Indian Journal of Communication Engineering and Systems (IJCES)
Vol.3.No.1 2015 pp 118-126.
Available at: www.goniv.com
Paper Received: 04-04-2015
Paper Accepted: 14-04-2015
Paper Reviewed by: 1. R. Venkatakrishnan 2. R. Marimuthu
Editor: Prof. P.Muthukumar

EFFICIENT TEXT AND WORD LEVEL RECOGNITION FROM NATURAL IMAGES

S.Chithira Priya[1] and R.Vijay[2]
PG Student[1], Assitant Professor[2]
Department of Computer Science and Engineering[1]
MNSK College of Engineering[1]
Pudukkottai (TN), India.
aishu.may7@gmail.com

ABSTRACT

Extracting text from photographs and videos is a challenging problem that has received a significant amount of attention. Scene text recognition has inspired great interests from the computer vision community in recent years. It is difficult to extract text from images and videos with text character and interference at the background. Recognizing text from natural image is a difficult task, even more so than the detection of text from the scanned documents. To evaluate the performance of recent algorithms in detecting and recognizing text from complex images in this proposed paper two methods are implemented, text detection and text recognition. In text detection, contrast map is binaries through median filter and merged with Canny's edge map to spot the text stroke edge pixels with feature extraction. In Text Recognition, Text understanding and Text Retrieval schemes are used. First step is to Coach Character recognizer to know the class of a character class in picture patch. The features detectors such as Harris-Corner, Maximal Stable Extremal Regions (MSER) and dense sampling and Histogram of Oriented Gradients (HOG) descriptors are used. Second step is to generate a binary classifier for each character class in text retrieval. Segmentation based word level recognition is also implemented with the help of lexicon analysis with best results. Then finally, recognized text is converted into voice format.

Index Terms— Text detection; text recognition; text understanding; text retrieval.

I. INTRODUCTION

Research in document analysis and recognition has traditionally focused on processing and analyzing scanned documents. Detecting text in natural images, as different to scanning of printed pages, faxes and business cards, is a main step for an amount of Computer Vision applications, such as automated aid for visually impaired, regular geocoding of businesses, and robotic map-reading in urban environments. Mobile visual search has gained popular interest with the increasing availability of high-performance, low-cost camera-phones. In recent years, image search systems have been developed for applications such as product

recognition and landmark recognition .In these cases, text can be extracted as a high-level feature to complement low-level visual features in content-based image/video retrieval systems. It can also be used in a wide range of other multimedia applications such as mobile image search, sign conversion and business name investigate in street-level images. The difficulties come from the fact that characters embedded in the scene can appear in some fonts, with some colors, and on messy backgrounds. Natural scene display board images contain text information which is often required to be automatically recognized and process. Scene text may be any textual part of the outlook images such as names of streets, institutes names, names of shops, construction names, corporation names, street signs, traffic information, notice signs etc. Researchers have alerted their attention on development of techniques for understanding text on such display boards. MSERs denote a set of distinguished regions, which are defined by an Extremal property of its intensity function in the region and on its outer boundary. In addition, MSERs have all the properties required of a stable local detector. Recently, Maximally Stable Extremal Regions (MSERs) based text detection has been widely explored. The main advantage of these approaches over other component based approaches is focused in the effectiveness of using MSERs as character/component candidates. It is based on the observation that text components usually have higher color contrast with their backgrounds and tend to be form homogenous color regions, at least at the text level. The MSER algorithm adaptively detects stable color regions and provides a good solution to localize the components without explicit binarization. Text detection and recognition in natural scene images has recently received increased attention of the computer vision community. Text is a pervasive element in many environments, solving this problem has potential for significant impact.

