Text detection and recognition using enhanced MSER detection and a novel OCR technique

Conference Paper · May 2016

DOI: 10.1109/ICIEV.2016.7760054

CITATIONS
16

READS
5,704

4 authors, including:

Md rabiul Islam
Carl von Ossietzky Universität Oldenburg
76 PUBLICATIONS **487** CITATIONS
SEE PROFILE

Chayan Mondal
Bangabandhu Sheikh Mujibur Rahman Science & Technology University
10 PUBLICATIONS **61** CITATIONS
SEE PROFILE

Abu Syed Md. Jannatul Islam
Khulna University of Engineering and Technology
32 PUBLICATIONS **138** CITATIONS
SEE PROFILE

Text Detection and Recognition Using Enhanced MSER Detection and a Novel OCR Technique

Md. Rabiul Islam*, Chayan Mondal, Md. Kawsar Azam, and Abu Syed Md. Jannatul Islam
Department of Electrical and Electronic Engineering (Khulna University of Engineering & Technology, Bangladesh)
* E-mail: rabiulkuet@hotmail.com

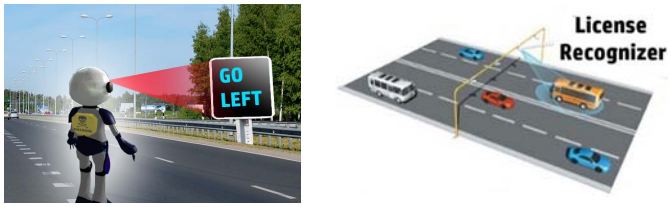
Abstract—**Detection and recognition of text from any natural scene image is challenging but essential extensively for extracting information from the image. In this paper, we propose an accurate and effective algorithm for detecting enhanced Maximally Stable Extremal Regions (MSERs) as main character candidates and these character candidates are filtered by stroke width variation for removing regions where the stroke width exhibits too much variation. For the detection of text regions, firstly some preprocessing is applied to the natural image and then after detecting MSERs, an intersection of canny edge and MSER region is produced to locate regions that are even more likely to belong to text. Finally, the selected text region is taken as an input of a novel Optical Character Recognition (OCR) technique to make the text editable and usable. The evaluation results substantiates 77.47% of the *f*-measure on the ICDAR 2011 dataset which is better than the previous performance 76.22%.**

Keywords—Text detection; optical character recogntion; preprocessing; MSER & canny edge; stroke width.

I. INTRODUCTION

Image contains many technical and digital information used in the different fields of computer vision. In recent years, visual detection and recognition of text from image is claimable because of its application in content based image searching, robotic navigation (Fig. 1a), automatic car number plate recognition (Fig. 1b), extracting passport or business card or bank statement information, converting handwriting to real time control of computer, making editable the text of any image etc. But owing to irregular background, variations of font style, size, color, orientation, geometric and photometric distortion, text must have to be robustly detected from any natural scene image.

Text detection has been considered in several current studies in various competitive methods. Existing methods of the detection of text can be classified into three groups: (1) texture based method [3]; (2) Connected Component (CC) based method [4],[5]; and (3) hybrid method [1]. In texture based method text is considered as exclusive texture that is dissimilar from the irregular background. A trained classifier is engaged on the job of identification of existence of text on the image. The segmented texture is filtered with nonlinear portion. The output of the filter at $tanh(\alpha t)$, where $\alpha = 0.25$. If images are more complicated, texture segmentation scheme is not sufficient [6]. Connected component based method find out character candidates from image by connected component analysis which is followed by grouping character candidates to text. Supplementary checks have to be performed to remove false position. In hybrid method text is detected by extracting connected components as character candidates by binarization. Non characters are erased by Conditional Random Fields (CRFs).



(a) (b)

Figure 1. Applications of text detection and recognition. (a) Robotic navigation; (b) Automatic license plate checking system.

Recently, MSER based text detection has become a heart among all types of text detection method [2]. Maximally Stable Extremal Regions (MSER) works with the intensity management of a digital image. The term "extremal region" represents connected component which differentiate the higher or lower intensity of a pixel to the outer boundary pixel. Although MSER is actually in the family of connected component based method, different new projects have taken a great importance on it because of its promising performance.

