Text Based Pattern Recognition

Rupali B. Telkikar[1], S. G. Kejgir[2]
Shri Guri Gobind Singhji Institute of Engineering and Technology Vishnupuri, Nanded, Maharashtra, India

Abstract—Detecting and recognizing text in computer vision is one of the difficult tasks and there is a lot of research going in technical field. To know the information contents of an image or the valuable data, there is need of analyzing the text appears in it. In this paper, for extracting the text regions Maximally Stable Extremal Region technique is used. Non-text region is removed by using geometric properties objects. Canny edge detector is used for obtaining edges of character component. Output of canny edge detector is then given to Optical Character Recognition for recognition of text present in the scene image.

***Index Terms*— Maximally Stable Extremal Region, Textbox, Canny Edge Detection, Optical Character Recognition**

I. INTRODUCTION

Now a days as technology is growing rapidly, and so digital camera, smartphones have become widely available to us. Most of the human works are done by machines. One of the crucial problems of computer vision is to read, detect and recognize text in natural scene images. Text detection plays an important role in many document image understanding tasks, such as Optical Character Recognition (OCR). There are so many applications of text-based pattern recognition such as automatic visual classification, multi-language translation, and text detection plays crucial role for blind people. When they want to read text present in scene image, such as aided navigation, wearable or portable computers, content-based document coding, license plate recognition, text- based image indexing, industrial automation, etc.

Scene text in real world may usually contain different colors, font size, scales, geographical configuration, languages, orientations (direction) etc. even in same image. Sometimes background of image may be very tangled with some materials such as, bricks, nails or nut-bolds on boards, grasses, fences which might easily get confused between text and non-text region and error are almost indifferentiable from actual text. Then, there are several disturbing factors such as- uneven illumination source, noise, less resolution, slant occlusion which effect on accuracy of detection.

Accurate detection of scene texts with such a complex background is extremely challenging problem until now and there is much need for further research and improvement.

The paper is specially focused on text detection problems. Because the text submerges in natural scene image for purpose of recognition, firstly it must be detected. Text based pattern recognition steps are as follows: Firstly, extract the text regions from whole image frame using MSER (Maximally Stable Extremal Regions). Then non-text region is removed using geometrical properties of text and filter it out using stroke width transform. The next step is of bounding box creation over the text region. After that by using OCR the text present in input image is reconstructed.

The block diagram of text-based pattern recognition is as shown in figure 1.

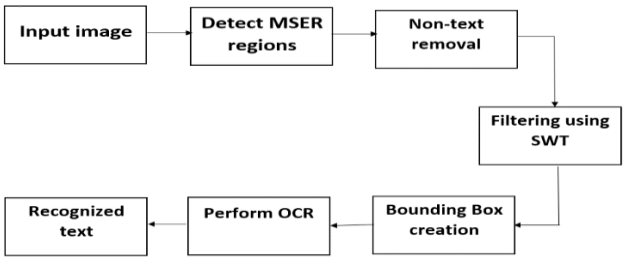

Figure 1: Block diagram of text-based pattern recognition

II. LITERATURE SURVEY

2.1 Sliding-window method

The sliding-window technique is one of the region-based strategy, which focuses on deciding if a given fix is a piece of a content locale or not. This calculation recognizes content information by sliding a multi-scale sub-window all through all the potential areas in the picture. From that point forward, a prepared classifier is utilized to decide the nearness of content data. The text localization and recognition method proposed by Neumann *et al*.

2.2 Connected component method

Firstly, the associated segments are separated from given information picture. Furthermore, to smother the associated segments, the rehashed segments are expelled, and the CC's are consolidated by bunching calculations, at that point content lines are drawn. At long last, non-content lines are erased by a content line classifier. What's more, this method utilizes an associated segment investigation, which comprises of breaking down the spatial game plan of edges or homogeneous shading and grayscale segments that have a place with characters. As delineated by Cai et al. have exhibited a content identification approach which depends on character highlights, for example, quality of edge, thickness and even dispersion of edges. To start with, they apply a shading edge location calculation in YUV shading space and evacuated non-content edges utilizing a low limit calculation. From that point forward, a neighborhood thresholding method is connected to keep low-differentiate message and improve the foundation. At last, projection profiles are breaking down to confine content regions.

