A Review on Optical Character Recognition Techniques

Hiral Modi
P. G. Scholar
CSE Department
Gujarat Technological University
Ahmedabad, India

M. C. Parikh, PhD
Associate Professor
CSE Department
Gujarat Technological University Ahmedabad,
India

ABSTRACT
At present scenario, there is growing demand for the software system to recognize characters in a computer system when information is scanned through paper documents. This paper presents detailed review in the field of Optical Character Recognition. Various techniques are determined that have been proposed to realize the center of character recognition in an optical character recognition system. OCR (Optical Character Recognition) translates images of typewritten or handwritten characters into the electronically editable format and it preserves font properties. Different techniques for pre-processing and segmentation have been surveyed and discussed in this paper.

General Terms
Pattern Matching.

Keywords
Character Recognition System, Image Segmentation, OCR, Preprocessing, Skew correction, Classifier.

1. INTRODUCTION
OCR (Optical Character Recognition) translates images of typewritten or handwritten characters into machine editable format. OCR reads damaged or low-quality codes and returns the best guess at what the code is. It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static data, or any suitable documentation. OCR does not deal with quality and sharpness of characters. To overcome the limitations of OCR a new approach comes into picture which is OCV.

Projection Profile-based methods used makes segmentation easy to separate the text in document image into lines, words, and characters independent of the Language in the Text. Different methods are used at each intermediate stage of OCR. Text Segmentation is done using Projection Profile method. They proposed an algorithm for correction of the skew angle of the text document [1]. Blur is the important factor that damages OCR accuracy. In this paper prediction method based on a local blur estimation is proposed. The relation between blur effect and character size is investigated which is useful for the classifier. Classifier separates the given document into three classes: readable, intermediate, non-readable classes [2].

The grading system is used to evaluate the performance of printed text using various quality measures. The recognition results showed high recognition rate as the system was able to perform a recognition rate of 98.69 % along with a precision of 0.9857 and a sensitivity of 1 [3]. This paper presents complete OCR (Optical Character Recognition) system for camera captured image/graphics embedded textual documents for handheld devices [4]. Paper [5] describes the skew detection and correction of scanned document images written in Assamese language using the horizontal and vertical projection profile analysis OCR consists of many phases such as Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition [6].

1.1 Digitization
Digitization is the process of converting a paper-based handwritten document into electronic format. Here, each document consists of only one character. The electronic conversion is accomplished by using a method whereby a document is scanned and an electronic representation of the original document as an image file format is produced. The author used various scanners for digitization, and the digital image was going for next step that is a preprocessing phase.

1.2 Pre-processing
In The pre-processing phase, there is a series of operations performed on the scanned input image. It enhances the image rendering it suitable for segmentation the gray-level character image is normalized into a window sized. After noise reduction, a bitmap image is produced. Then, the bitmap image was transformed into a thinned image.

1.3 Segmentation
The Segmentation phase is the most important process. Segmentation is done by separation from the individual characters of an image. Segmentation of handwritten characters into different zones (upper, middle and lower zone) and characters is more difficult than that of printed documents that are in standard form. This is mainly because of variability in a paragraph, words of line and characters of a word, skew, slant, size and curved. Sometimes components of two adjacent characters may be touched or overlapped and this situation creates difficulties in the segmentation task. The touching or overlapping problem occurs frequently because of modified characters in upper-zone and lower-zone.

1.4 Feature Extraction
In this phase, features of individual character are extracted. The performance of an each character recognition system that depends on the features that are extracted. The extracted features from input character should allow classification of a character in a unique way. Different types of features are available like diagonal features, intersection, open-ended features, zoning features.

