An Enhanced MSER Pruning Algorithm for Detection and Localization of Bangla Texts from Scene Images

Rashedul Islam, Rafiqul Islam, and Kamrul Talukder
Computer Science and Engineering Discipline, Khulna University, Bangladesh

Abstract: *Text detection and localization have great importance for content based image analysis and text based image indexing. The efficiency of text recognition depends on the efficiency of text localization. So, the main goal of the proposed method is to detect and localize text regions with high accuracy. To achieve this goal, a new and efficient method has been introduced for localization of Bangla text from scene images. In order to improve precision and recall as well as f-measure, Maximally Stable Extremal Region (MSER) based method along with double filtering techniques have been used. As MSER algorithm generates many false positives, we have introduced double filtering method for removing these false positives to increase the f-measure to a great extent. Our proposed method works at three basic levels. Firstly, MSER regions are generated from the input color image by converting it into gray scale image. Secondly, some heuristic features are used to filter out most of the false positives or non-text regions. Lastly, Stroke Width Transform (SWT) based filtering method is used to filter out remaining non-text regions. Remaining components are then grouped into candidate text regions marked by bounding box over each region. As there is no benchmark database for Bangla text, the proposed method is implemented on our own prepared database consisting of 200 scene images of Bangla texts and has got prominent performance. To evaluate the performance of our proposed approach, we have also tested the proposed method on International Conference on Document Analysis and Recognition(ICDAR) 2013 benchmark database and have got a better result than the related existing methods.*

Keywords: *MSER, scene image, ICDAR, aspect ratio, euler number, bangla text.*

Received July 27, 2017; accepted June19, 2018
https://doi.org/10.34028/iajit/17/3/11

1. Introduction

Text in scene images provides important information about semantic of the images [33] that helps people to understand the meaning of the images. In some cases, text may be the main component of scene images. Automatic detection and extraction of text from scene images have drawn the attention of the researchers due to its wide application areas like content-based image retrieval, text-based image indexing, automatic annotation of the image, robotics, document analysis, keyword-based image search, etc. It is also useful for those who have language barrier like visually impaired persons and foreigners [3]. It will assist in establishing a text editable method from a scanned document. The efficiency of text extraction depends on proper detection. So at first, an algorithm has to be designed that will be capable of detecting and localizing texts from the input image with high accuracy.

The existing text detection method can be divided into three groups [31]: sliding window-based methods [5, 31], Connected Component (CC) based methods [21, 31] and hybrid methods [22, 25]. In sliding window-based methods, a multi-scale fixed sized sliding window is moved through the possible locations of an image to detect text candidates; machine learning technique is then used to identify text. Although the method is robust to noise and blur, it is a slow process, as it has a large search space. Connected component based methods are used at first to extract CCs as character candidates from images based on intensity, color, stroke width etc., Then non-text CCs are removed by using different properties of the texts. The hybrid methods exploit advantages of both the sliding window and CC based methods in order to increase robustness for detecting texts from the scene images. Recently, Maximally Stable Extremal Region (MSER) based method [17], which belongs to CC based method, won the first place in both International Conference on Document Analysis and Recognition (ICDAR) 2011 and ICDAR 2013 competitions [15, 28]. Though MSER based method is a successful method for detection of scene texts in comparison with existing methods, it produces a large number of false positives, as a result, decreases precision as well as f-measure. In this paper, a robust text detection and extraction method is proposed to work with scene images containing Bangla texts. Bangla language has different types of characters (vowel, consonant, joined letter, having full ‘matra’, half ‘matra’ and no ‘matra’) and characters are divided into different zones. Details are discussed in section 2. The above mentioned characteristics of Bangla texts make it complex and challenging task to detect them and extract from scene images. To achieve the goal, MSER algorithm [19] has been applied along with two-stage filtering technique to get better accuracy than existing methods. The MSER algorithm is applicable to an image of low quality. This algorithm does not depend on a language while detecting text from scene images [33]. The complexity of the algorithm proposed by Matas *et al*. [19] Is $O(n\log(\log(n)))$ where n is the number of pixels in the image. MSER algorithm produces a large number of false positives. To eliminate these false positives, rule-based filtering technique and SWT have been applied. After that, all the individual text characters have been merged to form single rectangular bounding box around individual words.

The remainder of the paper is organized as follows. Characteristics of Bangla texts are described in section 2. Section 3 reviews recent text detection methods. Section 4 describes an overview of text detector. The proposed method is described in section 5, and the experimental results are shown in section 6.

2. Characteristics of Bangla Text

In Bangla alphabet, there are 39 consonants, 11 vowels, and 10 numerals. Modifiers are used for writing texts in Bangla language. Two types of modifiers are used such as vowel and consonant modifiers and they are used only with consonants. A modifier can be used on any side of a Bangla character. The presence of headline is a unique property of Bangla text. It is a horizontal line always located at the upper portion of a character. Among the 50 basic characters, 10 do not have headline, 8 have half headlines and the rest have full headlines. There are some characters that contain curve line above them. Such characters are ই,ঈ,উ,ঊ,ঐ,ঔ,ট,ঠ. A Bangla text may be partitioned into three zones. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic characters or compound characters below the headline and lower zone is the portion where some of the modifiers can reside. The imaginary line separating middle and lower zone is called base line. Figure 1 shows the zones of Bangla text.

Figure 1. Different zones of Bangla text.

The concept of uppercase and lowercase is absent in Bangla script and writing style of Bangla is from left to right in a horizontal manner. Due to the presence of some unique features as mentioned above, detection and recognition of Bangla texts have become more challenging than English texts.

