A Technical Review on Text Recognition from Images

Pratik Madhukar Manwatkar
Department of Computer Technology, YCCE, Nagpur (M.S.),
441 110, India.
pratikmm@ymail.com

Dr. Kavita R. Singh
Department of Computer Technology, YCCE, Nagpur (M.S.),
441 110, India.
singhkavita19@yahoo.co.in

***Abstract*—Text recognition in images is an active research area which attempts to develop a computer application with the ability to automatically read the text from images. Nowadays there is a huge demand of storing the information available on paper documents in to a computer readable form for later use. One simple way to store information from these paper documents in to computer system is to first scan the documents and then store them as images. However to reuse this information it is very difficult to read the individual contents and searching the contents form these documents line-by-line and word-by-word. The challenges involved are: font characteristics of the characters in paper documents and quality of the images. Due to these challenges, computer is unable to recognize the characters while reading them. Thus, there is a need of character recognition mechanisms to perform document image analysis which transforms documents in paper format to electronic format. In this paper, we have reviewed and analyzed different methods for text recognition from images. The objective of this review paper is to summarize the well-known methods for better understanding of the reader.**

***Keywords*: Document Image Analysis (DIA), electronic format, text recognition, font characteristics.**

I. Introduction

Now-a-days, there is a growing demand for the software systems to recognize characters in computer when information is scanned through paper documents as we know that there are number of historical, mythological books and newspapers which are in printed format. Day by day due to atmospheric changes or due to improper handling they get damaged. Therefore, nowadays there is a huge demand in “storing the information available in these paper documents in to a computer storage disk and then later reusing this information by searching process”. One simple way to store information in these paper documents in to computer system is to first scan the documents. Whenever we scan the documents through the scanner, the documents are stored as images in the computer system. These images contain text that cannot be edited by the user. But to reuse this information it is very difficult for the computer system to read the individual contents and search the contents form these documents line-by-line and word-by-word. The reason for this difficulty is the font characteristics of the characters in paper documents are different to font of the characters in computer system. As a result, computer is unable to recognize the characters while reading them. This concept of storing the contents of paper documents in computer storage place and then reading and searching the content is called document processing. Sometimes in this document processing we need to process the information that is related to languages other than the English in the world. This process is also called Document Image Analysis (DIA). To handle DIA in recent years many approaches have been proposed by researchers, each approach has its own advantages and limitation which is discussed in detail in forthcoming section of this paper.

The paper is organized as follows: in section 2, overview of text recognition system. Section 3, presents the literature review on text recognition. Section 4, discusses about the applications of text recognition and finally section 5 summarizes the paper.

II. Text Recognition System

In this section we briefly describe the overall architecture of text recognition system as shown in figure 1. A text recognition system receives an input in the form of image which contains some text information. The output of this system is in electronic format ***i.e.*** text information in image are stored in computer readable form. The text recognition system can be divided in following modules: (*A*) pre-processing (*B*) text recognition (*C*) post-processing. Each module is further described in detail as bellow:

A. Pre-processing Module

The paper document is generally scanned by the optical scanner and is converted in to the form of a picture. A picture is the combinations of picture elements which are also known as pixels. At this stage we have the data in the form of image and this image can be further analyzed so that’s the important information can be retrieved. So to improve quality of the input image, few operation are performed for enhancement of image such as noise removal, normalization, binarization ***etc*.**