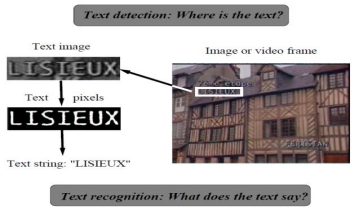

Fig 1. Text Detection and Recognition

II. RELATED WORK

In this section, we present a review of previous works involved in text recognition. Text detection aims to filter out nontext outliers to localize text regions from cluttered background [3],[7], text recognition is to convert to transform image-based text information in the detected regions into readable text format. [1] In this paper, we introduce a new skeleton pruning method based on contour partitioning. Any contour partition can be used, but the partitions get by Discrete Curve Evolution (DCE) yield best results. Again, many existing methods displace skeleton points in order to introduce pruned skeletons.

Fig 2. Examples of Natural Images with text

In scene images, the best recognition rate was only about 41.2%. optical character recognition (OCR) systems [2], can achieve almost perfect recognition rate on printed text in scanned documents, but cannot accurately recognize text information directly from camera-captured scene images and videos, and are usually sensitive to font scale changes and background interference which widely exists in scene text. Although some OCR systems have started to support scene character recognition,

the recognition performance is still much lower than the recognition for scanned documents. Moreover, as OCR is not considered as a black box, several outputs are taken into account to intermingle recognition and correction steps. Based on a public database of natural scene words, detailed results are also presented along with future works. Many algorithms were proposed to improve scene-image-based text character recognition. Text characters from categories are distinguished by boundary shape and skeleton structure.

In [4], we apply methods recently developed in machine learning - specifically, large-scale algorithms for learning the features automatically from unlabeled data -- and show that they allow us to construct highly effective classifiers for both detection and recognition to be used in a high accuracy end-to-end system. In [5], the question of feature set of object recognition, adopting linear SVM based human detection. After reviewing existing edge and gradient based descriptors, grids of Histograms of oriented gradient (HOG) descriptors significantly outperform existing feature sets for human detection. [15] adopted conditional random field to combine bottom-up character recognition and top-down word-level recognition. [13] modeled the inner character structure by defining a dictionary of basic shape codes to perform character and word retrieval without OCR on scanned documents.

III. TEXT LEVEL RECOGNITION

A. *TEXT DETECTION*

1) Text Localization:

a) *Color Decomposition:* To decompose a scene image into several color-based layers, we have designed a boundary clustering algorithm based on bigram color uniformity. We define the uniformity of their color difference as bigram color uniformity. Text information is generally attached to a plane carrier as attachment surface with uniform colors respectively. Color difference is related to the character boundary, which serves as a border between text strokes and the attachment surfaces. We then model color difference by a vector of color pair, obtained by cascading the RGB colors of text and attachment surfaces. Each boundary can be explained by a color-pair, and we cluster the boundaries with similar color pairs into the sample layer. The boundaries of text characters are separated from those of background outliers, as shown in Fig. 3.The colored background of the image is separated from the image and the text characters are highlighted.

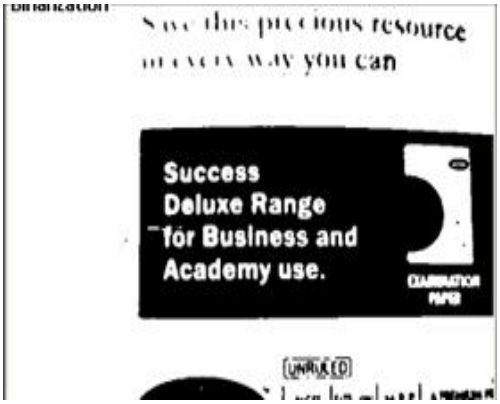

Fig 3. Color Decomposition of scene image

b) Horizontal Alignment: Analyze geometrical properties of the boundaries to detect the existence of text characters in each color layer. Thus we design an adjacent character grouping algorithm to search for image regions containing text strings. In order to extract text strings in slightly non-horizontal orientations (Fig.3), we search for possible characters of a text string within a reasonable range of horizontal orientation. When estimating horizontal alignment, we do not require all the characters exactly align in horizontal orientation, but allow some differences between neighboring characters that are assigned into the same string. In our system we set this range as $\pm\pi/6$ degrees relative to the horizontal line. This range could be set to be larger but it would bring in more false positive strings from background. In addition, our scene text detection algorithm can handle challenging

font variations, as long as the text has enough resolutions.

