higher (bright extremal areas) or lower (dark extremal areas) intensity than the pixels on the boundary of the regions. Since text typically has a different background and relatively uniform color or intensity, MSER is a viable choice for detection [15]. It works well on images containing uniform regions with individual boundaries, however it is sensitive to image blur. We can see in Fig. 3(b) that MSER is not good with complex background images (blur, shadow, and low contrast). Several characters are overlapped and shapes also change, therefore it is difficult to identify text regions. To solve this problem, we need to use edge information to enhance the extracted CER.

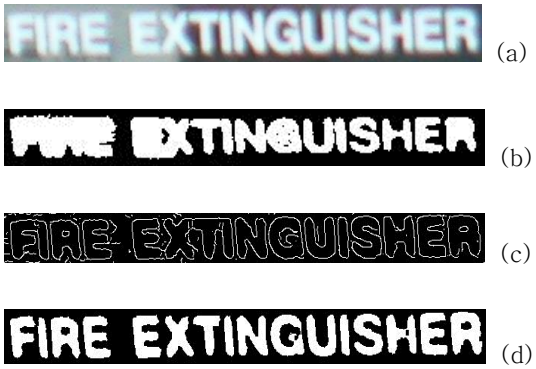


Fig. 3. (a) Original image, (b) binary image of MSER detection, (c) Edge detection, (d) Edge enhanced CERs detection

CER step, these characters are very clean and separate from each other. Fig.3(d) demonstrates the edge enhanced CER binary image, which provides an improved representation of the text where distinct characters are separated.

2.2 Text region filtering

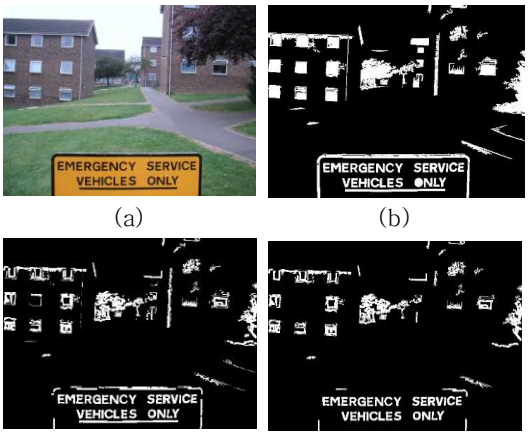
In addition to text regions, non-text regions are also detected by edge enhanced CER, which will be removed. These connected components are defined as sets of connected region pixels from the binary image in the initial step. Some connected components of non-text areas are rejected due to weak heuristic rules based on geometrical properties (aspect ratio, area, height, width) as in equation (1). The connected components in region R that do not satisfy one of the following conditions are eliminated. H_r and W_r represent the height and width of region R respectively. Let us indicate W and H as the width and height of the original image. The text regions-filtered image is shown in the Fig. 4(d).

$$\begin{cases} H_r < H/3 \\ W_r < W/3 \\ \dfrac{H_r}{W_r} < 4 \\ \dfrac{W_r}{H_r} < 4 \end{cases} \qquad (1)$$

2.3. Stroke width transform

The SWT calculates the width of the most likely stroke covering each pixel. We can define a stroke as connecting part of an image with an approximately

constant width value [9].



(a) (b)

(c) (d)

Fig.4. Edge enhanced CER detected process. (a) Original image, (b) MSER detection, (c) binary image of edge enhanced CER detection, (d) Text regions-filtered image

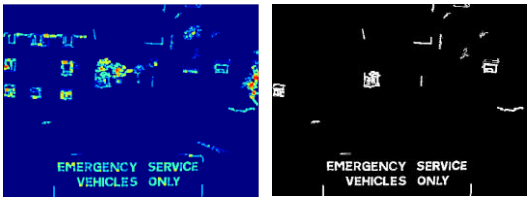
Text candidate regions are variations in the stroke width of images. Characters have similar stroke widths or thickness in most languages. Therefore, non-text regions can be removed where stroke width show too much variation as in equation (4) with T_{SWT} = 0.45. In many cases, we cannot eliminate non-text regions because they have constant width, which represent the same characters. Therefore, we need to use tensor voting to identify text and non-text regions. This approach works well to remove non-text areas. The distance map image is applied to label each candidate text pixel. We can see the stroke width transform image in Fig. 5.

$$Std = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}|M_i - \mu|^2} \quad (2)$$

$$\mu = \frac{1}{T}\sum_{i=1}^{T} M_i \quad (3)$$

$$\frac{Std}{\mu} > T_{SWT} \quad (4)$$

, where μ is the mean of stroke width variation in each character and the standard deviation (Std) is the square root of the variance. M_i represents the stroke width variation number i in each character. T is the total number of stroke width variations in each character.