3. Related Work

Scene images may contain useful information that would be helpful for different categories of people. From this point of view, many researchers are working in this field. Earlier works in this field were confined in scanned documents only where the texts were presented in black color under white background [1]. Bhattacharya *et al*. [3] proposed an efficient method for extraction of Bangla and Devanagari texts from scene images by analyzing connected components obtained from the binary image. They did it by using the common feature of these two scripts i.e., presence of headline. The authors calculated height, mean, and standard deviation of every candidate headline component and used morphological operations to detect true headlines. Here, morphological operation (opening an object A by linear structuring element B) helps to identify headlines. They performed the experiment on their set of 100 images and got precision 68.8% and recall 71.2%. Their algorithm fails if there is small sized curved text present in the image. Another morphological approach was proposed by Ghoshal *et al*. [10]. In their proposed method, at first, they detected unattached text components and then segmented connected components from an image containing Bangla/Devanagari characters. The restriction of the proposed approach is that it can capture only highlighted texts. Uniform stroke thickness and presence of headline are the two major characteristics of Bangla and Devanagari scripts. Islam *et al.* [14] described the method of text detection from scene images using these two major characteristics of Bangla text. In the preprocessing stage, they extracted connected components by using morphological closing operation along with canny edge detector. Then they selected candidate text regions by computing stroke thickness. To perform the operation, the authors calculated mean (μ) and standard deviation (σ) of the local stroke thickness values. If $\mu > 2\sigma$, the thickness of the underlying stroke is nearly uniform and sub-image S will be selected as candidate text regions. The selected candidate text regions are filtered by some geometric properties of such regions. The rules are explained below :

- Aspect ratio: Range of aspect ratio of a text region is in between 0.1 and 10.
- *w, h factor:* If the size of an input image is l, then $w \le \frac{l}{2}$ and $h \le \frac{l}{2}$ are valid for a text region.

Here, w and h are width and height of a candidate text region respectively.

- For a candidate text region, height>10 pixels.
- If a_1 and a_2 are the area of two adjacent candidate text regions and r is the overlap area between them, then $r \leq 50\%\ of\ a_1\ and\ a_2$.

A list of regions will be selected after applying the above rules [30]. But there may also be some non-text regions. After that, the authors tested each selected regions to see whether they produced headline or not. A selected region will be treated as text region if it produces a headline. The authors also used similarity measures for detecting the text regions that missed out from the above operation. Finally, Bangla and Devanagari text regions were selected. The authors tested the algorithm on 100 sample images of their database and obtained precision 72% and recall 74%.

An image, captured by a digital camera may have perspective distortion. The above algorithms cannot handle the images having perspective distortion. Ghoshal *et al*. [11] have corrected the distortion by using homographic transformation. After correcting the perspective distortion, the authors detected headlines by applying morphological operation. Then the authors separated the components those are attached with the headline. The components selected by the above procedure may include texts as well as non-texts. The authors have separated headline attached text components by applying some characteristics of text like elongation, holes, aspect ratio, object to background pixel ratio. There are some text symbols those are not attached to headlines. For separation of these texts, the authors took the measure of increasing the area of the bounding box enough so that the text components lie inside it and thus separated. Then they removed the headlines and separated the text components. Another approach of scene text detection is to use the stroke width feature. Epshtein *et al.* [8]. estimated this stroke width by dense calculation of “stroke width transform” in a bottom-up approach from pixel level. They proposed a novel CC-based text detection algorithm [8]. Islam and Mondal [13] used the texture based approach in which, an image is decomposed and pre-processed to extract features. Then they used these features to train two classifiers that are based on Artificial Neural Network (ANN) and Support Vector Machine (SVM) classifiers. They used 56 features from which 8 are calculated from the mean, second order and third order central moments and the other 48 are calculated from Wavelet Histogram Energy (WHE). The overall technique included an iterative scan through the input by a fixed block size and defined whether it was text or non-text. The authors used a 16×16 block size and a scan interval of 4 pixels. Their system is divided into the following phases:

- Image decomposition.
- Feature extraction and selection.
- Text detection.

Banik *et al*. [2] proposed a method of segmentation of characters or their parts from Bangla texts extracted from scene images. The proposed algorithm can detect background and texts by combining unsupervised learning k-means clustering and Otsu’s threshold selection. In order to select optimal K value for K-means clustering, the authors proposed criteria to select. The segmentation is based on region growing and extraction of both headline and baseline of Bangla texts. The algorithm proposed by the authors work in the following steps as shown in Figure 2.

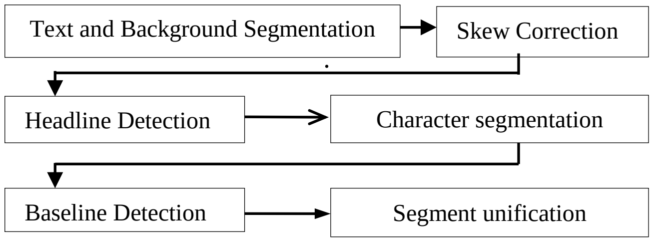


Figure 2. Basic steps of the algorithm [2].

As the proposed approach belongs to CC-based methods, so we also reviewed some of the existing CC-based methods. Yin *et al*. [32] described a robust method for the text detection in the natural scene. They proposed a novel MSER based scene text detection method. Shi *et al*. [29] proposed a novel scene text detection approach using graph model built upon MSERs which outperforms state-of-the-art methods both in recall and precision.

