# Maximally Stable Extremal Region Approach for Accurate Text Detection in Natural Scene Images

[1]Ms. Uma B. Karanje, [2]Prof. Rahul Dagade, [3]Prof. Sankirti Shiravale

Assistant Professor
Department of Computer Engineering
Marathwada Mitra Mandals College of Engineering
Pune, India

***Abstract* – Text detection method discovers the presence of text in images, videos, etc. This technique is very needful for many applications, based on content based image analysis, such as web image search, map analysis, video information retrieval, etc. It is very challenging to detect the text from natural scene images due to its complex background and noise.**

**The objective is to design a text detection system which will be able to detect maximum characters from natural scene images.**

**In this text detection system, first, the input natural scene image is pre-processed. The input color image is converted into grey and then Otsu binarization algorithm is applied on grey scale image. Then, in the next stage, character candidates are extracted from binarized image using the MSERs region detector algorithm. MSER will extract various features from input image and then divide it into number of different regions. Then, morphological filter is applied to remove noise and unwanted regions detected by MSER. After morphological operation some heuristic rules are applied to these regions; which removes non-text candidates. The probability of text is estimated by calculating features like height, width and area of contours. Finally, text candidates corresponding to true texts are constructed and displayed using rectangle.**

***Keywords*—Text detection, Maximally Stable Extremal Regions, Morphological filter, text construction.**

## I. Introduction

Text detection from image is one of the important tasks for image processing and computer vision. Image can have needful information and this information is important for fully understanding of the image. Text detection is useful in many content-based image and video applications, such as content-based web image search, video information retrieval, and mobile based text analysis and recognition. As in recent years, image, as the visual basis for perceiving the world, is the key media for information acquisition, expression and transmission. The text in natural scene images has to be robustly detected before being recognized and retrieved. Many problems need to be solved in order to read text in natural images including text localization, character and word segmentation, recognition, integration of language models and context, etc.[8]. Text detection and localization in natural scene images is important for content-based image analysis. This problem is challenging due to its complex background, and variations of font, size, color and orientation, etc.[2].

If the text can be automatically detected, extracted and recognized by the computers, then, more reliable content-based access to the image data can be achieved. Therefore, how to locate and extract textual information quickly and accurately from videos and images becomes a hot topic area in the world today.

## II. Liturature Survey

### *A. Different types of images that contains text in it :*

In general, the images which contains texts are divided into three different types : document image, scene image and born-digital images[1][5][14][20]. And the new type of image is also used for today’s multimedia systems i.e. heterogeneous images [3].

*1. Document images :* Document images are nothing but image-format of the document. Image format of any document is created by scanners or cameras. In which, the image is transformed from paper-based documents into image-format for electric read. In the early stage of text extraction, there is only focus on document images[27].

*2. Scene images :* Scene images contain the text, such as the advertising boards, banners, which is captured by the cameras; therefore scene text appears with the background part of the scene. These types of images are very challenging to detect and recognize, because the backgrounds are complex, containing the text in different sizes, styles and alignments. Also, scene text is affected by lighting conditions and perspective distortions. The current OCR software cannot handle complex background interferences and non-orienting text lines[27].

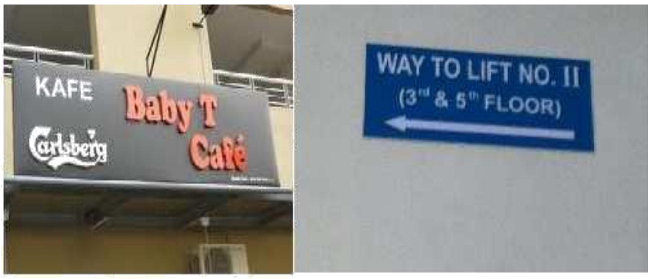


Figure 1. Sample scene images.

3. *Born-digital images :* Born-digital images are generated by computer software and are saved as digital images. Compared with document images and scene images, there are more defects in born digital images, such as more complex foreground/background, low resolution, compression loss, and severe edge softness. Therefore, during text extraction, it is difficult to distinct the text from the background[27].

4. *Heterogeneous images :* This type of image contains the combination of all above given images i.e. it can have a digital image with scene text and document text[27].

### B. *Different methods used to detect text from scene images :*

Many methods for text detection from scene images have been proposed over the past years; by various authors. This section gives brief review on methods based on connected components[2][9], edges[2][15], colors[2], combination of edges and colors[2], textures[2], corners[2], semiautomatic ground truth generation[7], strokes[1], etc.