(a) (b)

Fig.5. Stroke width transform process. (a) Distance transform image. (b) text candidate regions after stroke width filtering

2.4. Tensor voting process

Tensor voting is intended to be flexible enough to solve several problems with computer vision, especially mid-level vision applications. It can be used in problems that can be posed as perceptual organization problems. In 2-D [18], each token belongs to one of the following perceptual structures: curve, region, junction, or termination (e.g. a curve endpoint).

A tensor is divided into stick components and ball components. The input tokens are initialized based on their individual information. An input token that is an elementary curve with a normal is denoted by a stick tensor parallel to the normal. An un-oriented token is encoded by a ball tensor.

Each token in the voting field receives many votes from its neighboring tokens [16]. Vote gathering is implemented by tensor addition or consistently by the addition of 2 by 2 matrixes, generating generic tensors. A generic tensor is separated into stick and ball components with the following equation.

$$T = \lambda_1 \hat{e}_1 \hat{e}_1^T + \lambda_2 \hat{e}_2 \hat{e}_2^T \quad (5)$$

$$T = (\lambda_1 - \lambda_2)\hat{e}_1 \hat{e}_1^T + \lambda_2 (\hat{e}_1 \hat{e}_1^T + \hat{e}_2 \hat{e}_2^T) \quad (6)$$

, where λ_i are the corresponding eigenvalues in decreasing order and e_i define the eigenvectors.

Tensors support information related to smoothness and proximity of continuity by a voting process. Tensors have smooth salient features (i.e. curves) that allow them to support each other. Each tensor votes for its neighboring tensors and receives votes from them. The size and shape of this neighborhood and the vote orientation and strength are

summarized in predefined kernels or voting fields. Each feature requires a voting field. All voting fields are created from a fundamental stick voting field. The vote orientations correspond to the smoothest local continuation from the voter to the recipients. The vote strength at each recipient is computed by a decay function that is inversely proportional to the length of the curve from the voter to the recipient.

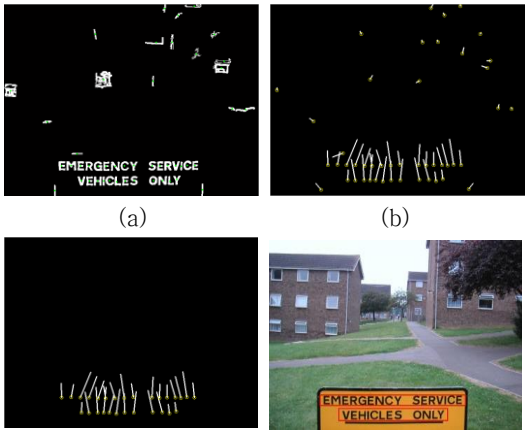
We can use center points to reject non-text areas from candidate text areas, which are created by CCs in the previous step. The input tokens corresponding to a center point in a text area are typically adjacent and nearly aligned on a line or smooth curve. Thus, the curve saliency of a token in a non-text area has lower curve saliency than that of a token in a text area, which we can see clearly in Fig. 6(b) and Fig. 6(c). The curve normal of a token in a text area represents the normal vector of the text line. Furthermore, based on the assumption that text lines are nearly aligned horizontally (the angle between the text line and the horizontal line can be up to 300), the other condition helps to remove the set of center points that are closely aligned vertically as in equation (7).

$$\begin{cases} |X| < \frac{1}{T}\sum_{i=1}^{T} |X_i| \\ |\cos(X, U)| < \sqrt{3}/2 \end{cases} \quad (7)$$

, where X is the normal vector at the center point, $U = (0, 1)$ defines a unit vector representing the vertical projection, and T represents the total center point numbers.

