Text Retrieval from Natural and Scanned Images

Jayshree Ghorpade-Aher
MIT College of Engineering, Pune

Sumeet Gajbhar
MIT College of Engineering, Pune

Amey Sarode
MIT College of Engineering, Pune

Piyush Daund
MIT College of Engineering, Pune

Govardhan Gayake
MIT College of Engineering, Pune

ABSTRACT
Digital documents are easy to handle, share and store than hard copy of documents. These made people to prefer digital document over hard copy of documents. Digital documents are nothing but scanned images of a document or natural images of notice boards, traffic signs. Text detection is an important process required to extract text from images. Text from images can be extracted using Optical Character Recognition (OCR). OCR works in three phases as pre-processing, segmentation, character recognition. Pre-processing is the first phase which uses different techniques for making text easy to extract from images. In segmentation phase, each character is isolated. Then this will be given as input to OCR recognition phase which will compare it with training data-set and will recognize character. In this survey paper, different techniques for OCR are discussed.

Keywords
Optical Character Recognition (OCR), Pattern Recognition, Automatic Number Plate Recognition (ANPR)

1. INTRODUCTION
The world is moving very fast and human efforts are reducing day by day as there are new technologies emerging out. Now days, use of camera is increased drastically. Various Books, documents are converted into digital form using scanner as well as camera. Sometimes it requires extracting this text data from images in order to store it digitally. To do so Optical Character Recognition technique can be used. There are several pre-processing techniques for images from which text can be extracted. Accuracy and efficiency of further two phases depends totally on first pre-processing phase. Different techniques are required as there are different kind of images such as scanned images and Natural images. A scanned image has simple background and text can be easily recognized. But in case of natural images which are taken by camera such as road signs, it may have different contrast, may have complex background, and may contain non textual part. It is difficult to recognize the text from natural images.

The present paper depicts the picture of different Text detection algorithm for natural images. Natural images may contain a lot of non-text objects. To exclude this non-text object and detect text from a natural image text detection algorithm with Maximally Stable Extremal Regions can be used. This algorithm uses Maximally Stable Extremal Regions as basic letter candidates. Then further candidate is processed with some geometric constraints and stroke width transformation to remove non text objects. Automatic Number Plate Recognition (ANPR) system is used to detect the numbers from Number Plate. First it localizes the number plate and then apply OCR on that region.

The i-novel algorithm is one of the fundamental algorithms for the text extraction from images. OCR usually needs training data set and it is font dependent. It requires large amount of storage and computation time thus increases the complexity of the system. The i- novel algorithm is font independent and do not require any training data set. It uses feature extraction technique to recognize the text.

This paper is divided into different sections. The second section presents Text Detection algorithm and in third section the i-novel algorithm is discussed. Finally the conclusion is presented in last section.

2. TEXT DETECTION ALGORITHM
2.1 MSER
In traffic sign detection system, the traffic signs are detected from the captured image using the well-defined structure or block that is Maximally Stable Extremal Regions (MSER). Individual character detected from MSERs and then grouped into lines for recognition through OCR. MSER consist of two parts i.e. detection and recognition. In detection a particular structure or frame is set in the scene and the character which is present in the region is detected or the region where the recognition algorithm will be applied is located. Once text region is located the recognition phase is performed and text is extracted. Figure 1 shows the Detection and Recognition phase of MSER [1].

2.1.1 Detection stage
This stage consists of three parts namely

- Determination of search region (where the traffic sign is expected to found)
- Detection of candidates from the region
- Reduction of candidate using contextual constraints

The dimension of these search regions is given explicitly to the system which depends upon the position of the signs on the roads. Then reduction of candidates using temporal and contextual information is done. Each candidate is matched with the previously defined frame. If match is found then use previously extracted text else treat it as a new traffic sign.

2.1.2 Recognition stage
This is the next stage that recognizes text content from the detected candidate region. Different perspective transformations are applied to align it vertically and then individual character are segmented, formed into words and

then sent to OCR for recognition. This transformation is performed by fitting the quadrilateral to detected traffic signs. Then within the candidate regions, lines of text in traffic signs are located. Lines are aligning vertically, unwanted text characters removed from line and then it passes to OCR. Detected text lines in the form of gray scale image are given to open source Tesseract tool which recognize the text.

