Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 322–330

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

Multi-Oriented Text Detection in Natural Scene Images Based on the Intersection of MSER With the Locally Binarized Image

Anurag Agrahari[a], Rajib Ghosh[b]

[a]National Institute of Technology Patna, Ashok Rajpath, Patna-800005, India
[b]National Institute of Technology Patna, Ashok Rajpath, Patna-800005, India

Abstract

The problem of text extraction is an interesting area of research in computer vision domain. In the recent years, emergence of various applications on smart hand-held devices such as translation of text from one language to another in real time, computerized aid for visually impaired, user navigation & traffic monitoring and driving assistance systems, has stimulated the renewed research interest in this domain. Retrieving text directly from natural scene images or videos is a challenging task due to variant patterns and orientations of scene text. Although various research investigations are available on horizontal oriented text detection in natural scene images, a little number of studies exist on text detection of multiple orientations. In this article, a robust method has been proposed for scene text detection having multiple orientations. The text in the images are horizontally, non-horizontally and curve oriented. The proposed method contains two main stages—maximally stable extremal region (MSER) computation and stroke width transformation (SWT). MSERs have been used to detect maximally stable extremal regions in the images. It is followed by implementing Canny edge detector for enhancement of edges in the images. To remove the non-text regions in the images, the combination of SWT image and geometric information has been used. The performance of the proposed method has been assessed on the IITR text datasets and it produces very encouraging results.

Keywords: Multi-oriented text detection; Natural scene image; MSER; Stroke width transform;

1. Introduction

Investigations to develop systems for detection and recognition of text in natural scene images have gained momentum in recent past due to the vast applications of these systems in this digital era. The main goal of these investigations

* Rajib Ghosh. Tel.: +91-8084023813.
E-mail address: rajib.ghosh@nitp.ac.in

10.1016/j.procs.2020.04.033

Fig. 1. Few instances of horizontal, non-horizontal and curve oriented text in scene images from IITR text datasets.

is to develop a system for automatic recognition of the text present in any natural scene image. The need to develop these types of systems has highly increased with the advent of smartphones. We find various roadside displays in different countries containing several useful textual information written in known/unknown script. If the person is unfamiliar with the script, it becomes impossible for him/her to understand those useful information. If a system can be developed which will automatically detect and recognize the text present in natural scene images then one can just grab the image of the roadside displays using his/her smartphone camera and can fed it to this system to understand the text after translating the recognized text using some standard online translator. Apart from this utility, the useful information obtained from scene image text is used in various content based image and video applications like image text retrieval, video text retrieval, and mobile text analysis. Complex and variable-colored backgrounds, various font sizes, and text orientations prompt the detection of text from scene images prior recognizing it.

Several investigations [3][5][6][7][24] have been reported till date on scene text detection from images and videos. However, a very little number [34][35] of investigations exist on detection of scene text of multiple orientations from images as well as videos. The present article proposes a new method to detect text of multiple orientations from natural scene images. The text in the images are horizontally, non-horizontally and curve oriented. Few instances of each type of orientation from IITR text datasets are shown in Fig. 1.

Connected component (CC) based text detection technique has been explored in the present investigation, which is again dependent on MSERs [1] as basic character candidates. In spite of some favorable properties, MSERs are sensitive to image blur. In order to extract the small characters from the images of low resolution, the canny edges and MSERs have been combined in the proposed method after performing the intersection between MSERs and locally binarized image created during preprocessing stage. The geometric and stroke width transformation (SWT) have then been applied for filtering and pairing of CCs. Ultimately, a second intersection is performed with the locally binarized image created during the preprocessing stage to prevent the false detection of text. The novelty of the proposed method lies in applying the intersection operation between the MSER image with the locally binarized image created during preprocessing stage and the generation of the SWT image of MSERs using the distance transformation technique. The comprehensive block diagram of the proposed investigation is presented in Fig. 2.

The remaining sections are organized as follows: Section 2 presents the literature survey in this problem area. Section 3 discusses the proposed method of scene text detection. Experimental results and analysis are presented in Section 4. Finally, the conclusion and future work of this article is discussed in Section 5.