Chen *et al*. [4] proposed an algorithm that used a combination of MSER and Canny edge detector for detecting text candidates. Image blur is efficiently handled by this combination because close symbols are distinguished by the canny detector. This is achieved by removing the MSER pixels outside the boundary formed by the canny edges. Different types of filtering techniques like size, aspect ratio, the number of holes, stroke width, were used by the authors to filter out the non-text regions. For the construction of text candidates, the authors used the single-link clustering algorithm. Main parameters used for this task are spatial distance, width, height, and aspect ratio. There is an additional check after text candidates are built. There is a probability that a line will be a text line if it contains three or more text objects. A text line is rejected if a significant portion of the objects is repetitive. The text lines are then split into individual words using Otsu’s method [24]. He *et al*. [12] proposed a novel text attentional convolutional neural network (text CNN) for extracting text related regions from the image components. They developed new learning mechanism to train the text CNN with multi-level and rich supervised information that includes character label, text region mask, and binary text/ nontext information. Using the rich supervision information enables the Text-CNN for discriminating ambiguous texts, and also increases its robustness against complicated background components. The main task of text/non-text classification is facilitated by low-level supervised information. They also developed a powerful low-level detector called CE-MSER which extends the widely used MSER by enhancing intensity contrast between text pattern and background. Neumann and Matas [23] proposed an unconstrained end-to-end scene text localization and recognition method by using MSER as initial text candidates. Then they introduced a novel feature based on character stroke area estimation. This feature is efficiently computed from a region distance map, which is invariant to scaling and rotations and allows to detect text regions efficiently. Overview of their proposed method is shown in Figure 3.

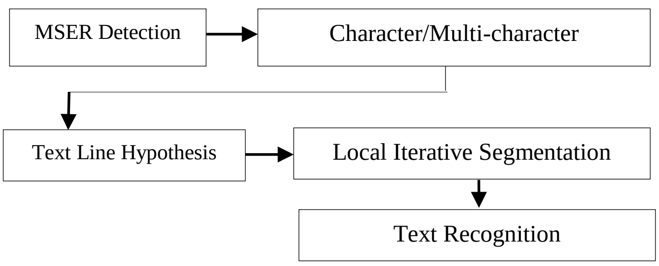


Figure 3. Overview of the method. Initial text hypotheses efficiently generated by an MSER detector are further refined using a local text model, unique to each text line [23].

Nahar [20] used ANN to recognize Arabic hand-written characters with the Genetics Algorithm (GA). They used GA to search for the best ANN structure.

In order to compare the performance of our proposed method, we have also implemented the edge-based algorithm and connected component based algorithm and observed their performances regarding precision, recall, and f-measure on ICDAR 2013 dataset and also on our own database containing 200 scene images of Bangla text.

4. Text Detector Overview

In this section, we have given an overview of text detector.

4.1. Benchmark for Text Detection

To achieve better recall rate, at first we have to identify the general criteria that should be considered in the detection of text from scene images. These are listed below:

- *Precision*: To increase precision, it should be ensured that the detected results should not contain non-text regions if possible.
- *Recall*: The text detection algorithm should localize as many text regions as possible so that it increases recall rate.

To incorporate the above criteria, we have developed an enhanced MSER pruning algorithm for effective scene text detection.

4.2. Why MSER?

MSER deals with the pixels of an image that has a uniform and stable intensity over its background. And it is assumed that text characters have significant intensity contrast to its background as well as have uniform intensity or color. Considering all these into account, MSER is a natural choice for effective scene text detection [4].

4.3. Pre-Processing of MSER

The MSER extraction implements the following steps:

- At first, a simple luminance thresholding is performed by the sweeping threshold of intensity from black to white.
- Extract Extremal Regions (ERs) by the way mentioned in (1).

$$S_r = \{p | I(p) > I(q) \quad \forall p \in S_r, \forall q \in B(S_r)\} \tag{1}$$

Where I is the gray scale image and p and q are pixel indices of I. *R* is a threshold value used for extracting the region, and $B(Sr)$ is the set of boundary pixels of S_r.

- Approximate a region with an ellipse or colored pixel list (optional).Those region descriptors are used as features. If an ER is maximally stable, it might be rejected if the following conditions satisfied [26].

- It is too big.
- It is too small.
- It is too unstable.
- It is too similar to its parent MSER.

5. Proposed Approach

Proper detection is the key issue of text detection and localization tasks. The efficiency of these tasks much depends on the design of a suitable algorithm that is capable of detecting text with high accuracy. So our first target is to design a method to detect text regions from the input scene image with a better result than existing methods. As filtering or removing non-text is the key challenge of every text detection algorithm, we tried to use better filtering technique to achieve the goal. In our proposed approach MSER based algorithm that was proposed by Matas *et al*. [19] have been used for detecting candidate text regions. Figure 4 shows the steps of our proposed approach.

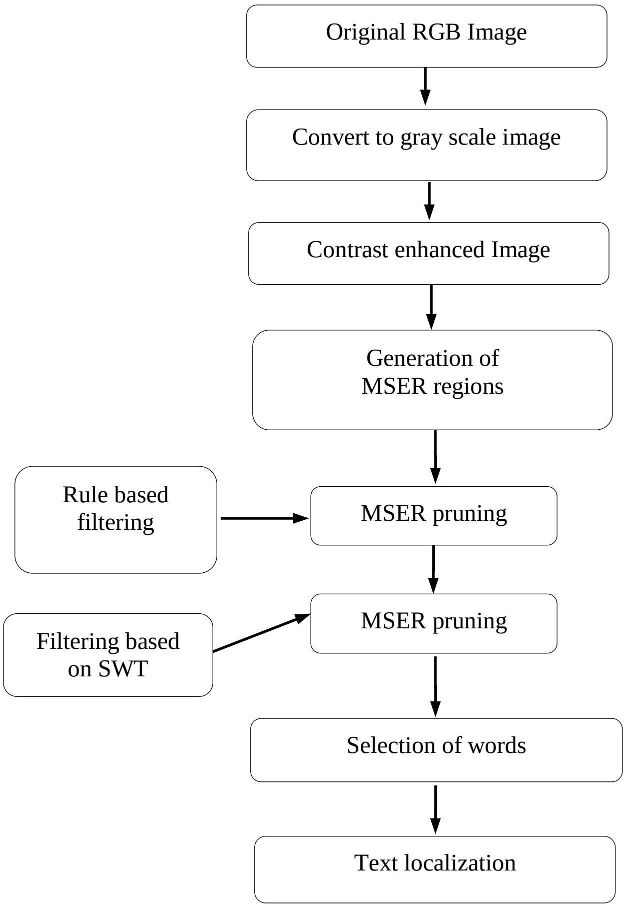


Figure 4. Flowchart of the proposed method.