In this paper, we propose an enhanced MSER based detection combining with text recognition. Firstly MSER approaches is applied to a scene image to detect a large region of non-character which employs MSER as basic connected component. Since natural scene images are normally affected by a wide variation of background, to detect small text or to detect text from blur image or limited resolution image canny edge and MSER have to be combined. For obtaining more reliable result we have used stroke width transformation of the image to perform filtering and joining connected components. Finally, an Optical Character Recognition (OCR) technique with clustering by Gaussian Mixture Model (GMM) is operated on the selected text region to recognize the actual text information of the natural scene image. Our method exhibits outstanding performance on natural scene detection and on ICDAR 2011 Robust Reading Competition dataset. The results rank on the first position for our best performance 77.47% in *f*-measure which is better than any other methods of the previous.

978-1-5090-1269-5/16/$31.00 ©2016 IEEE

The reminder of this paper are arranged as follows. Recent MSER based detection method are viewed in section II. The text detection method which we have proposed has been described in section III. In section IV recognition model of text is viewed and in section V results are shown with comparison to previous. Finally section VI concludes the paper.

II. RELATED WORK

Scene text detection and recognition is still an open challenging affair to be addressed. Scene text information extraction actually contents two unique process: text detection and text recognition. In this paper, we emphasize the detection technique. Although MSER based text detection has a promising performance, it is not totally free from some specific limitation. When image of a low quality or low resolution or strong noise affected or blur then most of the characters candidates of the text corresponds as non-character. Almost all the time non characters are also recognized as text region for the low value of character confidence.

In sliding windows based or region based method, a sliding windows is occupied to search for probable text in the image and a machine learning is used to determine texts. To do this multiple operation the process tends to be slow. In Neumann and Matas method [7] a two stage algorithm is used for Extremal Regions (ERs) pruning. Firstly, a classifier is trained to estimate the class-conditional probabilities of ERs. On the second stage these ERs are filtered by computable descriptors. So this methods is inconvenience for computational complexity. Carlos proposed a MSER pruning algorithm that contains two parts: (1) 'reduction of linear segmentation' which reduces linear segments in the MSER tree into one node and (2) 'hierarchical filtering' which eliminates the nodes by checking them considering the parameter size, aspect ratio, complexity, border energy and texture.

It is also difficult matter that how to group character candidates into text candidates. For the purpose, generally two types of approaches are used: rule based, clustering based. Neumann and Mata's gives their concern to group character candidates by using one and more top and bottom lines. Carlos made a complete connected graph over character candidates. Chen followed a way of clustering constrains stroke width and height difference. The above rule based methods work well but not reliable or not sufficient. Furthermore, it requires hand tuning parameter is required which is actually time consuming. On the other hand we firstly detect MSER region and then by different steps of filtering and intersection we get the region of text. Each steps are not so complicated and there is no hand changing parameter described in the next portion. Since there is no hand tuning parameter and Maximally Stable Extremal Region (MSER) is filtered with canny edge & stroke width configuration; it is claimed 'enhanced' MSER. After detecting the text region from the natural scene image a Gaussian Mixture Model (GMM) is used to recognize character candidates.

III. ENHANCED MSER BASED SCENCE TEXT DETECTION

Our text detection method is slightly different from the traditional Maximally Stable Extremal Regions (MSERs) method. We proposed an enhanced MSER detection technique to locate the position of text in the image and Optical Character Recognition (OCR) technique is applied to this selected text part of the image. The complete process of our detection and recognition algorithm is demonstrated in the following Fig. 2. The corresponding output of step by step techniques implemented on a natural scene image is displayed in Fig. 3. Thus incorporating these several key improvements, the proposed method is also sensitive to small letter, blur image, limited resolution image. The proposed enhanced MSER based method of text detection includes the following steps:

1. *MSER region detection:* Normally, text characters usually have consistent color. So we start to find the text by selecting the regions of similar intensities by using MSER region detector. Many non-text region is also detected and so further processing is applied.
2. *Intersection of canny edge with MSER region:* Canny edge detection algorithm performs a high response to edge detection. And intersection of MSER and canny edge produce the region that is likely be text. By using the region properties, some connected component can be removed. According to the variation of different font, image size, or languages the filtering thresholds are automatically detected in our proposed algorithm.
3. *Visualization of text candidate's stroke width:* Character in almost all language have a similar thickness throughout or stroke width. After this step the region where the stroke width contains too much variation is eliminated.
4. *Text candidate after stroke width filtering:* Non text region can be eliminated by determining a large variation in stroke width.
5. *Image region under mask created by joining individual characters:* Then the individual component is merged to compute a bounding box of text region. Morphological closing is done here.

