REVIEW ON OPTICAL CHARACTER RECOGNITION

Muna Ahmed Awel[1], Ali Imam Abidi[2]

[1]*Computer Science and Engineering, Sharda University, Greater Noida, India*
[2]*Assistant Professor, Computer Science and Engineering, Sharda University, Greater Noida, India*

Abstract - *Optical Character Recognition is the area of Pattern Recognition that has a topic of studies over the past some decades. Optical character recognition is technique of automatically identifying of different character from a record picture additionally provide full alphanumeric recognition of printed or handwritten characters, text numerical, letters, and symbols in to a computer process able layout including ASCII, Unicode and so forth. Optical character recognition is the bottom for many distinct styles of programs in diverse fields, a lot of which we use in our daily lives. Cost effective and less time consuming, corporations, submit offices, banks, security systems, and even the field of robotics hire this system as the base in their Operations. These days, there are numerous portions of research and making use of OCR technology. These OCR technologies help to examine unique documents written in English, Chinese, Hindu, Arabic, Russian, and others languages. On This paper present review of some researches has been made in English, Arabic and Devanagari characters. And explained the methodology they use and challenge they face during development of Optical character recognition.*

Key Words: *OCR, optical character recognition, character recognition, handwriting character recognition.*

1. INTRODUCTION

Character recognition, usually abbreviated to optical character recognition or shortened OCR, is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text [4]. It is a field of research in pattern recognition, artificial intelligence and machine vision. Though academic research in the field continues, the focus on character recognition has shifted to implementation of proven techniques. Optical character recognition technology was invented in the early 1800s, when it was patented as reading aids for the blind. In 1870, C. R. Carey patented an image transmission system using photocells, and in 1890 P.G. Nipkow invented sequential scanning OCR. However, the practical OCR technology used for reading characters was introduced in the early 1950s as a replacement for keypunching system [2]. A year later, D.H. Shephard developed the first commercial OCR for typewritten data. The 1980's saw the emergence of OCR systems intended for use with personal Computers. Nowadays, it is common to find PC-based OCR systems that are commercially available. However, most of these systems are developed to work with Latin-based scripts. Optical character recognition systems for Latin characters have been available for over a decade and perform well on clear typed text. There are research has also been directed at other non-Latin scripts such as Arabic, Japanese, Chinese, Hindu, Tibetan. In order to develop an OCR system it requires the development and integration of many sub systems. The first step is preprocessing such as skew detection and correction, noise detection and removal, binearization, thinning, and normalization. Then segmentation of document images into line, word and characters. This is followed by feature extraction for representing character images and a classification module that label characters to their proper class. Finally, post processing i.e. applying

2. LITERATURE REVIEW

Character recognition technique has been completed through studies on different characters for example, English, Arabic, Chinese, Devanagari, Bangla, Farsi and Kannada and so on. Totally, the complete method is carried out in three phase Preprocessing, Feature extraction and recognition[5]. In this paper only cover the study has been done on English, Arabic and Devanagari scripture.

2.1 In English Scripter Character Recognition

In 2004 N. M. Noor, M. Razaz and P. Manley-Cooke Proposed system using global geometrics feature extraction and geometric density classifier for feature extraction then neural fuzzy logic used for classification. Evaluation of the system has achieve for Geometric Density 77.89% and Geometric Feature 76.44% accuracy rate [6]. In 2010 Dewi Nasien, Habibollah Haron and Siti Sophiayati Yuhaniz This studies Take three datasets from NIST database considered lowercase letters 189,411, uppercase letters 217,812 and combination of uppercase and Lowercase letters 407,223 sample are used. Those Samples are divided into 80% for training and 20% for testing. For feature extraction used Freeman chain code (FCC). Support vector system (SVM) is selected for recognition step. The method recognize for the first dataset 86% accuracy, second dataset 88% of accuracy and third dataset 73% accuracy achieved [7]. In 2011 Vijay Patil and Sanjay Shimpi develop system that recognize handwritten English character using neural network for feature extraction system they used Character Matrix And for recognition back propagation neural network used. The result indicate that hand back propagation network provide more than 70% of accuracy rate[3]. 2015 M. S. Sonawane and Dr. C.A. Dhawale this study compare and evaluate two classifier which is artificial neural network and nearest neighbor. Used grid method to extract feature and the result show nearest neighbor achieve 61.53% accuracy when neural network gives 57.69%. Math lab tool was used for features extracted and recognition. The evaluation outcome suggests Nearest Neighbor is a better recognizer comparing with artificial neural network when implemented to English Characters[8].

