Text Detection in Multi-Oriented Natural Scene Images

M. Fouzia[1], C. Shoba Bindu[2]

[1] P.G. Student, Department of CSE, JNTU College of Engineering, Anantapur, Andhra Pradesh, India
[2] Associate Professor, Department of CSE, JNTU College of Engineering, Anantapur, Andhra Pradesh, India

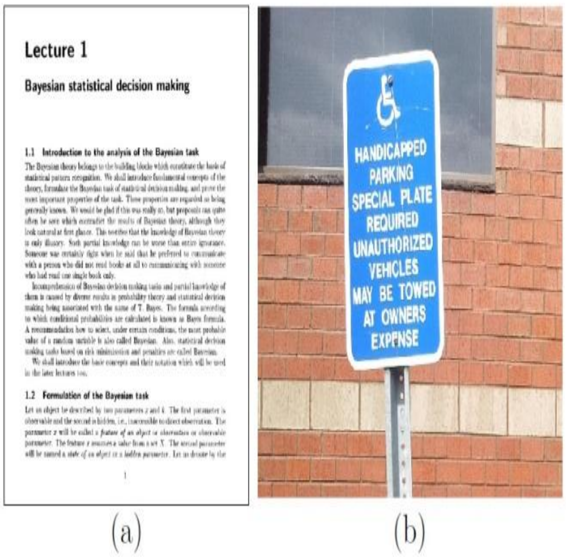
---***---

Abstract - *With the growing number of digital multimedia libraries, the need to efficiently index, browse and retrieve multimedia information is increased. Text embedded in images and video frames can help to identify the image information (e.g. somebody's name appearing in an image) or to display information which is independent of the image (e.g. important news during the transmission of a movie). In common, text in images can be categorized into two groups: artificial text and scene text. Scene text is part of the image, and appears unintentionally, like in traffic signs etc. whereas artificial text is created separately from the image and is laid over it in a later stage, like the name of a journalist at some point in a news program. Artificial text is usually a very good key to index image or video databases. This paper aims to detect text in scene images of multi-orientations, i.e., the images containing text in multiple text lines and the images with problems of skewed angle and distorted images. The preprocessed algorithm uses skew correction and image enhancement algorithms to pre-process the images for better recognition accuracy. Every image has to be pre processed before further processes. The pre processing technique is also tested for the accuracy and time taken to perform the detection.*

Key Words: Text Detection, Multi-orientation, scene images, OCR, Skew-correction, Cropping, and Enhancement.

1. INTRODUCTION

Methods for scene text localization and recognition aim to find all areas in an image (or a video) that would be considered as text by a human, mark boundaries of the areas (usually by rectangular bounding boxes) and output a sequence of (Unicode) characters associated with its content. A new MSER based scene text detection method is used to detect MSER regions which also can used to detect the text in the image [1]. Below fig. shows the difference between printed document and scene text localization and recognition.



Lecture 1

Bayesian statistical decision making

1.1 Introduction to the analysis of the Bayesian task

The Bayesian theory belongs to the building blocks which constitute the basis of statistical pattern recognition. We shall introduce fundamental concepts of the theory, formulate the Bayesian task of statistical decision making, and prove the most important properties of the task. These properties are regarded as being generally known. We would be glad if this was really so, but proposals can quite often be seen which contradict the results of Bayesian theory, although they look natural at first glance. This testifies that the knowledge of Bayesian theory is only illusory. Such partial knowledge can be worse than entire ignorance. Someone was certainly right when he said that he preferred to communicate with a person who did not read books at all to communicating with someone who had read one single book only.

Incomprehension of Bayesian decision making tasks and partial knowledge of them is caused by diverse results in probability theory and statistical decision making being associated with the name of T. Bayes. The formula according to which conditional probabilities are calculated is known as Bayes formula. A recommendation how to select, under certain conditions, the most probable value of a random variable is also called Bayesian. Also, statistical decision making tasks based on risk minimisation and penalties are called Bayesian.

We shall introduce the basic concepts and their notation which will be used in the later lectures too.

1.2 Formulation of the Bayesian task

Let an object be described by two parameters x and k. The first parameter is observable and the second is hidden, i.e., inaccessible to direct observation. The parameter x will be called a *feature of an object* or *observation* or *observable parameter*. The feature x assumes a value from a set X. The second parameter will be named a *state of an object* or a *hidden parameter*. Let us denote by the

1

Fig -1: (a) a scanned book page. (b) a sample image from dataset[2].