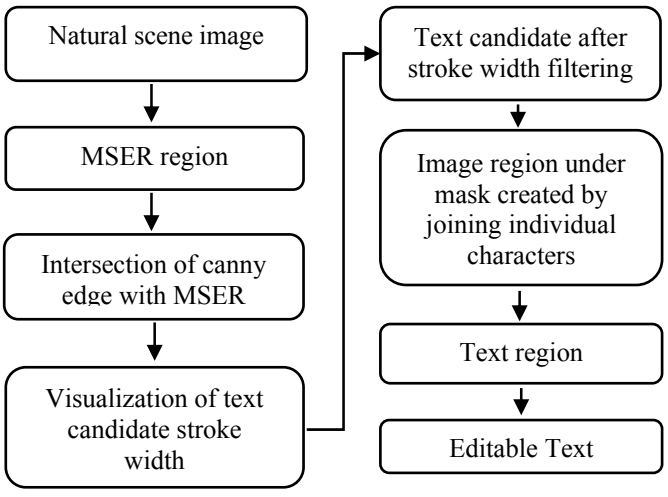


Figure 2. Step by step processing flowchart by our Enhanced MSER detection process.

Figure 3. Corresponding pictures of different steps according to the flowchart.

6. *Text region:* Finally text region of the image is detected efficiently. In Fig. 3 text part of the natural scene image is shown which is found by our algorithm.
7. *Optical character recognition technique applied on the text part:* To recognize the text Optical Character Recognition technique is activated on the achieved text region.

A. Variation and its regularization

Originally MSERs are controlled by a single parameter, which controls how the stability is calculated. The intensity variation of extremal region can be defined as follows. Consider, stability of an extremal region R is the inverse of the relative area variation of the region R; if the intensity level is increased by the value Δ. Then variation can be shown as:

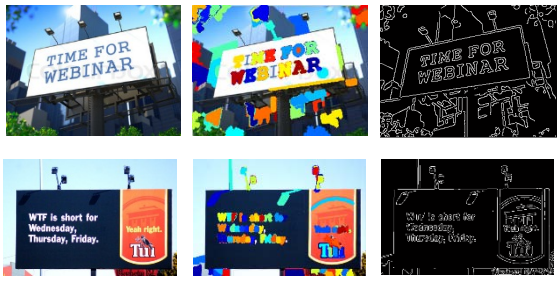
$$v(R_x) = \frac{|R(+\Delta) - R|}{|R|} \quad (1)$$

Where, $|R|$ is the area of the extremal region R, $R(+\Delta)$ denotes the extremal region which is Δ levels up which contains R, and $|R(+\Delta) - R|$ is the area difference of the two regions.

The proposed algorithm selects regions which are "maximally stable", that means they have the lower variation than the regions of one level above or below. Natural image contains some region which seemed to be maximal but actually not. So geometric filtering, masking and preprocessing are used to determine the text part from the image.

B. Canny edge detection

After getting MSER detected binary image, it is then intersect with canny edge. By using Gaussian filter or derivative gradient is formed. Canny edge is the weak edge of local maxima of the gradient of natural image (Fig. 4). As many noisy region of the image we perform a few number of simple and reliable geometric checks on each connected component to filter out the non-text object. Intuitively, all the character follows an aspect ratio in the range of 5:3 to 1:1 belonging to its size. We reject the connected component larger or smaller than the aspect ratio. Finally the regions of containing hole are also eliminated because they are unlikely to be a character.



(a) (b) (c)

Figure 4. Examples of MSER and canny edge detection detection. (a) original image; (b) MSER region; (c) Canny edge.

C. Non-text region elimination:

It is difficult to train an efficient classifier using unbalanced dataset. So a character classifier is employed to predict the posterior probabilities of text candidates corresponding the region without text.

The stroke width variation, smoothness i.e. average difference of the adjacent pixels, length width, height, aspect ratio etc. features are used to train character classifier. Consider 'O' be the observation where there are 'm' character candidates, among which 'n' are classified as non-character. Consider, 'O' is a function of m, n and the range of m($m \in \mathbf{N}, m \geq 2$) and n($n \in \mathbf{N}, n \leq m$). Now, the probability of text $P[O(m,n;p)|text]=p^{m-n}(1-p)^n$ and the probability of non-text $P[O(m,n;p)|non\_text]=p^n(1-p)^{m-n}$.

