Text Detection via Edgeless Stroke Width Transform

Anhar Risnumawan and Chee Seng Chan

Center of Image and Signal Processing
Faculty of Computer Science & Information Technology
University of Malaya, 50603 Kuala Lumpur, Malaysia
Email: anhar@siswa.um.edu.my; cs.chan@um.edu.my

***Abstract*—Text detection in scene images has gained widespread interests. A notable work, which is the Stroke Width Transform (SWT), has been attracting much interests due to its simplicity and efficiency. However, the SWT has difficulty in situations such as blur, low contrast, and illumination change images since it highly relies on the outcome from the edge detector. In this paper, a novel method is proposed to obtain stroke width image without the edge detectors. In particular, we replace the edge detector algorithm with the Extremal Regions (ERs) and propose a novel weighted Markov Random Field (MRF) method with three properties to construct a finer stroke width image. Experiment results on ICDAR datasets and a comparison with the state-of-the-art methods have shown the efficiency of the proposed method.**

I. Introduction

Text detection on scene images has gained much interest due to the emerged of wearable computing; as well as its usefulness in many real world applications, such as assisting visually impaired people, tourists navigation, enhancing safe vehicle driving, etc. [1], [2]. Due to the complex background and high variation of fonts, sizes, and color, text on scene images have to be robustly detected and one of notable works on the scene text detection is the Stroke Width Transform (SWT) [3].

The SWT is attracting much interest due to its simplicity and efficiency. The simplicity can be seen from which only the edge image is used. In particular, for each edge pixel, it traverses based on its gradient orientation till another pixel is encountered. This creates a path with its length traversed. Then, the path is saved by assigning the length value to the respective pixels of path in an image. This image is called the stroke width image, and an example is shown in Fig. 1. It's efficiency could be seen as most of the characters have uniform stroke width value, thus helps to easily form connected components using simple rule to eliminate the non-text components.

Yao et al. [4] has adopted the SWT to detect non-horizontal text lines and this increases the flexibility of SWT to detect text lines of any orientations (i.e. multi-oriented). Recently, Chen et al. [5] proposed a variant of the SWT with the employed of Maximally Stable Extremal Regions (MSER), and incorporating the Canny edge detector to obtain the connected components by pruning on binary image. The stroke width for each component then is extracted using the distance transform. Though successful, the aforementioned solutions often suffer from difficulties of detecting texts in blur, low contrast, and illumination images as depicted in Fig. 1 due to the limitation of the edge detectors. A possible solution for the aforementioned methods is by manually setting the edge detector threshold. However, such an attempt is impractical for a real time application and cumbersome.

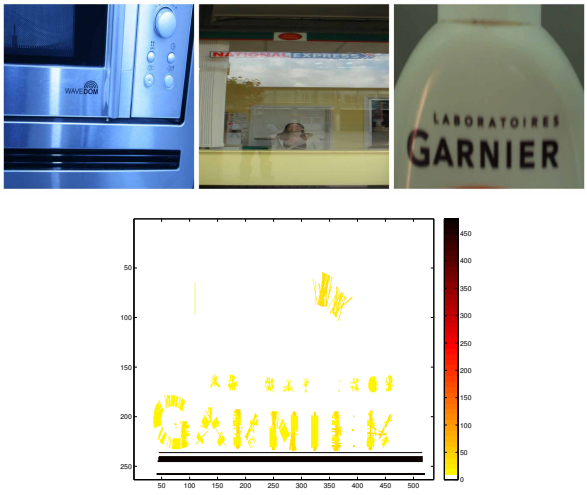


Fig. 1. Example of low contrast images (top-left), illumination images (top-middle), and blur images (top-right). An example of the stroke width image (bottom), where the color indicates stroke width value. Note that the generated stroke width image appears to be broken, causing partly connected components for single character. (Best viewed in color)

In this paper, we propose a novel method to obtain finer stroke width image without utilizing any edge detectors in order to handle the state-of-the-art limitations. To facilitate this, the Extremal Regions (ERs) [6] is adopted to replace the edge detector. Given an input image, regions are extracted using the ERs. A novel weighted Markov Random Field (MRF) with three properties is then proposed, which will be minimised using the α-expansion algorithm [7], [8], to accurately integrate the extracted regions into the finer stroke width image. Following this, the accurate connected components can be formed from the stroke width image using the rules as suggested by [3]. To remove the non-text components, functional max-margin [9] with decision tree is used as strong classifier in false-positives elimination stage. Finally, bounding box of the text components are formed.