2.2. Arabic Scripter Character Recognition

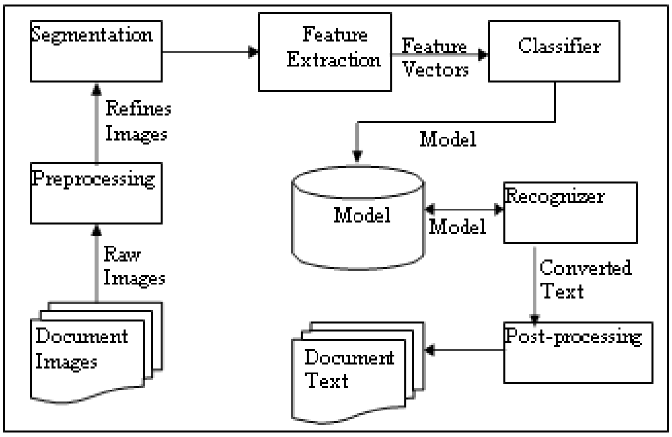
In 2002 Majid M. Altuwaijri and Magdy A. Bayoumi They develop system to recognize Arabic text using neural network used set of moment invariants descriptors (under shift, scaling and rotation) and artificial neural network (ANN) used for classification The study has shown 90% of a high accuracy rate [9]. In 2015 Ashraf Abdel Raouf, Colin A. Higgins, Tony Pridmore and Mah-moud I. Khalil Haar studied approach for recognizing Arabic character using Haar Cascade Classifier (HCC) These classifiers were trained and tested on some 2,000 images. To extract feature Haar-like feature extraction used and boosting of a classifier cascade. The system was tested with real text image and produces 87% accuracy rate for Arabic character recognition[10]. In 2017 N. Lamghari, · M. E. H. Charaf and · S. Raghay On this research the data are divided into three parts. From 34,000 character 70% are used for training, 15% for testing phase and 15% for validation. To extract feature hybrid feature extraction used (pixel density, resize, freeman code, structural features, invariant) for recognition used feed forward-back propagation neural network. The system has achieved 98.27% high recognition rate[11]. In 2018 Noor A. Jebrila, Hussein R. Al-Zoubib and Qasem Abu Al-Haijac In addition to the preprocessing step includes in particular three levels. In the primary section, they employed word segmentation to extract characters. In the second one section, Histograms of Oriented Gradient (HOG) are used for feature extraction. The very last phase employed Support Vector Machine (SVM) for classifying characters. They have carried out the proposed method for the recognition of Jordanian metropolis, city, and village names as a case examine, similarly to many other phrases that offers the characters shapes that aren't included with Jordan cites. The set has cautiously been selected to include each Arabic character in its all forms. To the conclusion, they have got constructed their own dataset inclusive of greater than 43.000 handwritten Arabic phrases (30000 used for training and 13000 used for testing stage). Recognition result show 99% rate of accuracy[12].

2.2 Devanagari Scripter Character Recognition

In 2011 Gyanendra K. Verma, Shitala Prasad, and Piyush Kumar Curvelet present in approach for Hindi handwritten character recognition using curvelet transformer. The study are used dataset that contain 200 images of character (each image contains all Hindi characters). Feature extract using curvelet transform and for recognition k-nearest neighbor the experiment result show more than 90% accuracy [13]. In 2013 Divakar Yadav, Sonia Sánchez-Cuadrado and Jorge Morato develop optical character recognition system using neural network for Hindi characters and trained with 1000 dataset. Feature extraction technique is histogram of projection based on mean distance, on pixel values and vertical zero crossing. Then classify using back-propagation neural network with two hidden layers. Experimental result show 98.5% correct recognition[14]. In 2015 Akanksha Gaur and Sunita Yadav this system extract feature using k-means clustering and classified used support vector machine using linear kernel and Euclidean distance. The evaluation show that SVM has better results using linear kernel than Euclidean distance. Maximum achieved using Euclidean distance is 81.7% accuracy. Using linear kernel giving 95.86% result[5]. In 2018 Nikita Singh present system with the title "An Efficient Approach for Handwritten Devanagari Character Recognition Based on Artificial Neural Network" for recognition hind character. For feature extraction they used histogram oriented gradients (HOG) and recognition used artificial neural network (ANN) classifier. The system get 97.06% high accuracy [15].