1. *Connected component (CC) based method :*
The method consists of two steps. The first step is to draw CC from images using a specific method and the second step is to estimate whether the CC is text CC or not based on CC feature and CC relative feature.

2. *Sliding window based method :*
Sliding window based methods, also known as region-based methods, use a sliding window to search for possible texts in the image and then use machine learning techniques to identify text. These methods are slow as the image has to be processed in multiple scales.

3. *Hybrid method :*
The hybrid method presented by Pan et al exploits a region detector to detect text candidates and extracts connected components as character candidates by local binarization; non-characters are eliminated with a Conditional Random Fields model, and characters can finally be grouped into text[1].

4. *Edge based method :*
This method is based on the factor like edge of character; edge is reliable feature of the text regardless of color/intensity, layout, orientations, etc. As the text region has high contrast to its background, the edges of character can be easily detected. There are two steps used in this method: first, an edge extraction algorithm (such as canny edge detector) is used to get the edges and second, smoothing algorithm or morphology is used for edges connections and obtaining a full character boundary. The main disadvantage of this method is that small image regions and stroke may be misidentified. Therefore this method needs to be verified using other methods.

5. *Color based method :*
In this method, color clustering is done by categorizing the pixels with same or similar colors and forming a candidate region. Then the candidate regions are analyzed and the CC is estimated. The main challenge of this method is the degree of clustering. If the data is over clustered, the background and text region may be mixed together. And if the data is under clustered, the number of clustering will be increased and the system performance will be degraded.

6. *Combination of edges and colors :*
Some methods combine Method 1 and Method 2, which detects both edges and colors of the text. This method has achieved better results by combining both features together than using these features separately.

7. *Texture based method :*
This method deals with text regions as a special texture. The region is identified as text region or not according to the extracted relevant texture of the candidate regions. To overcome the disadvantages mentioned above, hybrid approach is presented, which takes the advantages of both texture-based and CC-based methods, to robustly detect and localize texts in natural scene images. In this method, a text region detector is designed which is based on the texture. This can be used to estimate the probabilities of the position and the scale of the text and then it is analyzed to be text region or not.

8. *Corner based method :*
This approach is inspired by the observation that the characters in the text, usually contains multiple corner points. The method is to describe the text regions formed by the corner points using several discriminative features. The research on the method based on corners is still in the early stage. Compared with texture based method, this method is faster but the performance is less satisfied.

9. *Stroke based method :*
As a basic element of text strings, strokes provide robust features for text detection in natural scene images. Text can be modeled as a combination of stroke components with a variety of orientations, and features of text can be extracted from combinations and distributions of the stroke components.
One feature that separates text from other elements of a scene is its nearly constant stroke feature like stroke width. This can be utilized to recover regions that are likely to contain text. For

stroke-based methods, text stroke candidates are extracted by segmentation, verified by feature extraction and classification, and grouped together by clustering. These methods are easy to implement on specific applications because of the intuition and simplicity. However, complex backgrounds make text strokes hard to segment and verify.

*10. Semiautomatic Ground Truth Generation method :*

The semiautomatic ground truth generation system for text detection and recognition includes text with different orientation and language. In this method, the system allows user to manually correct the ground truth if the automatic method produces incorrect results. This method uses eleven attributes at the word level, namely: line index, word index, coordinate values of bounding box, area, content, script type, orientation information, type of text (caption/scene), condition of text (distortion/distortion free), start frame, and end frame to evaluate the performance of the method.

### C. *Summary of literature survey :*

Table below shows the comparison between different methods with the accuracy in their results; and datasets used for evaluating the performance of the method.