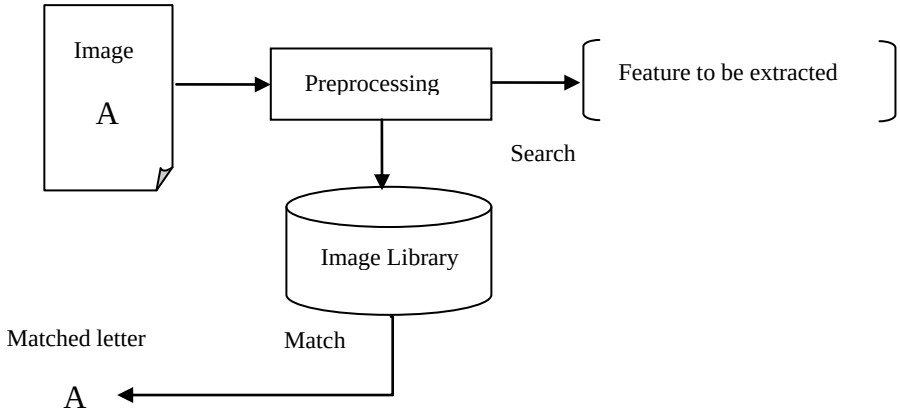

Figure 1 Phases of General Character Recognition System

2. METHODOLOGY

One of the most important steps of offline character recognition system is skew detection and correction which has to be used in scanned documents as a pre-processing stage in almost all document analysis and recognition systems. This paper describes the skew detection and correction of scanned document images written in Assamese language using the horizontal and vertical projection profile analysis [5]. Documents with background images in OCR cause an error. A non-linear transformation is used to enhance the contrast of each channel image. The method was tested using Tesseract (an open source OCR engine) and compared with two commercial OCR software ABBYY Finereader and HANWANG (OCR software for Chinese characters). The experimental results show that the recognition accuracies are improved significantly after removing background images [7]. For pre-processing Fourier Transform is used which decomposes an image into sine and cosine components with increasing frequencies. Fourier transform converts spatial domain onto frequency domain which is easily used for further processing [1]. Reading text from photographs is a challenging problem. They applied recently developed machine learning algorithms for learning the features automatically from unlabeled data. They proposed text detection and recognition system based on a scalable feature learning algorithm and applied it to images of text in natural scenes [8]. Since past few years, research has been performed to develop machine printed Chinese/English characters. In this paper, they described the search and fast match techniques. High-performance Chinese/English OCR engine is used to construct a large vocabulary. They have collected 1862 text lines from varied sources such as newspapers, magazines, journals, books, etc [9].

H. Wang and J. Kangas [10] proposed a method of identifying character- like regions in order to extract and recognize characters in natural color scene images automatically. Connected component extraction is used to check the block candidates. Priority adaptive segmentation (PAS) is implemented to obtain accurate foreground pixels of the character in each block. Paper [11] presented a system for text extraction based on the open-source OCR algorithm. The system is used for functional verification of TV sets. J. Diaz-Escobar [12] proposed a new method for recognition of content-less characters in degraded images using the phase congruency and local energy model. The suggested phase features are invariant to non-uniform illumination and slight geometric distortions. Degraded images were compared with that of the SIFT method in terms of recognition metrics. Another approach in the paper [13] Hauling the scene text from image and video is challenging due to the complex background, changeable font size, dissimilar style, unknown layout, poor resolution and blurring, position, viewing angle and so on. For text extraction region and connected component based methods are used. Artificial Neural network (ANN) is used as the classifier to filter out the text and non-text components.

There are natural variations in human writing so designing a reliable OCR system is a challenging task. An algorithm based on Kohonen Neural Network is presented in this paper. Kohonen algorithm that is one of Artificial neural network The experiments also demonstrated that system complexity can be reduced significantly without degrading performance by considering two-layered neural network rather than multiple layered neural networks [14]. In this paper [15] a complete OCR methodology for recognizing historical documents, either printed or handwritten without any knowledge of the font, is presented. The pre-processing and segmentation approach is used in order to detect text lines, words, and characters. Yaeger [16] has proposed a handwritten character recognition system. The proposed system works by using the neural network techniques. For the recognition of characters, a multi-layer perceptron is used by this system and it gives better results. J. Hu et. al [17] proposed a system in which high-level features are combined with low-level features on simple points and these are able to cover a huge amount of input patterns. Also, these features have invariance property which is used for normalizing the curvature of features. Funanda [18] has proposed a system which uses the HMM for the recognition of the online handwritten recognition. The proposed system reduces the usage of memory and also it improved the recognition rate of online handwritten characters. In paper [1] Horizontal Projection Profile and Vertical Projection Profile methods are used for segmentation. Different methods are used at each intermediate stage of OCR. Text Segmentation is done using Projection Profile method. They proposed an algorithm for correction of the skew angle of the text document.