Scene text localization and recognition (also known as text recognition and localization in real-world images, environment scene OCR or text-in-the-wild problem) is an open problem, unlike printed document recognition where widely used systems are capable of recognize more than 99% of characters [3] correctly. Text localization techniques can be classified into two important groups - techniques based on a sliding window and techniques based on regions (characters) grouping. Techniques in the first category [4] utilize a window which is moved over the image and the existence of text is estimated on the origin of local image features.

2. LITERATURE SURVEY

Text detection and recognition in natural scene images is, however, a challenging, unsolved computer vision problem. Scene text has complex background, image blur, partially occluded text, variations in font-styles, image noise and varying illumination as illustrated in Figure 1 (b).

The end-to-end scene text recognition problem is divided into a text-detection and text recognition task. Text-detection is a preprocessing step for the text-recognition task. The text detector has to locate words in natural scene images. The text-recognizer predicts a word shown in a cropped image patch which is retrieved by the detector. Several competitions within the scope of the ICDAR have been organized to assess the state-of-the-art. Hence, there are publicly available datasets like the ICDAR 2003 dataset [7] or the Street View Text (SVT) dataset [7], on which objective comparisons of state-of-the-art methods can be done.

2.1 Text Detection Methods

Text-detection methods can be divided into

- Region based and
- Texture based methods [9, 10].

Region based methods rely on image segmentation. Pixels are grouped to CCs (Character Candidates). These candidates are further grouped to candidate words and text lines based on geometric features.. Text confidence maps are created, which are post-processed and converted into word level bounding boxes. Furthermore there exist hybrid approaches which group connected components into word candidates and use a texture based classifier to validate these candidates. Since region based approaches rely on image segmentation they are subject to segmentation errors.

Texture based methods use sliding windows, if its step size or the image scales are not estimated correctly regarding the text size, text detection errors can occur. Candidate text lines are generated by a MSER segmentation and grouped to text line hypotheses, which are verified by a texture based method. Second entry is the text detector proposed by Yi et al. [11] (F-score: 62.32%). This method generates candidate patches with a segmentation method presented in [11]. Based on Canny edges and k-means color clustering text components are segmented. Candidate patches are classified by an AdaBoost texture classifier based on gradient features in block-patterns proposed by Chen et al. [12].

A) Region Based Methods:

The SWT is an image operator which assigns a stroke width to each pixel of an image (see Figure 2). The stroke width is determined by shooting rays on edges in the direction of the gradient. If the ray hits an edge with the same gradient direction modulo 180 degrees, the length of the ray determines the stroke width of the underlying pixels. Pixels which are adjacent to each other belong to the same character if their stroke width is similar. Hence, CCs are formed by segmenting the stroke width map. Adjacent CCs are grouped to words if the median stroke width ratio of two CCs does not exceed2:0, the height ratio of the two components is smaller than2:0 and distance and average color difference are within Thresholds learned from the training set.

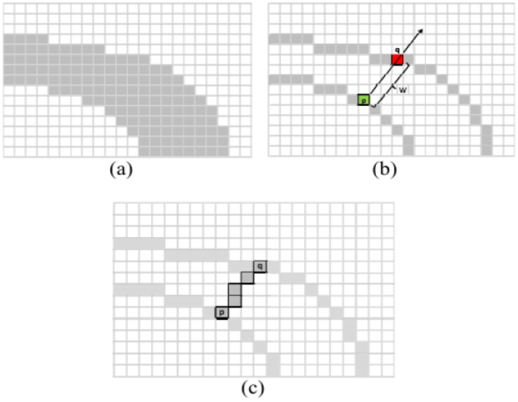


Fig 2: The SWT for a stroke shown in (a) is computed by shooting a ray from the edges in the gradient direction (b). If another edge with the same gradient direction is found along the ray, the width between the start- and end-point is assigned as stroke width to the underlying pixels (c).

Chen et al. [12] propose a text detection method using MSER. The outlines of MSER are enhanced by Canny edges. This makes MSER less sensitive to blur. Based on geometric cues these candidate character regions are then grouped to words and text lines. Neumann et al. [13] propose ERs for segmenting regions. ERs are extracted on the RGB, HSI and gradient images to retrieve character candidate regions. Instead of using heuristics as Epshtein et al. [9] for labeling text, an AdaBoost classifier based on geometric features is used. Text-CCs are then grouped to words. Yao et al. [14] generalizes the SWT for arbitrary oriented text. Camshift is used to find character orientation, length of major and minor axis and barycenter.