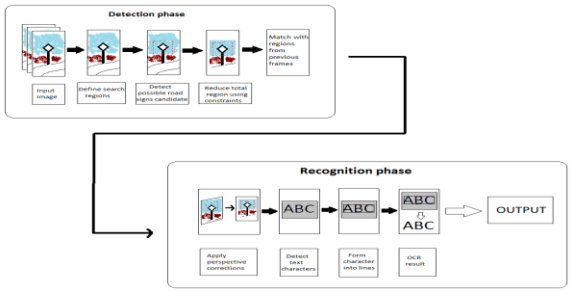

Figure 1: Detection and Recognition phase

2.2 Texture Based Techniques And Cc-Based Technique

In visual search systems, image features are extracted and then these features are compared with large dataset to recognize text from images. In such visual search system, the text is required to be located robustly. This is difficult task as there are number of various distortions, different font styles and different lightening conditions. There are two categories for text detection techniques as: Texture-based and Connected Component (CC)-based.

2.2.1 Texture-based Technique

Texture based techniques uses special texture which easily gets separate from its background. Then feature extraction can be done and classifier can be used to identify text in images.

2.2.2 CC-based Technique

In CC-based techniques, it uses geometric constraints to remove non text objects. Then using geometric properties text line can be formed[2].

First step is to adjust contrast for input images to enhance the contrast. Further step involves extracting MSER regions from image and edges are obtained from gray scale images. Then the resulting CCs are filtered using geometric constraints on properties like aspect ratio and number of holes. The stroke width information is robustly computed using a distance transform and objects with high variation in stroke width are rejected. Text candidates are grouped pairwise and form text lines. Finally, words within a text line are separated, giving segmented word patches at the output of our system.

2.3 ANPR

Automatic Number Plate Recognition system works with three modules:

- Number Plate Localization (NPL)
- Character Segmentation(CS)
- Optical Character Recognition(OCR)

In the first module, the number plate is localized in the image. Then in second module, pre-processing techniques is applied to localized region of image. The OCR stage includes character segmentation and recognition. Researcher have chosen DSP (Digital Signal Processors) or FPGA (Field Programmable Gate Arrays) for implementing ANPR systems to get real time and improved performance [3].

2.3.1 NPL Module

This NPL module uses different morphological operations for feature extraction. In these operations, feature extraction and removal of noise is done on the NP region. Morphological operations decrease computational intensity of system. The performance increases with respect to other algorithms such as edge detection algorithm. And it will recognize text more accurately as compared to other algorithms.

2.3.2 CS Module

In this module, basically pixel projection operation is carried out in different phases. Output image from the NPL module is binarised and inclined image is rotated before it is fed to CS module. In first phase, the unnecessary parts of NP are firstly removed by an NP height optimization step. After this the vertical positions of the characters and the horizontal positions of the characters is recognized by vertical and horizontal projection step.

2.3.3 OCR Module

OCR module uses NN (Neural Network) to translate the scanned character into text. Binary image matrix is given in the form vector to NN as an input. Initially NN was trained with training characters and then it is tested against the input characters.

FPGA development board is used as hardware to implement OCR system which gives better throughput and accuracy as com-pared to other systems. The overall accuracy rate of ANPR system is 93 % approx.

3. i - NOVEL ALGORITHM FOR OCR

An image from which text is to be extracted is given as input. Image can be a scanned document as well as camera image i.e. natural images. Algorithm assumes that input images are in Y Cb Cr color format. This algorithm will concentrate on Y part which is nothing but gray scale images [4].

Initially using an appropriate threshold gray level image is thresholded to obtained black and white image which is often called as binary image. The next step is to detect edge using edge detection algorithm. After edge detection, image is segmented. A segment is a sequence of pixel having color black in an edge-detected image. After segmentation process, feature extraction is done which will produce the final extracted text.

Edge detection is the process which identifies the boundaries of an object. Here the aim is to detect the sharp variation in the image brightness so as to distinguish between the text objects. This edge is nothing but the small number of pixels having high intensity difference than other neighboring pixels. The edge detection can be done by sobel or laplace method [5], [6]. The Sobel edge detector uses a pair of 3x3 convolution masks, one estimating the gradient in the x-direction (columns) and the other estimating the gradient in the y-direction (rows). Figure 2 shows the 3x3 convolution mask matrix.

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Figure 2: 3x3 convolution mask matrix

The magnitude of gradient can be calculated as:

$$|G| = \sqrt{(Gx)^2 + (Gy)^2}$$

In Laplace method, calculate 5*5 convoluted masks to approximate the second derivative, unlike the sobel method which approximates the gradient. The output of thresholding and edge detection is show in figure 3 [4].

Profiling of segments is just process of categorizing segments in different types such as short, long, curve. This profiling information is stored and used further. Every alphabet is made up of several segments.