Our contributions are 2-folds: Firstly, we propose a variant of the SWT that able to cope with blur, low contrast, and illumination images, which is critical for text detection on scene images. Secondly, we introduce a novel weighted MRF

with three properties, minimised using α-expansion algorithm to build the finer stroke width image.

The rest of the paper is organized as follows: In Section. II, we describe the proposed method. Section III provides the experimental results and Section IV concludes the paper.

II. Formulating the Proposed Method

The overall framework of the proposed method is shown in Fig. 2. Firstly, regions are extracted from the input image using the ERs [6] (which is reviewed in Section II-A). Secondly, a novel weighted MRF with three properties is proposed to accurately integrate the extracted regions into finer stroke width image (Section II-B). Thirdly, connected components analysis is performed to get the components candidate. Then, the non-text components are eliminated using the functional max-margin [9] with decision tree (Section II-C). Finally, text lines formation is performed to obtain the bounding box of text components (Section II-D).

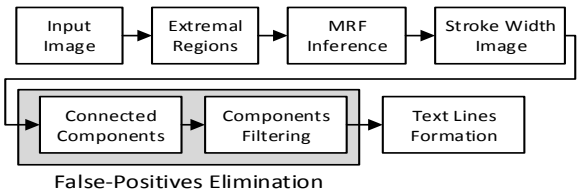


Fig. 2. The overall framework of the proposed method.

A. Extremal Regions

Given an input image, regions are extracted using the ERs such that the region r is a set of pixels whose intensity is less than a threshold value. To compute the regions, the threshold value is increased from 0 to 255 of each channel. Pixels that below the threshold value and correspond to local intensity minima, will appear and eventually grow larger. The regions then are represented by the set of connected components in the sequence.

For every threshold, descriptor of each region $\Theta_{r_{(th)}}$ is incrementally computed, that is, by combining the descriptor of the region from the previous threshold, $\Theta_{r_{(th)}} = \Theta_{r_{(th)}} \cup \Theta_{r_{(th-1)}}$. This can be done in $O(1)$. The descriptor is used as feature for a classifier that estimate the probability of region is character $P(r = char)$. If the probability $P(r = char)$ does not satisfy the global limit, $P(r = char) < p_{min}$, the region is eliminated.

In practice, we employed the RGB, HSI, gradient intensity channels and the following descriptor: area, bounding box, perimeter, Euler number, and horizontal crossings, as suggested by [6].

Let defines the output of ERs is set of regions $\mathcal{R} = \{r_1, r_2, r_n, \ldots, r_N\}$ and its probability $P(r_n = char)$, where N is the total of extracted regions from all the channels.

B. Weighted MRF inference

Conventionally in ERs, the regions that do not satisfy the global limit p_{min} will be eliminated. However, defining correct threshold value (p_{min}) is surprisingly difficult and at the same time we also need to ensure that the eliminated regions are all non-text. A possible solution would be selecting low global limit p_{min} so that truly non-text regions are eliminated. Empirically, $p_{min} = 0.1$ is selected.

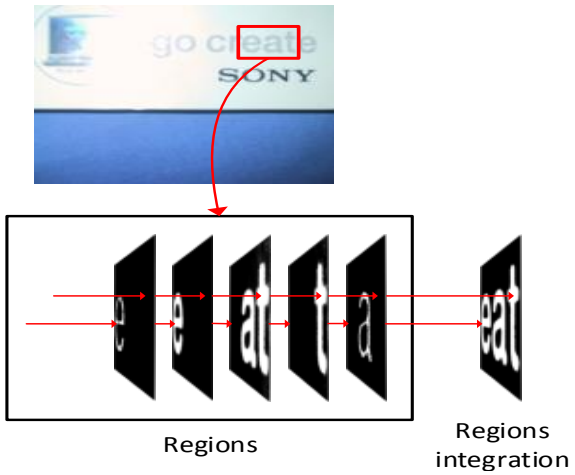


Fig. 3. Illustration of regions integration. In this example, the regions of red box are extracted using ERs. Then, for every pixel (red arrows), it selects the pixel from the best region and maps it onto the stroke width image.

