Optical Character Recognition Techniques: A Survey

Sukhpreet Singh
M.tech Student, Dept. of Computer Engineering, YCOE Talwandi Sabo BP. India.
Sukhpreet_dandiwal28@yahoo.com

ABSTRACT

This paper presents a literature review on English OCR techniques. English OCR system is compulsory to convert numerous published books of English into editable computer text files. Latest research in this area has been able to grown some new methodologies to overcome the complexity of English writing style. Still these algorithms have not been tested for complete characters of English Alphabet. Hence, a system is required which can handle all classes of English text and identify characters among these classes.

Keywords: *OCR, templates, English-alphabets, alpha-numerics*

1. INTRODUCTION

Optical Character Recognition [1] – [5] is a process that can convert text, present in digital image, to editable text. It allows a machine to recognize characters through optical mechanisms. The output of the OCR should ideally be same as input in formatting. The process involves some pre-processing of the image file and then acquisition of important knowledge about written text.

That knowledge or data can be used to recognize characters. OCR [1] is becoming an important part of modern research based computer applications. Especially with the advent of Unicode and support of complex scripts on personal computers, the importance of this application has increased.

The current study is focused on exploration of possible techniques to develop an OCR [2] system for English language when noise is present in the signal. A detailed analysis of English writing system has been done in order to understand the core challenges. Existing OCR systems are also studied to know the latest research going on in this field. The emphasis was on finding workable segmentation technique and diacritic handling for English strings, and built a recognition module for these ligatures. The complete methodology is proposed to develop an OCR system for English and a testing application is also made. Test results are reported and compared with the previous work done in this area.

2. MOTIVATION

In the last years the trend to digitize (historic) paper based documents such as books and newspapers, has emerged. The aim is to preserve these documents and make them fully accessible, searchable and process able in digital form. Knowledge contained in paper based documents is more valuable for today's digital world when it is available in digital form.

The first step towards transforming a paper based archive into a digital archive is to scan the documents. The next step is to apply an OCR (Optical Character Recognition) process, meaning that the scanned image of each document will be translated into machine process able text. Due to the print quality of the documents and the error-prone pattern matching techniques of the OCR process, OCR errors occur. Modern OCR processors have character recognition rates up to 99% on high quality documents. Assuming an average word length of 5 characters, this still means that one out of 20 words is defect. Thus, at least 5% of all processed words will contain OCR errors. On historic documents this error rate will be even higher because the print quality is likely to be of lower quality.

After finishing the OCR process several post-processing steps are necessary depending on the application, e.g. tagging the documents with meta-data (author, year, etc.) or proof-reading the documents for correcting OCR errors and spelling mistakes. Data which contains spelling mistakes or OCR errors is difficult to process. For example, a standard full-text search will not retrieve misspelled versions of a query string. To fulfill application's demanding requirements toward zero errors, a post-processing step to correct these errors is a very important part of the post-processing chain.

A post-processing error correction system can be manual, semi-automatic or fully automatic. A semi-automatic post-correction system detects errors automatically and proposes corrections to human correctors who then hive to choose the correct proposal. A fully-automatic post-correction system does the detection and correction of errors by its own. Because semi-automatic or manual corrections require a lot of human effort and time, fully-automatic systems become necessary to perform a full correction.

3. DESIGN OF OCR

Various approaches used for the design of OCR systems are discussed below:

Matrix Matching [6]: Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

Fuzzy Logic [6]: Fuzzy logic is a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black/ white etc. An attempt is made to attribute a more human-like way of logical thinking in the programming of computers. Fuzzy logic is used when answers do not have a distinct true or false value and there is uncertainly involved.

Feature Extraction [3]-[6]: This method defines each character by the presence or absence of key features, including height, width, density, loops, lines, stems and other character traits. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images.

Structural Analysis [6]: Structural Analysis identifies characters by examining their sub features- shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.

Neural Networks [6]: This strategy simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize characters through abstraction is great for faxed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns.

3.1 Structure of OCR Systems

Diagrammatic representation of the structure of an OCR system is given in figure 1.