| Author's Name | Year | P | R | f | Methodology | DataSet Used |
|---|---|---|---|---|---|---|
| Lukas Neumann et.al.[1] | 2011 | 68.9 | 52.5 | 59.6 | MSER ++ | ICDAR 2011 |
| Lukas Neumann et.al.[18] | 2012 | 73.1 | 64.7 | 68.7 | 2 Stage Algorithm for ERs Pruning | ICDAR 2011,<br>Street View Text Dataset |
| Cunzhao Shi et.al.[19] | 2013 | 83.3 | 63.1 | 71.8 | Graph cut model with MSER | ICDAR 2011 |
| Xu-Cheng Yin et.al.[1] | 2013 | 86.29 | 68.26 | 76.22 | MSER as Character candidate | ICDAR 2011,<br>Multilingual DB,<br>Street view DB,<br>Multi-orientation DB. |
| Rodrigo Minetto et.al.[24] | 2014 | 0.74 | 0.63 | 0.68 | Use of snooper text,<br>toggle-mapping image segmentation,<br>HOG-based descriptor | ITW,<br>SVT,<br>EPS,<br>ICD DB. |
| B.H.Shekar et.al.[21] | 2015 | 0.84 | 0.79 | 0.82 | Obtaining skeleton using morphology | ICDAR 2003,<br>ICDAR 2011 |
| Xu-Cheng Yin et.al.[22] | 2015 | 0.81 | 0.63 | 0.71 | hierarchical clustering with a unified distance metric learning framework | USTB-SV1K DB,<br>MSRA-TD500 DB,<br>ICDAR 2011,<br>ICDAR 2013. |
| XiaobingWang et.al.[23] | 2015 | 0.81 | 0.68 | 0.74 | multi-layer segmentation,<br>higher order conditional random field(CRF),<br>Graph cuts | ICDAR2003,<br>ICDAR2011,<br>ICDAR2013 |
| Runmin Wang et.al.[25] | 2015 | 0.77 | 0.60 | 0.68 | confidence map and context information | ICDAR2005,<br>ICDAR2011,<br>ICDAR2013 |
| Bowornrat Sriman et.al.[26] | 2015 | 80.67% | | | SIFT algorithm,<br>trained patch models,<br>K-means clustering | TSIB Thai,<br>NECTEC Thai,<br>ICDAR. |

Table 1 : Comparative analysis of literature survey

## III. Text detection system

The text detection system has various stages for detecting text in natural scene images. These stages are shown in figure 5. There are five different stages, in first stage; system takes input image which is nothing but the natural scene image. This image is converted from RGB to gray and preprocessed using the Otsu binarization method. In next stage, the character candidates are extracted using the MSER approach and then noise and non-text removal is done using morphological filter. Finally detected text is constructed and displayed as an output.

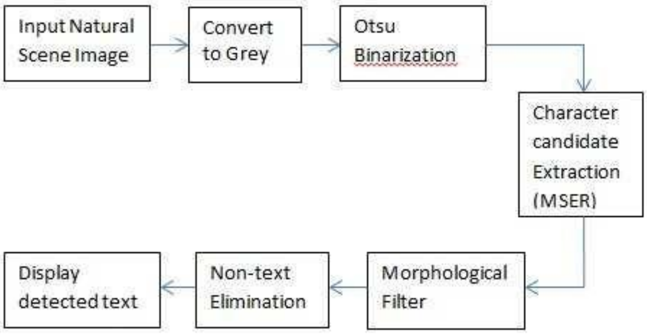


Figure 2. Architecture for text detection system.

### 1) *Pre-processing*

This is the primary phase of the image analysis, where binarization is done on the input natural scene image. The method for binarization is thresholding. Thresholding is important technique in image segmentation, enhancement and object detection [7]. In this stage, the natural scene RGB image shown in figure 3 (a) is taken as input and converted into the grey scale image (as shown in figure 3 (b)). It will enhance the result of the text detection system by converting input pixels into only two colors i.e. black and white. After that, Otsu binarization is applied as shown in figure 3(c). Otsus binarization method is a parameter-less, global thresholding binarization method[7][17]. It assumes the presence of two distributions (one for the text and another for the background), and calculates a threshold value to minimize the variance between the two distributions. It also removes noise from the natural scene image.

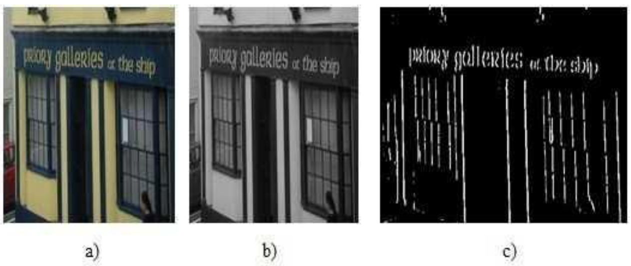


Figure 3. a) Input image; 2) output of gray scale; and 3) output of Otsu binarization.