J. r´ı Matas [19] presented an end-to-end real-time scene text localization and recognition method. In the first stage of the classification, the probability of each ER being a character is estimated using novel features calculated with O(1) complexity. In second stage only ERs with locally maximal probability are selected.

Huei-Yung Lin and Chin-Yu Hsu [20] presented neural network based approach which reduces the training time and maintains the high recognition rate. Multi-stage approach and

pre-processing are done for the experiment. Preprocessing is performed to partition the training data prior to training stage. In this paper [21], a computer vision and character recognition algorithm for a license plate recognition (LPR) is presented to be used as a core for intelligent infrastructure like electronic payment systems (toll payment, parking fee payment), freeway. Based on the connected component analysis and novel adaptive image segmentation technique is presented [21].

3. COMPARISON

Paper [5] presented that projection profile is used as a suitable feature for skew detection. Vertical Projection Profile Analysis allows small noise which produces error where Horizontal Projection Profile Analysis reduces the effect of noise. The time complexity of Vertical is high with compared to horizontal projection profile. In paper [7] author proposed a method which is used to remove the background image from pilling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes, document. The experiment is done using three OCR software tool: HANWANG OCR, ABBYY, and Tesseract. With compared to Tesseract OCR, HANWANG OCR, and ABBYY OCR better because there are built-in functions are available to preprocess image before text extraction

In paper [8] they trained their character classifier with features. They tested 5198 characters from 62 classes 26 upper- and 26 lower-case letters). Accuracy for the largest system (1500 features) is the highest, at 81.7% for the 62-way classification problem.

Segmentation is an important stage of OCR in image processing. In this paper [22] they surveyed different techniques which are available for segmentation. Most methods are categorized into three groups: the analytical, the empirical goodness and the empirical discrepancy groups. Segmentation algorithms can be evaluated analytically or empirically, so the evaluation methods can be divided into two categories: the analytical methods and the empirical methods. The analytical methods directly examine and assess the segmentation algorithms themselves by analyzing their principles and properties. The empirical methods indirectly judge the segmentation algorithms by applying them to test images and measuring the quality of segmentation results. Empirical methods are classified into two types: empirical goodness and empirical discrepancy method. In first method properties of segmented images are measured using"goodness" parameters. Where in the second type some references that present the ideal or expected segmentation results are first found.

.

Devices like Personal Data Assistants (PDA) which is pen input devices require good online handwriting character recognition algorithms. A. Funada et al. [18] proposed a new algorithm to recognize on-line handwriting and it utilize HMM(Hidden Markov Model). The memory reduction rate is a function of the matrix size and the number of states. They performed character segmentation, character classification which is fairly standard multilayer perceptron trained with error backpropagation provides the ANN character classify

4. APPLICATION

Optical Character Recognition is a vast field with a number of varied application which is described below [23]. For OCR enhanced image segmentation algorithm based on histogram equalization using genetic algorithms are used.

4.1 Captcha

A CAPTCHA is a program that can generate and grade tests that human can pass but current computers programmers' cannot. In CAPTCHA, an image consisting of series of letters of number is generated which is obscured by image distortion techniques, size and font variation, distracting backgrounds, random segments, highlights, and noise in the image. This system can be used to remove this noise and segment the image to make the image tractable for the OCR (Optical Character Recognition) systems.

4.2 Institutional Repositories and Digital Libraries

Institutional repositories are digital collections of the outputs created within a university or research institution. It is an online locale of intellectual data of an institution, especially a research institution where it is collected, preserved and aired. It helps to open up the outputs of an institution and give it visibility and more impact on worldwide level

4.3 Invoice Imaging

Invoice imaging is widely used in many businesses applications to keep track of financial records and prevent a backlog of payments from pilling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes.

4.4 Automatic Number Recognition

Automatic number plate recognition [6] is used as a mass surveillance technique making use of optical character recognition on images to identify vehicle registration plates. ANPR has also been made to store the images captured by the cameras including the numbers captured from the license plate.