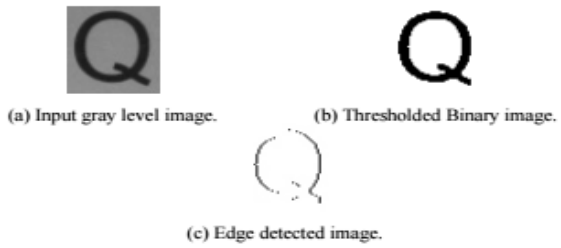

Figure 3: Output of threshold and edge detection block

For example, alphabet Q can be segmented as shown in table I and image has four zones as left and right top, left, right bottom. Segments matched with such types of profiling data for each alphabet and then corresponding ASCII value for text is returned

Table 1:Segment Profiles For The Alphabet Q

Sr. no	Profile
1	Left top zoned, long left-sided curve segment
2	Right top zoned long right-sided curve segment.
3	Right/left bottom zoned left-sided line segment

4. CONCLUSION

Text detection algorithm uses MSER as basic letter candidate. Maximally stable extreme regions are sensitive to blur and it is difficult to recognize font with smaller size for limited resolution. Thus edge-enhance MSER can be used to overcome sensitivity to image blur. Text detection algorithm with enhanced edge MSER can efficiently implemented for visual search engines. Usually OCR algorithm uses artificial neural networks and vector based training data sets to accurately recognize the text. These algorithms are font dependent and require redundant processing. The i-novel algorithm is very useful for recognizing multiple fonts. The i-novel algorithm doesn't use MSER for localization of blocks, instead it uses feature extraction on complete image. This algorithm is faster, memory efficient as it does not require any training data or mechanism, and it is independent of fonts.

5. ACKNOWLEDGMENT

We express our gratitude towards the department of Computer Engineering, MITCOE for giving us the opportunity to present the paper. A special thanks to our guide, Prof. Jayshree Ghorpade-Aher, whose help, stimulating suggestions and encouragement helped us to coordinate our paper.

6. REFERENCES

[1] Jack Greenhalgh and Majid Mirmehdi, Recognizing Text-Based Traffic Signs, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTA-TION SYSTEMS, VOL. 16, NO. 3, JUNE 2015.

[2] Huizhong Chen, Sam S. Tsai, Georg Schroth, David M. Chen, Radek Grzeszczuk and Bernd Girod, Robust Text Detection in Natural Images with egde-enhanced Maximally Stable Extremal Regions, ICIP 2011.

[3] Xiaojun Zhai, Faycal Bensaali and Klaus McDonald-Maier, Automatic Number Plate Recognition on FPGA, 978-1-4799-2452-3/13 IEEE, 2013.

[4] Sushruth Shastry, Gunasheela G, Thejus Dutt, Vinay D S and Sudhir Rao Rupanagudi, i - A novel algorithm for Optical Character Recognition (OCR), 978-1-4673-5090-7/13 IEEE, 2013.

[5] Julinda Gllavata, Ralph Ewerth and Bernd Freisleben, A Robust Algorithm for Text Detection in Images, University of Marburg, D-35032 Marburg, Germany.

[6] Jayshree Ghorpade, Raviraj Palvankar, Ajinkya Patankar and Snehal Rathi, EXTRACTING TEXT FROM VIDEO, Signal & Image Processing : An International Journal (SIPIJ) Vol.2, No.2, June 2011

[7] Mrunmayee Pati,Ramesh Kagalkar, An Automatic Approach for Translating Simple Images into Text Descriptions and Speech for Visually Impaired People, International Journal of Computer Applications (0975 8887)Volume 118 No. 3, May 2015.

[8] Binh Quang Long Mai, Tue Huu Huynh, Anh Dong Doan, A Study about the Reconstruction of Remote, Low Resolution Mobile Captured Text Images for OCR, 978-1-4799-6956-2/14 IEEE, 2014.

[9] Sushruth Shastry, Gunasheela G, Thejus Dutt, Vinay D S and Sudhir Rao Rupanagudi, i - A novel algorithm for Optical Character Recognition (OCR), 978-1-4673-5090-7/13 IEEE, 2013.

[10] Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat, Optical Character Recognition Implementation Using Pattern Matching, International Journal of Computer Science and Information Technolo-gies, Vol. 5 (2) , 2014.

[11] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, Andrew Y. Ng, Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning, ICDAR.

[12] Jisha Gopinath, Aravind S, Pooja Chandran, Saranya S S, Text to Speech Conversion System using OCR, IJETAE, Volume 5, Issue 1, January 2015.

[13] Yao Li and Huchuan Lu, Scene Text Detection via Stroke Width, 21st International Conference on Pattern Recognition (ICPR 2012), November 11-15, 2012, Tsukuba, Japan.

[14] Er. Kavneet Kaur, Vijay Kumar Banga, Number Plate Recognition using OCR Technique, International Journal of Research in Engineering and Technology eISSN: 2319-1163 — pISSN: 2321-7308.