Given the set of regions $\mathcal{R}$, the idea is to select which regions are the best to represent the text pixels. More specifically, for every pixel in image, the pixel from the best region is selected and mapped onto stroke width image, as illustrated in Fig. 3.

In practice, the set of regions $\mathcal{R}$ often contain multiple regions and different regions for a single character. For example, the character 'e' in Fig. 3 has two regions, thin and thicker characters. The best region therefore is highly necessary to be selected to get accurate representation of the text pixels and to reduce the number of regions.

We formulate the problem as a minimization of MRF algorithm and propose three properties to integrate the regions into finer stroke width image. The objective function $\mathcal{J}$ can be written as follows,

$$\mathcal{J} = \sum_{p \in \mathcal{D}} \psi_p(\ell(p)) + \lambda_1 \sum_{p \in \mathcal{D}} \psi'_p(\ell(p)) + \lambda_2 \sum_{p,q \in \mathcal{D}} \psi_{p,q}(\ell(p), \ell(q)) \tag{1}$$

where the thress properties are ψ_p, ψ'_p, and $\psi_{p,q}$ that capture unary and pairwise relation, respectively. The position of pixels $p, q \in \mathcal{D}$ are in spatial domain $\mathcal{D} \in \mathbb{R} \times \mathbb{R}$ and λ_1, λ_2 are weighted constants. Label of p-pixel and q-pixel are $\ell(p)$ and $\ell(q)$, respectively.

To select the best region, the index of the regions are represented by labels. Then, the number of labels are determined by the number of regions $|\mathcal{R}|$. By doing this, the output of the minimization problem (1) is that every pixel has its label which corresponds to the best region.

The three properties are described as follows,

Property 2.1: The property of character pixels, ψ_p is defined as,

$$\psi_p(\ell(p)=l)=\begin{cases}(1-P(r_l=char)) & ,p\in r_l\\ 1 & ,otherwise\end{cases} \quad (2)$$

This property is taken into account with the intention of capturing the pixels that are highly likely representing character.

Property 2.2: The property of character consistency. This property ψ'_p is inspired from MSER method [10] that detects regions of having maximally stable while increasing the threshold from 0 to 255. In this work, we capture the character consistency differently by counting the occurrences of pixels belonging to several regions. Recall that the regions $\mathcal{R}$ can contain multiple regions for the same character. Pixels which satisfy this property are highly likely contained in several regions.

This can be devised using histogram as follows: Firstly, for every pixel p, the occurrences value $\mathcal{O}\in\{o_1,o_2,\ldots,o_{10}\}$ are computed which is obtained from the histogram of $\mathbf{P}=\{P(r_n=char):p\in r_n\}$ using 10 bin, where $o_{1:10}$ are the frequency value of histogram. Secondly, we define a function $f_{hist}:\mathcal{Z}\rightarrow\mathcal{O}$ that maps scaled probability $z_p(l)=\text{floor}(P(r_l=char)/10)+1$, where $p\in r_l$ and $z_p\in\mathcal{Z}$, to the occurrences value. This property is defined as follows,

$$\psi'_p(\ell(p)=l)=-\log f_{hist}(z_p(l)) \quad (3)$$

Property 2.3: The property of pixels smoothness $\psi_{p,q}$ is defined to influence the neighbouring pixels into having the same label. This can be devised by taking the difference between the probability $P(r=char)$ of pixel p and q. Thus, the stronger the difference, the more likely the neighbouring pixels have different label. This property is defined as,

$$\psi_{p,q}(\ell(p),\ell(q))=\delta[\ell(p)\neq\ell(q)](P(r_l=char)-P(r'_l=char))^2 \quad (4)$$

where $p\in r_l$ and $q\in r'_l$.

