Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.314 – 323

RESEARCH ARTICLE

Effective Techniques for the Detection, Extraction and Conversion of Devanagari Text from Traffic Panels

Miss. Gayatri H. Khobaragade

Department of Computer science & Engg.

G.H.R.A.E.T. Nagpur, India

Prof. Deepak Kapgate

Department of Computer science & Engg.

G.H.R.A.E.T. Nagpur, India

Abstract: The main objective of this paper describes methods for panel detection and recognition from the road side authorized traffic panel board (Text). It detects exact Green & yellow traffic panel from all the ordinary panels especially Devanagari panels. Input images can be the no. of warning panels or green panels and it is invariant to size so that it is scaled. After that the whatever information contain on input panel will be separated along with bounding boxes so that system can be easily understand the textual and non textual part using object segmentation and object extraction as well, then using OCR text will be recognize and extract also construct small additional dictionaries to increase the accuracy of the OCR and finally with the help of N-gram concept we can able to convert the extracted Devanagari text (Hindi) into international language(English).

Keywords: panel detection, OCR, RGB, HSV, N-gram algorithm, object extraction, color segmentation

I. Introduction:

In image processing blob analysis often required to check some objects' shape and depending on the shapes it perform further object segmentation of a particular object or not. Identification of road side traffic panels correctly at the right place is very important for car drivers to make sure themselves and safe journey. But sometimes, due to the change of environmental condition or different viewing angles, traffic panels may be look difficult until it is too late. As there is increase the accident ratio due to the poor visibility and language problem of the road side traffic panels because in India most of the panels in Devanagari languages (like Tamil, Marathi, Hindi, Sanskrit, Telgu etc.) so that some of the peoples and especially foreigners cannot understand the meaning even can't read it therefore our system convert it into international language for their convenience. Once the correct panel detected by the system then using object extraction and segmentation all the objects

will be box out for the better understanding i.e. separate text from input image after that text will be recognize and extract the only textual part by using OCR as well as our small dictionary for the better accuracy. OCR analyzes a static text document and the overall process is actually "off-line" process. OCR can able to capture live motions like the direction, the order in which segments are drawn and the way of putting down the pen how we lifting it. Near-neighbour analysis" can make use of co occurrence frequencies to correct errors. At last final text convert it into English language using the concept of N-gram algorithm. Following fig shows the overall system architecture including all the phases.

The rest of the paper is organized as follows: literature survey is in section 2. Proposed techniques for panel detection, text detection, text recognition and extraction & conversion are in section 3. Section 4 presents output snaps of the overall project and result analysis. & finally section 5 presents conclusions and future work.

Following figure shows the overall system architecture:

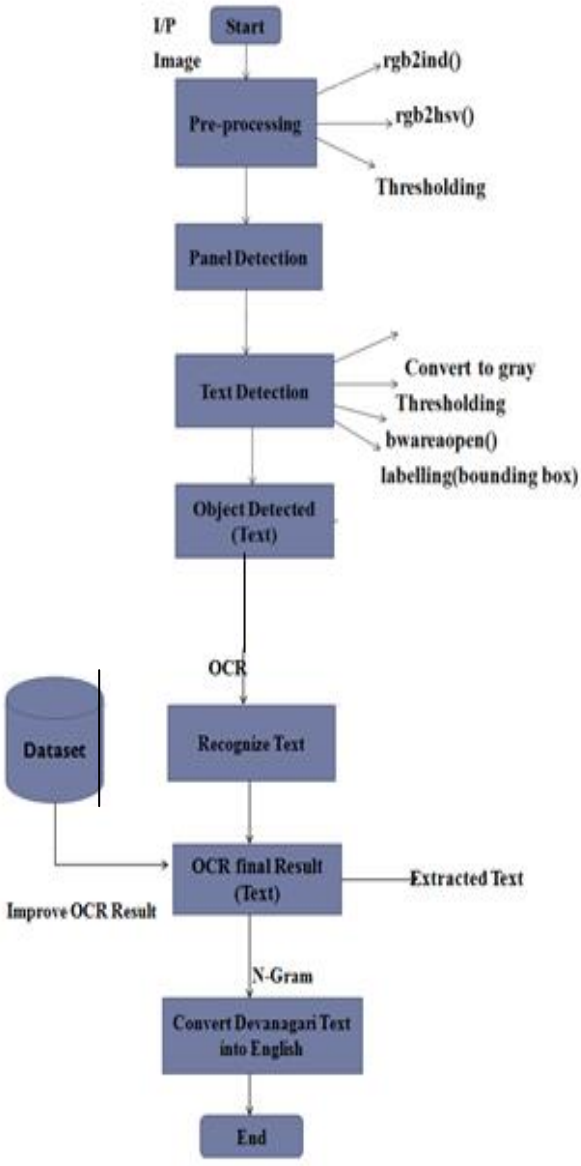


Fig: System Architecture

II. Literature Survey:

Off line Handwritten Character Recognition can be implementing by two ways that is, online character recognition and Offline character recognition and off-line character recognition will be further divided into machine printed and handwritten character recognition respectively. Off-line handwritten Devanagari character recognition method proposed by Mahesh Jangid et al. [4] that uses three feature extraction techniques based on recursive sub divisions of the character image, zone density of the pixel and leading distribution of neighbouring background to foreground pixels[14].

Optical Character Recognition is a system that provides a full alphanumeric recognition of handwritten and printed characters at electronic speed just by scanning the text document [9]. Text Documents can be scanned using a scanner or we can also use capture image then scan it using scanner and then given to the OCR systems. OCR then recognizes the words or characters in the scanned document & converts them into ASCII data. [9]

III. Proposed Approach:

Our Proposed methodology consists of four steps (1) Panel Detection. (2) Text Detection. (3) Text Recognition & Extraction. (4) Text Conversion.

3.1 Panel Detection:

The objective of this system is to detect the presence or the absence of green-background and yellow-background road traffic panels; it can be located on the right/left side of the highway road or above the road. System identify the traffic panels and warning panels using color segmentation(mapping).it divide the input colour image into three channels(R,G,B)of 8 mega-pixels of each especially for green-panels and HSV(Hue-saturation-Value) for warning panels. Color characterized by three quantities: Hue- Dominant color as perceived by an observer (red, orange, or yellow), Saturation- Relative purity of color; pure spectrum colors are fully saturated (it is 1/α to the white light), Brightness- Achromatic notion of intensity it includes Chromaticity and Tri stimulus values []. Pre-processing step is must in panel detection it includes following steps:

1. Rgb image will be converted into indexed image i.e. rgb2ind ()
2. Concatenate all the RGB channels of 8 mega-pixels of each.
3. Then again convert RGB image into HSV using rgb2hsv () function in matlab.
4. Set a particular threshold value of the both the HSV and RGB panels.

The thresholding is an important part of the color-level mapping class. It actually converts a grey image into a binary image that is an of 0/1 format. A binary image is usually stored as an unsigned byte of wave type. Sometimes this may appear to be wasteful, then it has advantages in terms of speed and also in allowing us to use some bits of each i.e bits can be turned on or off for binary masking. The threshold operation, producing the binary threshold image & correlation value for the threshold quality []. Following are some snaps of the panel detection stage in matlab:

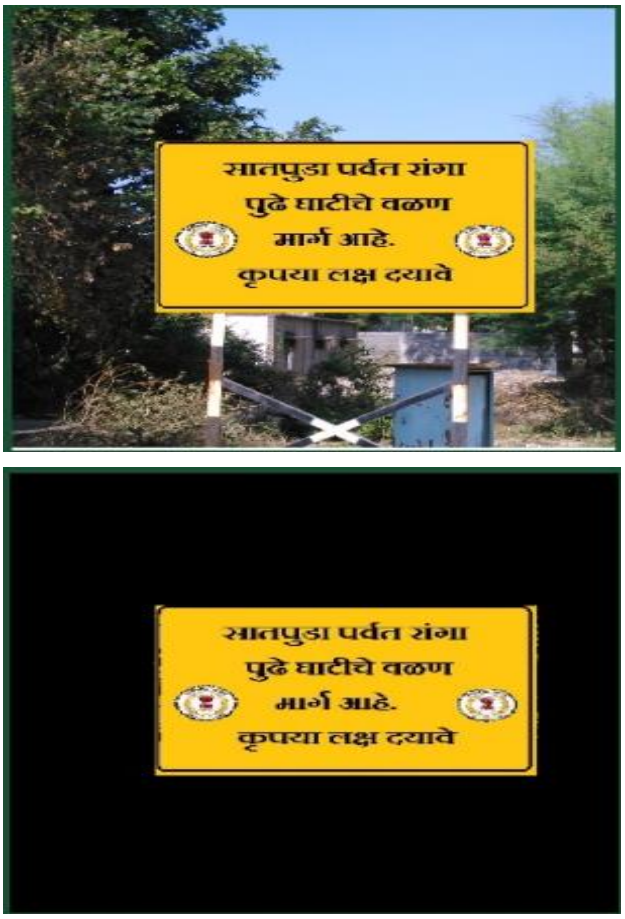


Fig1: (a) Input image (b) panel detected image

3.2 Text Detection:

K.P.Adhiya introduced Devanagari text detection methods from color image et.al [2]. Here we present a methodology for detecting text from printed colour document and recognize Devanagari printed Script (in Hindi language) from final detected text.

For that we used object segmentation & extraction approach for detecting the static text from image. The resultant text image document will be passing to OCR for (reorganization) Identification purpose. In text detection it actually checks the presence or absence of the text (object) for that here system use object segmentation and object extraction.

With the help of object segmentation we can separate all the objects (text) and using bwlabel () function labelled the connected components respectively. As well as we remove all the small objects using bwareaopen () function in matlab so that all the small objects which has less than 30 pixels will be automatically removed from the image. For that we have to convert the rgb image into gray image and using im2bw () convert it into black & white image as well [2].

Fig: 2 (a) Original image (b) Text Detected image

Following are the steps of Text Detection in matlab:

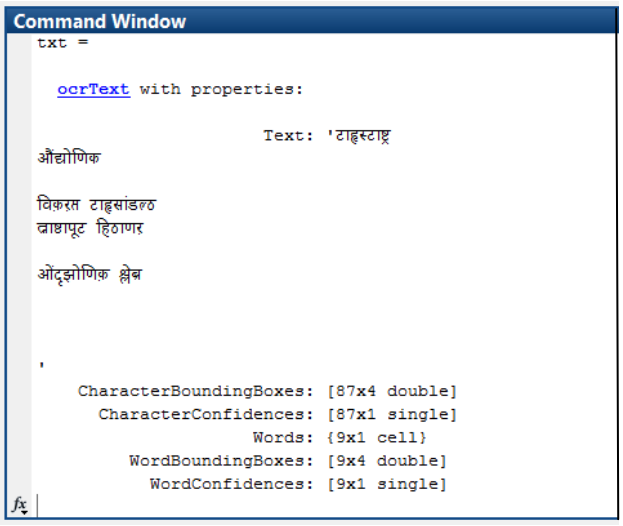
1. To detect text we perform object segmentation and object extraction for the separation of the entire object which in kind of text and image.
2. Then we convert the colored(rgb) image into grayscale using rgb2gray
3. Then we find the graythreshold for gray intensity and covert to black and white using im2bw.
4. Bwareaopen function can be use to remove the smaller objects in image less than 10 pixels or 30 pixels.
5. Then we label the image using bwlabel with 8-way connectivity i.e. to find all objects in image.
6. Then we plot the bounding box around those objects which comes under above condition only assume that it would be the text.

3.3 Text Recognition & Extraction:

Optical character recognition is the electronic conversion of binary images of printed as well as handwritten text into m/c-encoded text. It is mostly used as a type of data entry from the printed text documents, whether invoices, passport documents, computerized receipts, bank statements, printouts of static-data, mail, business cards.OCR is method of digitizing printed/handwritten texts therefore it can be easily searched, edited, stored more compactly, displayed off line or on-line, and used in machine processes like text-to-speech, machine translation, text mining and key data [5]. We can also increased the accuracy of OCR by making own small dictionary for particulars i.e. if the recognition output is constrained by a lexicon means [3] for example, all the words in the English language for a specific field. But sometimes this method can be difficult if the any printed or handwritten document contains words not in the lexicon means proper nouns.

We can use own handmade dictionary to influence the character or word segmentation step, for improved accuracy [3].OCR read each and every detected character bound by small rectangular boxes, after reading OCR compare each character with trained dataset and recognizes every character and places it along the original character in mat files, but due to variation in font, font size sometimes it fails to recognize in that case we can use our own Devanagari dictionary so that when OCR recognize false selection we compare those with our dictionary and replace it by using mat files. This helps to increase the accuracy of OCR.