2.3 Texture method

Texture based strategy manages text regions as a special surface. The area is indicated as text region or non-text locale relying upon the removed important texture of the text regions. At that point mixture approach is connected, which takes

the benefits of both textures based and CC-based techniques, to vigorously identify and restrict messages in question pictures. In this procedure, a text locale identifier is planned which depends on the textural highlight of content. This can be utilized to decide the probabilities of the area and the size of the content and afterward it is dissected to be text region or non-text region.

2.4 Corner method

The corner-based strategy is motivated by the perception that the characters in the text, for the most part contains various corner edges. The strategy is to follow the content regions framed by the corner focuses utilizing some discriminative highlights. The exploration on corner-based technique is further in the beginning period. Contrasted and surface-based strategy, this technique is quicker however the presentation is less fulfilled.

III. DATASET

3.1 ICDAR 2011

One of the widely used dataset is ICDAR 2011 (International Conference on Document Analysis and Recognition). This dataset is publicly available. for scene text detection purpose this ICDAR dataset was specially designed. In ICDAR 2011 robust reading competition this near-horizontal dataset is used. This dataset comprises of total 484 images. Among which 229 are used for training purpose and 255 images are used for testing purpose. The images in this dataset are fully colored images. The size of the images varies from 422 X 102 to 3380 X 2592 in pixels. English and numerical characters are included in this image. Some images of ICDAR 2011 are shown in figure 2.

Figure 2: ICDAR 2011 Dataset

3.2 SVT

From Google Street View, SVT (Street View Text Detector) is harvested. The picture content has exceptionally low resolution, it shows high changeability while managing outside road level view. are listed below:

1. Picture message regularly originates from business signage.
2. Business names are effectively accessible through geographic business look.

This makes street view text detector a unique set. This makes it suitable for word spotting in wild. Some images from SVT dataset are as shown in figure 3.

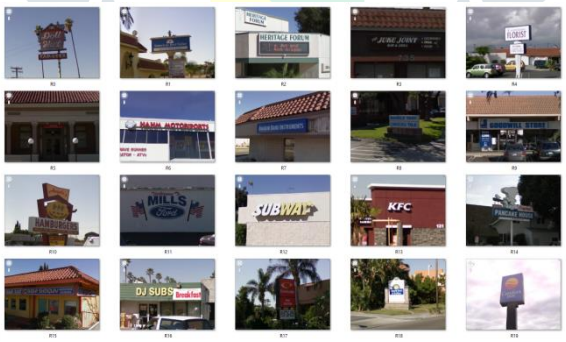

Figure 3: SVT Dataset

3.3 MSRA-TD 500

Despite of the fact that ICDAR 2011 dataset is broadly utilized in research areas, but there are many defects in it. For instance, most of the text lines appears in English language and horizontal way. As in real world text lines are not necessary, the text might be in any language and with the random orientation. There are total 500 natural images in MSRA-TD 500 dataset. And the image size varies from 1296*864 to 1920 X 1280 in pixels. This dataset contains text lines with different languages. For example, English, Chinese, or combination of both, with different colors, different sizes, and with different orientation. MSRA-TD 500 dataset has more challenges than ICDAR 2011 dataset because of orientation of text region. Some images from MSRA-TD 500 are shown in figure 4.

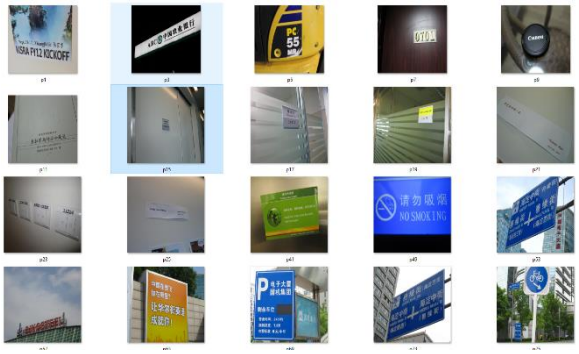

Figure 4: MSRA-TD 500 Dataset

IV. THE PROPOSED METHOD

In this paper we have compare the accuracy with two methods. The first approach is using canny text detector and another approach is by using the traditional method, which are illustrated below:

4.1First approach

The first approach contains following steps:

4.1.1 MSER Extraction

Matas *et.al.* proposed the Maximally stable extremal region detector method. MSER is portrayed by area of associated segment with comparable force esteems limited by differentiating foundation. It works under mass location strategy. It is a stable associated part of some dark level arrangements of the picture. MSER relies upon the edge of the picture, if some edge worth is chosen, at that point the pixels underneath that limit worth are white and every one of those above or equivalent of that edge worth are dark. In this work, least limit worth is in the middle of 0.8 to 0.9 are taken by experimentation premise. MSER identifies the items and every one of the articles can be loaded up with various hues. In this procedure a portion of the districts incorporate the additional foundation pixels these are expelled by vigilant edge location process.