### 2) *Character Candidate Extraction*

**Extremal region :** An extremal region is a connected component of an image whose pixels have either higher or lower intensity than its outer boundary pixels[1]. Extremal regions of the whole image are extracted as a rooted tree. An extremal region is in fact a set of pixels. Its variation is defined as follows[1][18]:

Let $R_l$ be an extremal region in the input image, then the branch of the tree rooted at $R_l$ will be as :

$$B(R_l) = (R_l, R_{l+1}, \ldots, R_{l+\Delta})$$

where, Δ is a parameter for region R; the variation (instability) of $R_l$ is defined as :

$$V(R_l) = | R_l + \Delta R_l | \div | R_l |$$

where, $| R |$= number of pixels in R. An extremal region $R_l$ is a maximally stable extremal region if its variation is lower and more stable than its parent $R_{l-1}$ and child $R_{l+1}$ [1].

MSER algorithm extracts the extremal region of image. The extremal region is defined by using the contours of input image. Then it calculates the number of different regions and gives different colors to each and every regions. So these regions are used to identify the connected component in natural scene image, which is useful to analyse the texts availability and background in that image.

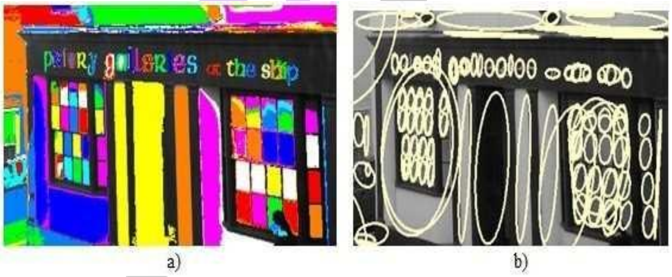


Figure 4. a) Output of detected regions as MSERs; b) Applying ellipses on detected region

Figure 4 (a) shows the detected extremal regions using MSER approach, and figure 4 (b) shows the applied ellipses on these regions for better understanding of the detected regions.

### 3) *Non-text elimination and text construction*

In this stage of the system, non-text regions or unwanted extremal regions detected by MSER is removed by defining the constraints on the contours. First the Morphology algorithm [5] is applied on the detected MSERs; which easily connect the very close regions together while leaving those whose positions are far away to each other isolated. And then again the contours are calculated. To reduce noise and connect the strokes in the binary image, here, morphological operator for closing operation is used. In low contrast images, a character sometimes might be broken to pieces. Thus, it is more likely that connected component analysis might make the wrong decision. So, we have to merge these regions first using morphological filter.

As shown in figure 5 (a), the output of the closed morphology has the joint area, also known as text blobs. Then this area is bounded by the rectangle after applying some geometrical constraints.

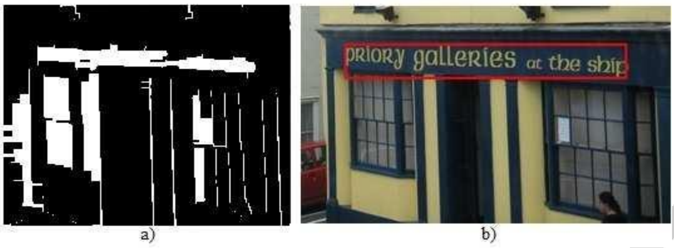


Figure 5. a) Output of morphology b) Output after the non-text elimination and text construction with rectangle.

Here, by observing the patterns of the scene text, some other heuristic rules are created and applied on the detected contours. As the normal texts feature has the width greater than its height, so, only the region satisfying this condition is selected. The size of detected contours must be greater than 150 and the height of contour must be greater than atleast15. After satisfying all these conditions, the region is selected and displayed using the rectangle. Here, the output of final clustering is displayed with the rectangle. But in some results, like as shown in figure 5(b), rectangle border crosses the text, so it will affect the output from the OCR; i.e. next stage of the detection system. The text should be properly detected and displayed for accuracy of OCR. The value of height, width, and the x, y co-ordinates of the connected component is first searched and then value of x is decreased by 10, y is decreased by 7, width is increased by 20 and height is increased by 10. So, the OCR will get the accurate bounding box of the texts, detected by text detection system as shown in figure 6 (b). It is more accurate bounding box than as shown in figure 6(a).

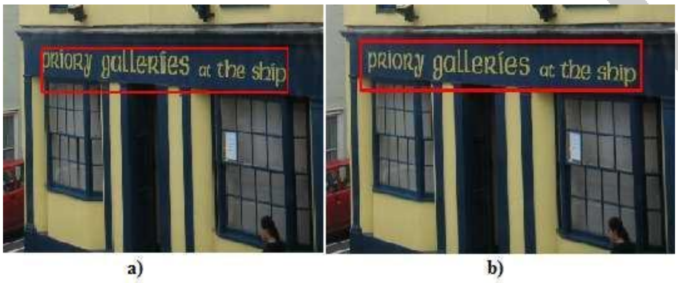


Figure 6. Bonding box results a) before increase in co-ordinates b)accuracy after increase in co-ordinates of the block.