We minimize the objective function (Eq. 1) using the α-expansion [7], [8] and the output is every pixel has its label. Recall that, the labels indicate the index to the computed regions $\mathcal{R}$. Using this information, we eliminate the regions that are not indexed by the labels, and we define these new regions as $\mathcal{R}'$.

From this step we have constructed the stroke width image, but the pixels value represent the labels that correspond to the index of regions $\mathcal{R}'$. The idea then is to convert the label from each pixel to stroke width value. Firstly, we extract the stroke width of $\mathcal{R}'$ using distance transform as proposed by [5], lets define this as $\mathcal{R}_{sw}$. Secondly, for every pixel and its label, the stroke width value is extracted from $\mathcal{R}_{sw}$. If we define this function as $\mathcal{SW}$, then it can be written as,

$$\mathcal{SW}:\{p\in\mathcal{D},\ell(p)\}\rightarrow\mathbb{R}^{M\times N} \quad (5)$$

where M and N are image width and height, respectively.

Note that the distance transform is used to compute the stroke width as it has the advantage of solving undesirable holes, which usually appear on larger font.

C. False positives elimination

With the stroke width image, connected components then are extracted using simple rule, that is grouping of neighbouring pixels if its stroke width ratio do not exceed an empirically determined threshold, which is suggested by [3]. Compared to [6] where there are high possibility of having many number of multiple components from many channels, the number of connected components of the proposed method are much lower, single component for single character, since we firstly formed it into stroke width image then only computed the connected components.

The resulting components can contain non-text, to eliminate non-text components we perform preliminary filtering using 3 features as suggested in [4]. These features, namely width variation, aspect ratio, and occupation ratio, have been proven to be both effective and efficient.

Functional max-margin [9] using decision tree then is employed as strong classifier to further remove non-text components. This classifier is employed since its ability to utilize the unary and pairwise energy. With this, the classifier can be more discriminant and can be solved efficiently since it uses the sub-gradient method without utilizing any quadratic programming. Note that, in Section II-B unary and pairwise are used in pixels level, while in this step in components level.

For the unary energy, the following features are used - the features as in the preliminary filtering, width variation which is the ratio between the component standard deviation and mean of stroke width, and histogram of oriented gradient (HOG) features using 4x4 window and 6 orientation bins [4].

For the pairwise features, these features are employed [11] - shape difference, spatial distance, overlap ratio, and stroke width mean ratio (the minimum ratio of the mean stroke width of two components).