4.1.2 Connected component extraction

In recent scene text detection methods, MSER algorithm is commonly used. This is used to extract CC (Connected Components). At the same time MSER algorithm extracts both the fine and large structures.

The set of extracted CCs' is denoted in equation (1).

$$C = \{c_p\} \quad (1)$$

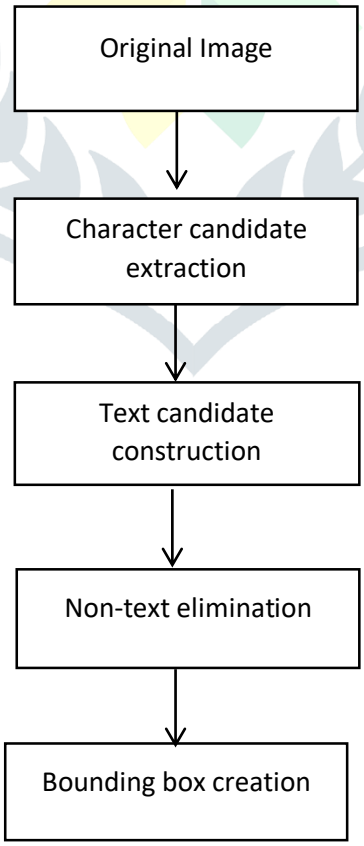

Figure 5: Block diagram of MSER extraction

Compute the center (x_p, x_y) and $\sum_p$ as the covariance matrix of the pixel positions. It is further representing the covariance matrix as is given in equation (2):

$$\sum_p = \sigma_1 v_1 v_1^T + \sigma_2 v_2 v_2^T \qquad (2)$$

Where,
$\sigma_1 \& \sigma_2 =$ eigen value
$v_1 \& v_2 =$ Corresponding eigen vectors
The worl flow of MSER extraction technique is as shown in figure 5.
The figure 6 shows input image taken from ICDAR 2011 dataset. This image contains text embedded with background elements such as trees, road, nails, pole, etc. Figure 7 shows extracted text using MSER technique. The different text elements are filled by different colors.

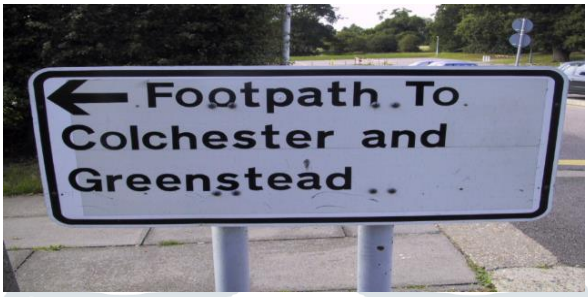

Figure 6: Input Image

Figure 7: Extracted text region by MSER technique

4.1.3 MSER Implementation

MSER implementation algorithm is given in figure 8.

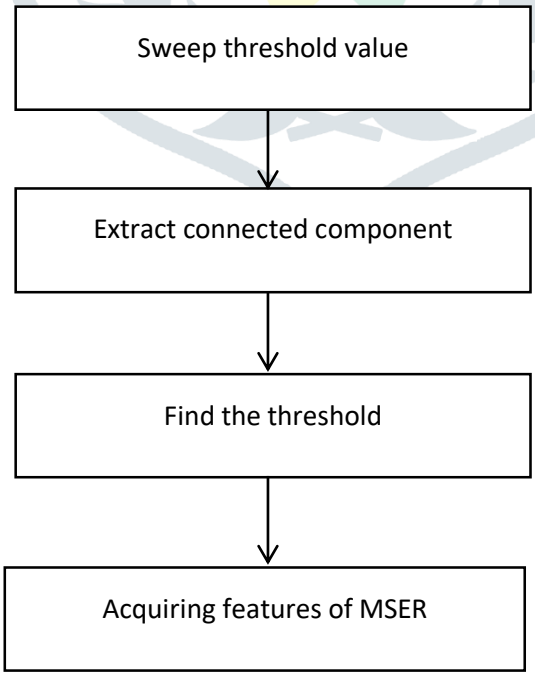

Figure 8: MSER implementation algorithm

To begin with, clear the limit force from dark to white district, to play out a straightforward luminance edge of picture. At that point remove the associated part that is extremal areas. After that discover the edge esteem when extremal area is maximally steady. At last, locale descriptor highlights are separated by MSER.

