

New Algorithm for Artificial Text Detection in News Videos

Matheel E. Abdulmunem¹, Eman Hato²

**¹Department of Computer Science, University of Technology-Iraq,
Baghdad, Iraq**

**²Department of Computer Science, Mustansiriyah University,
Baghdad, Iraq**

Abstract

Artificial Text in videos plays an important role for automatically indexing video content, since much semantic information is carried with them. In this paper an effective algorithm is presented to detect artificial text region in the frames of news videos by utilizing the geometric properties of MSER regions.

The video is segmented into a sequence of frames and then the frames that contain a caption at most are selected to be the input to proposed algorithm. Firstly the Maximally Stable Extremal Region (MSER) is applied for each frame to locate a large number of individual text characters as MSERs regions. Then, candidate's text regions are filtered to remove non-text regions by using geometric properties with thresholds technique. To accurately detect text in the video frame, three types of geometric properties are utilized: centered of regions, mean intensity and bounding box. Finally, text regions are merged, highlighted and recognized using Optical Character Recognition (OCR) system directly.

Experimental results indicate that proposed algorithm provides 0.9445 recall and 0.8095 precision accuracy rate for selected video frames. The performance of the proposed algorithm is satisfactory for artificial text detection under different size, color and background.

Keyword: Artificial text, Geometric properties, MSER, Scene text, OCR.

1. Introduction

Digital videos and images are one of the most important media in our daily lives. With the advancement of multimedia technologies and modern telecommunication, a huge amount of these multimedia resources are traveled over internet. It is very important to develop efficient and effective systems to retrieve and manage videos and images using their contents [1].

Among all the contents, text contains a high level of semantic information; it can be easily understood by both computer and human. This makes text detection and extraction useful to improve the retrieval system performance in real time applications. For example, text of web images can reflect the web pages topic. Subtitle headlines of news videos usually highlight information about the happened events. Text of sports videos give information about scores results [2]. The aim of text detection is to find regions of image regions that contain only text that can be recognized by the user directly or entered into the OCR system to recognize it [3].

The video frames consist of two categories of text: artificial text and scene text. Artificial texts (also called graphics or caption) are artificially superimposed to frames during editing process such as the subtitle of the news program. While scene text is part of the camera-captured scene such as billboards, road signs and writing on shirts. Scene text is unpredictable and less related to video content than artificial text, which is related to the subject of a video and more structured, making it extremely useful for building concept indexes [4,5]. An example of text categories occurrence in video frame is shown in Figure (1).



Figure (1): Categories of text occurrence of video frame.

Although text provides important information about video content, it is still difficult to detect and extract text embedded in images or video frames. The blurry and illumination changes of the frames, various fonts, color, size and alignment of text in the single frame and from frame to frame and the background change or movement behind the text are the common difficulties to accurately detect text [6,7].

The aim of this paper is to provide a new algorithm for artificial text detection of news videos with low computation cost and preserving the efficiency and accuracy of detection rate by utilizing the information of geometric properties of MSER regions.

2. Text Detection Methods

The methods of text detection can be classified into: connected component based methods and sliding window based methods (also known as region based methods). In connected component based methods, the text and non-text pixels are separated by applying a fast low-level filter followed by grouping the text pixels that have similar properties such as intensity, stroke width or color to create component candidates into text. The leverages of these methods are computational efficiency for detecting text regions however they have difficulty when text find in complex or low contrast background. Therefore, they usually perform well for artificial text that has regular spacing and uniform color [8].

The sliding window based methods detect text regions by moving a sub window in multiple scales through all image locations to search for possible texts regions and then use learned classifier to distinguish text and non-text regions. The classifier can use different features such as edges, gradients, texture, color and other region features. Unlike connected component, the sliding window based methods are more robust and efficient for the complex background, but the computational cost is high because there are a large number of training samples for the classification process [9].

For example, in [10] the text localization method is presented using the connected components. Firstly linear spatial filter is designed for image binarization and then a local classifier consisting of a cascade of three multilayer perceptron is employed to remove non-text components. MSER and Stroke Width Transform (SWT) are applied for scene text detection as low level filters in connected components based method with great success. For example authors in [11] are used MSER region to detect regions containing text and SWT with Canny operator as edge detection to filter and remove non-text regions.