3. MAJOR STEPS INVOLVE IN CHARACTER RECOGNITION

Building an OCR engine is not an easy thing to do as the main difficulty lies with – identifying each character and word. For making an OCR engine from scratch below are the steps which one can follow to make sure that the OCR meets the desired expectation of character recognition and this is the methodology and the steps most of researchers used.



3.1 Optical Scanning

To start with an OCR, image can be capture by digital camera also but after seeing the challenge been faced in privies work better to use scanner therefor consider first need putting together a good optical scanner. With the help of this scanner, an image of original file or document is captured. It is commanding to select scanner with a good sensing tool and transport mechanism.

3.2 Pre-processing:

Preprocessing is performing different operation on the scanned or input image. It helps to remove noise from image make character clear and It basically enhances the image rendering suitable for segmentation. Preprocessing has various task are such as converting gray scale, binarization, thinning, skewing and normalization.

3.3 Segmentation:

Once the preprocessing produces noise free clean character image, it's then segmented into several subcomponents. There are three steps of segmentation first line segmentation divide the character in image horizontal second word segmentation the divide words from line sentence last character segmentation divide the characters from word. Finally we get segmented characters those character help for feature extraction and recognition.

3.4 Feature extraction:

This is one of the riskiest components in an OCR development. The main aim is to extract important patter from characteristics. The selected features are expected to contain pattern that differentiate one character from other and relevant information from the input data, so that the classification can be performed by using those patter extract from segmented character this instead of the complete original data.

3.5 Training and recognition:

Investigation of OCR's pattern recognition can be done via template matching, statistical technique, syntactic or structural techniques, and artificial neural networks. The system also have to be learn in such a way that the problem associated to incomplete vocabulary is solved.

3.6 Post-processing:

In this final process, activities like grouping, error detection and correction take place. During grouping, symbols in the text are associated with strings. However, it's impossible to reach 100% accurate identification of characters, only some of the errors can be detected and deleted as per the context.

4. CHALLENGES OF OPTICAL CHARACTER RECOGNITION

For better and high character recognition accuracy there are so many OCR techniques but still difficult to achieve 100% correct recognition especially for character that has similarity. The challenges I observe during review is many of them related to the data collection and preprocessing if we can identify and rid of those challenges we can get high correct recognition. The following issues created due to collecting input data using digitals camera. Instead of using camera to capture characters or scripts prefer to scan the document but let's see what those challenges are.

4.1 Scene Complexity

Input data taken with camera may have other object is also for example building, homes, panting and other objects to separate those objects from text or character is very tough. The data that content non textual contents make preprocessing difficult there for affect the character recognition process.

4.2 Conditions of Uneven Lighting

Many times image taken from road or outdoor affected by light and shadows. This is another challenge for optical character recognition. It make difficult to detect and segment characters. This kind of issues makes scanning document more preferable than capturing it by camera. Camera light flash also may help for additional lighting and create shadows in images.

4.3 Skewness (Rotation)

Image taking using camera also disturb by this issue. The angle of the image incorrect therefor when we fed this data to optical character recognition system the outcome will be incorrect. But there are techniques to solve this problem like Fourier transformer, projection profile, Hough transform and so on.

4.4 Blurring and Degradation

This also caused by image taken with camera. This happen when images are taken from distance, trying to capture on movements and Lack of focusing. Image taking on this and other circumstance face blurring and degradation. For segmentation and accurate recognition sharpness of characters is needed.

4.5 Fonts and style

Characters that are connected each other like Arabic, Hindi and fonts style like Italic and other overlap each other this make difficult for optical character recognition system during segmentation process hard to detect and divide words in to character.

4.6 Multilingual Environments

Characters that have multi environment such as, language that has large number of character Ethiopian, Korean, Chinese, Japanese and other. Characters that written connectedly with each other Arabic language. Ethiopian language Amharic alphabet similarity of characters it's difficult for computer to see the difference between most of them. Therefor this kind of multi environmental characters are challenges for OCR to divide and extract individual characters and recognize correctly.

4.7 Damage documents

When the input document are very old and damage whether we take it in camera or scanned will be very difficult to observe the character, content many noise when we try to remove those noise sometime the data or image lose it necessary content or characters.