Fig 4.Adjacent character grouping system using Horizontal Alignment

2) Feature Extraction: Analyze text strokes using feature extraction algorithms such as Harris-Corner, Maximal Stable Extremal Regions (MSER), dense sampling and Random sampling. In computer vision and image processing the concept of feature detection refers to methods that aim at computing abstractions of image information and making local decisions at every image point whether there is an image feature of a given type at that point or not. The resulting features will be subsets of the image domain, often in the form of isolated points, continuous curves or connected regions. Feature detection is a low-level image processing operation. That is, it is usually performed as the first operation on an image, and examines every pixel to see if there is a feature present at that pixel. If this is part of a larger algorithm, then the algorithm will typically only examine the image in the region of the features. As a built-in pre-requisite to feature detection, the input image is usually smoothed by a Gaussian kernel in a scale-space representation and one or several feature images are computed, often expressed in terms of local derivative operations.

a) *Harris Corner:* A corner can be defined as the intersection of two edges or a point. It is junctions of curves. Generally corner points are more stable features over changes of viewpoint. Corner detection is widely used in computer vision application such as motion detection, image matching, tracking. Harris corner detector is used to extract the corner points. The Harris corner detector is a popular interest point detector. Because there is no effect of rotation, scale, illumination variation, and image noise on the performance of Harris corner detector. It is based upon the local auto-correlation function of a signal, where the local auto-correlation function measures the local changes of the signal with patches shifted by a small amount in different directions.

b) Maximally Stable Extremal Regions (MSERs): Extremal Regions Detector, here a new set of image elements that are put into correspondence, the so called Extremal regions. Extremal regions possess highly desirable properties: the set is closed under 1. Continuous transformation of image coordinates 2. Monotonic transformation of image brightness. Detecting Extremal regions: detect anchor points .Anchor points detected at multiple scales are local extremas of intensity .Explore image around rays from each anchor point. Go along every ray starting from this point until an extremum of function f is reached. It is found on the surveillance that text components typically have higher color contrast with their backgrounds and be predisposed to be type homogenous color area, at slightest at the character level. All points create some irregularly-shaped region. Approximately corresponding regions are obtained for affine-transformed regions. The MSER algorithm adaptively notice steady color regions and offers a fine explanation to focus the components with no unambiguous binarization. MSER regions are of all-purpose, data-dependent shape, i.e. composite adequate to offer enough restraint to describe affine frames. They are related, randomly shaped, probably nested, and do not cover the whole picture, i.e. they do not form a separation. The speedup of the computation time and the improvement of the detection and tracking stability are evaluated. The edge-enhanced MSER detected in the query image can be used to extract feature descriptors like for visual search.

c) Dense Sampling: Densely sampled image patches and then apply for each feature a local optimization of the position and scale within a bounded search area. This way, we get

dense coverage of the entire scene and clearly defined spatial relations, as in the case of dense sampling, yet with improved repeatability, as in the case of interest points.

d) Random Sampling: Random sample theory is an effective tool for detecting features in images. This paper presents an adaptive random sampling scheme that clusters random samples into candidate features. The required trial number is reduced by adaptive sampling, thereby reducing the run time of the algorithm.