Authors in [12] used the SWT algorithm to create a stroke width map that is clustered by the K-Means algorithm. After that several superpixel images are created depend on the stroke width map and connected components are extracted. The connected components are filtered by color, stroke width and geometrical features to construct text line. In [13] authors present method for text detection in natural images using sliding window methods. A discriminative Convolutional Neural Network (CNN) is trained with multi-level information as character candidate classifier. The image region is classified as non-text, weak text and strong text using double threshold filtering. The final filtering process of non-text regions is done through a recursive neighborhood search algorithm. The heuristic features (stroke width, spatial location, size and color) are used to group the characters into text lines.

The proposed method in [14] used Scale Invariant Feature Transform (SIFT) descriptors for text detection in the scene images. SIFT features is classified into text and non-text regions using neural network classifier. In [15] the pruning algorithm is designed to extract MSER character candidates using a strategy of minimizing regularized variations. These character candidates are clustered using the single link algorithm. The probabilities of text candidates corresponding to non-text are estimated to eliminate non-text and identify text with a text classifier.

In order to improve performance of detection, connected component and sliding window based methods are integrated. For example in [16] MSERs is detected in the original image as basic connected component. MSERs-based graph model is created and then labeled as text or non-text using graph cut algorithm. Lastly, heuristic rules are used to group the text lines. The proposed method presented in [17] employs corner features to detect the artificial text from videos. The algorithm also detects the moving text in videos by combining text features with motion features that extract by optical flow technique. To learn the classification criteria, the decision tree is adopted.

The proposed method in [18] used the gradient and smooth edge information to generate edge images, and the SWT edge in the analysis of connected component to remove intra and inter character errors. The Deep Belief Network (DBN) is then used to distinguish between character and non-character connected component.

3. Maximally Stable Extremal Region (MSER)

MSER is a local feature detector that extracts a number of co-variant regions from a gray scale image called MSERs. The term Extreme Region (ER) refer to

sets of connected pixels have intensity values higher (bright extremely regions) or lower (dark extremely regions) than their outer boundaries. An extremism region is maximally stable when it shows invariance to the transformation of image intensities over a large range of intensity threshold levels. MSER is very efficient for multi-scale detection, low computational complexity and it widely used for finding the correspondence between the two images [19].

4. Region Geometric Properties

A region in an image can have geometric properties, such as pixel intensity, orientation, area, center of mass and bounding box. The property values are used directly as features descriptors that are employed to remove non-text region candidates. The non-text regions are removed using thresholds techniques, whereby thresholds can be calculated or learned using machine learning techniques. These features are a fast and easy way to discern non-text regions from text features in images [20].

5. Proposed Text Detection Algorithm

In this paper, a text detection algorithm is proposed to detect regions of artificial text in frames of videos. Figure (2) illustrate the general structure of proposed algorithm. The operations sequence that describes the proposed algorithm is illustrated as follows:

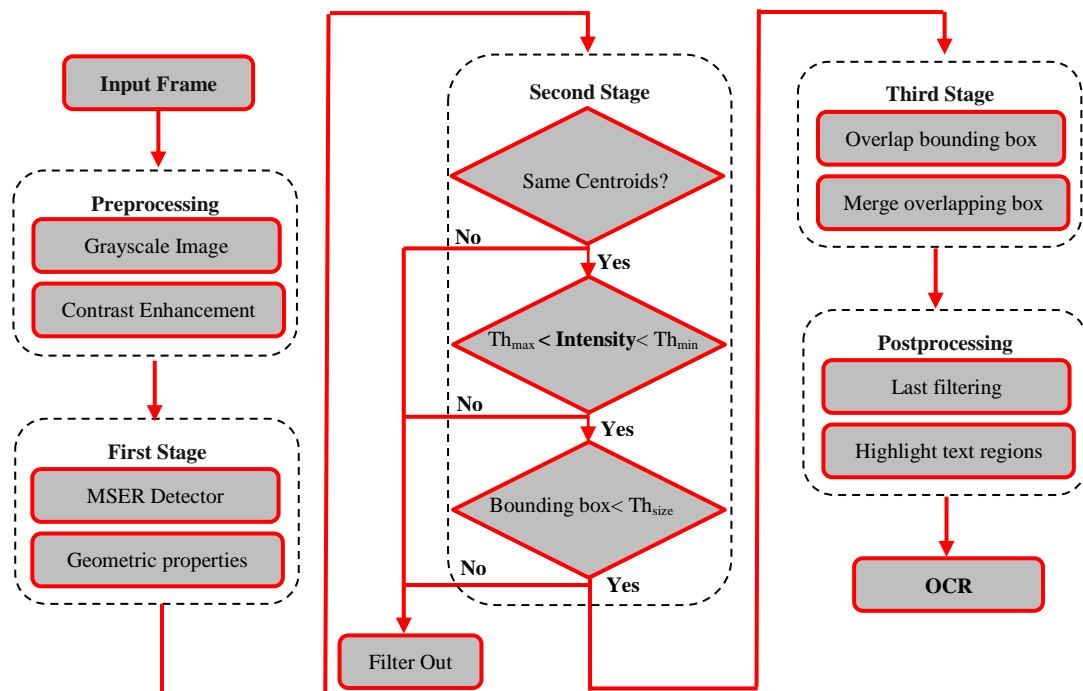


Figure (2): Flowchart of proposed algorithm.

Preprocessing

There is a necessity to select text frames before text detection to reduce computations and erroneous. So that a video is broken into a successive frames and then selected the frames that contain the artificial text at most, making the detection process more efficient and accurate. The input frame is converted into grayscale image and the local contrast is then enhanced to bring out details and increase the appearance of large-scale light-dark transitions of image leading to stable intensity profiles.

First Stage

MSER detector is computed to frame in order to obtain the MSER regions as connected components. These regions are used as preliminary detection of candidate text regions which are filter out again to remove non-text components using three types of geometric properties: centered of regions, mean intensity and bounding box. An example of applying the MSER detector for a selected frame is shown in Figure(3).

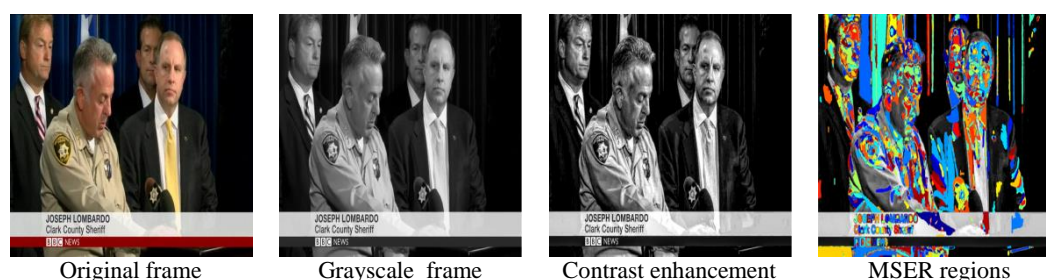


Figure (3): MSER detector for a selected frame.

Second Stage

Although MSER detects most regions of text, many non-text regions still exist with the candidate text region, therefore, geometric properties with thresholds technique are employed to filter out non-text regions.

It can be noted that artificial text characters have approximately the same size and spacing between them, as well as the text line is always aligned horizontally and in contrast to the background. The proposed algorithm exploits these properties by extracting the following geometric properties from MSER regions:

- 1. Centered of regions:** the coordinates of the center of region.
- 2. Mean intensity of regions:** the average value of the intensity values in the region.
- 3. Bounding box of regions:** the set of coordinates of the rectangular border that contain the fully region.

In this stage, the algorithm searches for sequence regions that have same x-coordinate from their center values to diagnose horizontal alignment regions.

These regions will remain and others regions discard. Then the mean intensity value of regions should be greater or less than specific thresholds (TH_{min} or TH_{max}) otherwise the regions is rejected. This inherently assumes that all text regions appear in the video frames in contrast to the background that mean, it will be close to the black or white color. The last filter is performed by considering bounding box size of the regions. Since the size of artificial text character is approximately equal, the regions that have bounding box size greater than specific thresholds (TH_{size}) are filtered out. It is worth to mention that the values of the thresholds (TH_{min} , TH_{max} , TH_{size}) are experimentally selected. An example of second stage processing is shown in Figure (4).