Let P(text) and P(non-text) are the prior probability of T for text and non-text respectively. Then according to Bayes function, the posterior probability of T sensitive to non-text can be expressed as

$$P(non\_text \mid O(m,n;p)) = \frac{P(O(m,n;p) \mid non\_text) * P(non\_text)}{P(O(m,n;p))} \quad (2)$$

Where the probabilities of observation =P(O(m,n;p));

$P(O(m,n;p))=P(O(m,n;p)|text)P(text)+P(O(m,n;p)non\_text)P(non\_text)$ (3)

Non-text candidate regions are eliminated under the condition that is P(non_text|O(m,n;p)) ≥ ε, Where ε is the threshold value. To train the character classifier 73% of ICDAR training dataset is used for the appropriate value of threshold ε and for the remaining 27% dataset, varying threshold value is used. Performance is varied with the variation of ε. In our experiment it is clear that when ε =0.996; almost 98.1% of non-text are eliminated that means almost all text are saved.

IV. TEXT RECOGNITION

Recognition process is employed to our selected text region from the natural scene image. We train a character recognizer which can recognize 10 digits (0 to 9), 52 English letters (26 for upper case and 26 for lower case) total 62 characters. Two dataset containing image patches of complete and full text characters is used to train. For different font style shown in Fig. 5, character is described helps to get more accurate results in recognition. For the description of text character a novel character descriptor model is proposed which combined MSER detector to find out stroke components and Random detector to extract preset number in random pattern.

A. Gaussian Mixture Model (GMM)

Gaussian mixture model is normally used to fit a vector of unknown parameter. In our algorithm, Dense Detector produces a uniform 8×8 key points array and Random Detector produces 64 key points randomly. We use 8 Gaussian distribution for clustering. For building GMM, firstly calculate K centers of the Histogram of Oriented Gradients (HOG) descriptor, where K means clustering (K=8).

S-th center is used as initial means μ_s of the s-th Gaussian in GMM. The values of co-variance σ_s and initial weight w_s are calculated form the value of means. Then, Expectation-Maximization (EM) algorithm is employed to estimate the the following three parameters: means, co-variance and weight.

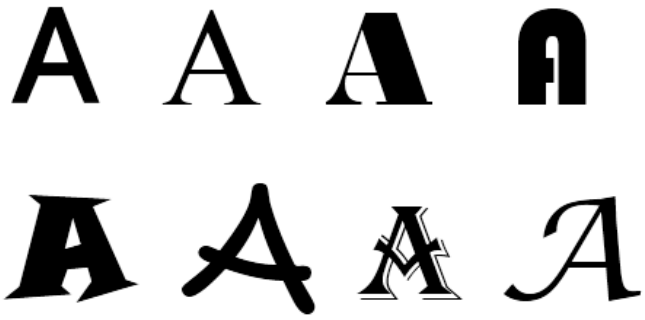


Figure 5. Different font styles are trained for getting more accuracy in character recognition.

A likelihood vector of all Gaussian is represented by the equation (4).

$$P_x = \langle w_s p_s(x \mid \mu_s, \sigma_s) \rangle \Big|_{s=1}^{K}$$

$$= \langle w_1 p_1(x \mid \mu_1, \sigma_1), w_2 p_2(x \mid \mu_2, \sigma_2), ..., w_k p_k(x \mid \mu_k, \sigma_k) \rangle \quad (4)$$

$$P_s(x \mid \mu_s, \sigma_s) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \quad (5)$$

Where, x represents the HOG based feature vector at a keypoints, P_x denotes the likelihood vector of feature vector x, and $P_s(x \mid \mu_s, \sigma_s)$ denotes the probability value of x at the s-th Gaussian. A character descriptor is used as a feature vector of character patch to train recognizer in SVM model.