CCs are filtered and grouped to chains. False positive chains are eliminated by a Random Forest (RF) classifier.

B) Texture Based Methods:

Chen et al. [12] propose an AdaBoost classifier based on x- and y-derivative features, intensity features and edge-linking features computed in block-patterns of the detector-window (see Figure 3). Text-patches have low-entropy in these block-patterns.

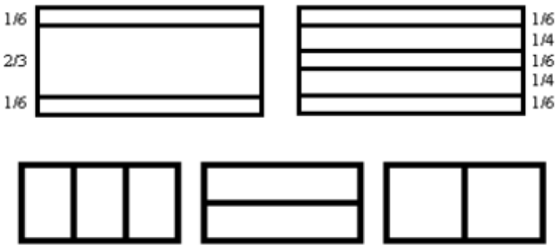


Fig 3: Block-patterns proposed from Chen et al. [12].

C) Hybrid approaches:

Minetto et al. [15] propose a hybrid approach where CCs are labeled with shape descriptors like zernike moments as text or non-text regions. CCs are then grouped to textline candidates. These candidates are then verified by discriminative classifiers using a novel Fuzzy HOG (F-HOG) feature set.

Gonzalez et al. [16] propose a hybrid localization method based on a combination of MSER and local adaptive thresholding for segmentation. CCs are filtered based on geometric features such as aspect ratio or occupy ratio. Hypotheses are created which are verified by a texture based classifier.

2.2 Text Recognition

Similar to text-detection systems, recognition systems can be divided into region-based and texture-based approaches. Texture-based approaches use low-level vision features to assign candidate labels to text-windows. On the contrary, region-based methods assign a candidate label to segmented CCs. Similar to texture-based methods; a language model is then used to predict the final word.

A) Region Based Methods

Neumann et al. [13] propose a recognition system which classifies MSER regions which are detected by a text-detection system as text-components. Boundaries of normalized MSER regions are inserted into separate images based on their orientation. In total 8 orientations are used. Each orientation image is filtered with a Gaussian and sub-sampled in a 5 X 5 image. Hence, a 5 X 5 X 8 = 200 dimensional feature vector is used for classifying MSER regions. To improve the prediction, a language model based on bigrams is used.

B) Texture Based Methods

Wang et al. [8] propose HOG features with a Random Ferns classifier to detect and classify text in an end-to-end setting. The multiclass-detector is trained on cropped synthetic and real-world letters. Non-maxima of the detector responses are suppressed. The remaining letters are then combined in a Pictorial Structure framework, where letters are parts of words. For each word in a dictionary, the most plausible character responses are found in the image. Detected words are then rescored based on geometric information and non-maxima suppression is done to remove overlapping word-responses.

Minetto et al. [15] propose, similar to Wang et al. [8], HOG features with SVMs to detect and classify text. Instead of using Pictorial Structures, CRFs are used to predict the final geometric information and classifier information to predict a candidate word. The final word prediction is done by retrieving the word in a dictionary with the shortest edit distance.

Yildirim et al. [11] propose Hough Forest (HF) for text recognition. Cross-scale binary features, which are thresholded differences of regions across different scales, are proposed for word recognition. For predicting the final word from candidate-responses a CRF energy function is minimized, which incorporates geometrical information and lexicon priors.

3. PRE PROCESSING TECHNIQUE

The pre-processing phase which is not defined in any other before techniques described in the above sections. The basic idea is to enhance the input image so that the accuracy of the text localization and recognition can be increased. Figure 4 represents the pre-processing flow.