5. CONCLUSION

In the research works revised in this paper, character recognition system use different approaches and many of them get good accuracy. What we can understand from this paper is feature extraction techniques should be choose according to the character you working because each scripts or alphabets has its own nature therefor need to find techniques which fit or suitable for characters. The better able to extract features from character more we can detect and recognize characters in highest accuracy.

ACKNOWLEDGMENT

First and foremost I thank my supervisor Dr. Ali Imam Abidi for his guidance in my research work. He gave me the best advice and helping me with provide me necessary document for my researches" **Review on Optical Character Recognition**" ,Who also helped me in the survey of the related work of different authors .I came to know about so many new things I am really thankful to them for his care and support.

REFERENCES

[1] J. Cowell and F. Hussain, “Amharic character recognition using a fast signature based algorithm,” *Proc. Int. Conf. Inf. Vis.*, vol. 2003–Janua, pp. 384–389, 2003.

[2] I. Stoianov, “Optical Character Recognition of Historical Documents,” *Clover.Slavic.Pitt.Edu*, 1995.

[3] V. Patil and S. Shimpi, “Handwritten English character recognition using neural network,” *Elixir Comp. Sci. Engg*, vol. 41, no. 3, pp. 5587–5591, 2011.

[4] K. A. Okrah, “Nyansapo (the wisdom knot): Toward an African philosophy of education,” *Nyansapo (The Wisdom Knot) Towar. an African Philos. Educ.*, no. 224, pp. 1–121, 2003.

[5] A. Gaur and S. Yadav, “Handwritten Hindi character recognition using k-means clustering and SVM,” *2015 4th Int. Symp. Emerg. Trends Technol. Libr. Inf. Serv. ETTLIS 2015 - Proc.*, pp. 65–70, 2015.

[6] N. M. Noor, M. Razaz, and P. Manley-Cooke, “Global geometry extraction for fuzzy logic based handwritten character recognition,” *Proc. - Int. Conf. Pattern Recognit.*, vol. 2, pp. 513–516, 2004.

[7] D. Nasien, H. Haron, and S. S. Yuhaniz, “Support Vector Machine (SVM) for english handwritten character recognition,” *2010 2nd Int. Conf. Comput. Eng. Appl. ICCEA 2010*, vol. 1, pp. 249–252, 2010.

[8] M. S. Sonawane and C. A. Dhawale, “Evaluation of Character Recognisers: Artificial Neural Network and Nearest Neighbour Approach,” *2015 IEEE Int. Conf. Comput. Intell. Commun. Technol.*, pp. 129–132, 2015.

[9] M. M. Altuwaijri and M. A. Bayoumi, “Arabic text recognition using neural networks,” pp. 415–418, 2002.

[10] A. AbdelRaouf, C. A. Higgins, T. Pridmore, and M. I. Khalil, “Arabic character recognition using a Haar cascade classifier approach (HCC),” *Pattern Anal. Appl.*, vol. 19, no. 2, pp. 411–426, 2016.

[11] N. Lamghari, M. E. H. Charaf, and S. Raghay, “Hybrid Feature Vector for the Recognition of Arabic Handwritten Characters Using Feed-Forward Neural Network,” *Arab. J. Sci. Eng.*, vol. 43, no. 12, pp. 7031–7039, 2018.

[12] N. A. Jebril, H. R. Al-Zoubi, and Q. Abu Al-Haija, “Recognition of Handwritten Arabic Characters using Histograms of Oriented Gradient (HOG),” *Pattern Recognit. Image Anal.*, vol. 28, no. 2, pp. 321–345, 2018.

[13] R. Rani, R. Dhir, and G. S. Lehal, “Information Systems for Indian Languages,” *Commun. Comput. Inf. Sci.*, vol. 139, no. January 2016, pp. 174–179, 2011.

[14] D. Yadav, S. Sánchez-Cuadrado, and J. Morato, “Optical character recognition for Hindi language using a Neural-network approach,” *J. Inf. Process. Syst.*, vol. 9, no. 1, pp. 117–140, 2013.

[15] N. Singh, “An Efficient Approach for Handwritten Devanagari Character Recognition based on Artificial Neural Network,” *2018 5th Int. Conf. Signal Process. Integr. Networks, SPIN 2018*, pp. 894–897, 2018.