Following are some snaps of matlab which shows how OCR works:



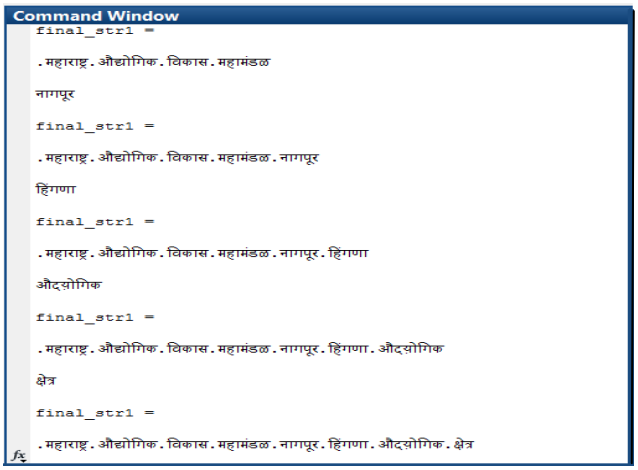


Fig: 3(a) OCR recognize text 3(b) Using own dictionary OCR modified text

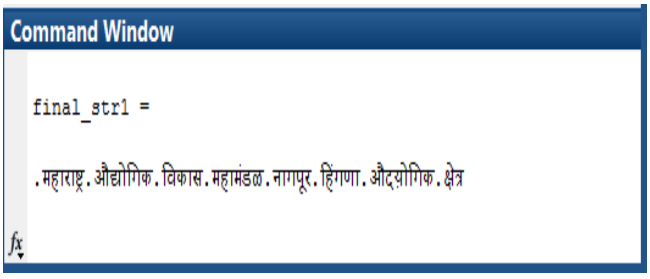


Fig: 3 (c) Text Extractions.

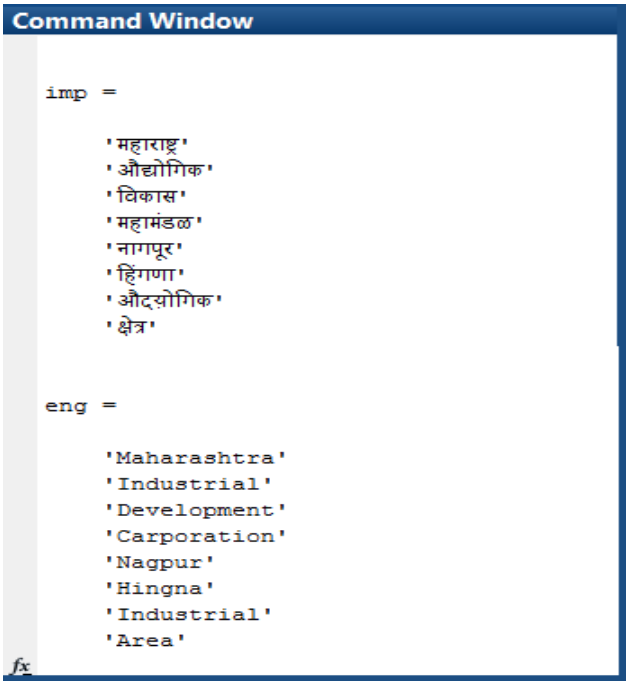
3.4 Text conversion:

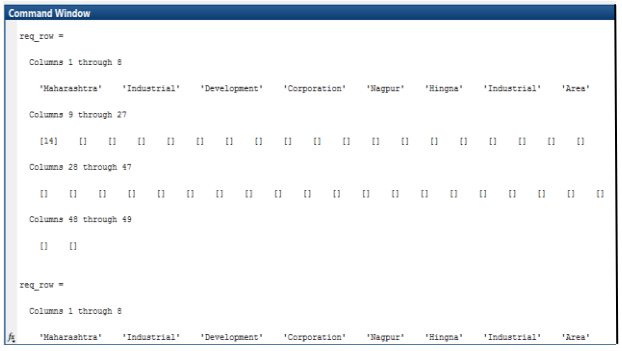
An n-gram is nothing but the sub-sequences of n elements from a particular given sequence. It is used in different statistical NLP's and genetic sequence analysis. The items in query may be words, characters or base pairs according to the application. For example, the sequence of characters "Indian army " has a 3-gram of ("Ind", "ndi", "dia", "ian ", "an, n", ...), and has a 2-gram of ("In", "nd", "di", "ia", "an ", " n", ...).Output of n-gram can be uses for Statistical machine translation and Spell checking. N-gram can be defining the tail of the list of sentences or words used by every recursive call, i.e. thus we can determine each and every tail of words list as well as complete list. Using N-Gram we can extract 1st 'n' element or item from every words list. Hence it is also called mapping over the data, size 1 of n-gram can be referred as unigram, size 2 referred as bigram,3 referred as trigram,4 will be four gram and so on.