2.5. Grouped connected components

The text candidates are collected by using the text line information to find the text areas. All connected components with non-zero heights that are close together in the horizontal direction are gathered together through a morphological process. Those grouped connected components that either do not include center points or have only a few center points will be removed. Fig. 6(d) shows the final text detection result.



(a) (b) (c) (d)

Fig.6. Extract text region using tensor voting. (a) Input tokens, (b) resulting tensors, (c) tensor voting of text line location, (d) text detection

III. EXPERIMENTAL RESULTS

In our method, we apply the Edge enhancement CER, SWT and Tensor voting algorithms. We will present the experimental results from calculating the components of the proposed method. To evaluate the accuracy of our method, input images captured from 2003 ICDAR and 2013 ICDAR contest test images are used. We also apply different font styles, sizes, color, highlight, blur and low contrast to the images. The algorithm is executed in Matlab 2014 on an Intel Core i7-3770 CPU at 3.40 GHz, with 8GB RAM on a Windows 7 system.

Fig. 7 show the experimental results for text detection in images from Chen et al.'s [15], Toan et al.'s [16] and our method. Chen et al.'s [15] method extracts the text using MSER and SWT. Toan et al.'s [16] paper presents a new idea using tensor voting. The two methods have a lot of false positives and false negatives in several cases.

For the initialized information, we set up parameters for Edge enhancement CER as the range [100 5000] for all cases. The existing methods compared to our proposed method were from Chen et al. [15], Toan et al. [16], Rodrigo et al. [21], Yao et al. [22], Xiaobing et al. [23], CASIA_NLPR et al. [24], Huang et al. [11], and Lei et al. [25]. In addition, Huang et al.'s method used a trained convolution neural network classifier and CASIA NLPR was the winner in the ICDAR 2013 competition.

On the ICDAR 2003 dataset, the proposed method

achieves the best performance with the best precision (90.3%) and best recall (89.5%) as shown in Table 2. On the ICDAR 2013 dataset, our method has the highest precision (93.2%) and high recall (91.5%) as shown in Table 3. Our recall is lower than that of Lei et al. [25] (92.3%). Lei's method used a neural network to classify text and non-text regions. This method also has high performance but it requires a lot of time for training and needs to be trained with many images. Our method used tensor voting, which is useful for eliminating non-text regions in complex background cases (highlight, blur). Therefore our false positives were fewer than for any of the other methods.

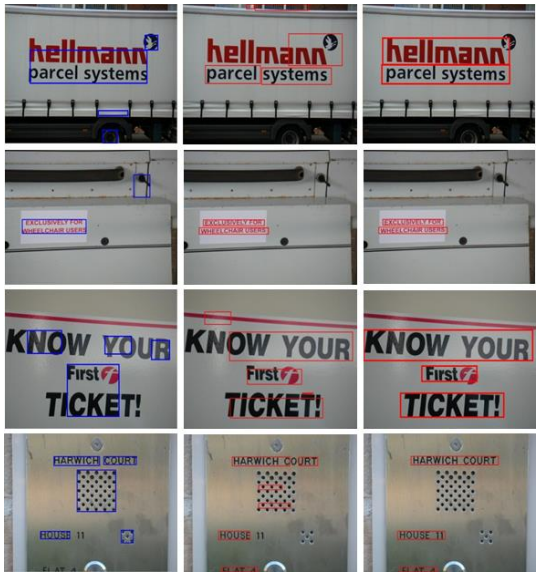
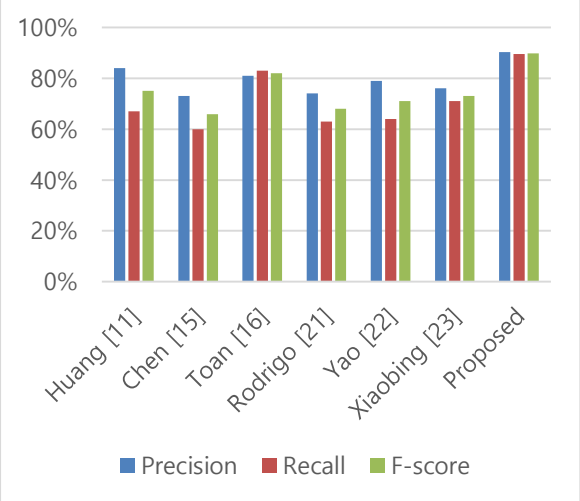