B. HISTOGRAM OF ORIENTED GRADIENTS

Histogram of Oriented Gradients (HOG) is attribute descriptors utilized in computer vision and image processing for the principle of entity discovery. The method calculates occasion of grade direction in restricted segment of a picture. Local thing facade and form inside a picture can be explained by the allocation of concentration gradients or edge directions. The execution of these descriptors can be attained by separating the picture into tiny related regions, called cells. For every cell accumulating a histogram of gradient directions or edge orientations for the pixels inside the cell. The grouping of these histograms then symbolizes the descriptor. For enhanced accurateness, the local histograms can be contrast-normalized by manipulative a measure of the passion across a larger region of the picture, called a block. Now utilizing this value to regularize all cells inside the block. These normalization consequences in improved invariance to modify in clarification or shadowing.

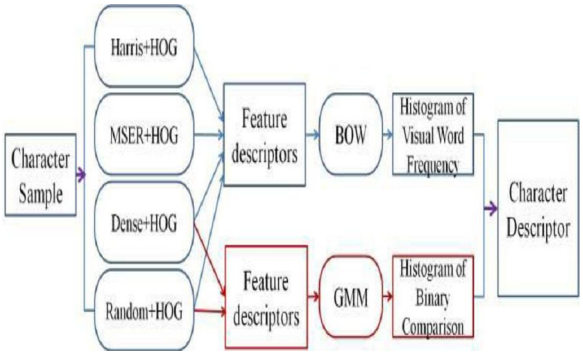

Fig 5. Usage of HOG in all Detectors

C. TEXT RECOGNITION

After detecting text region in the image, from that text region text is extracted from the image using character descriptors and structure configuration. These methods used to convert images with text into editable formats and processes input images with text and get editable documents like TXT file. It employs four types of keypoint detectors, Harris detector (HD) to extract keypoints from corners and junctions, MSER detector (MD) to extract keypoints from stroke components, Dense detector (DD) to uniformly extract keypoints, and Random detector (RD) to extract the preset number of keypoints in a random pattern. At each of the extracted keypoints, the HOG feature is calculated as an observed feature vector in feature space. HOG is selected as local features descriptor because of its compatibility with all above keypoint detectors. In the process of feature quantization, two models are used to aggregate the extracted feature. They are,

- Bag-of-Words Model (BOW)
- Gaussian Mixture Model (GMM)

1) ***Bag-of-Words Model (BOW):*** BOW is applied to keypoints from all the four detectors. At each feature detector, build a vocabulary of 256 visual words. This number is experimentally chosen to balance the performance of character recognition and the computation cost. At a character patch, the four detectors are applied to extract their respective keypoints, and then their corresponding HOG features are mapped into the respective vocabularies, obtaining four frequency histograms. Each histogram has 256 dimensions. Cascade the four histograms into BOW-based feature representation in $256 \times 4 = 1024$ dimensions.

2) ***Gaussian Mixture Model (GMM):*** GMM is applied to those only from DD and RD, because GMM-based feature representation requires fixed number and locations of the keypoint all character patch

samples, while the numbers and locations of keypoints from HD and MD depend on character structure in the character patches. The combination of BOW-based and GMM-based feature representations improve the performance on scene character recognition, which is prepared for text understanding

In text retrieval application, the query character class is considered as an object with fixed structure, and we generate its binary classifier according to structure modeling. Character structure consists of multiple oriented strokes, which serve as basic elements of a text character. From the pixel-level perspective, a stroke of printed text is defined as a region bounded by two parallel boundary segments. For a character patch, we generate its visual word histogram as BOW-based feature representation, and binary comparison histogram as GMM-based feature representation. Then the two feature representations are cascaded into the character descriptor of the patch. It is used as a feature vector of character patch to train text character recognizer in SVM model. The combination of BOW-based and GMM-based feature representations improve the performance on scene character recognition, which is prepared for text understanding. Their orientation is regarded as stroke orientation and the distance between them is regarded as stroke width. Histogram as feature representation. Then their corresponding character classifiers are invoked to confirm the character classes. If most of the queried characters exist, the text retrieval application will provide positive response, otherwise provide negative response.