N-gram language model proposed by Michael Flor et al. [4] .which is very efficient technique for pattern (text, word, letters etc) matching. It uses pattern matching and extraction algorithm. It check for a particular word or we can say letter picked from base dictionary check it in every dictionary row wise, and increase the counter as he found some matches by checking rows. Whose count having highest count that line will pick and consider as final match found string. Also make Devanagari to English conversion dictionary also, check the index number of final string and pick it from conversion dictionary.

In such way we can convert the any kind of Devanagari text into English just we should maintain proper dictionary for every panels or we can say set of data. Various types of n-gram first one is 'unigram' in this we check single item with every one, second is 'bigram 'which operate on two letters and last one is 'trigram' uses three letters,' fourgram'uses four letters and so on[4].

Following snapshots will clear the idea of n gram n conversion technique.





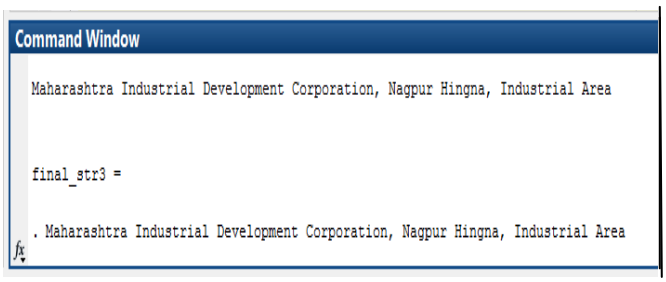


Fig: 4 (a) Dictionary to English conversion dictionary (matlab). (b) Sorted string with meaning using N-gram and dictionary approach (matlab). (c) Final string of conversion (matlab). (d) System extraction output (window). (e)Result of text conversion (Devanagari to English conversion).

IV. Result Analysis:

Input Images	OCR accuracy (%)	Proposed technique (Improve OCR accuracy (%)
Image1	20%	80%
Image2	80%	90%
Image3	42.85%	57.14%
Image4	8.33%	85.91%
Image5	68.29%	87.31%
Image6	46.92%	88.52%

Table 1: Recognition & extraction ratio for OCR

Results of Identified Devanagari (Hindi Words) script from Images:

Dataset Name	Parameters	Hindi Word
Image 1	Correct Classification	20%
	Misclassification	80%
	Rejection	0%
Image2	Correct Classification	94.44%
	Misclassification	5.66%
	Rejection	0%
Image3	Correct Classification	57.44%
	Misclassification	13.99%
	Rejection	28.57%
Image4	Correct Classification	65.58%
	Misclassification	21.86%
	Rejection	12.55%
Image5	Correct Classification	33.33%
	Misclassification	66.66%
	Rejection	0.00%
Image6	Correct Classification	38.28%
	Misclassification	57.14%
	Rejection	3.57%

Table 2: OCR correct classification & rejection ratio

V. Conclusion and Future Work:

In this paper, we presented new techniques for the traffic panel's detection, extraction and not least for conversion also. Many works has been done in the same research but no one make the complete whole package of all this things. In this paper we use color segmentation for panel detection either green or yellow from the background image which gives near about 95% accuracy. Next is text detection for that we use object extraction and segmentation concept which smartly separate all the objects (text) gives 90% accuracy. Text recognition and Extraction gives 99% result but using our own dictionary only. By using OCR trained dataset it gives 80% accuracy but by applying concept of additional dictionary it gives finally 99% result. For text conversion we use N-gram algorithm which checks the overall probability of single letter throughout the whole dictionary, it gives 99 % result. In future we can improve the dictionary data for more accurate result as the more efficient dictionary will gives more correct results. We will focus on text extraction and detection for better n accurate accuracy as well as on area also in an automatic way.

References:

[1] A. Gonzalez, L.M. Bergasa, J. Javier Yebes, M.A. Sotelo," Automatic Information Recognition of Traffic Panels using SIFT descriptors and HMMs," *International IEEE Annual Conference on Intelligent Transportation Systems, pp. 1289-1294, September 2010.*

[2] SATISH R. DAMADE," IDENTIFICATION OF DEVANAGARI SCRIPT FROM IMAGE DOCUMENT", International Journal of Computer Engineering and Technology (IJCET), Vol.4 No.5, pp.224-231, sep-oct (2013).