Fig. 7. Results from Chen[15], Toan[16], and our method

Table 1. Comparison of methods from 2003 ICDAR dataset



The comparison is supported by precision and recall, which are represented by the following equations:

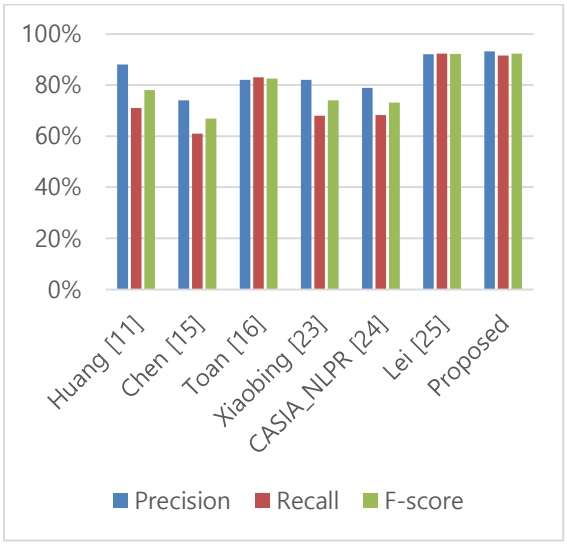
$$F\text{-}score = 2 * \frac{\text{Precision}*\text{Recall}}{(\text{Precision}+\text{Recall})} * 100\% \quad (8)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} * 100\% \quad (9)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} * 100\% \quad (10)$$

Where TP is the number of characters identified correctly, FP is the number of texts inaccurately detected, and FN is the missing number of characters.

Table 2. Comparison of methods from 2013 ICDAR



In our method, most text regions are identified properly, however some regions are not. We observe some failure with natural scene images in Fig. 8. Because of the light, spot, some characters from "Engineering" and "Distributed" in Fig. 8(a) are covered. Our method could not handle these text regions. In Fig. 8(b), all letters are rejected because of a similar color between the texts and background. We cannot extract the connected components of text in the CER step. In Fig. 8(c), the image contains a rare font and the illumination change. There are a few cases of failure in the dataset, which do not affect the result performance. The experimental results depend on the potential text regions extracted in the first step and in the final step. In the first step, CERs are not extracted, so text regions

cannot be detected. Tensor voting is used in the final step, which will eliminate non-text regions, therefore it has the effect of reducing false positives.

(a) (b) (c)

Fig. 8. Failure cases

IV. CONCLUSION

In our paper, an efficient text detection method for natural images is introduced based on Edge enhancement CER, SWT, and especially tensor voting. 2D tensor voting is applied when candidate text regions corresponding to input tokens are used to compute curve salience values to remove false positives. Several unsolved problems in this area such as highlighting, blur, various sizes and fonts for texts are suitably resolved. Edge enhanced contrasting extremal regions are effective in extracting candidate text connected components. Moreover, tensor voting works very well for removing non-text regions. The experiment results indicate that our method is useful for identifying text and has high accuracy with complex natural scene images. In future works, we will consider using other methods to identify text in more difficult cases.

REFERENCES

[1] G. Zhou, Y.Liu, Q.Meng, Y.Zhang, "Detecting multi lingual text in natural scene," International Symposium on Access Spaces, pp.116-120, 2011

[2] S.A. Angadi, M.M.Kodabagi, "A texture based methodology for text region extraction from low resolution natural scene images," *Advance Computing Conference*, pp.121-128, 2010

[3] Xiaoqing Liu, J. Samarabandu, "Multiscale Edge-Based Text Extraction from Complex Images," ICME, 2006

[4] Qixiang Ye, Jianbin Jiao, Jun Huang, Hua Yu, "Text detection and restoration in natural scene images," *Journal of Visual Communication and Image Representation*, Vol. 18, pp.504-513, 2007

[5] Hyunsoo Choi, GueeSang Lee, "Natural Scene Text Binarization using Tensor Voting and Markov Random Field," *KISM Smart Media Journal*, Vol. 4, pp.18-23, 2015

[6] R. Jiang, F.Qi, L.Xu, G.Wu, "Using connected components features to detect and segment text," *Journal of Image and Graphics*, 2006