4.1.4 Canny Edge Detection

The Canny edge finder is a standout amongst the most generally utilized edge discovery calculation. It was created by John F. Canny in 1986. In this work, multi arrange calculation is utilized that joins the adequacy of scene content recognition.

4.1.1.1 Specification for text detection

A general criterion in text detection is as follows:

- **Recall** – Text detection must be able to localize lot of text regions.
- **Precision** – Non-text regions should not present in the detected results.
- **Uniqueness** – Each detected character should be marked only once.
- **Compactness** – It should accurately be able to localize the corresponding character with extra margin.

4.1.5 Processing of Canny Edge Detector

It is multi stage algorithm. Figure 9 is the input image taken from ICDAR 2011 dataset, used for applying multistage algorithm as follows:

Figure 9: Input image for analysis of canny edge detection

- **Noise Removal:**

As Edge Detection techniques are susceptible to the noise in image. For removing noise first step is to apply the Gaussian filter. Figure 10 shows image after applying Gaussian Filter on figure 9.

Figure 10: Gaussian smooth image

- **Intensity Gradient:**

Second step is to apply sobel kernel on horizontal and vertical Gaussian smoothened image, to get first derivative G_x and G_y in respective directions.

$$Edge\ Gradient\ (G) = \sqrt{G_x^2 + G_y^2} \qquad (3)$$

$$Angle\ (\theta) = \tan^{-1}\left(\frac{G_y}{G_x}\right) \qquad (4)$$

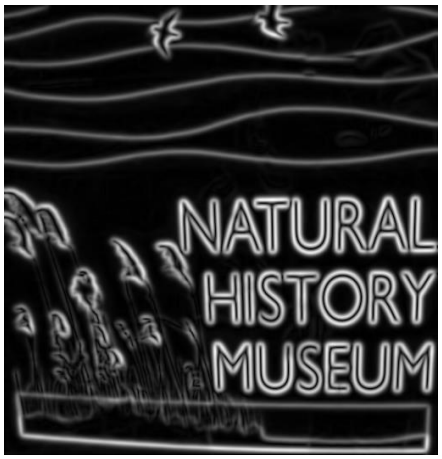

Figure 11: Gradient magnitude image

- **Non-Maximal Suppression:**

For discarding any unnecessary pixels which might not consist the edge, a full scan of picture is done. This step is carried out once we get the gradient magnitude and direction.
As shown in above figure, 12. There is a point A on the edge in vertical direction. Normal to the edge is gradient direction. Point B and C are in direction of the gradient.

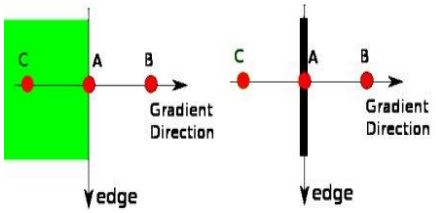

Figure 12: Non-maximal suppression

Then we check for the point A with points B and C to see whether it establish a local maximum or not. If yes, then we proceed for next step otherwise it put to zero or suppressed. And the result we get from this non-maximal suppression stage is a binary image which has "thin edges".

Algorithm:

If $magn(i,j) < magn(i_1,j_1)$ or $magn(i,j) < magn(i_2,j_2)$
Then $I_N(i,j) = 0$
Else $I_N(i,j) = magn(i,j)$

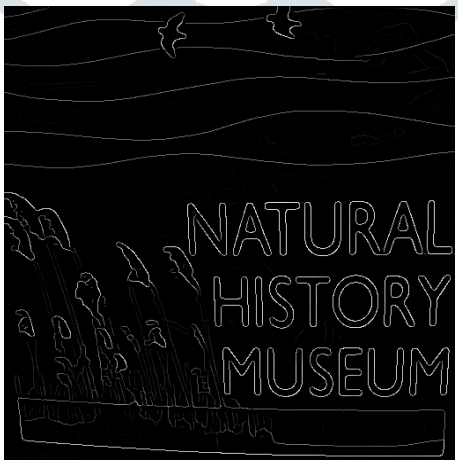

Figure 13: Image after non-maximal suppression

- **Hysteresis Thresholding:**

In hysteresis thresholding, here in this stage can decides among all edges which are really edges and which are not. For elaborating this we need two thresholds one is lower threshold and other is higher threshold. An edge with intensity gradient value more than higher threshold is definitely an edge. And the edges which are having intensity gradient value less than lower threshold are definitely not edge. And those which lie between these two thresholds depending on their connectivity they are classified edges or non-edges based. If they are connected to sure-edge pixel then they are considered as edge. Otherwise they are also rejected or eliminated.