B. Character stroke configuration

In the view of pixel-level prospective, stroke is defined as a region bounded by two parallel boundary segments. A character structure consists of multiple strokes. The distance between the two parallel boundaries is called stroke width. Same class character has the same stroke. We use a stroke alignment method for estimating the average value of stroke configuration Fig. 6. Thus our proposed method can detect text for any kind of fonts, styles, and sizes. The stroke alignment is defined by the following equations.

$$E = \sum_i (D(\bar{S}, T_i(S_i)) + g(T_i)) \quad (6)$$

$$D(S_m, S_n) = \sum_i (|S_m(p) - S_n(p)|)^2 \quad (7)$$

Where, S indicates the mean value of stroke configuration, S_i indicates i-th stroke configuration. T_i represents the transformation of the i-th stroke. D represent the distance between the strokes of adjacent two character stroke. The letter p is for peak value.

The overall program procedure can be described as following procedure and my detection software interface is shown in Fig. 6.

```
a= MSER region
b= canny edge
c= a ∩ b
d= character confidence
        if (d ≥ threshold value)
        return [t];  %t=text
        else
        return [nt]; %nt=non-text

similarly check stroke width
boxes = round (vertcat (stats> areaThreshold). BoundingBox))
for i=1:size(boxes,1)
        imshow(imcrop(colorImage, boxes(i,:)));
        title('Text region')
        end

recognition process
outout.
```

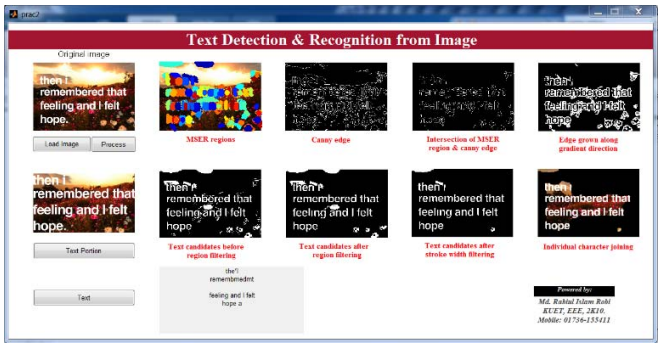


Figure 6. Our developed text detector and recognizer software interface.

V. EXPERIMENTS RESULT

The proposed theme of scene text detection is employed on the publicly available dataset ICDAR 2011 Robust Reading Competition Dataset as Fig. 7 [9]. The dataset having 229 training image and 255 testing image is more perplexing but offers several enhancement over the previous scheme. Basically the value of recall, precision and f measure is defined by

$$recall = \frac{\sum_{i=1}^{|G|} match_G(G_i)}{|G|} \quad (8)$$

$$precision = \frac{\sum_{j=1}^{|D|} match_D(D_j)}{|D|} \quad (9)$$

$$f = 2\frac{recall.precision}{recall + precision} \quad (10)$$

Here, D= set of detected rectangles

G= set of ground truth rectangles

In previous manipulation scheme, matching function are defined as follows.

$$match_G(G_i) = \max_{j=1...|D|} \frac{2.area(G_i \cap D_j)}{area(G_i) + area(D_j)} \quad (11)$$

$$match_D(D_j) = \max_{i=1...|G|} \frac{2.area(D_j \cap G_i)}{area(D_j) + area(G_i)} \quad (12)$$

From ICDAR 2011 competition result, the performance of our method and very recent top scoring methods (like Xu-Cheng,

Figure 7. Text detection example on ICDAR 2011 dataset. Text is detected from different irregular background and font variations.

TABLE I. COMPARISON OF PERFORMANCE OF TEXT DETECTION ON ICDAR 2011 ROBUST READING COMPETITION DATASET

Name of the method	Recall (%)	Precision (%)	*f* (%)
Our proposed method	68.4	89.3	77.47
Xu-Cheng Yin method	68.26	86.29	76.22
Shi method [10]	63.1	83.3	71.8
Kim's method [11]	62.47	82.98	71.28
Neumann and matas [7]	64.7	73.1	68.7
Epshtein [12]	60.0	73.4	66
Yi's method	58.09	67.22	62.32
TH-TextLoc system	57.68	66.97	61.98
Neumann's method	52.54	68.93	59.63
HWDavid	46	44	45
Fabrizio	39	46	43

Yin Method, Shi Method, and Kim's Method) is shown in the Table I. Our proposed method has achieved a better recall 68.4 and better precision 89.3 rather than the highest of the previous methods of 68.26 and 86.29 respectively.