2. Literature survey

Several research investigations have already been reported on scene text detection and recognition. Matas et al. [1] presented one algorithm to detect an affinely invariant stable subset of extremal regions, MSER. Anthimopoulos et al. [2] reported one investigation on detection of scene text in both images and video frames. This investigation has used one machine learning technique, Random Forest classifier to which a set of feature vectors generated using the local binary pattern were fed. In another study [3], convolutional neural network (CNN) based end-to-end, lexicon driven method has been proposed. Chen et al. [4] presented a MSER based scene text detection algorithm where MSERs have

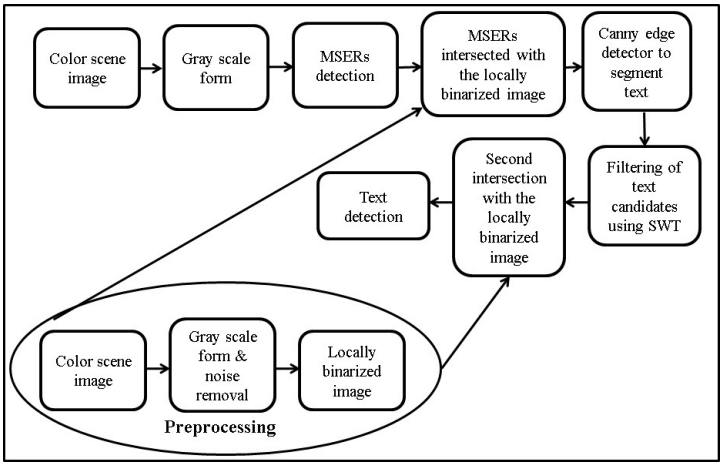

Fig. 2. Comprehensive block diagram of the proposed method.

been employed as preliminary letter candidates and the candidates have been filtered using geometric information to remove the non-text regions. Structure based partitioning and grouping of text candidates to detect text strings have been reported in another investigation [5].

Yin et al. [6] presented a pruning algorithm to determine MSERs in order to detect character candidates for text extraction. Neumann et al. [7] reported one research finding where text have been located in natural scene images using pruned exhaustive search. A novel selector of MSER exploiting region topology has been included in the method. Raj et al. [8] reported one investigation to extract scene text by analyzing CCs. In this investigation, mathematical morphological operations have been used to extract the headlines of Devanagari text. In another study [10], an investigation was reported to recognize the scene text in order to provide aid to the visually impaired persons.

Bhattacharya et al. [11] reported a scheme to extract text from scene images based on analysis of CCs. Another CC based investigation [12] has been reported to detect scene text where the scene color images have been separated into homogeneous layers based on the similarity of colors. Jain et al. [14] reported one investigation on locating text in scene images with complex background. The investigation proposed one method applicable to numerous applications in real-life such as transformation of offline document into electronic version, internet searching, color image indexing, etc.

Hanif et al. [15] presented one text detection method using a cascade of boosted ensemble. In this study, different feature values were extracted from different regions of the scene images using a small set of feature values and these features were studied using linear discriminant function to detect the text regions. In another study [16], AdaBoost classifier was adopted for fusion of various features in order to extract the scene text. Mishra et al. [18] presented two different frameworks for scene text recognition. One of those, known as the bottom-up method was proposed based on the piece-wise character detections in the image. The detected text was then recognized using conditional random field (CRF) model based classifier. Gllavata et al. [19] reported one investigation for detection of text in video frames based on the unsupervised learning using k-means algorithm. This study has divided the entire image into three different clusters—text, simple and complex background. Epshtein et al. [21] presented one method using canny edge detection and SWT to detect text in scene images. In another study [22], CRF model based classifier has been applied to classify connected components as ”text” or ”non-text”.

Maruyama et al. [30] reported one investigation to detect signboard characters in scene images. In this investigation, Harr wavelet, histogram of oriented gradient (HOG), and moment statistics were used separately and the features were studied using support vector machine (SVM) classifier. This problem has also been addressed using Gabor filters to recognize scene characters [31]. A multi-level MSER technique has also been proposed to detect the text candidates based on the measurement of text probability [32]. Jaderberg et al. [33] presented an end-to-end system for extracting and recognizing text in scene images. Region growing mechanism was used in this system for text detection, whereas recognition task was accomplished using deep convolutional neural network (DCNN). Yin et al. [34] proposed a method for multi-orientated scene text detection, where text candidates have been constructed through clustering

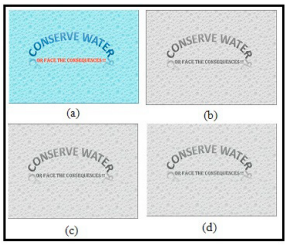

Fig. 3. (a) Original color image, (b) Gray scale form, (c) Image after noise removal, and (d) Contrast enhanced image.

of characters based on the adaptive hierarchical clustering technique. This clustering technique has used an unified distance metric learning framework to learn the similarity weights and clustering threshold simultaneously. Another investigation [35] has proposed a scheme for multi-oriented scene text detection where the text detection method has relied upon Fourier-Laplacian filtering technique.