978-1-4799-6480-2/15/$31.00 ©2015 IEEE

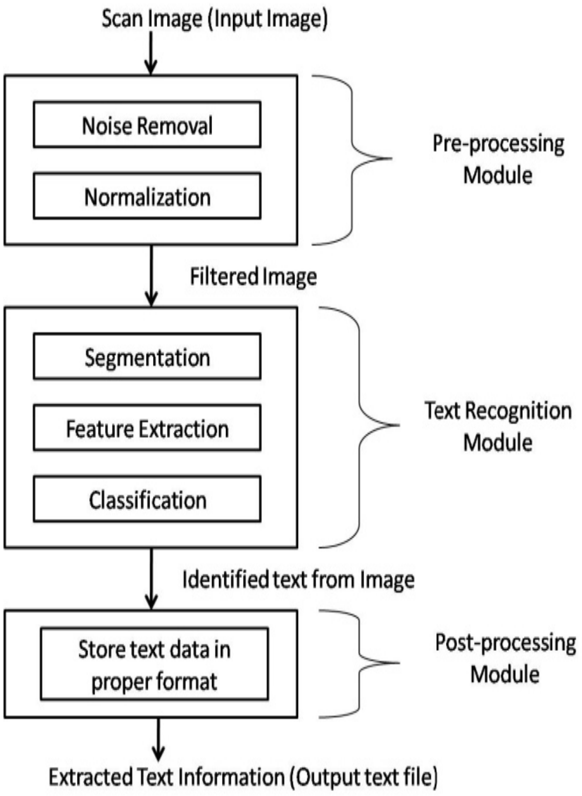


Figure 1: Architecture of text recognition

1) Noise Removal

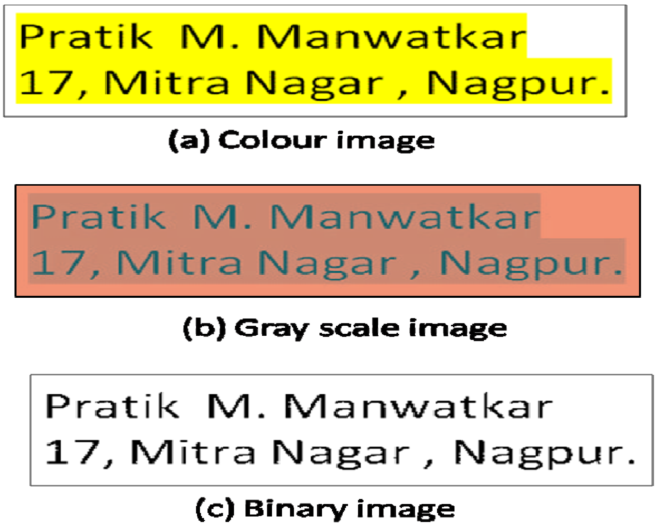
Noise removal is one of the most important process. Due to this quality of the image will increase and it will effect recognition process for better text recognition in images. And it results in generation of more accurate output at the end of text recognition processing. There are many methods for image noise removal such as mean filter, min-max filter, Gaussian filter ***etc.***

2) Normalization

Normalization is one of the important pre-processing operation for text recognition. The normalization is applied to obtain characters of uniform size, slant and rotation.

3) Binarization

Binarization is one of the important pre-processing operation for text recognition. A printed document is first scanned and is converted into a gray scale image. Binarization is a technique by which the gray scale images are converted to binary images. This separation of text from background that is required for some operations such as segmentation. Figure 2 shows a colour image (a) gray image (b), and binary image (c) of a text image.


Pratik M. Manwatkar
17, Mitra Nagar , Nagpur.
Pratik M. Manwatkar
17, Mitra Nagar , Nagpur.
Pratik M. Manwatkar
17, Mitra Nagar , Nagpur.

(a) Colour image

(b) Gray scale image

(c) Binary image

Figure 2: Binarization process in text recognition

B. Text Recognition Module

This module can be used for text recognition in output image of pre-processing model and give output data which are in computer understandable form. Hence in this module following techniques are used.

1) Segmentation

In text recognition module, the segmentation is the most important process. Segmentation is done to make the separation between the individual characters of an image.

2) Feature Extraction

Feature extraction is the process to retrieve the most important data from the raw data. The most important data means that's on the basis of that's the characters can be represented accurately. To store the different features of a character, the different classes are made. There are many technique used for feature extraction like Principle Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), zoning, Gradient Based features, Histogram ***etc.***

3) Classification

The classification is the process of identifying each character and assigning to it the correct character class, so that texts in images are converted in to computer understandable form. This process used extracted feature of text image for classification i.e. input to this stage is output of the feature extraction process. Classifiers compare the input feature with stored pattern and find out best matching class for input. There are many technique used for classification such as Artificial Neural Network (ANN), Template Matching, Support Vector Matching (SVM) *etc.*

C. Post-processing Module

The output of text recognition module is in the form text data which is understand by computer, So there need to store it in to some proper format(***i.e.*** text or MS-Word)for farther use such as editing or searching in that data.