Here, we first briefly introduce the extremal regions. An extremal region is such a region where the intensity of all the pixels inside the region is either larger or lower than the intensity of its boundary pixels. Extremal regions of the entire image are extracted as a rooted tree. An extremal region is nothing but a set of pixels. A maximally stable extremal region is the extremal region whose variation is less and has more stability than its parent and child [16]. Let Ω_i be an extremal region, $B(\Omega_i) = (\Omega_i, \Omega_{i+1}, ...\Omega_{i+\Delta})$ be the branch of the tree having root at Ω_i. The variation of Ω_i is defined in (2) as

$$v(\Omega_i) = \frac{|\Omega_{i+\Delta} - \Omega_i|}{|\Omega_i|} \tag{2}$$

Here, i is the current intensity level and Δ is the increment, $|\Omega|$ denotes the number of pixels in Ω. From (2) we can say that an extremal region Ω_i is a maximally stable extremal region if the value of $v(\Omega_i)$ is lower and it is more stable than its parent Ω_{i-1} and child Ω_{i+1}.

MSER method can extract not only text components but also many non-text components [31]. For removing these huge number of non-text components we have developed a robust MSER pruning algorithm to improve precision as well as f-measure. Two stage (double) filtering technique has been introduced in this proposed approach for removing non-text regions.

5.1. Image Acquisition

The image was captured by using a digital still camera (OLYMPUS VH-210 (14 MP).

5.2. Pre-Processing

In this step, the input color image is converted to gray scale image (0 to 255-pixel values). Then the gray scale image is enhanced i.e., contrast of the image is enhanced. This is done because MSER has some problems to detect the characters of an image with very low contrast [31]. This approach will help us to improve recall rate by selecting more texts as MSER regions.

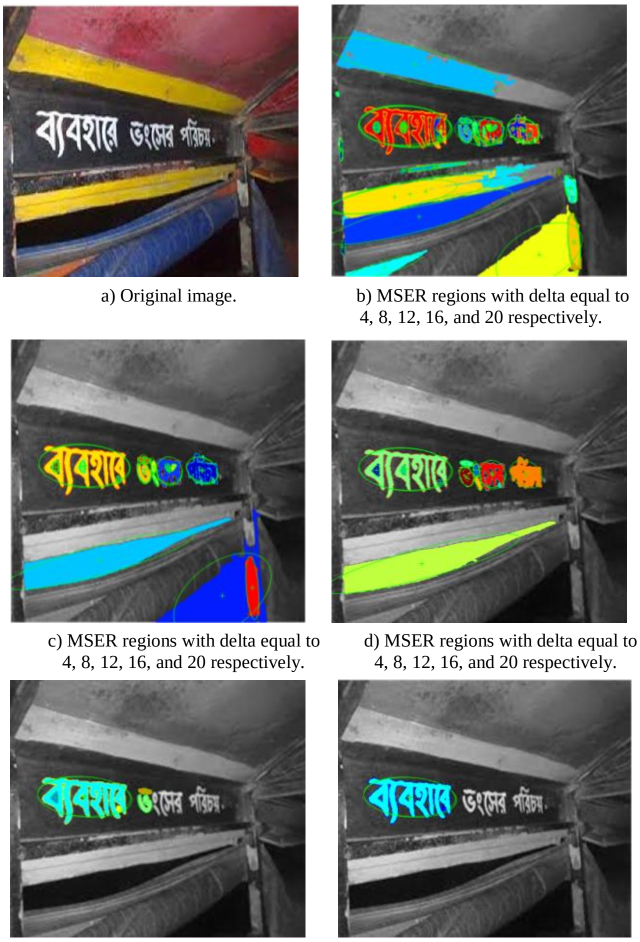
5.3. Generation of Candidate Text Regions

Candidate text regions have been generated by the MSER algorithm. Texts in scene images have consistency in color and high contrast that leads to stable intensity profiles. For this reason, MSER algorithm is used to select candidate text regions from the obtained gray scale image. It can be noticed that many non-text regions also detected along with the text regions. MSER regions of an Image are shown in Figure 5.

a) Original image. b) Projected MSER regions.

Figure 5. MSER Regions of an image.

In order to increase the recall rate, selection of all the text regions of the input image is our main target. To do so, we have to consider different values of delta(Δ) from (2). Because from (2) it is clear that, when the delta is too small, the stable regions will merge with their neighbor and become less stable and if it is too large, many text regions could be missed. Therefore by varying the delta from 2 to 20 with interval 2, we can greatly improve the recall rate. Figure 6 shows the MSER regions of the input image with the different values of delta.



a) Original image.

b) MSER regions with delta equal to 4, 8, 12, 16, and 20 respectively.

c) MSER regions with delta equal to 4, 8, 12, 16, and 20 respectively.

d) MSER regions with delta equal to 4, 8, 12, 16, and 20 respectively.

e) MSER regions with delta equal to 4, 8, 12, 16, and 20 respectively.

f) MSER regions with delta equal to 4, 8, 12, 16, and 20 respectively.

Figure 6. Detection of MSER regions with varying delta.

5.4. MSER Pruning

In order to increase the efficiency of the MSER algorithm, proper pruning of non-text regions is a key issue. To perform the task the following two-stage (double) filtering techniques have been introduced.