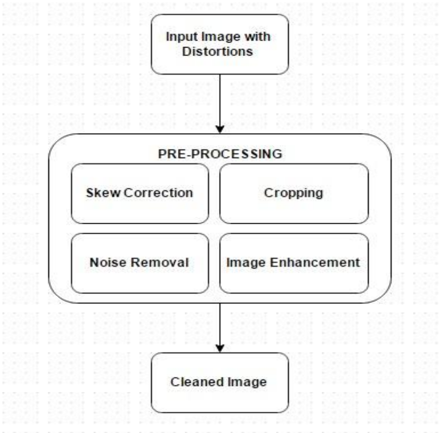


Fig 4: Pre Processing Steps

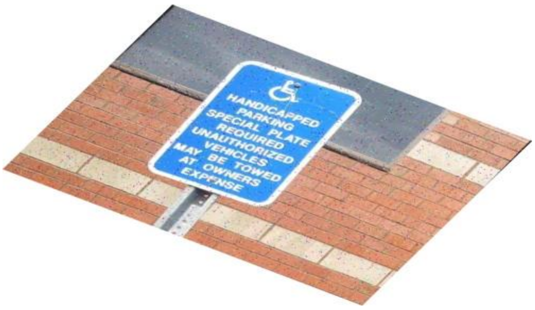


Fig 5: Input distorted image with skew and Noise

Skew Correction:

An important part of any document recognition system is detection and correction of skew in the image of a page. Page layout analysis and preprocessing operations used for character recognition depend on an upright image or, at least, knowledge of the angle of skew. One example of a process which is spoilt by skew is the use of horizontal and vertical projection profiles.

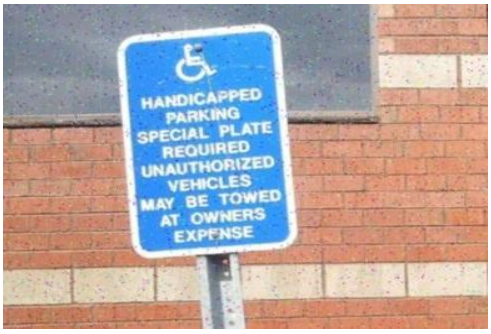


Fig 6: Image after Skew Correction

Noise Removal:

Digital images are prone to a variety of types of noise. Noise is the result of errors in the image acquisition process that result in pixel values that do not reflect the true intensities of the real scene. There are several ways that noise can be introduced into an image, depending on how the image is created. For example:

- If the image is scanned from a photograph made on film, the film grain is a source of noise. Noise can also be the result of damage to the film, or be introduced by the scanner itself.
- If the image is acquired directly in a digital format, the mechanism for gathering the data (such as a CCD detector) can introduce noise.
- Electronic transmission of image data can introduce noise.

After the pre-processing is completed the enhanced image is fetched as input to the next processes. The further stages can be processed for the enhanced image so that the recognition accuracy can be improved compared to the existing techniques. Figure 7 shows the enhanced image.

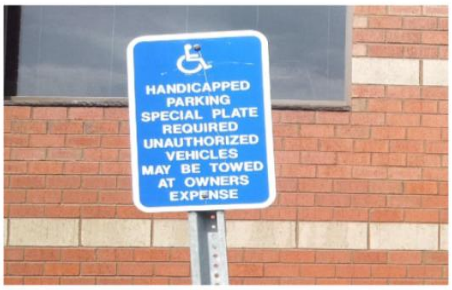


Fig 7: Enhanced Image after Pre-Processing

The image is then processed under various stages to get the desired output all those steps are depicted in the fig 8.

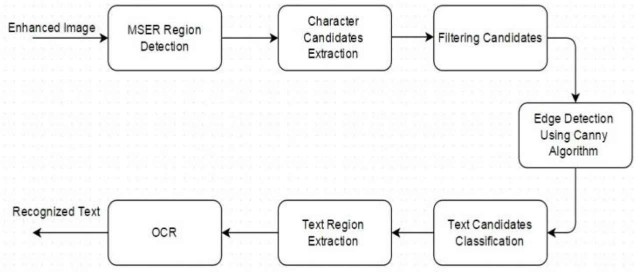

Enhanced Image
MSER Region Detection
Character Candidates Extraction
Filtering Candidates
Edge Detection Using Canny Algorithm
Text Candidates Classification
Text Region Extraction
OCR
Recognized Text

Fig 8: Flow of the Detection & Recognition Procedure

3.1 MSER Region Detection

As the intensity difference of text to its background is naturally considerable and a uniform intensity or color within every letter can be understood, MSER is a usual choice for text detection. While MSER has been recognized as one of the finest region detectors because of its robustness against scale, view point and lighting changes, it is perceptive to image blur. Thus, small letters may not be distinguished or detected in case of defocus or motion blur by using plain MSER to images of restricted resolution.