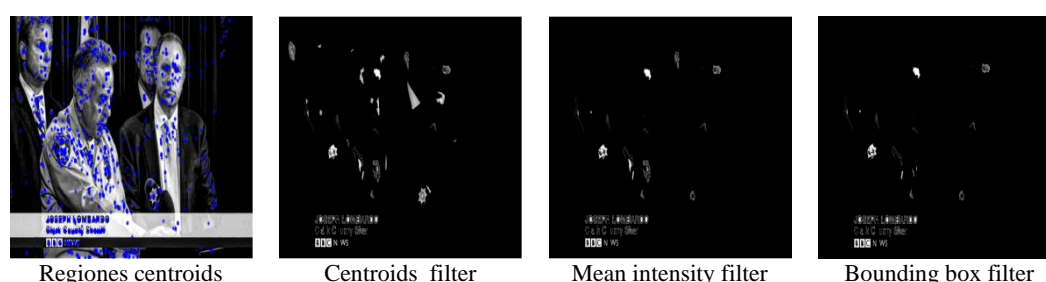


Figure (4): Non-text region filtering.

Third Stage

The individual text characters are integrated into words or text lines at this stage in order to enables the OCR to recognize more meaningful information than just individual characters.

To merge single text regions into text lines, the bounding boxes are expanded by overlap ratios. This lead to overlap the bounding boxes of neighboring text regions where the text regions that are part of the same word or text line form a series of nested bounding boxes. The overlapping bounding boxes are merged into a single bounding box around text lines.

Postprocessing

To minimize false detection as much as possible, the regions that have bounding box size smaller than specific thresholds (TH_{size}) are discarded. This is as opposite to the last step in second stage, because here the bounding box content multiple text regions, in other words, the regions that not overlap with any other regions are removed (non-text regions). Figure (5) shows an example of applying the third stage and postprocessing process for a selected frame.

The OCR system is used to recognize the text within each bounding box after the text regions are detected.

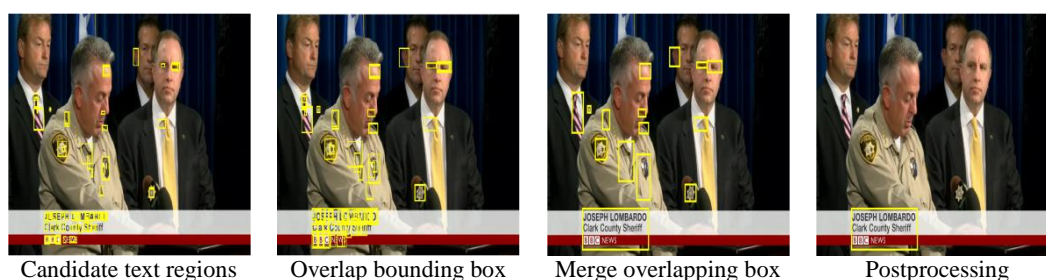


Figure (5): Lines text detection.

6. Experiments Results

The performance of proposed algorithm is evaluated using selected video frames obtained from the BBC news videos as test material. The optimal thresholds are obtained for text detection experimentally by changing their values until the best performance is reached.

Table (1) present the experiments results for each frame, where the text regions are highlighted and the text obtained from OCR is detailed. It is clear from the Table (1) the high performance of proposed algorithm; it is able to correctly detect text regions in the selected frames. One can notice that proposed algorithm handles different size text and different colors very well without false detection. Thus the proposed algorithm has some robustness for size and color changes of text.















Although the MSER is used in proposed algorithm to handle the all most text regions, some of text regions are still undetected due to low frame contrast and color instability as shown in table (2) for frame No. (1) and frame No.(2). Furthermore some of non-text regions are detected due to the illumination condition as shown in table (2) for frame No. (2), (3), (4).

To judge the correctness of the text regions detected in this work, the Recall (R) and Precision (P) measures used in this work to as defined in the following equation:

$$\text{Recall(R)} = \frac{\text{True}}{\text{True+Miss}} \quad (1)$$

$$\text{Precision(P)} = \frac{\text{True}}{\text{True+False}} \quad (2)$$

Table (1): The result of proposed text detection algorithm.