1. Filtering using the geometric property of text: In order to filter out non-text regions from selected MSERs, following geometric properties of text can be used [18].

   - Aspect ratio.
   - Eccentricity.
   - Extent.
   - Euler number.
   - Solidity.

Brief explanations of the above rules are stated below.

- *Aspect Ratio*: Ratio of width and height of a candidate region is known as the aspect ratio. Since most text characters have aspect ratio being close to 1, we can eliminate CCs with very large and very small aspect ratio [4]. Here we have considered the region as text region if the corresponding aspect ratio is less than or equal to 2. If it is more than 2, the region is considered as a non-text region and thus eliminated. It is shown in (3).

$$AR = \frac{max\ (width\ ,\ height\ )}{min\ (width\ ,\ height\ )} \tag{3}$$

- *Eccentricity*: Ratio of the distance between the foci of an ellipse and its major axis length is termed as eccentricity. The range of the value is in between 0 and 1. An ellipse having the value of eccentricity = 0 represents a circle, while an ellipse with eccentricity = 1 is a line segment. As every MSER region (text and non-text) is surrounded by ellipses, we must select the threshold value in between 0 and 1. Eccentricity (E) is shown in (4).

$$E = \frac{D}{L} \tag{4}$$

Where D is the distance between the foci of the ellipse and L is the major axis length of the ellipse. To perform the experiment, we have taken the value equal to 0.995 to eliminate non-text regions. This value may be changed for different images.

- *Euler number*: The numeric value obtained by subtracting the number of holes in the objects of a region from the number of objects in that region is known as Euler number. We can calculate Euler Number (EN) of an MSER region by the following way shown in (5).

$$EN = num\_obj - num\_hole \tag{5}$$

Since text characters have less number of holes than that of non-text regions, the value of Euler number will be in the range of 0 to -4 for text regions. But for non-text regions, it may be less than -4. In our experiment, we have used the value of EN is less than -4 to eliminate an MSER region.

- *Extent*: The ratio of the area of a region to the area of the bounding box is called extent.

  Extent (ET) is shown in (6)

$$ET = \frac{area}{area\ of\ bounding\ box} \tag{6}$$

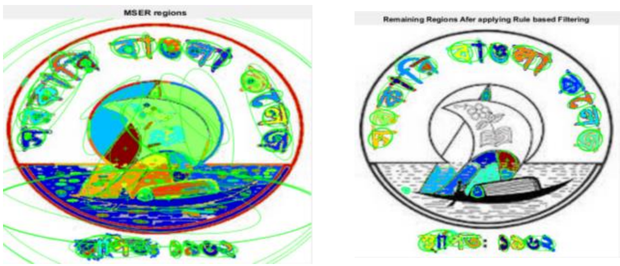
If ET of a candidate region is in between 0.2 and 0.9, it is considered as a text region, otherwise eliminated from the MSER tree considering it as a non-text region.

- *Solidity*: Solidity of a region can be defined as the proportion of the pixels in the convex hull that is also in the region. Solidity (S) is shown in (7).

$$S = \frac{area}{convex\_area} \tag{7}$$

In order to determine the value of S for pruning non-text regions, we have used some training image containing only text regions and found that the values of S for all the regions is more than 0.3. From this viewpoint, we came to a decision that an MSER region having the value of S is less than 0.3 will be considered as a non-text region and eliminated.

The above geometric properties are good for pruning out non-text regions from the MSER regions of an image [18]. For successfully filtering out non-text regions, these properties have been used and then remove regions based on their property values. These property values can be tuned up for different types of images [7]. Figure 7 shows the remaining MSER regions after applying the above geometric rules to filter out non-text MSER regions on the image shown in Figure 7-a).



a) Image with MSER regions. b) Remaining MSER regions after applying geometric properties of the text.

Figure 7. Filtering of MSER regions.

2. Filtering in terms of variation of stroke width: The length of a straight line from one text edge pixel to another along its gradient direction is known as stroke width of the particular text. It is evident that stroke width of a single character almost remains same and this is true for texts of any language like English, Bangla, Devanagari, Arabic etc. So, no special care is required for detection of Bangla text. However, there is a substantial change in stroke width of a non-text character. As a result, stroke width variation can be used as an effective metric to differentiate between text and non-text. The stroke width property has been exploited by several researchers, such as in the work [8, 33]. The researchers calculated stroke width from a stroke boundary to another along gradient direction. The skeleton of the region is an effective tool to represent the structure of a region. As mentioned in [30], we can take advantage of the skeleton to extract stroke width of a region. Binary thinning operation and distance transform are used to measure stroke width of a region as mentioned in [16]. Suppose a candidate text region has very small stroke width variation than another region, the region will be considered as text region because the line and curve that make up the region have a similar width. All text regions bear this common feature and it helps the user to separate them from non-text. Following steps must be followed to calculate stroke width of a region for removing non-text regions from the candidate MSERs.

- Compute the stroke width image: to compute stroke width image, at first we have got the skeleton of MSER regions by using binary thinning operation. Then on every foreground pixel of the skeleton, a distance transform is applied to every foreground pixel on the skeleton to compute the Euclidean distance from this pixel to the nearest boundary of the corresponding MSER. We have got the distance image and skeleton image which are used in the following section. Figure 8 shows MSER region (region image) and corresponding stroke width image.
- Compute stroke width variation metric: Stroke width variation helps to remove non-text regions using a threshold value. The stroke width variation over the entire region is measured by the following way as shown in (8) and (9).