The survey shows that the most of the investigations have focused to develop horizontal text detection schemes, whereas the present investigation proposes a method to detect multi-oriented text in scene images.

3. Proposed method

Before applying the proposed method for text detection, a set of basic preprocessing steps have been applied on one copy of each raw scene image to ease the implementation of the subsequent stages of the proposed method on the another copy of the image.

3.1. Preprocessing

During preprocessing, initially, adaptive contrast enhancement has been applied on the input image for image enhancement after converting the input color image into its gray scale form and removal of noise. Image binarization technique has then been applied on this enhanced form of the image in order to separate the background and foreground portions of the image. Fig. 3 shows the resultant images after color to gray scale conversion, noise removal from the gray scale form and applying contrast enhancement technique on the filtered image.

The global binarization methods like the one using Otsu's technique [36] are not usually suitable for camera captured images since the gray level histograms of such images are not bimodal in nature. Due to the binarization using a single threshold value the text information are often lost in such images. In order to avoid this problem, local binarization method has been used in the present investigation which is usually a window based method and the selection of window size hugely affects the outcomes this binarization method. In local binarization method, adaptive thresholding technique has been used in which the simple average gray level value around a pixel within a 30x30 sized window has been considered as the threshold for that pixel.

3.2. Text detection

In the present investigation, the proposed algorithm for text detection contains the following steps:

1. Computation of regions of similar intensities in the gray scale form of the other copy of the original color image using the MSER region detector.
2. Conversion of the MSER regions to its equivalent binary mask.
3. Intersection of locally binarized image created during preprocessing stage with the binary mask.
4. Application of Canny edge detector for further segmentation of the text.
5. Filtering of text candidates using CC analysis.
6. Filtering of text candidates using the SWT of MSER regions.

7. Resultant image obtained from step 6 is again intersected with the locally binarized image created during pre-processing to extract the desired text string from the image.

Fig. 4 shows the outcomes after applying various steps of the proposed algorithm. The said steps are detailed below.

Fig. 4. The outcomes after applying various steps of the proposed algorithm on a scene image containing horizontal and curve oriented text—(a) Original color image, (b) Gray scale form, (c) Extraction of MSERs (d) After first intersection between MSERs and locally binarized image, (e) Applying Canny edge detector, and (f) Text detected in various boxes after applying SWT of MSERs and applying second intersection.

3.2.1. MSER

MSER posses several favourable properties like affine transformation of image intensities, stability, etc., to locate the text regions in any scene image. However, it cannot show good performance in the case of image blur. MSER is known for its popularity in object detection in images [37]. The MSER process finds a number of variable regions from images known as MSER regions. The MSER is a connected component which is a collection of gray scale pixels of similar intensity values. Since the text portions in any image usually share the similar gray scale intensity levels and its intensity is quite often different from that of the background, so MSER can locate the text regions quite efficiently. The ”extremal” word in MSER signifies that the intensity levels inside the MSER region differ highly with respect to the outer regions. If intensity levels vary immensely among different regions then better MSERs are obtained. The repeatability of MSER is high; filter invariance and implementation are faster. So, MSER is a faster region detector.

3.2.2. SWT

Stroke width is nothing but the length of the connecting line from one pixel on the edge of the text region to another on the same edge. This length is calculated along the gradient direction. The basic reason of stroke width extraction is we get almost the same stroke width in a single character appearing in different positions of the text, however, this width varies significantly in non-text regions due to its irregular nature. Initially, skeletons of MSERs is considered to extract the stroke width. Next, the distance transform is applied on each foreground pixel of the skeleton by computing the distance between that pixel to the closest boundary of the respective MSER. It is computed for each skeleton of MSERs. Thus, we get a skeleton-distance map. Variance on the skeleton-distance of each CC is computed to measure the difference between text and non-text regions. It may be noted that the text characters usually posses smaller variances as compared with the non-text regions. As a consequence, CCs with large variance have been removed. The threshold of the variance value has been decided empirically. In this manner, few non-text regions have been eliminated.