No.	Original Frame	Original Text	Detected Text	Recognized Text
1		JOSEPH LOMBARDO Clark County Sheriff BBC NEWS		JOSEPH LOMBARDO Clark County Sheriff EIEI NEWS
2		BREXIIIT NEXT STAGE AGREED Britain due to leave the EU by end of March 2019 BBC NEWS		BREXIIIT NEXT STAGE AGREED Britain due to leave the EU by end of March 2019 EEI NEWS
3		BREXI NEXT STAGE AGREED EU gives go-ahead for next phase of talks BBC NEWS		BREXI NEXT STAGE AGREED EU gives go-ahead for next phase of talks
4		NEW PAKISTAN PM Parliament due officially to elect new PM BBC NEWS		NEW PAKISTAN PM Parliament due officially to elect new PM Em NEWS
5		RUSSIA COURTROOM SHOOTING Shooting began after defendants tried to escape BBC NEWS		RUSSIA COURTROOM SHOOTING Shooting began after defendants tried to escape HE NEWS
6		MANBIJ FFIEED FROM IS Alliance push militants out of key Syrian town BBC NEWS		MANBIJ FFIEED FROM IS Alliance push militants out of key Syrian town EEI NEWS
7		NEW PAKISTAN PM Nawaz Sharif stepped down after court decision BBC NEWS		NEW PAKISTAN PM Nawaz Sharif stepped down after court decision EIE NEWS










8		VEGAS SHOOTING INVESTIGATION Police: Gunman may have planned earlier attack BBC NEWS		VEGAS SHOOTING INVESTIGATION Police: Gunman may have planned earlier attack Em NEWS
9		ANNA HOLLIGA BBC LEARNING ENGLISH BBC NEWS		ANNA HOLLIGA EEI LEARNING ENGLISH BBC NEWS
10		NEWSROO\ LIVE NEW PAKISTAN PM Majority party nominated Shahid Khaqan Abbasi BBC NEWS		NEWSROO\ LIVE NEW PAKISTAN PM Majority party nominated Shahid Khaqan Abbasi EIEI NEWS
11		TODAY WORLD NEWS		TODAY WORLD News
12		BBC WEATHER		BE! WEATHER

Table (2): Example of wrong detection.

No.	Original Frame	Detected Text	Notes
1		TRUMP FIRES COMMUNICATIONS CHIEF President Trump fires Anthony Scaramucci BBC NEWS 20:17 COMMUNICATIONS DIRECTOR BREAKING PRESIDENT TRU	<ul style="list-style-type: none"> Not all part of text is detected.
2		CARRACCI IS WHITE HOUSE COMMUNICATIONS DIRECTOR	<ul style="list-style-type: none"> Not all part of text is detected. Non-text region is detected.

3			<ul style="list-style-type: none"> • Non-text region is detected.
4			<ul style="list-style-type: none"> • Non-text region is detected.

Where True, Miss, False is the number of correctly detected text regions, missed text regions and falsely detected text regions respectively. High precision (low rate of false detection) and high recall (low rate of miss in detection) indicate perfect text detection. The number of correctly, missed and falsely detected text regions is manually counted for all selected video frames (frames in Table (1) and (2)), then the recall and precision are computed and presented in Table (3).

Table (3): Proposed algorithm performance.

Total text regions	True	Miss	False	Recall(R)	Precision(P)
18	17	1	4	0.9445	0.8095

It is clear from Table (3) that the effect of the proposed algorithm is good and it is able to correctly detect the whole text regions. The recall and precision rate in Table(3) are high, indicating the high performance of the proposed algorithm.

Computational complexity is an important aspect for efficient detection techniques. It is noticeable that the proposed algorithm has low computation cost because it uses only the MSER detector with geometric features and it can still predict the text regions with satisfactory performance.

7. Conclusions

In this paper, an artificial text detection algorithm is presented using MSER detector and geometric properties. The MSER detector is employed in the proposed algorithm to locate a large number of individual text characters as MSERs regions. While the geometric properties are used to filter the MSER text regions in order to remove non-text regions. Three types of geometric properties are utilized: centered of regions, mean intensity and bounding box. The text

regions are highlighted and recognized using Optical Character Recognition (OCR) system directly.

The experimental results demonstrate the efficiency of the proposed algorithm that can achieve a high accuracy rate according to recall and precision measures. In addition to, the proposed algorithm has the ability to efficiently detect artificial text regions under different size, color and background.

Although the use of MSER detector has desirable properties for text detection such as invariance to intensity transformation and low computational cost, some text regions are not detected due to illumination changes and bad contrast therefore further work is still needed.