IV. VIDEO CONVERSION

Video is a rich information source than images. Videos are made of frames .Each Video has its own Characteristics such as Commercials, News, Sports. Video processing systems require a stream processing architecture, in which video frames from a continuous stream are used one or more at a time. The user uploads the video. The Video can be obtained for lesions of any size, shape, and composition in an acceptable amount of time. Motion of both Object and Camera create more difficulties. Noise Removal Filter is used to remove the noise to improve the video quality and segment the video based on similarities. Each frame of the video is converted into individual images using video file reader. Each frame has specific size. Dynamic frames are created at video uploaded time. Frames are used to create datasets. Text from the frames is extracted by text recognition technique.

V. WORD LEVEL RECOGNITION

In Segmentation based Word level recognition using lexicon Analysis, word image is first segmented into individual character images. Features are extracted from the segmented character images and represented by feature vectors. These character feature vectors are then concatenated and matched with similar feature vectors for lexicon words. Hence the character features are compared under the contextual constraints represented by a lexicon. The features can be as simple as pixel values. Each segmented character is normalized to a 24x24 grid. The pixel values of each segmented character are then concatenated to form a word feature vector. Thus a word segmented into 4 Characters has a feature vector of 24x24x4=2304 elements. In the matching process, a distance measure is computed for word feature vectors of the same length. To allow for segmentation errors, lexicon words with lengths one more or one less than the input image are also matched. This is done by deleting one character at a time from different positions of the longer vector. The distance is the number of different elements divided by the length of the words compared. Since character segmentation is correct in many cases, a weight is to decrease the distance scores for equal length vectors and increase the scores for unequal length vectors. A ranking of the lexicon is then produced by sorting the words in order of increasing distance.

VI. EXPERIMENTAL RESULTS

Our experimental system in java platform gives us some insight into algorithm design and performance improvement of scene text extraction. First, the assumptions of horizontal alignment in text layout analysis make sense in applications. The accuracy of scene text detection could be improved by using the intersections of extracted text regions from consecutive frames captured by the camera at an identical scene. We can acquire text information from natural scene captured by camera images or videos to understand surrounding environment and objects. To evaluate the performance using following rates such as accuracy rate (AR) and false positive rate (FPR), are calculated to evaluate the performance of queried character classification.AR represents the ratio between the number of correctly recognized text characters and the total number of text characters. FPR represents the ratio between the number of incorrectly predicted negative samples and the total number of negative samples.

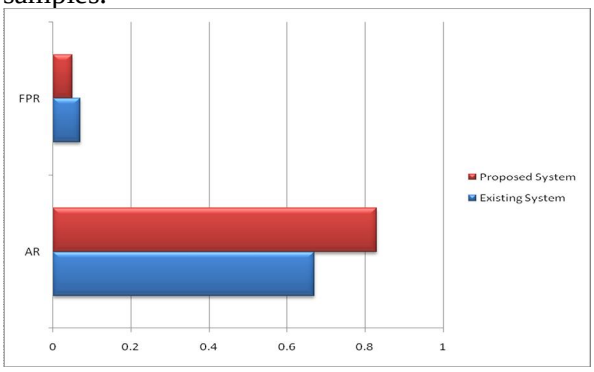

Fig 6. Performance Comparison between Existing & Proposed System

VII. CONCLUSION

Text detection in natural scene images and videos remains a challenging problem due to complex background, low image quality and/or variation of text appearance. In proposed presented a technique of scene text recognition from identify text regions, which is well-matched with mobile applications. It identifies text regions from image or video and distinguishes text information from the identify text regions. In scene text detection, describe analysis of color disintegration and horizontal alignment is performed to search for image regions of text strings. In scene text recognition, two methods, text understanding and text retrieval, are correspondingly proposed to take out text information from surrounding location. In videos, video file reader is used to convert the video frames into images. Recognized text are converted into editable documents such as word or text documents and then converted into voice format .Using lexicon analysis, word level recognition from real time images, videos and browsed images are also implemented with best results.