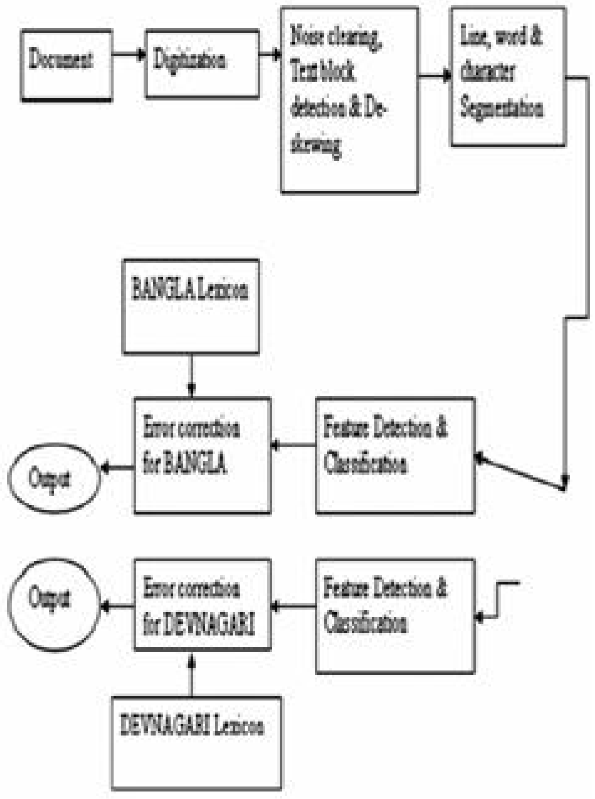
Document
Digitization
Noise clearing, Text block detection & De-skewing
Line, word & character Segmentation
BANGLA Lexicon
Error correction for BANGLA
Feature Detection & Classification
Output
Output
Error correction for DEVNAGARI
Feature Detection & Classification
DEVNAGARI Lexicon

Fig 1: Diagrammatic Structure of the OCR System

(adapted from [6])

3.2 Stages in Design of OCR Systems

Various stages of OCR system design are given in figure 2.

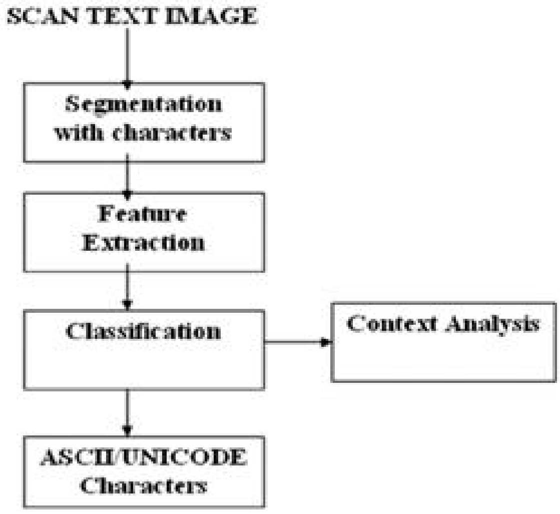

Fig 2: Stages in OCR Design (adapted from [6])

3.3 Reasons for Poor Performance of OCR Systems

Existing OCR systems generally show poor performance for documents like old books: print and paper quality inferior due to aging, Copied Materials: documents like photocopies or faxed documents, where print quality is inferior to the original, Newspapers: generally printed on low quality paper etc.

For such degraded documents, the system recognition accuracy comes down to 80- 90%. But if we want to use the OCR system for Banking and Corporate sector, this accuracy rate is not up-to-mark.

4. RELATED WORK

Claudiu et al. (2011) [1] has investigated using simple training data pre-processing gave us experts with errors less correlated than those of different nets trained on the same or bootstrapped data. Hence committees that simply average the expert outputs considerably improve recognition rates. Our committee-based classifiers of isolated handwritten characters are the first on par with human performance and can be used as basic building blocks of any OCR system (all our results were achieved by software running on powerful yet cheap gaming cards).