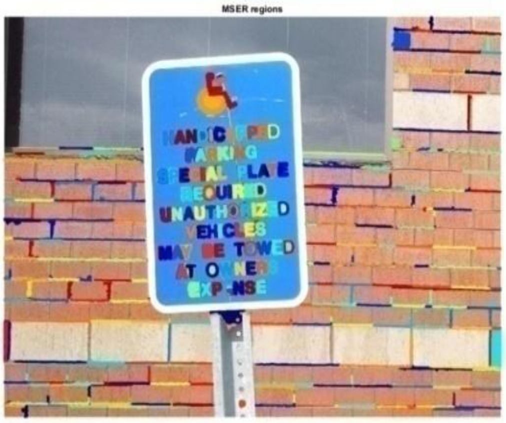


Fig 9 shows the MSER Regions.

3.2 Character Candidates Extraction:

Fortunately, rather than identifying the character, we can simply choose the one that is more possible to be characters in a parent-children relationship. Claimed that such pair wise relationships may not be sufficient to eliminate non-character MSERs, and pruning should exploit some complicated higher-order properties of text. Alternatively, this indicates that this probability can be fast estimated using our regularized variation scheme with reasonable accuracy. Figure 10 depicts the output of this stage.

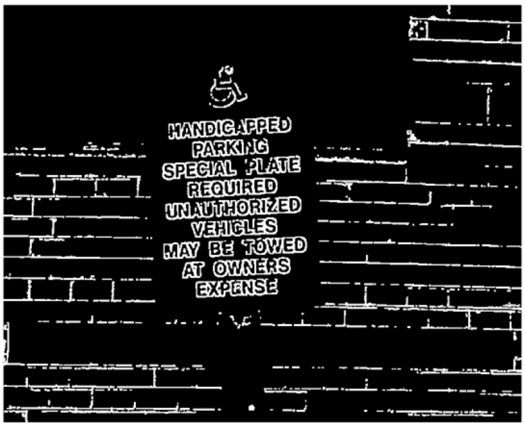


Fig 10: Character Candidates Extraction

3.3 Connected Component Analysis:

CCA is a well utilized algorithm in image processing that gathers and scans image pixels in labeled components depending on pixel connectivity [16]. An eight-point CCA step is used to find all the objects inside the binary image produced from the earlier stage. The output of this step is an array of N objects.

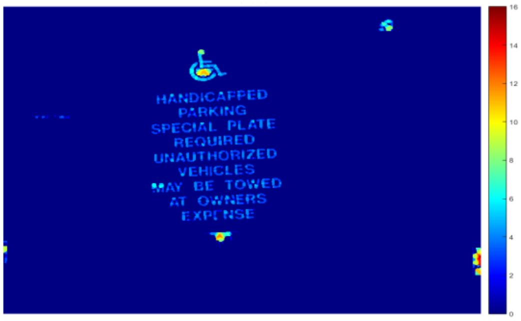


Fig 11 depicts the output of the CCA stage.

3.4 Edge Detection:

As hand written text is basically placed on white background, it has a propensity to give high reaction to edge detection. Additionally, a meeting point of edges with the MSER regions is going to give regions that are even more possible to fit in to text. Figure 12 depicts the edge detection.

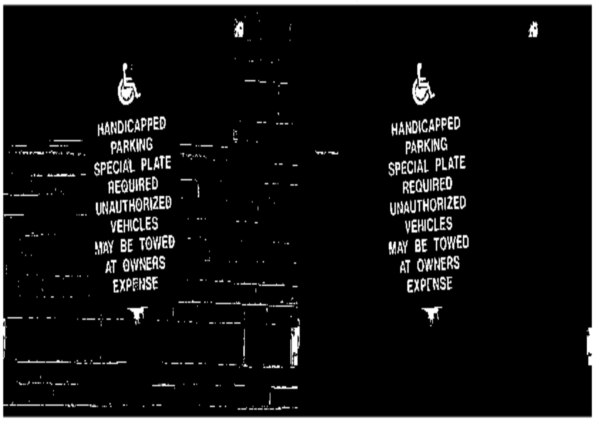


Fig 12: Edge Detection

3.5 Text Candidates Classification:

As it is hard to train an effective text classifier using such an unbalanced database, most of the non-text candidates need to be removed before training the classifier. The proposed method uses a character classifier to approximate the posterior probabilities of text candidates related to non-text and get rid of text candidates with high non-text probabilities. The following features are used to train the character classifier: text region height, width and aspect ratio, smoothness (defined as the average difference of adjacent boundary pixels' gradient directions) and stroke width features (including mean and variance of character stroke widths). Characters with small aspect ratios such as “i”, “j” and “l” are labeled as negative samples, as it is very uncommon that some words comprise many small aspect ratio characters .Figure 13 depicts the extracted characters.