III. Literature Review

As discussed earlier text recognition from images is still an active research in the field of pattern recognition. To address the issues related to text recognition many researchers have proposed different technologies, each approach or technology tries to address the issues in different why. In forthcoming section we present a detailed survey of approaches proposed to handle the issues related to text recognition.

Yang et al.[1] has proposed a novel adaptive binarization method based on wavelet filter is proposed. This approach was processes faster, so that it is more suitable for real-time processing and applicable for mobile devices. They evaluated this adaptive method on complex scene images of ICDAR 2005 database. Sankaran et al. [2] has proposed a novel recognition approach that result in a 15% decrease in word error rate on heavily degraded Indian language document images.

Gur et al. [3] has discussed some problems in text recognition and retrieval. Automated optical character recognition(OCR) tools do not supply a complete solution and in most cases human inspection is required. They suggest a novel text recognition algorithm based on usage of fuzzy logic rules relying on statistical data of the analyzed font. The new approach combines letter statistics and correlation coefficients in a set of fuzzy based rules, enabling the recognition of distorted letters that may not be retrieved otherwise. They focused on rashi fonts associated with commentaries of the bible that are actually handwritten calligraphy.

Rhead et al. [4] has considered real world UK number plates and relates these to ANPR. It considers aspects of the relevant legislation and standards when applying them to real world number plates. The varied manufacturing techniques and varied specifications of component parts are also noted. The varied fixing methodologies and fixing locations are discussed as well as the impact on image capture.

Badawy, W. et al. [5] has discussed the Automatic license plate recognition (ALPR) is the extraction of vehicle license plate information from an image or a sequence of images. The extracted information can be used with or without a database in many applications, such as electronic payment systems (toll payment, parking fee payment), and freeway and arterial monitoring systems for traffic surveillance. The ALPR uses either a color, black and white, or infrared camera to take images.

Jawahar et al. [6] has proposed a recognition scheme for the indian script of devanagari. They used approach does not require word to character segmentation, which is one of the most common reason for high word error rate. They have been reported a reduction of more than 20% in word error rate and over 9% reduction in character error rate while comparing with the best available OCR system.

Ntirogiannis et al. [7] has studied that the document image binarization is of great importance in the document image analysis and recognition pipeline since it affects further stages of the recognition process. The evaluation of a binarization method aids in studying its algorithmic behavior, as well as verifying its effectiveness, by providing qualitative and quantitative indication of its performance. They proposed a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images.

Malakar et al. [8] has described that extraction of text lines from document images is one of the important steps in the process of an Optical Character Recognition (OCR) system. In case of handwritten document images, presence of skewed, touching or overlapping text line(s) makes this process a real challenge to the researcher.

Tirthraj Dash et al have discussed HCR using associative memory net (AMN) in their paper [9]. They have directly worked at pixel level. Dataset was designed in MS paint 6.1 with normal Arial font of size 28. Dimension of image was kept 31 X 39. Once characters are extracted, their binary pixel values are directly used to train AMN.

Pradeep et al. [10] have proposed neural network based classification of handwritten character recognition system. Each individual character is resized to 30 X 20 pixels for processing. They are using binary features to train neural network. However such features are not robust. In post processing stage, recognized characters are converted to ASCII format. Input layer has 600 neurons equal to number of pixels. Output layer has 26 neurons as English has 26 alphabets. Proposed ANN uses back propagation algorithm with momentum and adaptive learning rate.