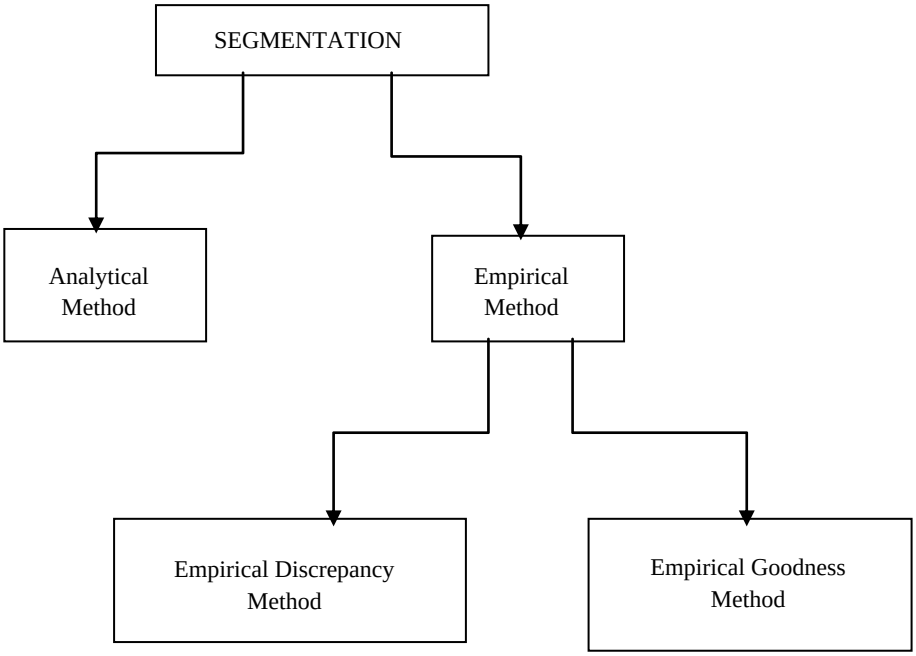

Figure 2 Evaluation of Segmentation Algorithm

Table 1. Comparison of the different OCR Techniques

Author(s)	Data set	Method	Recognition rate
Yaeger et al. [16] (1998)	(A-Z) characters, (0-9) digits, 23 symbols with writer independent system	Multi-Layer Perceptron	21.3%
Hu et al.[17] (2000)	(a) 500, 1000 and 2000 unipen database. (b) 5000, 10000 and 20000 unipen database.	Hidden Markov Model	91.8%, 90.5% and 87.2%for (a) dataset and 83.2%,79.8% and 76.3% for (b) dataset.
Funanda et al. [18] (2004)	Kanji, Katakana, Hirangana, Western alphabets and symbols with writer Independent system.	Hidden Markov Model	91.34%
A. F. Mollah et al. [4] (2011)	Set of 100 business cards images.	Segmentation using Vertical Projection Profile	92.74%
M. Shen [7] (2015)	1160 images with various resolutions, font sizes and noise levels.	Image Enhancement using non-linear Transformation	-
J. B. Pedersen et al. [3] (2016)	100 images with a total of 840 characters	Character based segmentation and Nearest Neighbour Classifier	98.69%
V. Kieu et al. [2] (2016)	IPAD contains 297 document images and PME contains 1998 document images	Fuzzy-C-Means clustering method	90.57%
C. N. E. Anagnostopoulos et al. [21] (2006)	1334 natural-scene gray-level vehicle images	probabilistic neural network (PNN)	96.5% (Segmentation) 89.1%(Entire Plate Recognition)
A. Coates *et al.* [8]	ICDAR data set 5198 test characters	Machine Learning Algorithm	85.5%

4.5 Legal Industry

The legal industry is also one of the beneficiaries of the OCR technology. OCR is used to digitize documents and directly entered into a computer database.

4.6 Banking

Another important application of OCR is in banking, where it is used to process cheques without human involvement

Cheque can be inserted into a machine where the system scans the amount to be issued and the correct amount of accessed as necessary

4.7 Healthcare

Healthcare has also seen an increase in the use of OCR 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.

5. CONCLUSION

"This paper elaborated survey of disparate techniques for OCR" has been studied. Handwritten character, natural scene images, business cards and TV set images are selected for experimentation. A systematic flow of OCR system is discussed. In this paper projection profile based method for segmentation, fourier transform technique is for pre-processing, and nearest neighbour classifier for classification are described. This paper can be helpful to the researcher for selecting most appropriate techniques to achieve optimum results for application according to a different parameter described in the previous section.