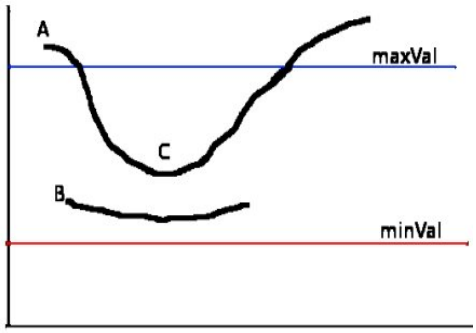

Figure 14: Graph for Hysteresis thresholding

Algorithm for applying Hysteresis thresholding is as follows:

Let, t_l be the lower threshold
t_h be the higher threshold

$||\nabla f(x,y)|| \geq t_h$ definitely an edge
$t_l \geq |||\nabla f(x,y)|| < t_h$ maybe an edge depending on context

$||\nabla f(x,y)|| < t_h$ definitely not an edge

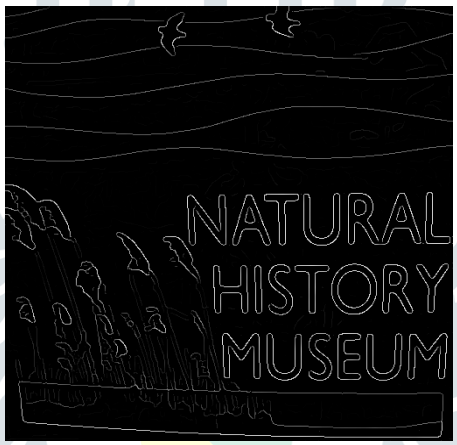

Figure 15: Image after Hysteresis thresholding

For "maybe" edges, decides on the edge of neighboring pixel is strong edge.
After hysteresis thresholding the output image obtained is as shown in figure 15.

4.2 Second approach

Like the first approach, the extraction of character candidate process is similar using MSER. Let's consider figure 16. For performing the non-text region removal filtering algorithm.

Figure 16: Input image for analysis of removal of non-text method

4.2.1 Non-text Removal

Following are the two techniques which are used to remove non-text region. For illustration of these methods consider the image 17 on which MSER extraction is done.

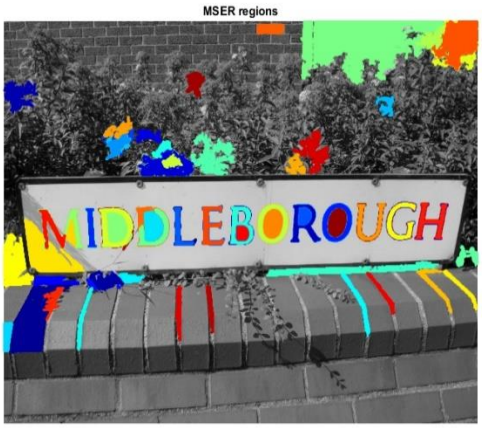

Figure 17: MSER extracted of input image

4.2.1.1 Using geometric properties

- **Aspect ratio:**

Aspect ratio is a mathematical parameter which describes the relationship between the adjacent sides of mechanical objects. Here we consider the ratio of width and height of bounding box. Here we take aspect ratio greater than 3 for better performance.

$$Aspect\ Ratio = \frac{L}{W} \quad (5)$$

Where,

L = Length of the object

W = Width of the object

For non-rectangular objects, aspect ratio can be defined using characteristic dimensions of the object. For example, for an ellipse, the aspect ratio can be defined as the ratio of the major to minor axis.

- **Eccentricity:**

The measure of the aspect ratio is eccentricity. It is used to measure the circular nature of the given region. It is given by the ratio of the length of major axis to minor axis. Usually the eccentricity is greater than 0.995. It can also be defined as the distance between foci and/or the ellipse and its major axis.