Georgios et al. (2010) [2] has presented a methodology for off-line handwritten character recognition. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the character image so that the resulting sub-images at each iteration have balanced (approximately equal) numbers of foreground pixels, as far as this is possible. Feature extraction is followed by a two-stage classification scheme based on the level of granularity of the feature extraction method. Classes with high values in the confusion matrix are merged at a certain level and for each group of merged classes, granularity features from the level that best distinguishes them are employed. Two handwritten character databases (CEDAR and CIL) as well as two handwritten digit databases (MNIST and CEDAR) were used in order to demonstrate the effectiveness of the proposed technique.

Sankaran et al. (2012) [3] has presented present a novel recognition approach that results in a 15% decrease in word error rate on heavily degraded Indian language document images. OCRs have considerably good performance on good quality documents, but fail easily in presence of degradations. Also, classical OCR approaches perform poorly over complex scripts such as those for Indian languages. Sankaran et al. (2012) [3] addressed these issues by proposing to recognize character n-gram images, which are basically groupings of consecutive character/component segments. Their approach was unique, since they use the character n-grams as a primitive for recognition rather than for post-processing.

By exploiting the additional context present in the character n-gram images, we enable better disambiguation S between confusing characters in the recognition phase. The labels obtained from recognizing the constituent n-grams are then fused to obtain a label for the word that emitted them. Their method is inherently robust to degradations such as cuts and merges which are common in digital libraries of scanned documents. We also present a reliable and scalable scheme for recognizing character n-gram images. Tests on English and
Malayalam document images show considerable improvement in recognition in the case of heavily degraded documents.

Jawahar et al. (2012) [4] has propose a recognition scheme for the Indian script of Devanagari. Recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts. This is mainly due to the complexity of the script, writing style etc. Our solution uses a Recurrent Neural Network known as Bidirectional Long- Short Term Memory (BLSTM). Our approach does not require word to character segmentation, which is one of the most common reason for high word error rate. Jawahar et al. (2012) [4] has 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.

Zhang et al. (2012) [5] has discussed the misty, foggy, or hazy weather conditions lead to image color distortion and reduce the resolution and the contrast of the observed object in outdoor scene acquisition. In order to detect and remove haze, this article proposes a novel effective algorithm for visibility enhancement from a single gray or color image. Since it can be considered that the haze mainly concentrates in one component of the multilayer image, the haze-free image is reconstructed through haze layer estimation based on the image filtering approach using both low-rank technique and the overlap averaging scheme. By using parallel analysis with Monte Carlo simulation from the coarse atmospheric veil by the median filter, the refined smooth haze layer is acquired with both less texture and retaining depth changes. With the dark channel prior, the normalized transmission coefficient is calculated to restore fogless image. Experimental results show that the proposed algorithm is a simpler and efficient method for clarity improvement and contrast enhancement from a single foggy image. Moreover, it can be comparable with the state-of-the-art methods, and even has better results than them.

Badawy, W. et al. (2012) [6] 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.

Ntirogiannis et al. (2013) [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. This paper addresses a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images. In the proposed evaluation scheme, the recall and precision evaluation measures are properly modified using a weighting scheme that diminishes any potential evaluation bias.

Yang et al. (2012) [8] has proposed a novel adaptive binarization method based on wavelet filter is proposed in this paper, which shows comparable performance to other similar methods and processes faster, so that it is more suitable for real-time processing and applicable for mobile devices. The proposed method is evaluated on complex scene images of ICDAR 2005 Robust Reading Competition, and experimental results provide a support for our work.

Sumetphong et al. (2012) [9] has proposed a novel technique for recognizing broken Thai characters found in degraded Thai text documents by modeling it as a set-partitioning problem (SPP). The technique searches for the optimal set-partition of the connected components by which each subset yields a reconstructed Thai character. Given the non-linear nature of the objective function needed for optimal set-partitioning, we design an algorithm we call Heuristic Incremental Integer Programming (HIIP), that employs integer programming

(IP) with an incremental approach using heuristics to hasten the convergence. To generate corrected Thai words, we adopt a probabilistic generative approach based a Thai dictionary corpus. The proposed technique is applied successfully to a Thai historical document and poor quality Thai fax document with promising accuracy rates over 93%.