## IV. RESULTS AND DISCUSSIONS

### 1) *DATASETS*

Accuracy of the text detection system is calculated by three parameters : Recall, Precision and f- measure. This text detection system is evaluated using two different datasets: 1. Own Dataset; 2. ICDAR 2011. Also the experiments on multilingual, street view, multi-orientation, blur, similar foreground and background images are demonstrated, it shows the effectiveness of this method.

#### *1.1 Own Dataset*

Own Dataset contains 30 different scene images, captured in various places, in different light effects and font sizes, styles, etc. These images tested on the text detection system and results are shown using recall, precision and f measure values.

#### *2.1 ICDAR 2011 dataset*

The ICDAR 2011 Robust Reading Competition (Challenge 2: Reading Text in Scene Images) database is a widely used database for benchmarking scene text detection algorithms. The database contains 229 training images and 255 testing images. Here, 130 scene images from ICDAR 2011 datasets are tested on the MSER text detection system, using recall, precision and f-measure values.

Number of texts in input image is considered as the relevant text, and the number of text detected by system is defined as retrieved text. If the system detects non text regions as text, then it is considered as the false positives in the output. And if the system does not detect the text region available in input image, then it will be considered as false negative. The value for the same is calculated using the equations below:

**Precision = (No. of text in input image ∩ No. of detected text in output image) ÷ No. of detected text in output image ;**

**Recall = (No. of text in input image ∩ No. of detected text in output image) ÷ No. of text in input ima ge;**

**f-measure = 2 (Precision × Recall ) ÷ (Precision + Recall).**

### 2) *SNAPSHOTS OF THE TEXT DETECTION SYSTEM*

Different stages to detect the text region from natural scene image are as shown in figure 7. Input image in figure (a) is pre-processed by converting it into gray scale (figure (b)) and then Otsu binarization is applied as in figure(c). The result of MSER algorithm is shown in figure(d), and then ellipses are applied as shown in figure (e). The morphology will connect the detected region as shown in figure (f), so it is easy to construct it with the rectangle as shown in figure (g).

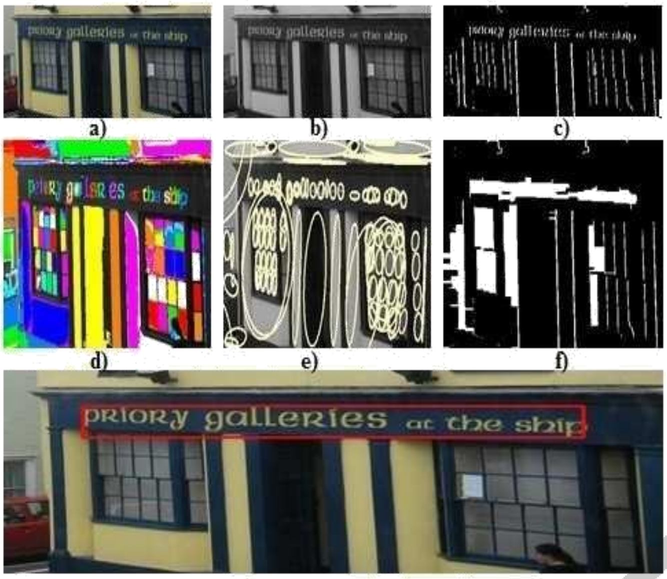


Figure 7. Stages of text detection system

### 3) COMPARATIVE RESULT OF DIFFERENT METHODS

As shown in graph below (figure 8), result of the system is good compare to some other methods. This system gives 84.7 precision, 85.59 recall and 85.12 f-measure (Table 2). Figure 8 shows comparison graph of this system with some well-known text detection methods :