- **Extent:**

The extent may be defined as the size and location of rectangle that encloses text.

- **Solidity:**

Solidity describes extent to which the shape is concave or convex. Solidity is also a ratio of pixels in convex hull area that are also given in region. Here we have considered the solidity less than 0.3.

$$solidity = \frac{A_s}{H} \quad (6)$$

Where,

A_s= Area of shape region

H = Convex Hull

- **Euler Number:**

The relationship between the number of holes and the contiguous parts on a shape is described by Euler number. The Euler number is feature of binary image. It is calculated by subtracting connected components and number of holes. The Euler number should be less than -4.

Mathematically,

$$Eul = S - N \quad (7)$$

Where,

S = Number of contiguous parts.

N = Number of holes.

Figure 18. shows the output image after performing non-text region removal using geometric properties.

Figure 18: Non-text region removed using above geometric properties

4.2.2 Stroke Width Transform

A computer vision algorithm called Stroke Width Transform (**SWT**) is one which can be used for the task of detecting text in images. This is one of the non-trivial tasks which is especially used for camera pictures, but SWT performs well in this field. In this paper our aim is to detect segments in an image of a natural scene. This can be done by using the modified version of the Stroke Width Transform by detecting text segments in an image of a natural scene.

The application gets an RGB image to look in and restores another picture where the found content sections are checked. Because of the highlights of the SWT, the subsequent framework can recognize message paying little heed to its scale, heading, text style, and language.

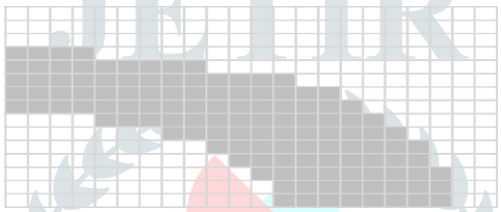

Figure 19: Typical stroke with darker foreground

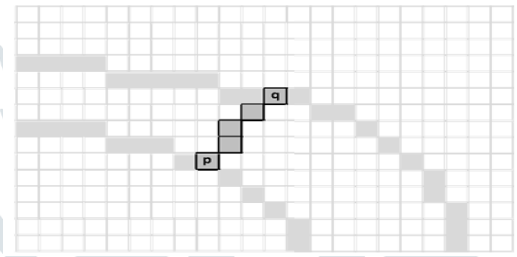

Figure 20: P is boundary pixel

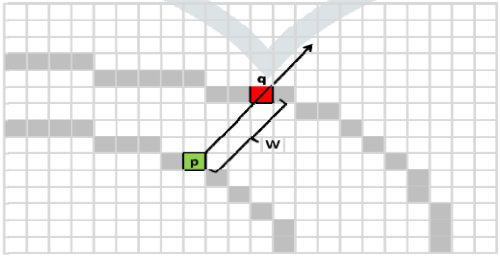

Figure 21: Each pixel through the line is allotted with minimum value to get stroke width.

Binary image can be transformed into stroke width images (skeleton images) and the stroke width variation can be calculated by these skeleton images.

Stroke width variation metric is given by equation 8, as follows:

$$\frac{std.dev\ of\ stroke\ width\ values}{Mean\ of\ stroke\ width\ values} \qquad (8)$$

Where,

$std.dev$ = standard deviation

Figure 22: Non-text removal using stroke width variation

4.3 Bounding Box

For better localization of text region, this step is essential. Here as shown in figure 23, The bounding box around text region are marked in yellow color with line width of 3 pixels. There are three words in image which are highlighted by the bounding box separately.

Depending on desirable output visualization one can vary the pixel width of bounding box from 2 to 5 pixels. As one chooses less value the bounding box will appear very thin otherwise broader outline.

Figure 23: Bounding box creation over text region

4.4 OCR

Optical character acknowledgment or optical character Recognition (OCR) is the mechanical or electronic transformation of images, handwritten document or printed data into machine-encoded information, a photograph of a report, a scene-photograph (for instance the content on signs and boards in a scene photograph) or from caption content superimposed on a picture (for instance from a transmission). By and large used as a sort of information section from printed paper data records – paying little heed to whether worldwide ID reports, requesting, bank clarifications, robotized receipts, business cards, mail, printouts of static-data, or any sensible documentation – it is an average technique for digitizing printed messages with the objective that they can be electronically modified, looked, set away progressively basic, demonstrated on the web, and used in machine systems, for instance, mental enlisting, machine understanding, (isolated) substance to-talk, key data and information mining. OCR is a field of research in model affirmation, man-made awareness, and PC vision.