$$sw\_value = distanceimage(skeletonimage) \tag{8}$$

$$sw\_metric = \frac{std(sw\_value)}{mean(sw\_value)} \tag{9}$$

- Threshold the stroke width metric: A threshold is applied to remove non-text regions by the way mentioned in Equations (10) and (11)

$$sw\_threshold = 0.4 \tag{10}$$

$$sw\_filindex = sw\_metric > sw\_threshold \tag{11}$$

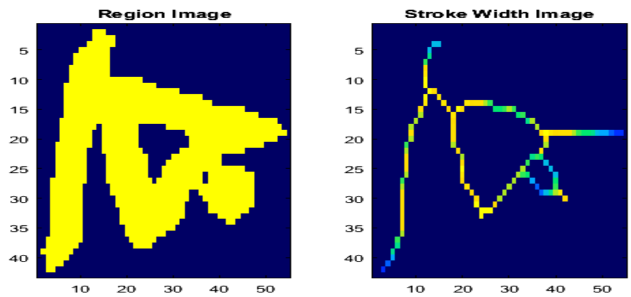


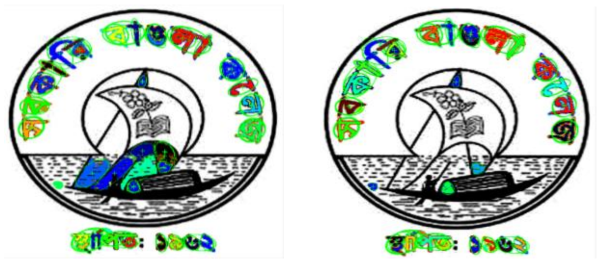
Figure 8. Stroke width image of an MSER region.

From the experiment, it is seen that the value of sw_threshold plays an important role in removing non-text regions. The range of this value is in between 0.1 and 0.9. Effect of changes of this threshold is shown in Figure 9.

a) When sw_threshold=0.2. b) When sw_threshold=0.3. c) When sw_threshold=0.7.

Figure 9. Detected text regions.

The effect of applying stroke width variation in Figure 10-a is shown in Figure 10-b.



a) After applying the geometric rule. b) After applying stroke width variation.

Figure 10. MSER regions.

5.5. Identification of Words

Every single character or region has been marked by small rectangular boxes. In order to get individual word as a single unit, these small bounding boxes have been merged by following the procedures as stated below.

- Expand the area of each small bounding box by 2%
- Find out neighboring regions.
- A series of coinciding bounding boxes will be produced.
- Compute overlap ratio between all the bounding box pairs.
- Merge these boxes to get a single rectangular box around individual words.

Finally, an output image has been obtained by applying the above procedure in an image. Figure 11 shows it where red colored rectangular boxes indicate the single region of text.

Figure 11. Detected text regions.

6. Experimental Results

The proposed method has been implemented on Windows platform using MATLAB r2016a. As there is no publicly available standard database of scene images containing Bangla text, we have developed one such database captured by a still camera from different places. We have an intention to make this database available to the researchers. Two hundred sample images have been taken from the said database to collect simulation result of the proposed method. A few of these sample images with detected texts are shown in Figure 12.

Figure 12. A few sample images from our database. Detected texts are marked by red color rectangular area.

The simulation results have been summarized by Correctly Detected Characters (CDC), False Positives (FP), and False Negatives (FN). This process of counting is known as ground truth [27]. We have calculated precision, recall and f-measure defined in Equation (12), (13), and (14) respectively.

$$P = \frac{CDC}{CDC + FP} \times 100\ \% \tag{12}$$

$$R = \frac{CDC}{CDC + FN} \times 100\ \% \tag{13}$$

$$f\_measure = \frac{2 \times P \times R}{(P + R)} \times 100\ \% \tag{14}$$

Where P is the Precision and R is the Recall.

The proposed method provides precision 76.19%, recall 89.16%, and f-measure 80.19% on our own database consisting of 200 scene images of Bangla texts. We have implemented the method proposed by Chowdhury *et al*. [6] and has got precision 63.27%, recall 85.90% and f-measure 70.18% using the same dataset (Our own database). It is mentioned in [6] that, their algorithm performs better than the proposed algorithm of [3] using the same database. From this point of view, it is clear that our proposed method is better than the existing methods. Table 1 shows the experimental results.

Table 1. Experimental results on the 200 sample images of our own database.

Method	Precision (%)	Recall (%)	F-measure (%)
Proposed	76.19	89.16	80.19
Chowdhury *et al.* [6]	63.27	85.90	70.18
Bhattacharya *et al.* [3]	68..8	71.2	69.98

To observe the result of the proposed method, images from ICDAR 2013 database were taken and got precision 83.84%, recall 85.88% and f-measure 83.07%. Figure 13 shows a few samples of the images of ICDAR 2013 with detected texts marked by red colored rectangular boundaries.

Figure 13. A few samples from ICDAR 2013 database of scene images where detected texts are marked by red colored bounded boxes by the proposed method.

Table 2 shows the comparison of the proposed method with existing methods in [12, 23, 34] on ICDAR 2013 benchmark database.

Table 2. Experimental results on the ICDAR 2013 dataset.

Method	Precision (%)	Recall (%)	F-measure (%)
Proposed	83.84	85.88	83.07
He *et al.* [12]	93	73	82
Zhang *et al.* [34]	88	74	80
Neuman and Matas [23]	81.8	72.4	77.1

Table 3 shows the comparison result of the proposed method with existing edge based method [27], and connected component based method [9] and hybrid method [14] on 200 sample images of our database. Graphical representation of the comparison of the three methods is shown in Figure 14.

Table 3. Experimental results on the 200 sample images of our own database.