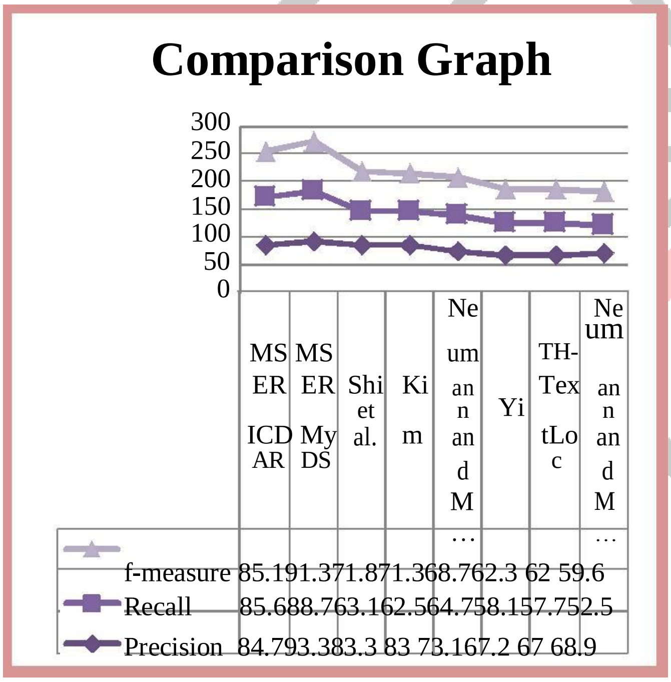


Figure 8. Comparison of system with other methods by recall, precision and f- measure.

Table 2. Result analysis of ICDAR 2011 and Own Dataset

| Dataset | Precision | Recall | F-measure |
|---|---|---|---|
| ICDAR 2011 | 84.66 | 85.59 | 85.12 |
| Own Dataset | 93.25 | 88.65 | 91.34 |

***Result of different images :***

As shown in figure below, the system is tested using images with different features like font style, orientation, blur, etc. From figure 9-14, figure (a) shows original natural scene image taken as input to the system, figure (b) shows the result of images with grey scale and figure (c) shows the output of Otsu binarization. Then, Figure (d) shows result of the character candidate extraction with MSER and figure (e) shows the region bounded with ellipses. Figure (f) shows the result of the morphology under the previous image, and by using the size of contours the final rectangle is applied as shown in (g).