4. Experimental results and analysis

The performance of the present investigation has been assessed using the IITR text datasets [38] of natural scene images where samples with multiple oriented (horizontal, non-horizontal and curve) text are available. The dataset statistics are presented in Table 1.

Table 1. Dataset statistics.

Text orientation	Samples
Horizontal	100
Non-horizontal	84
Curve	50

4.1. Text detection results

The proposed method of text detection has shown satisfactory results for scene images having both uniform and non-uniform backgrounds. Fig. 5 shows correct detection of text having multiple orientations in two scene images from IITR text datasets. Detected text have been enclosed within various boxes in this figure.

Fig. 5. Detected text within various boxes in two scene images from IITR text datasets—(a) Horizontal and curve oriented text, and (b) Non-horizontal oriented text.

Text detection accuracy of the proposed system has been calculated manually where if all of the text portions present in any image sample has been confined within the boxes properly then it has been considered as an accurate detection of the text. The accuracy of the proposed system is presented in Table 2. The precision (P), recall (R) and F1-score (F) of the proposed system are presented in Fig. 6, Fig. 7 and Fig. 8 respectively.

Table 2. Accuracy of the proposed text detection system.

Text orientation	Accuracy
Horizontal	92%
Non-horizontal	89.28%
Curve	88%

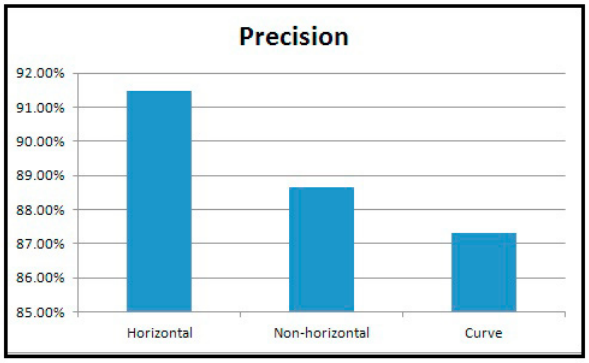

Fig. 6. Precision of the proposed scene text detection system.

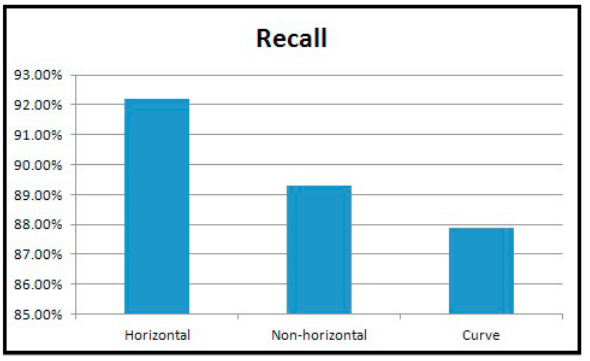

Fig. 7. Recall of the proposed scene text detection system.

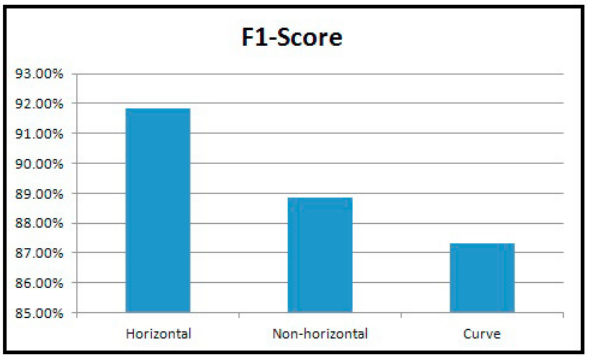

Fig. 8. F1-Score of the proposed scene text detection system.

4.2. Error analysis

It can be seen from Table 2 that 100% correct text detection accuracy could not be obtained from the proposed system. It has been found by analysis that erroneous outcomes have been provided either due to the complex background of the images or similarity in the intensity levels between the foreground text regions and the background non-text regions. Error has also occurred in few samples due to the similarity in the stroke width of text and non-text portions. Few erroneous outcomes are shown in Fig. 9.

Fig. 9. Few erroneous outcomes.

4.3. Comparative performance analysis

As the performance of the proposed method has been evaluated using publicly available IITR text dataset, so the performance of the present investigation has been compared with one existing study [35] which has only used the same dataset to evaluate its text detection performance. The performance comparison is presented in Table 3.