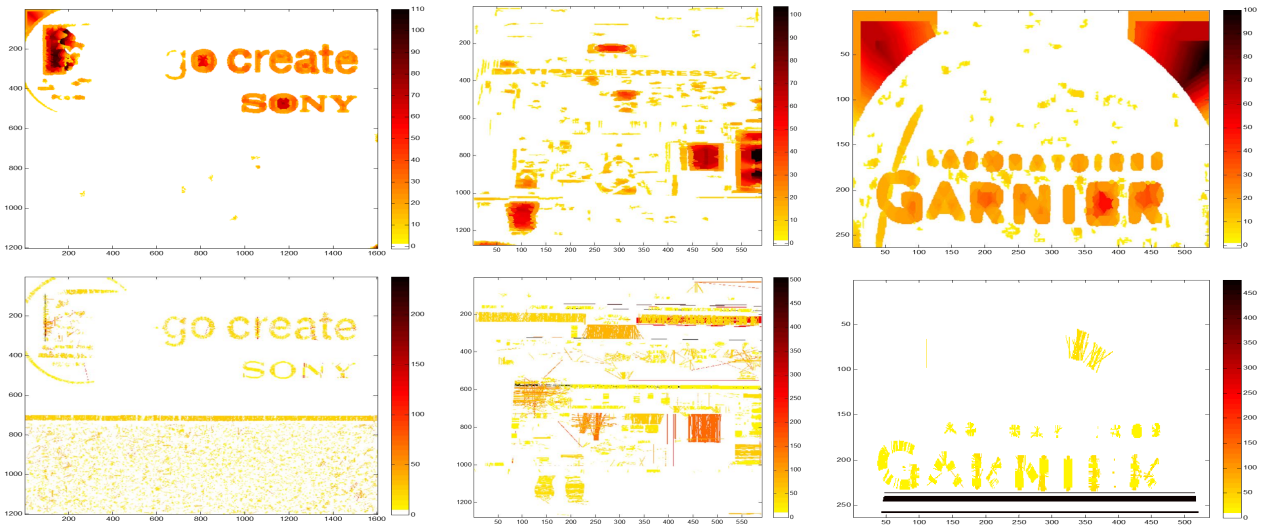
D. Text lines formation

From the previous step, most of the non-text components have been eliminated resulting only text components. In this step, the text components then are grouped and bounding boxes are formed. We adopt grouping method [12] due to simplicity to setup for horizontal text lines formation and it is parameter free. More importantly, stroke width value is used as feature for grouping the text components. Thus, helping to differentiate between group of characters since most likely a group of characters have similar stroke width value.

III. Experimental results

We evaluate the performance of the proposed method using ICDAR2005 [13] for comparison with the conventional SWT [3], [4] and MSER [5] methods, as well as ICDAR2011 [14] since it is widely used for scene text detection.

In order to find the optimal weighted parameters λ_1 and λ_2 in Eq. 1, we select 100 samples images and manually labelled every pixels of the text regions as ground truth, as shown in



(a) (b) (c)

Fig. 4. Comparison of the stroke width images of (a) low contrast from Fig. 3, (b) illumination from Fig. 1, and (c) blur from Fig. 1 between the proposed method (1st row) and the conventional SWT method (2nd row). The x and y axis represents pixel position and the colorbar indicates stroke width value. Note that the proposed method produces finer stroke width image while the conventional SWT produces many undesirable holes and broken components. (Best viewed in color)

Fig. 5. Pixels level ground truth for parameters evaluation.

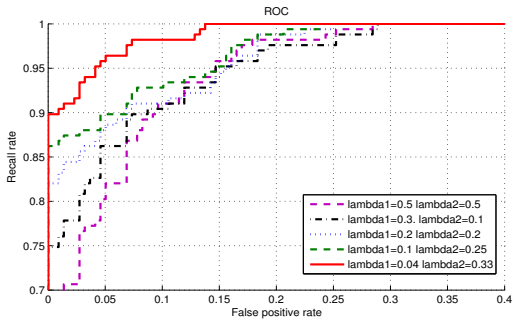


Fig. 6. The ROC curves for different λ_1 and λ_2.

Fig. 5. Then, we compute the scores using pixels intersection between the stroke width images (binary image is taken by setting $SW(p) = 1$ if $SW(p) > 0$, otherwise $SW(p) = 0$) from the proposed method and the pixels ground truths. This result is shown in Fig. 6 which suggests $\lambda_1 = 0.04$ and $\lambda_2 = 0.33$ as optimal weighted parameters.

Qualitative comparison of the stroke width images between the proposed method and the conventional SWT [3] is shown In Fig. 4. It is clearly showed that the proposed method produces finer stroke width image as compared to the conventional SWT. It is also worth noting that the conventional SWT produces many undesirable holes and broken components. This can cause partly connected components for a single component using the simple rule by grouping of neighbouring pixels if its stroke width ratio do no exceed the threshold. Moreover, the bottom image results of Fig. 4 are computed by manually setting Canny low threshold so that most of the characters can be detected, but this also can produce many noises.

To evaluate the overall performance of the proposed method, experiment on ICDAR2005 is conducted. This dataset contains 258 training images and 251 testing images. The images are in color and resolution vary from 307x93 to 1280x960 pixels including low contrast, illumination change, and blur images. For evaluating the score of ICDAR dataset, bounding box matching is used [15], that is the intersection between bounding box of the proposed method and the ground truth is taken. For K images, we compare set of bounding boxes ground truth G^k with set of detected bounding boxes D^k. Precision and recall then are formally defined as follows,

$$\begin{aligned} \text{Precision} &= \frac{\sum_k \sum_j m(G_j^k, D^k, t_r, t_p)}{\sum_k |G^k|} \\ \text{Recall} &= \frac{\sum_k \sum_j m(D_j^k, G^k, t_r, t_p)}{\sum_k |D^k|} \end{aligned} \tag{6}$$

where m is a function that gives a value 1 for correctly detected bounding box and 0 if does not match. This function also taking into account for one-to-many and many-to-one matches [15]. We employed threshold of precision $t_p = 0.4$ and threshold of recall $t_r = 0.8$ as suggested for standard ICDAR evaluation. The performance score then is formally defines as F-score=2.Precision.Recall/(Precision+Recall).