Rajib et al[11] have proposed Hidden Markov Model based system for English HCR. They have employed global as well as local feature extraction methods. Global feature involves four gradient features, six projection features and four curvature features. And to extract local features, image is divided in to nine equal blocks and 4 gradient features are calculated from each block, so total of 36 features are extracted. So overall feature vector contains 50 features per character. O = [G(4) P(6) C(4) L(36)], where G, P, C and L represents global gradient, projection, curvature and local gradient features respectively. Number in parenthesis represents number of respective features. HMM is trained using these feature and experiment is carried out. Post processing is also applied after recognition phase of HMM to highly confused group of characters like N and M, O and Q, C and O etc. For each group new feature is calculated to discriminate characters within the group.

In literature [12], T.Som et al have discussed fuzzy membership function based approach for HCR.Character images are normalized to 20 X 10 pixels. Average image (fused image) is formed from 10 images of each character. Bonding box around character is determined by using vertical and horizontal projection of character. After cropping image to bounding box, it is resized to 10 X 10 pixels size. After that, thing is performed and thinned image is placed in one by one raw of 100 X 100 canvas. Similarity score of test image is matched with fusion image and characters are classified.

In literature [13], Rakesh kumar et al has proposed single layer neural network based approach for HCR to reduce training time. Characters are written on A4 size paper in uniform box. Segmented characters are scaled to 80 X 80 pixels. Each 0 is replaced by -1 for better training.

Kim [14] has proposed an approach in which LCQ (Local Color Quantization) is performed for each color separately. Each color is assumed as a text color without knowing whether it is real text color or not. To reduce processing time, an input image is converted to a 256-color image before color quantization takes place. To find candidate text lines, the connected components that are extracted for each color are merged when they show text region features. The drawback of their proposed method is the high processing time since LCQ is executed for each color.

Agnihotri and Dimitrova [15] have proposed an algorithm which uses only the red part of the RGB color space, with the aim to obtain high contrast edges for the frequent text colors. By means of a convolution process with specific masks they first enhance the image and then detect edges. Non-text areas are removed using a preset fixed threshold. Finally, a connected component analysis (eight-pixel neighborhood) is performed on the edge image in order to group neighboring edge pixels to single connected components structures. Then, the detected text candidates undergo another treatment in order to be ready for an OCR.

Cai et al.[16] have presented a text detection approach which is based on character features like edge strength, edge density and horizontal distribution. First, they apply a color edge detection algorithm in YUV color space and filter out non-text edges using a low threshold. Then, a local thresholding technique is employed in order to keep low-contrast text and simplify the background. Finally, projection profiles are analyzed to localize text regions.

Jain and Yu [17] first perform a color reduction by bit dropping and color clustering quantization, and afterwards, a multi-value image decomposition algorithm is applied to decompose the input image into multiple foreground and background images. Then, connected component analysis combined with projection profile features are performed on each of them to localize text candidates. This method can extract only horizontal texts of large sizes.

IV. Application

Text recognition technology may be apply throughout the entire spectrum of industries, revolutionizing the document management process. This technology enable scan documents to become more than just image files, turning into fully searchable documents with text content that is recognized by computers. With the help of this technology, people no longer need to manually retype important documents when entering them into electronic databases. Instead, Text recognition system extracts relevant information and enters it automatically. The result is accurate, efficient information processing in less time. In the following, we overview some applications of text recognition system

A. Banking[18]

The uses of image text recognition vary across different fields. One widely known application is in banking, it is used to process checks without human involvement. A check can be inserted into a machine, the writing on it is scanned instantly, and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for handwritten checks as well, though it occasionally requires manual confirmation. Overall, this reduces wait times in many banks.

B. Legal[18]

In the legal industry, there has also been a significant movement to digitize paper documents. In order to save space and eliminate the need to sift through boxes of paper files, documents are being scanned and entered into computer databases. Image text recognition further simplifies the process by making documents text-searchable, so that they are easier to locate and work with once in the database. Legal professionals now have fast, easy access to a huge library of documents in electronic format, which they can find simply by typing in a few keywords.

C. Healthcare[18]

Healthcare also use of image text recognition technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. By using image recognition technology they are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded. As a result, healthcare providers can focus on delivering the best possible service to every patient.