V. EXPERIMENT RESULT

5.1 Dataset and evaluation process

To completely examine the proposed strategy with the best scene text detection, we evaluate our technique on datasets ICDAR 2011 and MSRA-TD 500 which are publicly available.

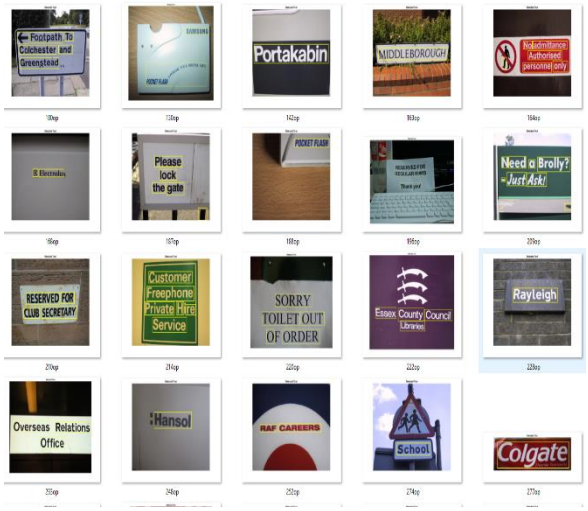

Figure 24: Output of text detection sample on ICDAR 2011 dataset

Moreover, for validating the generalization capacity of evaluated method, we apply the method on natural images captured which contains two different languages such as Marathi and English language text in same image, as shown in figure 25. The image is available at (lasttraintopanvel.blogspot.com)

Figure 25: Text detection sample on multi-language mixed text detection

The formulae for calculating accuracy, precision and recall are as given in equations number 9,10 and 11 respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

Where,

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Comparison of proposed method with Zhandong Liu *et al.* is as shown in table number 1.

Table 1: Comparison evaluation table

Method	**Accuracy**
Zhandong Liu *et al.*	82.94%
Proposed Method	86.98%

Table 2: Performance evaluation on ICDAR 2011 dataset

Threshold	TP	FP	TN	FN
0.0	95	195	21	126
0.1	125	89	22	99
0.2	225	77	28	86
0.3	168	74	32	79
0.4	189	69	33	61
0.5	149	65	36	54
0.6	238	59	39	41
0.7	249	44	41	35
0.8	296	39	44	30
0.9	374	36	47	27

VI.CONCLUSION

The method for text-based pattern recognition is proposed in this paper. The method consists of five steps: Text extraction by MSER technique, Non-text region removal using two filtering techniques using geometrical properties and stroke width transform, Canny edge detection for edge estimation, bounding box creation for text localization, and OCR for text recognition. Non-text region removal is a key element in this method.

The implemented method in this paper is examined on three public datasets i.e. ICDAR 2011, MSRA-TD500 and SVT dataset. On ICDAR 2011 dataset the method achieves 86.98 % accuracy.

REFERENCES

[1] Zhandong Liu, Yong Li, "Method for unconstrained text detection in natural scene image" *IET Computer Vis.*, 2017, Vol. 11 Iss. 7, pp. 596-604 © The Institution of Engineering and Technology 2017.

[2] X.chen and A.yuille. Detecting and reading text in natural scenes. In proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) volume 2, pages II-366. IEEE, 2004.

[3] W. Wu, X. Chen, and J. Yang, "Detection of text on road signs from video," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, No. 4, pp. 378–390, Dec. 2005.

[4] Kethineni Venkateswarlu, Sreerama Murthy Velaga, "Text detection on scene images using MSER" *International Journal of Research in Computer and Communication Technology*, Vol 4, Issue 7, July-2015.

[5] L.Neumann and J.Matas. "Real-time scene text localization and recognition", In *proceedings, Eighth International conference on Document Analysis and Recognition (CVPR)*, 2012.

[6] Hyung II Koo, "Text-Line Detectionin Camera-Captured Document Images Using the State Estimation of Connected Components" *IEEE Transaction on Image Processing*, Vol. 25, No. 1.

[7] (http://robustreading.opendfki.de/wiki/Scenetext)

[8] http://www.iaprtc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)