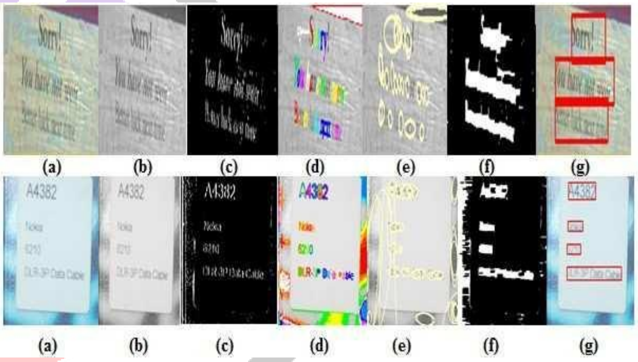


Figure 9. Sample result of blur images

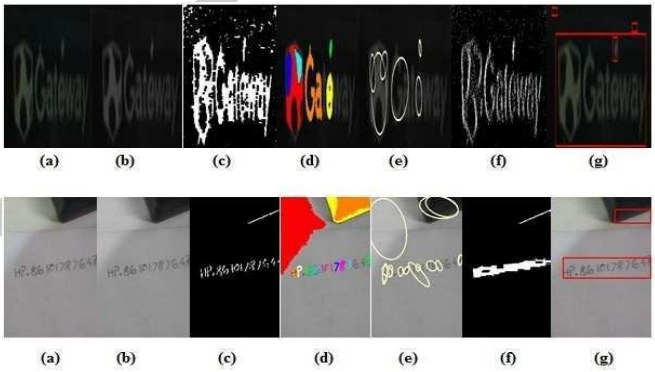


Figure 10. Sample result of low contrast images

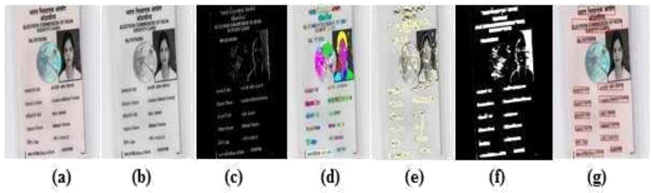


Figure 11. Sample result of Multi-orientation and multi-lingual image

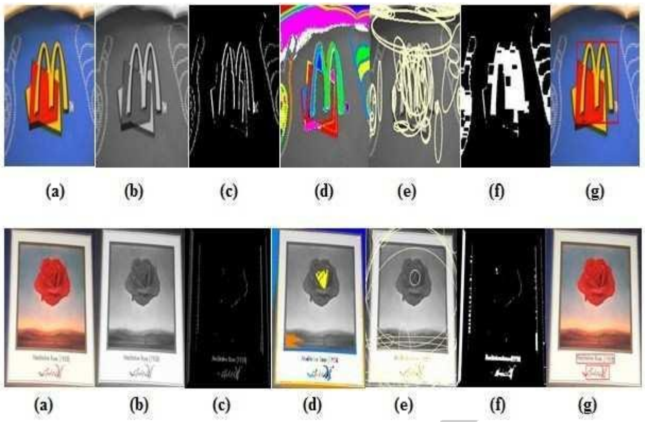


Figure 12. Sample result of different font style images

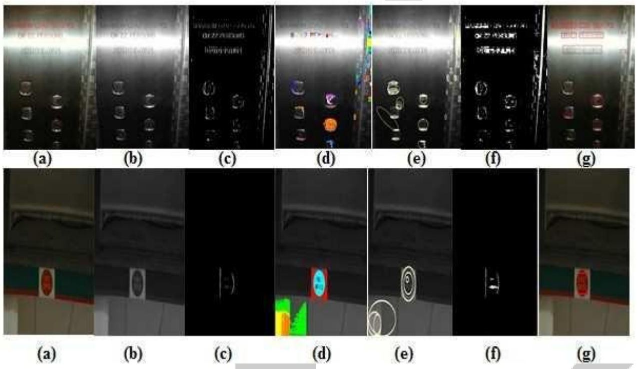
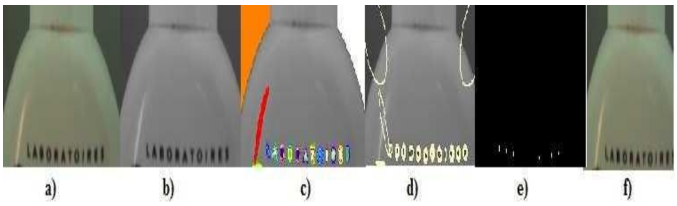


Figure 13. Sample result of uneven lightning images

There are some cases where, this text detection system generates false alarms due to unsuccessful in bounding detected connected components in the morphological closing operation. Some sample images which generates false alarms are shown in figure 14 below.



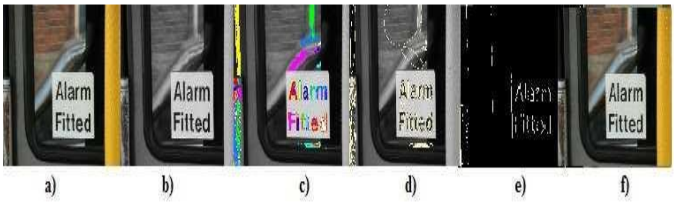


Figure 14. Images where system generated false alarms.

## V. CONCLUSION

This text detection system with MSER and morphological algorithm enables to detect most characters from the natural scene images. The MSER method is proven better results from various techniques, because it easily finds the extremal regions from the input image. It finds most of the characters from natural scene images, although having noise, blur, etc. The heuristic rules increase the clarity and accuracy of the detected text. The future enhancement of this system is to conduct more experiments to examine the text detection procedure with more algorithms and for specific applications. Also to detect highly blurred texts in low-resolution natural scene images.

## VI. REFERENCES

1] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao, "Robust Text Detection in Natural Scene Images, IEEE transaction on Pattern Analysis And Machine Intelligence, Vol:36, 2013, pp. 970-983.

[2] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, "A Hybrid Approach to Detect and Localize Texts in Natural Scene Images, IEEE Transactions On Image Processing, Vol. 20, 2011.

[3] Huizhong Chen, Sam S. Tsai, Georg Schroth, David M. Chen, Radek Grzeszczuk and Bernd Girod, "Robust Tex t Detection In Natural Images With Edge-Enhanced Maximally Stable Extremal Regions, 18th IEEE international conference on Image Processing, pp.2609-2612, 2011.

[4] Jian Zhang, Renhong Cheng, Kai Wang, Hong Zhao, "Research on the text detection and extraction from complex images, Fourth International Conference on Emerging Intelligent Data and Web Technologies, Vol. 10, 2013. pp. 708-713.

[5] Thuy Ho, Ngoc Ly, "A Scene Text-Based Image Ret rieval System, IEEE interna- tional symposium on Signal Processing and Information Tech., pp. 79-84, 2012.

[6] Honggang Zhang, KailiZhao, Yi-ZheSong, JunGuo, "Text extraction from natural scene image: A survey",Els evier journal on Neurocomputing ,pp.310-323, 2013.