AlSalman et al. (2012) [10] has proposed that Braille recognition is the ability to detect and recognize Braille characters embossed on Braille document. The result is used in several applications such as embossing, printing, translating...etc. However, the performance of these applications is affected by poor quality imaging due to several factors such as scanner quality, scan resolution, lighting, and type of embossed documents.

Mutholib et al. (2012) [11] has proposed that Android platform has gained popularity in recent years in terms of market share and number of available applications. Android operating system is built on a modified Linux kernel with built-in services such as email, web browser, and map applications. In this paper, automatic number plate recognition (ANPR) was designed and implemented on Android mobile phone platform. First, the graphical user interface (GUI) for capturing image using built-in camera was developed to acquire car plate number in Malaysia. Second, the preprocessing of raw image was done using contrast enhancement, filtering, and straightening. Next, an optical character recognition (OCR) using neural network was utilized to extract texts and numbers. The proposed ANPR algorithm was implemented and simulated using Android SDK on a computer. The preliminary results showed that our system is able to recognize most of the plate characters by almost 88%. Future research includes optimizing the system for mobile phone implementation with limited CPU and memory resources, and geo-tagging of the image using GPS coordinates and online database for various mobile applications.

Chi et al. (2012) [12] has proposed that because of the existence of possible carbon and seals, it's quite often that images of financial documents such as Chinese bank checks are suffered from bleed-through effects, which will affect the performance of automatic financial document processing such as seal verification and OCR. Chi et al. (2012) [12] has presented an effective algorithm to deal with bleed-through effects existing in the images of financial documents. Double-sided images scanned simultaneously are used as in-puts, and the bleed-through effect is detected and removed after the registration of the recto and verso side images.

Ramakrishnan et al. (2012) [13] has proposed that many machine learning algorithms rely on feature descriptors to access information about image appearance. Using an appropriate descriptor is therefore crucial for the algorithm to succeed. Although domain- and task-specific feature descriptors may result in excellent performance, they currently have to be hand-crafted, a difficult and time-consuming process. In contrast, general-purpose descriptors (such as SIFT) are easy to apply and have proved successful for a variety of tasks, including classification, segmentation, and clustering. Unfortunately, most general-purpose feature descriptors are targeted at natural images and may perform poorly in document analysis tasks. Ramakrishnan et al. (2012) [13] has proposed a method for automatically learning feature descriptors tuned to a given image domain. The method works by first extracting the independent components of the images, and then building a descriptor by pooling these components over multiple overlapping regions. We test the proposed method on several document analysis tasks and several datasets, and show that it outperforms existing general-purpose feature descriptors.

Chattopadhyay et al. (2012) [14] has worked on a low complexity video OCR system has been presented, that can be deployed on an embedded platform. The novelty of the proposed method is the use of low processing cycle and memory and yet getting a recognition accuracy of 84.23% which is higher than the usual video OCR recognition accuracy. Moreover, the proposed method can recognize about 180 characters on average per frame in 26.34 milliseconds.

Malakar et al.(2012)[15] 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. The present technique extracts 87.09% and 89.35% text lines successfully from the said databases respectively.

Sankaran et al. (2012) [16] has proposed a recognition scheme for the Indian script of Devanagari. Recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts. This is mainly due to the complexity of the script, writing style etc. Our solution uses a Recurrent Neural Network known as Bidirectional Long Short Term Memory (BLSTM). Our approach does not require word to character segmentation, which is one of the most common reason for high word error rate. Sankaran et al. (2012) [16] has 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.

Gur et al. (2012) [17] has discussed that text recognition and retrieval is a well known problem. Automated optical character recognition(OCR) tools do not supply a complete solution and in most cases human inspection is required. In this paper the authors 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. The authors focused on Rashi fonts associated with commentaries of the Bible that are actually handwritten calligraphy.

Devlin et al. (2012) [18] has discussed that when performing handwriting recognition on natural language text, the use of a word-level language model (LM) is known to significantly improve recognition accuracy. The most common type of language model, the n-gram model, decomposes sentences into short, overlapping chunks. In this paper, we propose a new type of language model which we use in addition to the standard n-gram LM. Devlin et al. (2012) [18]'s new model uses the likelihood score from a statistical machine translation system as a reran king feature. In general terms, we automatically translate each OCR hypothesis into another language, and then create a feature score based on how "difficult" it was to perform the translation. Intuitively, the difficulty of translation correlates with how well-formed the input sentence is. In an Arabic handwriting recognition task, Devlin et al. (2012) [18] were able to obtain a 0.4% absolute improvement to word error rate (WER) on top of a powerful 5-gram LM.