REFERENCES

[1] X. Bai, L. J. Latecki, and W.-Y. Liu, “Skeleton pruning by contour partitioning with discrete curve evolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 3, pp. 449–462, Mar. 2007.

[2] R. Beaufort and C. Mancas-Thillou, “A weighted finite-state framework for correcting errors in natural scene OCR,” in Proc. 9th Int. Conf. Document Anal. Recognit., Sep. 2007, pp. 889–893.

[3] X. Chen, J. Yang, J. Zhang, and A. Waibel, “Automatic detection and recognition of signs from natural scenes,” IEEE Trans. Image Process., vol. 13, no. 1, pp. 87–99, Jan. 2004.

[4] A. Coates et al., “Text detection and character recognition in scene images with unsupervised feature learning,” in Proc. ICDAR, Sep. 2011, pp. 440–445.

[5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2005, pp. 886–893.

[6] T. de Campos, B. Babu, and M. Varma, “Character recognition in natural images,” in Proc. VISAPP, 2009.

[7] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in Proc. CVPR, Jun. 2010, pp. 2963–2970.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models,"IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[9] T. Jiang, F. Jurie, and C. Schmid, "Learning shape prior models for object matching," in Proc. CVPR, Jun. 2009, pp. 848–855.

[10] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2594–2605, Sep. 2011.

[11] C. Yi and Y. Tian, "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4256–4268, Sep. 2012.

[12] Y. Liu, J. Yang, and M. Liu, "Recognition of QR code with mobile phones," in *Proc. CCDC*, Jul. 2008, pp. 203–206.

[13] S. Lu, L. Li, and C. L. Tan, "Document image retrieval through word shape coding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1913–1918, Nov. 2008.

[14] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. Int. Conf. Document Anal. Recognit.*, Aug. 2003, pp. 682–687. 2982 IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 7, JULY 2014.

[15] A. Mishra, K. Alahari, and C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1063–6919.

[16] N. Nikolaou and N. Papamarkos, "Color reduction for complex document images," *Int. J. Imag. Syst. Technol.*, vol. 19, no. 1, pp. 14–26, 2009.

[17] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3538–3545.

[18] E. Ohbuchi, H. Hanaizumi, and L. A. Hock, "Barcode readers using the camera device in mobile phones," in *Proc. Int. Conf. Cyberworlds*, Nov. 2004, pp. 260–265.

[19] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1491–1496.

[20] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene text recognition using part-based tree-structured character detection," in *Proc. CVPR*, Jun. 2013, pp. 2961–2968.

[21] P. Shivakumara, W. Huang, and C. L. Tan, "An efficient edge based technique for text detection in video frames," in *Proc. IAPR Workshop Document Anal. Syst.*, Sep. 2008, pp. 307–314.

[22] D. L. Smith, J. Feild, and E. Learned-Miller, "Enforcing similarity constraints with integer programming for better scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 73–80.

[23] R. Smith, "An overview of the tesseract OCR engine," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2007, pp. 629–633.

[24] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[25] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2010.

[26] K. Wang, B. Bbenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011,pp. 1457–1464.

[27] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1733–1746, Oct. 2009.

[28] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1083–1090.

[29] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2594–2605, Sep. 2011.

[30] C. Yi and Y. Tian, "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4256–4268, Sep. 2012.

[31] C. Yi, X. Yang, and Y. Tian, "Feature representations for scene text character recognition: A comparative study," in *Proc. 12th ICDAR*, Aug. 2013, pp. 907–911.

BIOGRAPHY

S.Chithira Priya is a Currently pursuing M.E in the Department of Computer Science and Engineering, MNSK College of Engineering, Pudukkottai, Tamil Nadu, India. She received Bachelor of Technology in Information Technology in 2011 from Sudharsan Engineering College,Pudukkottai,TamilNadu, India. Her research interests are Image Processing and Networking.She had attended National and International Conferences and got Best Paper Award too