In this dataset, the performance of the proposed method is significantly achieved (precision = 76%, recall = 64% and

Methods	Precision (%)	Recall (%)	F-score (%)
Proposed method	**76**	64	**69**
Epshtein et al. [3]	73	60	66
Chen et al. [5]	73	60	66
Yao et al. [4]	69	**66**	67

TABLE I. ICDAR2005 RESULTS

Methods	Precision (%)	Recall (%)	F-score (%)
Proposed method	82	**69**	75
Neumann & Matas [16]	**85.4**	67.5	**75.4**
Neumann & Matas [6]	73.1	64.7	68.7
Shi et al. [17]	83.3	63.1	71.8
Kim method	83	62.5	71.3
Koo et al. [18]	79.1	62	69.5

TABLE II. ICDAR2011 RESULTS

F-score = 69%) as compared to the conventional SWT [3], [4] and edge-enhanced MSER [5]. The better performance of the proposed method could be attributed to the ability to detect blur, low contrast and illumination images in the dataset.

In Table II, the comparison of the proposed with the state-of-the-art methods on ICDAR2011 is conducted. The same evaluation is used as in Eq. 6. In this result, the recall of the proposed method outperforms the state-of-the-art methods since the proposed method able to cope with blur, low contrast and illumination images. However, the low precision could be attributed to the characters which have high stroke width variation. Thus, it is likely that the resulting connected component of a character is broken. The classifier then in turn could not correctly classify as text component. This will be investigated in our future works.

IV. CONCLUSION

In this paper, we have presented a novel method to construct finer stroke width image without using any edge detectors. It is known that using edge detectors could be prone of missing important edges (which are text information) because of low contrast, illumination, and blur images. To deal with this, Extremal Regions has been adopted and proposed a novel 3 properties in MRF to accurately integrate the regions into finer stroke width image. The experiments show encouraging results that it would be beneficial for the future works on scene text detection using stroke width properties.

ACKNOWLEDGMENT

This research is supported by the Fundamental Research Grant Scheme (FRGS) MoE Grant FP027-2013A, H-00000-60010-E13110 from the Ministry of Education Malaysia.

REFERENCES

[1] K. Jung, K. In Kim, and A. K Jain, "Text information extraction in images and video: a survey," *Pattern recognition*, vol. 37, no. 5, pp. 977–997, 2004.

[2] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: a survey," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 2-3, pp. 84–104, 2005.

[3] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR*. IEEE, 2010, pp. 2963–2970.

[4] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *CVPR*. IEEE, 2012, pp. 1083–1090.

Fig. 7. Sample of final results of the proposed method with the generated bounding boxes of text components.

[5] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 2609–2612.

[6] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *CVPR*. IEEE, 2012, pp. 3538–3545.

[7] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1124–1137, 2004.

[8] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 2, pp. 147–159, 2004.

[9] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional max-margin markov networks," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 975–982.

[10] 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.

[11] Y.-F. Pan, X. Hou, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *Image Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 800–813, 2011.

[12] L. Gomez and D. Karatzas, "Multi-script text extraction from natural scenes," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 467–471.

[13] S. M. Lucas, "Icdar 2005 text locating competition results," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005, pp. 80–84.

[14] A. Shahab, F. Shafait, and A. Dengel, "Icdar 2011 robust reading competition challenge 2: Reading text in scene images," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1491–1496.

[15] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 8, no. 4, pp. 280–296, 2006.

[16] L. Neumann and J. Matas, "On combining multiple segmentations in scene text recognition," *ICDAR 2013*, 2013.

[17] 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*, 2012.

[18] H. Koo, D. Kim *et al.*, "Scene text detection via connected component clustering and non-text filtering." *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 2013.