In the point of fact our enhance MSER and stroke width based algorithm for locating the text is almost 77.47% successful in f measure. An Optical Character Recognition (OCR) technique performs a part of the code can successfully be operated on the selected text region of the natural scene image to recognize the text. Our system achieve higher performance because:

1. In MSER extraction algorithm the default parameter setting which is the major cause of degrading the recall value of the previous detection technique can be solved in our technique by changing the default setting. From the parameter set, Δ controls the variation, maximal variation v_+ omits too unstable MSER, and the minimal diversity parameter d_+ eliminates duplicate MSERs. The default parameter set *(Δ=5, v_+=0.25, d_+=0.2)* is changed to a new parameter set *(Δ=1, v_+=0.55, d_+=0.12)* due to detect some low contrast character and not to miss the region which is more likely to text.

2. The proposed method may be a great choice of modern technology because of its fast speed. A computer of Intel(R) Core(TM) i5 CPU M370, 2.40GHz properties takes 0.45s per image to process the algorithm. Whereas the time is 1.5s for Shi method with a PC of Intel(R) Core (TM)2 Duo 2.33 GHz CPU [10].

Among the different types of text detection method, our proposed method achieved more value for different performance measuring papameter shown in Fig. 8.

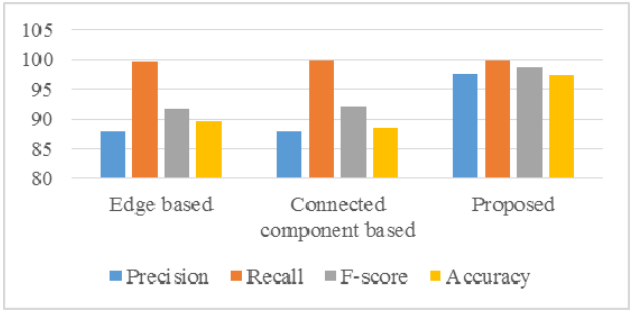


Figure 8. Comparison of performance among different types of method.

VI. CONCLUSION

This paper presented an enhanced MSER based scene text method. It is capable of differentiating the text part from the natural scene image and can recognize the text from the selected text region. To overcome the complexity of image blur and small letter, enhanced MSER had been developed with complementary properties of MSER and Canny edge. We proposed a unique image operator to determine accurately the stroke width of binary connected components. Finally, OCR with intersecting character description had been applied to the selected text part to recognize text. Our proposed system exhibits exclusive performance over state of the art methods on ICDAR 2011 competition.

REFERENCES

[1] G. Zhou, Y. Liu , Z. Tian, and Y. Su, "A new hybrid method to detect text in natural scene," *18th IEEE Int. Conf. Image Processing (ICIP)*, pp. 2605 – 2608, Sep. 2011.

[2] L. Gómez and D. Karatzas, "MSER-Based Real-Time Text Detection and Tracking," *22nd International Conference on Pattern Recognition (ICPR)*, pp. 3110-3115, 2014.

[3] 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. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1631 – 1639, Dec. 2003.

[4] H. I. Koo, and D. H. Kim, "Scene text detection via connected component clustering and nontext filtering," *IEEE Trans. Image Processing*, vol. 22, no. 6, pp. 2296 – 2305, Feb. 2013.

[5] A. Srivastav, and J. Kumar, "Text detection in scene images using stroke width and nearest-neighbor constraints," *TENCON IEEE Region 10 Conf.,* pp 1-5, Nov. 2008.

[6] V. Wu, R. Manmatha, and E. M. Riseman, "TextFinder: An automatic system to detect and recognize text in images*," IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 21, no.11, pp. 1224-1229, Nov. 1999.

[7] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," *Computer Vision–ACCV 2010. Springer Berlin Heidelberg*, vol. 6494, pp. 770-783, 2011.

[8] C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration," *IEEE Trans. Image Processing*, vol. 23, no. 7, pp. 2972 – 2982, Apr. 2014.

[9] www.cvc.uab.es/icdar2011competition/

[10] C. Shi, C. Wang, B. Xiao, Y. Zhang, and S. Gao, "Scene text detection using graph model built upon maximally stable extremal regions," *Pattern Recognition Letters*, vol. 34, no. 2, pp. 107–116, Jan. 2013.

[11] K. Junga, K. I. Kimb, and A. K. Jainc, "Text information extraction in images and video: a survey," *Pattern Recognition*, vol. 37, no. 5, pp. 977–997, May 2004.

[12] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," *IEEE Conf. Computer Vision and Pattern Recognition(CVPR)*,pp.2963–2970,Jun.2010.