Method	Precision (%)	Recall (%)	F-measure (%)
Proposed	76.19	89.16	80.19
Samarbandu and Liu [27]	64.51	73.64	65.66
Gllavata *et al.* [9]	66.45	80.64	70.95
Islam *et al.* [14]	71.83	79.26	72.61

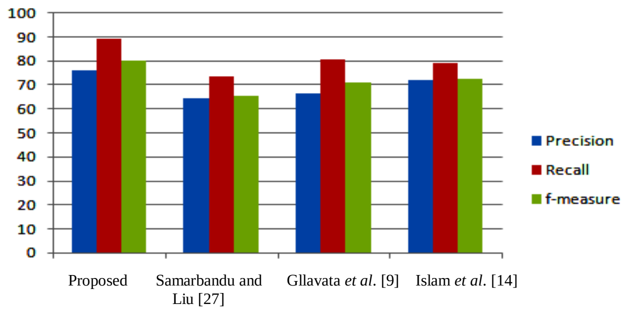


Figure 14. A comparison of the proposed method and existing methods on the basis of precision, recall, and f-measure.

7. Conclusions

Detection and localization of Bangla texts from the scene images are more challenging tasks to achieve satisfactory performance in all the applications due to large variations in character font, size, orientation, color etc., In order to increase the performance of the proposed method, two-stage (double) filtering technique has been used. The advantage of MSER algorithm is that it is invariant to language. Though it produces many false positives, the average precision, recall, and f-measure achieved were remarkable. The results show that our proposed method performs better than the existing related methods [9, 14, 27], on our own database and the proposed method also performs better than existing methods [12, 23, 34] on ICDAR 2013 database. From the experiment, it has observed that in some cases the proposed method generates more false positives. So our future plan is to use machine learning approach to enhance the task of filtering technique that may improve the performance of the proposed method. We also plan to create a database with scene images containing Bangla text and make it publicly available to the researchers.

References

[1] Asaduzzaman A., Molla K., and Molla G., "Printed Bangla Text Recognition using Artificial Neural Network with Heuristic Method," *in Proceedings of International Conference on Computer and Information Technology*, Dhaka, pp. 27-28, 2002.

[2] Banik P., Bhattacharya U., and Parul S., "Segmentation of Bangla Words in Scene Images," *in Proceedings of the 8th Indian Conference on Computer Vision, Graphics and Image Processing*, Mumbai, pp. 1-7, 2012.

[3] Bhattacharya U., Parui S., and Mondal S., "Devanagari and Bangla Text Extraction from Natural Scene Images," *in Proceedings of the 10th international Conference on Document Analysis and Recognition*, Barcelona, pp. 171-175, 2009.

[4] Chen H., Tsai S., Schroth G., Chen D., Chandrasekhar V., Takacs G., Vedantham R., Grzeszczuk R., and Girod B., "Robust Text Detection In Natural Images With Edge-Enhanced Maximally Stable Extremal Regions," *in Proceedings of IEEE International Conference on Image Processing*, Brussels, pp. 2609-2612, 2011.

[5] Chen X. and Yuille A., "Detecting and Reading the Text in Natural Scenes," *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, pp. 366-373, 2004.

[6] Chowdhury A., Bhattacharya U., and Parui S., "Text Detection of Two Major Indian Scripts in Natural Scene Images," *in Proceedings of International Workshop on Camera-Based Document Analysis and Recognition*, Beijing, pp. 42-57, 2012.

[7] Computer Vision Toolbox, available at: https://www.mathworks.com/help/vision/example, Last Visited, 2016.

[8] Epshtein B., Ofek E., and Wexler Y., "Detecting Text in Natural Scenes with Stroke Width Transform," *in Proceedings of Computer Society Conference on Computer Vision and Pattern*

Recognition, San Francisco, pp. 2963-2970, 2010.

[9] Gllavata J., Ewerth R., and Freisleben B., "A Robust Algorithm for Text Detection in Images," *in Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, Rome, pp. 611-616, 2003.

[10] Ghoshal R., Roy A., Bhowmik T., and Parui S., "Headline based Text Extraction from Outdoor Images," *in Proceedings of the 4th International conference Pattern Recognition, and Machine Intelligence*, Moscow, pp. 446-451, 2011.

[11] Ghoshal R., Roy A., and Parui S., "Recognition of Bangla Text from Scene Images through Perspective Correction," *in Proceedings of International Conference on Image Information Processing*, Shimla, pp. 1-6, 03 2011.

[12] He T., Huang W., Qiao Y., and Yao J., "Text Attentional Convolutional Neural Network for Scene Text Detection," *IEEE Transaction on Image Processing*, vol. 25, no. 6, pp. 2529-2541, 2016.

[13] Islam M. and Mondal A., "Towards a Standard Bangla PhotoOCR: Text Detection and Localization," *in Proceedings of 17th International Conference on Computer and Information Technology*, Dhaka, pp. 198-203, 2014.

[14] Islam R., Islam M., and Talukder K., "An Approach to Extract Text Regions from Scene Image," *in Proceedings of International Conference on Computing, Analytics and Security Trends*, Pune, pp. 1-6, 2016.

[15] Kartaz D., Shafait F., Uchida S., Iwamura M., BigordaL., Mestre S., Mas J., Mota D., Almazan J., and HerasL.," ICDAR 2013 Robust Reading Competition," *in Proceedings of the 12th International Conference on Document Analysis and Recognition*, Washington, pp. 1484-1493, 2013.

[16] Li Y. and Lu H., "Scene Text Detection via Stroke Width," *in Proceedings of 21st International Conference on Pattern Recognition*, Tsukuba, pp. 681-684, 2012.

[17] Maximally Stable Extremal Regions, available at: https://en.wikipedia.org/wiki/Maximally_stable_extremal_regions, Last Visited, 2016.

[18] Measure properties of image regions, available at: http://www.mathworks.com/help/images/ref/regionprops.html, Last Visited, 2016.

[19] Matas J., Chum O., Urban M., and Paul T., "Robust Wide Baseline Stereo From Maximally Stable Extremal Regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761-767, 2004.