6. REFERENCES

[1] A. S. Sawant, "Script Independent Text Pre-processing and Segmentation for OCR," *Int. Conf. Electr. Electron. Signals, Commun. Optim. - 2015*, pp. 1–5, 2015.

[2] V. Kieu, F. Cloppet, and N. Vincent, "OCR Accuracy Prediction Method Based on Blur Estimation," *2016 12th IAPR Work. Doc. Anal. Syst.*, pp. 317–322, 2016.

[3] J. B. Pedersen, K. Nasrollahi, and T. B. Moeslund, "Quality Inspection of Printed Texts," *IWSSP 2016- 23rd Int. Conf. Syst. Image Process. 23-25 May 2016, Bratislava, Slovakia*, pp. 6–9, 2016.

[4] A. F. Mollah, N. Majumder, S. Basu, and M. Nasipuri, "Design of an Optical Character Recognition System for Camera- based Handheld Devices," *IJCSI*, vol. 8, no. 4, pp. 283–289, 2011.

[5] B. Jain and M. Borah, "A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical," *IJSRP*, vol. 4, no. 6, pp. 4–7, 2014.

[6] E. N. Bhatia, "Optical Character Recognition Techniques : A Review," *IJARCSSE*, vol. 4, no. 5, pp. 1219–1223, 2014.

[7] M. Shen, "Improving OCR Performance with Background Image Elimination," *2015 12th Int. Conf. Fuzzy Syst. Knowl. Discov.*, pp. 1566–1570, 2015.

[8] A. Coates *et al.*, "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning."

[9] P. Road, "Confidence Guided Progressive Search and Fast Match Techniques for High Performance ChineseEnglish OCR *," *IEEE*, pp. 89–92, 2002.

[10] H. Wang and J. Kangas, "Character-Like Region Verification for Extracting Text in Scene Images," no. 11, 2001.

[11] I. Kastelan, S. Kukolj, V. Pekovic, V. Marinkovic, and Z. Marceta, "Extraction of Text on TV Screen using Optical Character Recognition," *IEEE*, pp. 153–156, 2012.

[12] J. Diaz-escobar, "Optical Character Recognition based on phase features," *IEEE*, 2015.

[13] A. Thilagavathy, K. Aarthi, and A. Chilambuchelvan, "A Hybrid Approach to Extract Scene Text from Videos," *ICCEET*, pp. 1017–1022, 2012.

[14] S. Goyal, "Optical Character Recognition," *IJARCSSE*, vol. 3, no. 11, pp. 982–985, 2013.

[15] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis, "A Complete Optical Character Recognition Methodology for Historical Documents," pp. 525–532, 2008.

[16] L. S. Yaeger, B. J. Webb, and R. F. Lyon, "Search for Online , Printed Handwriting N EWTON," *Am. Assoc. Artif. Intell.*, vol. 19, no. 1, pp. 73–90, 1998.

[17] J. Hu, S. G. Lim, and M. K. Brown, "Writer independent on-line handwriting recognition using an HMM approach," *J. PATTERN Recognit. Soc.*, vol. 33, pp. 133–147, 2000.

[18] A. Funada, D. Muramatsu, and T. Matsumoto, "The Reduction of Memory and the Improvement of Recognition Rate for HMM On-line Handwriting Recognition," *IEEE*, pp. 0–5, 2004.

[19] J. r´ı Matas, "Real-Time Scene Text Localization and Recognition," *IEEE*, pp. 3538–3545, 2012.

[20] H. Lin and C. Hsu, "Optical Character Recognition with Fast Training Neural Network," *IEEE*, pp. 1458–1461, 2016.

[21] C. N. E. Anagnostopoulos, I. E. Anagnostopoulos, V. Loumos, and E. Kayafas, "A License Plate-Recognition Algorithm for Intelligent Transportation System Applications," *IEEE*, vol. 7, no. 3, pp. 377–392, 2006.

[22] Y. J. Zhang, "A survey on evaluation methods for image segmentation," pp. 1–13.

[23] A. Singh, K. Bacchuwar, and A. Bhasin, "A Survey of OCR Applications," *Int. J. Mach. Learn. Comput.*, vol. 2, no. 3, pp. 314–318, 2012

.