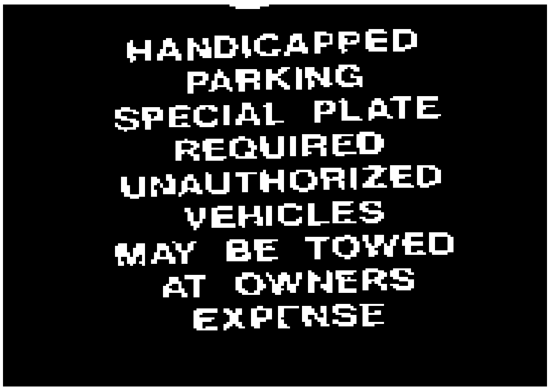


Fig 13: Classifying only Text

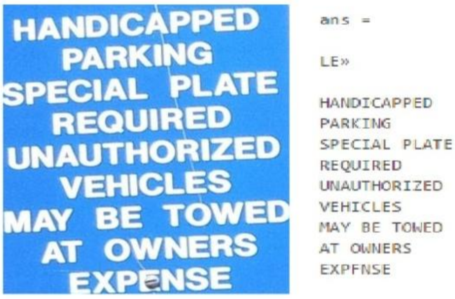
3.6 Text Region Extraction:

To calculate a bounding box of the text area, this paper initially merges the single characters into a unit connected element. This is completed using morphological closing. The region which contains the text will be outlined and extracted separately from the input image.

3.7 OCR:

The variation of text from a cluttered scene can efficiently improve OCR accuracies. Since the previous stages of the algorithm already extracted an appropriate segmented text region, this paper utilizes the binary text mask to further enhance the recognition accuracy

Figure 14 shows the extracted region with the appropriate OCR recognized characters.



```
ans =

LE»

HANDICAPPED
PARKING
SPECIAL PLATE
REQUIRED
UNAUTHORIZED
VEHICLES
MAY BE TOWED
AT OWNERS
EXPFNSE
```

Fig 14: Extracted Text Region and the Recognized Text

4. EXPERIMENTAL RESULTS

The pre processed technique is tested for two metrics which includes the recognition accuracy and the time taken to recognize the text. The testing is completely done on the images form ICDAR 2003 data set. For comparison we considered the traditional techniques. The below table shows the number of input images and the correctly recognized versus the faulty recognized inputs. The results are shown in both table 1 and figure 13.

TABLE -1 Comparison of Recognition Accuracies between Various Techniques.

Techniques Used	***No. Of Inputted Images***	***Recognized***	***Faults***
MSER Detection	***100***	***90***	***10***
SWT		***92***	***8***
Hybrid Approach		***95***	***5***
SVMs		***88***	***12***
Hough Forest		***87***	***13***
Pre processed method		***97***	***3***

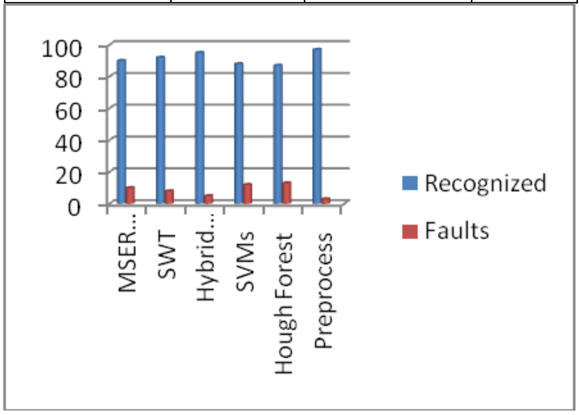


Fig 13: Faults Vs Recognition accuracies

Due the image pre processing stage the image is enhanced to an extent where the later processing stages have improved the accuracy, where as the existing techniques lacks this pre processing and the processing of the distorted images led them to deduce the recognition accuracy.