8. References

1. Yin, Xu-Cheng, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu. "Text detection, tracking and recognition in video: a comprehensive survey." *IEEE Transactions on Image Processing*, vol.25, no. 6 (2016): 2752-2773.
2. Zhu, Yingying, Cong Yao, and Xiang Bai. "Scene text detection and recognition: Recent advances and future trends." *Frontiers of Computer Science*, vol.10, no. 1 (2016): 19-36.
3. Shivakumara, Palaiahnakote, Weihua Huang, Trung Quy Phan, and Chew Lim Tan. "Accurate video text detection through classification of low and high contrast images." *Pattern Recognition*, vol.43, no. 6 (2010): 2165-2185.
4. Shivakumara, Palaiahnakote, Anjan Dutta, Chew Lim Tan, and Umapada Pal. "Multi-oriented scene text detection in video based on wavelet and angle projection boundary growing." *Multimedia tools and applications*, vol.72, no. 1 (2014): 515-539.
5. Anthimopoulos, Marios, Basilis Gatos, and Ioannis Pratikakis. "A two-stage scheme for text detection in video images." *Image and Vision Computing*, vol.28, no. 9 (2010): 1413-1426.
6. Shivakumara, Palaiahnakote, Trung Quy Phan, Shijian Lu, and Chew Lim Tan. "Gradient vector flow and grouping-based method for arbitrarily oriented scene text detection in video images." *IEEE transactions on circuits and systems for video technology*, vol. 23, no. 10 (2013): 1729-1739.

7. Kaur, Amandeep, Renu Dhir, and Gurpreet Singh Lehal. "A survey on camera-captured scene text detection and extraction: towards Gurmukhi script." *International Journal of Multimedia Information Retrieval*, vol. 6, no. 2 (2017): 115-142.
8. Ye, Qixiang, and David Doermann. "Text detection and recognition in imagery: A survey." *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 7 (2015): 1480-1500.
9. Huang, Weilin, Yu Qiao, and Xiaoou Tang. "Robust scene text detection with convolution neural network induced msr trees." In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) *European Conference on Computer Vision*, vol 8692, pp. 497-511. Springer, Cham, 2014.
10. Gironés, Xavier, and Carme Julià. "Real-Time Text Localization in Natural Scene Images Using a Linear Spatial Filter." In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1, pp. 1261-1268. IEEE, 2017.
11. Darshan, H. Y., M. T. Gopalkrishna, and M. C. Hanumantharaju. "Text Detection and Recognition Using Camera Based Images." In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*, pp. 573-579. Springer, Cham, 2015.
12. Liu, Shuang, Yu Zhou, Yongzheng Zhang, Yipeng Wang, and Weiyao Lin. "Text detection in natural scene images with stroke width clustering and superpixel." In *Pacific Rim Conference on Multimedia*, pp. 123-132. Springer, Cham, 2014.
13. Zhu, Wei, Jing Lou, Longtao Chen, Qingyuan Xia, and Mingwu Ren. "Scene text detection via extremal region based double threshold convolutional network classification." *PloS one*, vol. 12, no. 8 (2017): e0182227.
14. Mao, Junhua, Houqiang Li, Wengang Zhou, Shuicheng Yan, and Qi Tian. "Scale based region growing for scene text detection." In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 1007-1016. ACM, 2013.
15. Yin, Xu-Cheng, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. "Robust text detection in natural scene images." *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5 (2014): 970-983.

16. Shi, Cunzhao, Chunheng Wang, Baihua Xiao, Yang Zhang, and Song Gao. "Scene text detection using graph model built upon maximally stable extremal regions." *Pattern recognition letters*, vol. 34, no. 2 (2013): 107-116.
17. Zhao, Xu, Kai-Hsiang Lin, Yun Fu, Yuxiao Hu, Yuncai Liu, and Thomas S. Huang. "Text from corners: a novel approach to detect text and caption in videos." *IEEE Transactions on Image Processing*, vol. 20, no. 3 (2011): 790-799.
18. Xu, Hailiang, Like Xue, and Feng Su. "Scene text detection based on robust stroke width transform and deep belief network." In *Asian Conference on Computer Vision*, pp. 195-209. Springer, Cham, 2014.
19. Sharma, Nitin. "Text Extraction and Recognition from the Normal Images using MSER Feature Extraction and Text Segmentation Methods." *Indian Journal of Science and Technology*, vol. 10, no. 17 (2017).
20. Brooks, Teresa Nicole. "Exploring Geometric Property Thresholds For Filtering Non-Text Regions In A Connected Component Based Text Detection Application." *arXiv preprint arXiv:1709.03548* (2017).