[7] L. Neumann and J. Matas, "On combining multiple segmentations in scene text recognition," ICDAR, pp.523-527, 2013

[8] Li, Yao, and Huchuan Lu, "Scene text detection via stroke width," ICPR, pp.681-684, 2012

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

[10] Alvaro Gonzalez, Luis M. Bergasa, J. Javier Yebes, Sebasti´an Bronte, "Text location in complex images," ICPR, pp.617-620, 2012

[11] W. Huang, Y. Qiao, and X., Tang "Robust scene text detection with convolution neural network induced MSER trees," ECCV, pp.497-511, 2014

[12] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," ECCV, pp.512-528, 2014

[13] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao, "Scene text detection using graph model built upon maximally stable extremal regions," *Pattern Recognition Letters*, Vol. 34, pp.107-116, 2013

[14] L. Neumann and J. Matas, "Real-time scene text localization and recognition," CVPR, pp.3538-3545, 2012

[15] Chen Huizhong, et al., "Robust Text Detection in Natural Images with Edge-Enhanced Maximally Stable Extremal Regions," ICIP, 2011

[16] Toan Dinh Nguyen, Jonghyun Park, Gueesang Lee, "Tensor Voting Based Text Localization in Natural Scene Images,"

IEEE Signal Processing Letters, Vol. 17, pp.639 - 642 , 2010
[17] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L, "A comparison of affine region detectors," *International. J. Computer Vision,* Vol. 65(1), pp.43-72, 2005
[18] G. Medioni, M. S. Lee, and C.-K. Tang, "A computational framework for segmentation and grouping," Elsevier, 2000.
[19] S. Aja-Fernndez, R. d. L. Garca, D. Tao, and X. Li, "Tensors in image processing and computer vision," *Springer Publishing Co.* , 2009
[20] T.D. Nguyen, and G. Lee, "Color image segmentation using tensor voting based color clustering," in *Pattern Recogn. Letters*, pp.605-614, 2012
[21] Rodrigo Minetto , Nicolas Thome, Matthieu Cord, Neucimar J. Leite, Jorge Stolfi, "Snooper Text: A text detection system for automatic indexing of urban scenes," *Computer Vision and Image Understanding* 122, pp.92-104, 2014
[22] Yao Li, Wenjing Jia, Chunhua Shen, Van den Hengel, Anton, "Characterness: An Indicator of Text in the Wild," *IEEE Transactions on Image Processing*, Vol. 23, issue 4, pp.1666-1677, 2014
[23] Xiaobing Wang, Yonghong Song, Yuanlin Zhang, Jingmin Xin, "Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis," *Pattern Recognition Letters*, pp.41-47, 2015
[24] D.Karatzas, F.Shafait, S.Uchida, M.Iwamura, L.Bigorda, S.Mestre, J.Mas, D.Mota, J.Almazan, L.Heras, "Icdar2013 robust reading competition," ICDAR, pp.1484-1493, 2013
[25] Lei Sun, Qiang Huo , Wei Jia , Kai Chen, "A robust approach for text detection from natural scene images," *Pattern Recognition Letters*, pp.2906-2920, 2015

Authors

Van Khien Pham
He received B.S degree from Hanoi University of Sciences and Technology, Vietnam in 2012. He received M.S. degree in 2016 from Chonnam National University, Dept. of Electronics & Computer Engineering, Korea. His research interests are mainly in the field of Image Processing, Computer Vision.

Soo-Hyung Kim
He received his B.S. degree in Computer Engineering from Seoul National University in 1986, and his M.S. and Ph.D. degrees in Computer Science from Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. From 1990 to 1996, he was a senior member of research staff in Multimedia Research Center of Samsung Electronics Co., Korea. Since 1997, he has been a professor in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, document image processing, medical image processing, and ubiquitous computing

Hyung-Jeong Yang
She received her B.S., M.S. and Ph. D. from Chonbuk National University, Korea. She iscurrently an associate professor at Dept. of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. Her main research interests include multimedia data mining, pattern recognition, artificial intelligence, e-Learning, ande-Design.

Guee-Sang Lee
He received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, 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 in Chonnam National University, Korea. His research interests are mainly in the field of image processing, computer vision and video technology.