Table 3. Comparative performance analysis with existing studies.

Orientation	[35]	Proposed method
Horizontal	P=0.88, R=0.81, F=0.85	P=0.91, R=0.92, F=0.91
Non-horizontal	P=0.85, R=0.79, F=0.82	P=0.88, R=0.89, F=0.91
Curve	P=0.81, R=0.74, F=0.77	P=0.87, R=0.87, F=0.87

5. Conclusion and the future work

Detecting text directly from natural scene images is a challenging task because of variability in text patterns, various orientations of text, variant background inferences and similarity in intensity values between text and non-text regions. In this investigation, the proposed method is simple and novel to detect scene text having multiple orientations using MSER and SWT. The present work can be extended in future to determine the movement of the detection as well as the recognition of the detected scene text in Latin script. During recognition, features will be extracted from text of various shapes. Further investigations can also be carried out on detection and recognition of scene text having multiple orientations in various Indic scripts as well.

References

[1] J. Matas, O. Chum, M. Urban, and T. Pajdla. (2004) "Robust wide-baseline stereo from maximally stable extremal regions." *Image and vision computing* **22** (10): 761-–767.

[2] M. Anthimopoulos, B. Gatos, and I. Pratikakis. (2013) "Detection of artificial and scene text in images and video frames." *Pattern Analysis and Applications* **16** (3): 431-–446.

[3] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. (2012) "End-to-end text recognition with convolutional neural networks." *In Proc. of Intl. Conf. on Pattern Recognition*, Tsukuba Science City, Japan, 3304-–3308.

[4] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod. (2011) "Robust text detection in natural images with edge-enhanced maximally stable extremal regions." *In Proc. of 18th Intl. Conf. on Image Processing*, Brussels, Belgium, 2609-–2612.

[5] C. Yi, and Y. Tian. (2011) "Text string detection from natural scenes by structure-based partition and grouping." *IEEE Transactions on Image Processing* **20** (9): 2594-–2605.

[6] X.C. Yin, X. Yin, K. Huang, and H.W. Hao. (2014) "Robust text detection in natural scene images." *IEEE transactions on pattern analysis and machine intelligence* **36** (5): 970-–983.

[7] L. Neumann, and J. Matas. (2011) "Text localization in real-world images using efficiently pruned exhaustive search." *In Proc. of Intl. Conf. on Document Analysis and Recognition*, Beijing, China, 687–691.

[8] H. Raj, and R. Ghosh. (2014) "Devanagari Text Extraction from Natural Scene images." *In Proc. of Intl. Conf. on Advances in Computing, Communications and Informatics*, New Delhi, India, 513–517.

[9] L. Gomez, and D. Karatzas. (2013) "Multi-script text extraction from natural scenes." *In Proc. of Intl. Conf. on Document Analysis and Recognition*, Washington DC, USA, 467-–471.

[10] N. Ezaki, M. Bulacu, and L. Schomaker. (2004) "Text detection from natural scene images: towards a system for visually impaired persons." *In Proc. of 17th Intl. Conf. on Pattern Recognition*, Cambridge, UK, 683-–686.

[11] U. Bhattacharya, S. K. Parui, and S. Mondal. (2009) "Devanagari and bangla text extraction from natural scene images." *In Proc. of 10th Intl. Conf. on Document Analysis and Recognition*, Barcelona, Spain, 171-–175.

[12] K. Wang, and J. A. Kangas. (2003) "Character location in scene images from digital camera." *Pattern Recognition* **36** (10): 2287–2299.

[13] J. J. Weinman, Z. Butler, D. Knoll, and J. Feild. (2014) "Toward integrated scene text reading." *IEEE transactions on pattern analysis and machine intelligence* **36** (2): 375-–387.

[14] A. K. Jain, and B. Yu. (1998) "Automatic text location in images and video frames." *Pattern Recognition* **31** (12): 2055-–2076.

[15] S. M. Hanif, L. Prevost, and P. A. Negri. (2008) "A cascade detector for text detection in natural scene images." *In Proc. of 19th Intl. Conf. on Pattern Recognition*, Tampa, USA, 1-–4.

[16] S. M. Hanif, and L. Prevost. (2009) "Text detection and localization in complex scene images using constrained adaboost algorithm." *In Proc. of 10th International Conference on Document Analysis and Recognition*, Barcelona, Spain, 1-–5.