Al-Khaffaf et al. (2012) [19] has presented the current status of Decapod's English font reconstruction. The Pot race algorithm and its parameters that affect glyph shape are examined. The visual fidelity of Decapod's font reconstruction is shown and compared to Adobe Clears can. The font reconstruction details of the two methods are presented. The experiment demonstrates the capabilities of the two methods in reconstructing the font for a synthetic book typeset each time with one of six English fonts, three serif and three sans-serif. For both typefaces, Decapod is able to create a reusable TTF font that is embedded in the generated PDF document.

Rhead et al. (2012) [20] 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.

5. CONCLUSION

This paper has presented a related work on English OCR techniques. Various available techniques are studied to find a best technique. But is found that the techniques which provide better results are slow in nature while fast techniques mostly provide inefficient results. It is found that the OCR techniques based on neural network provide more accurate results than other techniques.

REFERENCES

[1] Dan ClaudiuCires¸an and Ueli Meier and Luca Maria Gambardella and JurgenSchmidhuber, “Convolutional Neural Network Committees for Handwritten Character Classification”, 2011 International Conference on Document Analysis and Recognition, IEEE, 2011.

[2] GeorgiosVamvakas, Basilis Gatos, Stavros J. Perantonis, “Handwritten character recognition through two-stage foreground sub-sampling” ,Pattern Recognition, Volume 43, Issue 8, August 2010.

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

[4] Naveen Sankaran and C.V Jawahar, “Recognition of Printed Devanagari Text Using BLSTM Neural Network”, IEEE, 2012.

[5] Yong-Qin Zhang, Yu Ding, Jin-Sheng Xiao, Jiaying Liu and Zongming Guo1, “Visibility enhancement using an image filtering approach”, Zhang et al. EURASIP Journal on Advances in Signal Processing 2012.

[6] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." (2012): 1-1.

[7] Ntirogiannis, Konstantinos, Basilis Gatos, and IoannisPratikakis. "A Performance Evaluation Methodology for Historical Document Image Binarization." (2013): 1-1.

[8] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu. "A fast adaptive binarization method for complex scene images." In Image Processing (ICIP), 2012 19th IEEE International Conference on, pp. 1889-1892. IEEE, 2012.

[9] Sumetphong, Chaivatna, and SupachaiTangwongsan. "An Optimal Approach towards Recognizing Broken Thai Characters in OCR Systems." Digital Image Computing Techniques and Applications (DICTA), 2012 International Conference on. IEEE, 2012.

[10] AlSalman, AbdulMalik, et al. "A novel approach for Braille images segmentation." Multimedia Computing and Systems (ICMCS), 2012 International Conference on. IEEE, 2012.

[11] Mutholib, Abdul, Teddy Surya Gunawan, and Mira Kartiwi. "Design and implementation of automatic number plate recognition on android platform." Computer and Communication Engineering (ICCCE), 2012 International Conference on. IEEE, 2012.

[12] Chi, Bingyu, and Youbin Chen. "Reduction of Bleed-through Effect in Images of Chinese Bank Items." Frontiers in Handwriting Recognition

(ICFHR), 2012 International Conference on. IEEE, 2012.

[13] Ramakrishnan, Kandan, and Evgeniy Bart. "Learning domain-specific feature descriptors for document images." Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on. IEEE, 2012.

[14] Chattopadhyay, T., Ruchika Jain, and Bidyut B. Chaudhuri. "A novel low complexity TV video OCR system." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

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

[16] Sankaran, Naveen, and C. V. Jawahar. "Recognition of printed Devanagari text using BLSTM Neural Network." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

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

[18] Devlin, Jacob, "Statistical Machine Translation as a Language Model for Handwriting Recognition." Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.

[19] Al-Khaffaf, Hasan SM, et al. "On the performance of Decapod's digital font reconstruction." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

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