5. CONCLUSION:

This paper presents a new scene text detection method with several novel techniques. First, this technique uses the pre-processing for enhancing the input image for better accuracies and then it utilized the fast and accurate MSERs pruning algorithm that enables us to detect most characters even when the image is in lesser quality. Second, we utilize a traditional self-training distance metric learning method that can learn clustering threshold and distance weights simultaneously; text candidates are constructed by clustering character candidates by the single-link algorithm using the learned parameters. Third, we put forward to use a character classifier to estimate the posterior probability of text candidate corresponding to non-text and eliminate text candidates with high non-text probability, which helps to build a more powerful text classifier. Finally, by integrating the above new techniques, we build a robust scene text detection system that exhibits superior performance over state-of-the-art methods on a variety of public databases. The system is evaluated on the ICDAR datasets, outperforming state-of-the-art methods in text detection, recognition and end-to-end dictionary driven scene text detection. Hence, the research question can be answered: it is possible to achieve competitive results with a combination of local features; region based approaches and learned local features.

REFERENCES:

[1] Xu-Cheng Yin, Xuwang Yin, Kaizhu Haung, and Hong-Wei Hao. Robust text detection in natural scene images, May 2014.

[2] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. ICDAR 2003 robust reading competitions. In ICDAR 2003, page 682, 2003.

[3] X. Lin. Reliable OCR solution for digital content re-mastering. In Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Dec. 2001.

[4] [4] X. Chen and A. L. Yuille. Detecting and reading text in natural scenes. CVPR, 2:366{373, 2004.

[5] Y.-F. Pan, X. Hou, and C.-L. Liu. A robust system to detect and localize texts in natural scene images. In Document Analysis Systems, 2008. DAS '08. The Eighth IAPR International Workshop on, pages 35 -42, sept. 2008.

[6] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In CVPR 2010, pages 2963 {2970, 6 2010.

[7] L. P. Sosa, S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. ICDAR 2003 Robust Reading Competitions. In Proceedings of the International Conference on Document Analysis and Recognition, pages 682–687. IEEE Press, 2003.

[8] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In Proceedings of the IEEE International Conference on Computer Vision, pages 1457–1464, Barcelona, Spain, 2011.

[9] B. Epshtein, E. Ofek, and Y.Wexler. Detecting Text in Natural Scenes with Stroke Width Transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2963–2970. IEEE, 2010.

[10] M. Anthimopoulos, B. Gatos, and I. Pratikakis. Detection of Artificial and Scene Text in Images and Video Frames. Pattern Analysis and Applications, 16(3):431–446, 2013.

[11] C. Yi and Y. Tian. Text String Detection from Natural Scenes by Structure-Based Partition and Grouping. IEEE Transactions on Image Processing, 20(9):2594–2605, 2011.

[12] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod. Robust Text Detection in Natural Images with Edge-Enhanced Maximally Stable Extremal Regions. In International Conference on Image Processing, pages 2609–2612. IEEE, 2011.

[13] L. Neumann and J. Matas. A Method for Text Localization and Recognition in Real-World Images. In Proceedings of the Asian Conference on Computer Vision, pages 770–783. Springer, 2011.

[14] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu. Detecting Texts of Arbitrary Orientations in Natural Images. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1083–1090, 2012.

[15] R. Minetto, N. Thome, M. Cord, J. Stolfi, F. Precioso, J. Guyomard, and N. J. Leite. Text Detection and Recognition in Urban Scenes. In International Conference on Computer Vision Workshops, pages 227–234. IEEE, 2011.

[16] Gonzalez, L.M. Bergasa, J.J. Yebes, and S. Bronte. Text Location in Complex Images. In International Conference on Pattern Recognition, pages 617–620, 2012.

Authors' Profiles:

M. Fouzia is currently pursuing her M.Tech degree in Computer Science and Engineering with specialization in Artificial Intelligence from Jawaharlal Nehru Technological University, Anantapur, India. She did her B.Tech Degree in Information Technology from KSRM college of engineering KADAPA, India in 2013.

C. Shoba Bindu is an Associate Professor of Computer Science and Engineering at Jawaharlal Nehru Technological University College of Engineering, Ananthapuramu. She obtained her Bachelor degree in Electronics and Communication Engineering, Master of Technology in Computer Science from Jawaharlal Nehru Technological University Hyderabad and Ph.D. in Computer Science and Engineering from Jawaharlal Nehru Technological University Anantapuramu. She has published several Research papers in National / International Conferences and Journals. Her research interests includes network security and Wireless communication systems.