D. Image text recognition in Other Industries[18]

Image text recognition technology is widely used in many other fields, including education, finance, and government agencies. This technology has made countless texts available online, saving money for students and allowing knowledge to be shared. Invoice imaging applications are used in many businesses to keep track of financial records and prevent a backlog of payments from piling up. In government agencies and independent organizations, image text recognition technology simplifies data collection and analysis, among other processes.

As the technology continues to develop, more and more applications are found for technology, including increased use of handwriting recognition.

V. Summary

In this paper we have reviewed and analyzed different methods to find text characters from scene images. We have reviewed basic architecture of text recognition from images. In which we discussed different techniques of image processing in particular sequence for text recognition from scan image. Also, we have discussed some application of text recognition system.

Reference

[1] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu, "*A fast adaptive binarization method for complex scene images*," 19th IEEE International Conference on Image Processing (ICIP), 2012.

[2] Shrey Dutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar, "*Robust Recognition of Degraded Documents Using Character N-Grams*," IEEE, 2012.

[3] Gur, Eran, and ZeevZelavsky, "*Retrieval of Rashi Semi-Cursive Handwriting via Fuzzy Logic,"* IEEE International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012.

[4] Rhead, Mke, "*Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems."* IEEE International Carnahan Conference on Security Technology (ICCST), 2012.

[5] Badawy, W. *"Automatic License Plate Recognition (ALPR): A State of the Art Review."* IEEE International Conference on Document Analysis and Recognition, 2012.

[6] Naveen Sankaran and C.V Jawahar, *"Recognition of Printed Devanagari Text Using BLSTM Neural Network,"* IEEE, 2012.

[7] Ntirogiannis, Konstantinos, Basilis Gatos, and Ioannis Pratikakis. "*A Performance Evaluation Methodology for Historical Document Image Binarization.,"* IEEE International Conference on Document Analysis and Recognition, 2013.

[8] Malakar, Samir, et al. *"Text line extraction from handwritten document pages using spiral run length smearing algorithm,"* IEEE International Conference on Communications, Devices and Intelligent Systems (CODIS), 2012.

[9] Tirtharaj Dash*, "Time efficient approach to offline hand written character recognition using associative memory net.,"* International Journal of Computing and Business Research, Volume 3, September 2012.

[10] J. Pradeepa, E. Srinivasana, S. Himavathib, *"Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten,"* International journal of Engineering, Volume 25, May 2012.

[11] Rajib Lochan Das, Binod Kumar Prasad, Goutam Sanyal, *"HMM based Offline Handwritten Writer Independent English haracter Recognition using Global and Local Feature Extraction,"* International Journal of Computer Applications, Volume 46, May 2012.

[12] T.Som, Sumit Saha*, "Handwritten Character Recognition Using Fuzzy Membership Function",* International Journal of Emerging Technologies in Sciences and Engineering, Volume 5, December 2011.

[13] Rakesh Kumar Mandal, N R Manna, *"Hand Written English Character Recognition using Row- wise Segmentation Technique,"* International Symposium on Devices MEMS, Intelligent Systems & Communication, pp. 5-9, 2011.

[14] P.K. Kim. "*Automatic Text Location in Complex Color Images Using Local Color Quantization*," IEEE TENCON, Vol. 1, pp. 629-632, 1999.

[15] L. Agnihotri and N. Dimitrova. "*Text Detection for Video Analysis,"* International Conference on Multimedia Computing and Systems, Florence, Italy, pp. 109-113, 1999.

[16] M. Cai, J. Song and M. R. Lyu*, "A New Approach for Video Text Detection,"* International Conference On Image Processing, Rochester, New York, USA, pp. 117-120, 2002

[17] K. Jain and B. Yu. *"Automatic Text Location in Images and Video Frames,"* International Conference of Pattern Recognition (ICPR), Brisbane, pp. 1497-1499, 1998.

[18] Application of OCR, from http://www.cvisiontech.com.