[7] Aroop Mukherjee, Soumen Kanrar, "Enhancement of Image Resolution by Bina- rization", International Journal of Computer Applications (0975 8887),Volume 10 No.10, 2010.

[8] Teofilo E. de Campos, Bodla Rakesh Babu, Manik Varma, "Character Recognition In Natural Images, Internati onal conf. on Intelligence Science and Big data Engg., pp. 193-200, 2011.

[9] Xiaobing Wang, Yonghang Song, Yuanlin Zhang, "Natural scene text detection in multi-channel conn ected component segmentation, 12th International conf. on Document Analysis and Recognition, pp. 1375-1379, 2013.

[10] Chucai Yi, Yingli Tian, "Text extraction from scene images by character appearance and structure modeling", Elsevier journal on Computer Vision and Image Understanding, 2013,pp. 182-194

[11] Chitrakala Gopalan, D.Manjula, "Sliding window approach based Text Binarization from Complex Textual images", International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, pp. 309-313

[12] Trung Quy Phan, Palaiahnakote Shivakumara, Souvik Bhowmick, Shimiao Li, Chew Lim Tan, Umapada Pal, "Semiautomatic Ground Truth Generation for Text Det ection and Recognition in Video Images", IEEE Trans. Circuits And Systems For Video Technology, VOL. 24, NO. 8, 2014

[13] Rong-Chi Chang, "Intelligent Text Detection an d Extraction from Natural Scene Images,15th North- East Asia Symposium on Nano, Information tech. and reliability, pp.23-28, 2011.

[14] Yao Li, Huchuan Lu, "Scene Text Detection via Stroke Width, 21st International Conference on Pattern Recognition, 2012. pp.681-684.

[15] Jonathan Fabrizio, Beatriz Marcotegui, Matthieu Cord , "Text detection in street level images, Journal on Pattern analysis applications, 2013, Vol 16, Issue 4, pp. 519-533.

[16] K. Wang, B. Babenko, and S. Belongie. "End-to- end scene text recognition. Inter- national conf. on Computer Vision, pp.1457-1464, Vol. 10, 2011.

[17] Anand Mishra, Karteek Alahari, C.V. Jawahar, " An MRF Model for Binarization of Natural Scene Text", IEEE conf on Document Analysis and Recognition, 2011,pp.11- 16.

[18] L. Neumann, J. Matas, "Real-time scene text lo calization and recognition, in Proc.IEEE Conf.on Computer Vision and Pattern Recognition, 2012, pp. 3538−3545.

[19] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao, "Scene text detection using graph model built upon maximally stable extremal region". vol 3 4, issue 2, 2013, page no. 107-116.

[20] C.P.Sumathi, T.Santhanam, G.Gayathri Devi, "A survey on various approaches of text extraction in images" , International journal of computer science and engineering survey, Vol 3, No.4, 2012, pp. 27-42.

[21] B.H.Shekar, Smitha M.L., "Skeleton Matching ba sed approach for Text Localization in Scene Images", arXiv:1502.03913v1, 2015, pp. 1-12.

[22] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang, and Hong-Wei Hao, "Multi-Orientation Scene Text Detection with Adaptive Clustering", IEEE Transactions On Pattern Analysis And Machine Intelligence, VOL. 37, NO. 9, 2015, pp. 1930-1937.

[23]Xiaobing Wang, Yonghong Song, Yuanlin Zhang, Jingmin Xin, "Natural scene text detection with mul ti-layersegmentation and higher order conditional random field based analysis", Elsevier publication on Patt ern Recognition Letters 60–61, 2015, pp. 41–47.

[24] Rodrigo Minetto , Nicolas Thome , Matthieu Cord, Neucimar J. Leite , Jorge Stolfi, "SnooperText: A t ext detection system for automatic indexing of urban scenes", Elsevier publication on Computer Vision an d Image Understanding 122, 2014, pp. 92–104.

[25] Runmin Wang, Nong Sang, "Changxin Gao, Text detection approach based on confidence map and context information", Elsevier publication on Neurocomputin g, 157, 2015, pp. 153–165.

[26] Bowornrat Sriman, Lambert Schomaker, "Object Attention Patches for Text Detection and Recognition in Scene Images using SIFT", ICPRAM_2015.

[27] Uma B Karanje and Rahul Dagade. Article: Survey on Text Detection, Segmentation and Recognition from a Natural Scene Images. International Journal of Computer Applications 108(13):39-43, December 2014.