[3] Shalin A. Chopra, Amit A. Ghadge," Optical Character Recognition", International Journal of Advanced Research in Computer and Communication Engineering,Vol.3 No. 1, January 2014.

[4] Michael Flor," A fast and flexible architecture for very large word n-gram datasets", Natural Language Engineering, vol.19 No.1, pp.61-93, 10 Jan 2012.

[5] Mahesh Jangid, "Devanagari Isolated Character Recognition by Using Statistical Features,*" International Journal of Computer Science and Engg. Vol. 3, No.2, pp. 2400 –2407, 6, June 2011.*

[6] Andrej Ikica, Peter Peer," An improved edge profile based method for text detection in images of natural scenes," *International Conference on Computer as a Tool (EUROCON), pp.1-4, 2011.*

[7] C.P. Sumathi, T. Santhanam, N. Priya," Techniques and challenges of automatic text extraction in complex image: a survey," *Journal of Theoretical and Applied Information Technology,Vol. 35 No.2, pp. 225-235,2012.*

[8] Sankaran, Naven, and C. V. Jawahar."Recogniton of printed Devanagari text using BLSTM Neural Network,*" Patern Recogniton (ICPR), 2012 21st International Conference on.IEEE, 2012.*

[9] Prof. Sheetal A. Nirve, Dr. G. S. Sable," Optical character recognition for printed text in Devanagari using ANFIS," *International Journal of Scientific & Engineering Research, Vol. 4, No. 10, PP.236-241, October-2013.*

[10] Dhore, M., Dixit, S. and Dhore, R., "Hindi and Marathi to English NE Transliteration Tool using Phonology and Stress Analysis," *Proceedings of coling Demonstration Papers, pp. 111-118, 2012.*

[11] Ashoka H.N., Manjaiah D.H., Rabindranath Bera," Feature Extraction Technique for Neural Network Based Pattern Recognition,*" International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 03, pp. 331-339, March 2012*

[12] Jonathan Fabrizio, Beatriz Marcotegui and Matthieu Cord," Text detection in street level images," *pattern anal, springer Verlag, pp. 519-533, 2013.*

[13] A. Gonzalez, L.M. Bergasa, J. Javier Yebes and J. Almazan," Traffic Panels Detection Using Visual Appearance," *Intelligent Vehicles Symposium (IV), pp. 1221 – 1226, June 2013.*

[14] Shruti Agarwal, Dr. Naveen Hemarjani," Offline Handwritten Character Recognition with Devnagari Script," *IOSR Journal of Computer Engineering (IOSR-JCE), Vol.12No.6, pp.82-86(May. - Jun. 2013.*

[15] Sanket Rege, Rajendra Memane, Mihir Phatak, Parag Agarwal," 2D GEOMETRIC SHAPE AND COLOR RECOGNITION USING DIGITAL IMAGE PROCESSING, *"International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering ,Vol. 2, No.6, pp. 2479-2487, June 2013.*

[16] Rajesh K. Bawa1 and Ganesh K. Sethi,"A binarization technique extraction for devanagari text from camera based images," *International Journal of Signal & Image Processing (SIPIJ) Vol.5, No.2, April 2014.*

[17] Ajay Garg, Simpel Jindal "To Extract Feature of Handwritten Devnagari Script," *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 3, No. 7, pp. 7501-7503, July 2014.

[18] Ms. Saumya sucharita Sahoo and Prof. Smita Tikar," Review methods of scene text detection and its challenges," *International Journal of electronics and communication Engineering (IJECET), Vol.5 No.01, pp. 74-81, January 2014.*

[19] Poonam M.Ingle, P. P. Gumaste," Handwritten Devnagari script recognition using phase correlation," *Proceedings of IRF International Conference, pp. 96-99, 30th March-2014.*

[20] *Samabia Tehsin, Asif Masood, Sumaira Kausar, and Fahim Arif,"* Fuzzy-Based Segmentation for Variable Font-Sized Text Extraction from Images/Videos,*" Hindawi Publishing Corporation Mathematical Problems in Engineering, pp. 1-10, 2014.*

[21] Kamaljeet Kaur, Parminder Singh," Review of machine transliteration systems," *International Journal of Advanced Engineering Applications, Vol. 7, No.3, pp.72-80, 2014.*