[17] M. S. Brown, and W. B. Seales. (2004) "Image restoration of arbitrarily warped documents." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26** (10): 1295–1306.

[18] A. Mishra, K. Alahari, and C. Jawahar. (2012) "Top-down and bottom-up cues for scene text recognition." *In Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, Providence, USA, 2687-–2694.

[19] J. Gllavata, R. Ewerth, and B. Freisleben. (2004) "Text detection in images based on unsupervised classification of high-frequency wavelet coefficients.", *In Proc. of 17th Intl. Conf. on Pattern Recognition*, Cambridge, UK, 425–428.

[20] Y. F. Pan, Y. Zhu, J. Sun, and S. Naoi. (2011) "Improving scene text detection by scale adaptive segmentation and weighted CRF verification." *In Proc. of 11th Intl. Conf. on Document Analysis and Recognition*, Beijing, China, 759–763.

[21] B. Epshtein, E. Ofek, and Y. Wexler. (2010) "Detecting text in natural scenes with stroke width transform." *In Proc. of Intl. Conf. on Computer Vision and Pattern Recognition*, San Francisco, USA, 2963–2970.

[22] H. Zhang, C. Liu, C. Yang, X. Ding, and K. Q. Wang. (2011) "An improved scene text extraction method using conditional random field and optical character recognition." *In Proc. of 11th Intl. Conf. on Document Analysis and Recognition*, Beijing, China, 708–712.

[23] Y. Zhong, K. Karu and A.K. Jain. (1995) "Locating Text in Complex Color Images." *In Proc. of 3rd Intl. Conf. on Document Analysis and Recognition*, Montreal, Canada, 146–151.

[24] C. Liu, C. Wang and R. Dai. (2005) "Text Detection in Images Based on Unsupervised Classification of Edge-based Features." *In Proc. of 8th Intl. Conf. on Document Analysis and Recognition*, Seoul, South Korea, 610–614.

[25] M. Cai, J. Song and M.R. Lyu. (2002) "A New Approach for Video Text Detection." *In Proc. of Intl. Conf. on Image Processing*, Rochester, USA, 117–120.

[26] E.K. Wong and M. Chen. (2003) "A new robust algorithm for video text extraction." *Pattern Recognition* **36**: 1397–1406.

[27] C.W. Lee, K. Jung and H.J. Kim. (2003) "Automatic text detection and removal in video sequences." *Pattern Recognition Letters* **24** (15): 2607–2623.

[28] Z. Liu, and S. Sarkar. (2008) "Robust outdoor text detection using text intensity and shape features." *In Proc. of 19th Intl. Conf. on Pattern Recognition*, Tampa, USA, 1–4.

[29] S. Lu, and C. L. Tan. (2006) "Camera text recognition based on perspective invariants." *In Proc. of 18th Intl. Conf. on Pattern Recognition*, Hong Kong, China, 1042–1045.

[30] M. Maruyama, and T. Yamaguchi. (2009) "Extraction of characters on signboards in natural scene images by stump classifiers." *In Proc. of 10th International Conference on Document Analysis and Recognition*, Barcelona, Spain, 1365–1369.

[31] J. Weinman, E. Learned-Miller, and A. Hanson. (2009) "Scene text recognition using similarity and a lexicon with sparse belief propagation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**: 1733–1746.

[32] S. Tian, S. Lu, B. Su, and C.L. Tan. (2014) "Scene text segmentation with multi-level maximally stable extremal regions." *In Proc. of 22nd Intl. Conf. on Pattern Recognition*, Stockholm, Sweden, 2703–2708.

[33] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. (2016) "Reading text in the wild with convolutional neural networks." *International Journal of Computer Vision* **116** (1): 1–20.

[34] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao. (2015) "Multi-orientation scene text detection with adaptive clustering." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** (9): 1930–1937.

[35] A. Sain, A. K. Bhunia, P. P. Roy, U. Pal. (2018) "Multi-oriented text detection and verification in video frames and scene images." *Neurocomputing* **275**: 1531–1549.

[36] N. Otsu. (1979) "A threshold selection method from gray level histograms." *IEEE Transactions on Systems, Man and Cybernetics* **9**: 62-–66.

[37] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. (2005) "A comparison of affine region detectors." *International Journal of Computer Vision* **65**: 43–72.

[38] https://sites.google.com/site/2partharoy/dataset-1.