[20] Nahar K., "Off-line Arabic Hand-Writing Recognition Using Artificial Neural Network with Genetics Algorithm," *The International Arab Journal of Information Technology*, vol. 15, no. 4, pp. 701-707, 2018.

[21] Neumann L. and Matas J., "On Combining Multiple Segmentations in Scene Text Recognition," *in Proceedings of the 12th International Conference on Document Analysis and Recognition*, Washington, pp. 523-527, 2013.

[22] Neumann L. and Matas J., "Scenetextlocalizationand Recognition with Oriented Stroke Detection," *in Proceedings of IEEE International Conference on Computer Vision*, Sydney, pp. 97-104, 2013.

[23] Neumann L. and Matas J., "Efficient Scene Text Localization and Recognition with Local Character Refinement," *in Proceedings of the 13th International Conference on Document Analysis and Recognition*, Tunis, pp. 746-750, 2015.

[24] Otsu N., "A Threshold Selection Method From Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979.

[25] Pan F., Hou X., and Liu L., "A Hybrid Approach to Detect and Localize Texts in Natural Scene Images," *IEEE Transaction on Image Processing*, vol. 20, no. 3, pp. 800-813, 2011.

[26] Region Detectors, available at: http://micc.unifi.it/dElbimbo/wpcontent/uploads/2011/03/slide_coroso/A34%20MSER.pdf, Last Visited, 2017.

[27] Samarabandu J. and Liu X., "An Edge-based text Region Extraction Algorithm for Indoor Mobile Robot Navigation," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 1, no. 7, pp. 2043-2050, 2007.

[28] Shahab A., Shafait F., and Dengel A., "ICDAR 2011 Robust Reading Competition Challenge 2: Reading Scene Images," *in Proceedings of International Conference on Document Analysis and Recognition*, Beijing, pp. 1491-1496, 2011.

[29] Shi C., Wang C., Xio B., Zhang Y., and Gao S., "Scene Text Detection Using Graph Model Built Upon Maximally Stable Extremal Regions," *Pattern Recognition Letters*, vol. 34, no. 2, pp. 107-116, 2013.

[30] Shivakumara P., Phan T., and Tan C., "A Laplacian Approach to Multi-Oriented Text Detection in Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 412-419, 2011.

[31] Sun L., Huo Q., Jia W., and Chen K., "A Robust Approach for Text Detection from Natural Scene Images," *Pattern Recognit*, vol. 48, no. 9, pp. 2906-2920, 2015.

[32] Yin X., Yin X., Huang K., and Hao H., "Robust Text Detection in Natural Scene Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970-983, 2014.

[33] Zarechensky M., "Text Detection in Natural Scenee with The Multilingual Text," *in the Proceedings of the 10th Spring Researcher's Colloquium on Database and Information Systems*, Veliky Novgorod, 2014.

[34] Zhang Z., Shen W., Yao C., and Bai X., "Symmetry-Based Text Line Detection In Natural Scenes," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, pp. 2558-2567, 2015.

Rashedul Islam received B.Sc. degree in Computer Science and Engineering (CSE) from Khulna University, Khulna, Bangladesh in 2002 and received M.Sc. degree in Computer Science and Engineering (CSE) from Uttara University, Dhaka, Bangladesh in 2011. He is working as an Assistant Professor in the Department of Information and Communication Technology (ICT) of Rajuk Uttara Model College, Uttara, Dhaka, Bangladesh. Currently, he is a Ph.D. student in the discipline of Computer Science and Engineering, Khulna University, Khulna, Bangladesh. He is an author of the book Higher Secondary Information and Communication Technology-for class XI-XII (English version). He has published four papers, which have been published in journals as well as in refereed international conference proceedings published by IEEE and other. Rashedul Islam is a member of the Institution of Engineers, Bangladesh (IEB). His present research interest includes Im age processing, Artificial Neural Network, Bio-metrics etc.

Rafiqul Islam obtained Ph.D. in Computer Science from Universiti Teknologi Malaysia (UTM) in 1999 and a combined Master (MS) and Bachelor Degree in Engineering (Computers) from Azerbaijan Polytechnic Institute (Azerbaijan Technical University at present) in 1987. He was a visiting fellow (a postdoctoral researcher) in Japan Advance Institute of Science and Technology (JAIST) in 2001. He worked as head of the Discipline of Computer Science and Engineering of Khulna University and as the Dean of the School of Science, Engineering and Technology of Khulna University. He worked as a Professor in the Department of Computer Science of American International University-Bangladesh (AIUB) from 2009 to 2015. Currently, he is a senior professor of Computer Science and Engineering Discipline of Khulna University, Khulna, Bangladesh. He has 27 years of teaching and research experiences. He has published more than 100 papers, which have been published in international and national journals as well as in refereed international conference proceedings published by IEEE, Springer and others. His research areas include design and analysis of algorithms in the area of image processing, secure cloud computing, external sorting, Information security, Network security data compression, bio-informatics, grid computing, cloud computing etc. Currently, he is doing researches on optimization process using metaheuristic algorithms. Recently his several papers have been published in the journals with high impact factors.

Kamrul Talukder has been a Professor in Computer Science and Engineering (CSE) Discipline of Khulna University (KU) in Bangladesh since 2011. Prof. Talukder was the head of CSE Discipline, KU for three years. He completed his B.Sc. in CSE with distinction from Khulna University in 1999, M.Sc. in Computer Science (by research) from National University of Singapore (NUS) in 2004 and D.Eng. from Hiroshima University in 2008. He was a recipient of the prestigious Japan Society for the Promotion of Science (JSPS) Postdoctoral fellowship for the duration of two years. Dr. Talukder is a life fellow of the Institution of Engineers, Bangladesh (IEB). His research interest is mainly focused on Image Processing, Formal Verification and Software Engineering. He has published more than 50 research publications in the various proceedings of international conferences and in journals.