A Study of Novel Optical Character Recognition Algorithms

*Roshan Suvaris[1] and Dr. S Sathyanarayana[2]
[1]*Research Scholar, Bharathiar University, Coimbatore*
[2]*Sreekantha First Grade Women's College, Mysore*
[1]*roshansuaris@gmail.com*, [2]*ssn_mys@yahoo.com*

Abstract: *The Optical Character Recognition has been the inseparable part of human life during everyday transaction. The OCR has extended its application areas in almost all fields viz. healthcare, finance, banking, entertainment, trading system, digital storage and so on. In the recent past, handwriting recognition is one of the hardeststudy areas in the area of image processing. In this paper the various techniques for converting textual content from number plates, printed, handwritten paper document into machine code have been discussed. The transforming method used in all these techniques is known as OCR. The English OCR system is necessary for the conversion of various published books and other documents in English into human editable computer text files. Latest researches in this area have included the methodologies thatidentify different fonts and styles of English hand written scripts. As of date, even though a number of algorithms are available, it has its own pros and cons. Since, recognition of different styles and fonts in machine printed and handwritten English script is a biggest challenge, this field is open for the researchers to implement new algorithms that would overcome the deficiencies of its predecessors.*

Keywords- **OCR, neural network, segmentation, hidden markov, fuzzy logic**

1. INTRODUCTION

The process that converts documents which are in image format into editable text format is known as Optical Character Recognition (OCR). This yields the same output as input image in formatting. The OCR has supported scanned document images to become more than just image files, changing into text documents which can be searchable and recognizable by computers.

The text can be mined from the input image with the OCRoperations and may be stored in electronic database. OCR is a three-step process. In the initial step, the document is scanned and converted into document image. In the next step some complex processing is applied to extract the characters from the image and converting it into editable ASCII characters. In the final step verification is performed to check the characters whether it is correct or not.

Mainly character recognition has been separated into two main categories machine printed character recognition and handwritten recognition. The handwritten character recognition once again separated into offline character recognition and online character recognition. The online character recognition performed while the user is writing the character in real time. The offline character recognition is performed on printed handwritten text images [1].

2. LITERATURE SURVEY

In this literature review, the latest research work which focuses on the area of optical character recognition with regard to different types of text documents such as printed, handwritten online and offline, have been reviewed.

The method proposed by Abdul Robby G, Antonia Tandra et al., [2] recognizes the characters using tesseract OCR tools. In this research non Latin characters that is Javanese characters are recognized. In this method the model has been trained using 5880 characters and the model has been implemented using

android application. They have attained a highest accuracy of 97.50% by uniting single boundary box for the full character and for the main body separate boundary boxes.

NabeelOudah. M, EstabraqAbdulredaa, Maher FaikEsmaile et al., [3] in their proposed method used weiner filter to filter the image and then active contour is used to segment the character in the image. In the final step correlation technique and template-equivalentor matching technique is used to recognize the characters. The Tesseract OCR is used evaluate the performance of the defined technique

Muhammad NaeemAyyaz, Waqar Mahmood and Imran Javed [4] proposed a feature extraction method in combination with SVM classifier, and it has offered good performance result on handwritten numbers and uppercase letters.

Xu-Yao Zhang, Cheng-Lin Liu and YoshuaBengio [5] in their research paper to achieve the maximum accuracy they have integrated the old-style normalization-cooperated direction-decomposed feature map and CNN to recognize the online and offline Chinese characters of HCCR and ICDAR -2013 database.

N. Venkata Rao, Dr. A.S.C.S.Sastry et al., [6] proposed method uses neural networks to recognize the handwritten offline character. It is an altered back propagation technique which produces good character recognition accuracy.

Christian BartzHaojin Yang ChristophMeinel [7] in their method spatial transformer network [STN] that can learn and can detect the text regions which are present in the image and it is capable of recognizing the text using text recognition network. It is a single deep NN semi supervised network. The network is capable of recognizing text in different datasets.

The method proposed by ZbigniewWojna, Alexander N Gorban et al., [8] uses convolutional neural network, Recurrent NN and novel attention mechanism to recognize French street name signs and achieved 84.2% accuracy. The proposed method also works well on google street view dataset as well as it is fast and accurate.

Vladimir Rybalkin, Norbert When et al., [9] in their method used hardware architecture of bidirectional LSTM networks along with connectionist temporal classification. They have also used FGPA hardware accelerator which produced 459 times higher throughput. Using this method online and offline handwritten characters and printed characters are recognized.

BingcongLi , Xin Tang et al., [10] proposed a lightweight Chinese and other language text recognition method named Hamming OCR. To identify or to encode the character hamming classifier used which uses locality sensitive hashing and the produced LSH code is straightly applied to substitute the output embedding. In this method cross layer parameter sharing method is used instead of feed forward network. Hamming OCR achieved good results on some competitive datasets.

GauriKatiyar and ShabanaMehfuz [11] used feed forward neural network with three layers and it combines four layers of feature extraction such as diagonal distance approach, box approach, gradient and mean operations to recognize the handwritten offline characters. The proposed method works well on CEDAR dataset and produces good recognition rate.

Durjoy Sen Maitra, Ujjwal Bhattacharya et al., [12] proposed method uses CNN and SVM. In this method CNN is used extract the features of 50 class bangla basic characters and different English numerals and the SVM used for the classification of the characters based on the features extracted by CNN.

Zheheng Rao, Chunyan Zeng et al., [13] proposed method uses extended nonlinear kernel residual network-based offline handwritten character recognition. Using this algorithm they have reduced the training time and recognition accuracy.

FeiXie , Ming Zhang et al., [14] proposed a new model which is a combination of feature extraction model and back propagation neural network which is used for number plate detection and character

recognition. The method has achieved a 97.7% accuracy and the method is tested on different images with complex backgrounds.

The method proposed by Shrinivas R. Zanwar ,Sandipann P. Narote et al., [15] uses independent component analysis, swarm intelligence and firefly algorithm together for feature extraction and to recognize the characters NN is used. The algorithm achieves high recognition for efficient classification.

NaumanSaleem, HassamMuazzam et al., [16] used vertical edge detection and the image is normalized using normalization technique. Using morphological and statistical technique the license plate region is extracted. Finally template matching is for recognizing the characters. The methodology is employed on different number plate images with variant illuminations and achieved an accuracy of 84.8% with less execution time.

Shrawan Ram, Shloak Gupta et al., [17] proposed method to recognize handwritten Devanagari characters in which they have used deep convolutional neural network with ReLU and dropouts and it produces decent results. The model is trained using UCI repository.

The method proposed by S´ergioMontazzolli Silva, and Cl´audioRosito Jung et al., [18] uses the novel convolutional neural network for detecting and correcting the number plates. Once this process is completed Optical character recognition is applied to the number plates and characters are identified. The method performs well on different types of number plates with many distortions.

NishaSharma ,Bhupendra Kumar and Vandita Singh [19] proposed method takes the segmented characters and applies statistical, directional and geometric methods have been applied to extract the character features. The multilayer perceptron neural net with back propagation and SVM is used for the classification of the characters. This method is applied for handwritten English upper and lower letters, numbers and for symbols. The proposed method achieves 98% accuracy for numerals, for special characters 96.5% accuracy, for uppercase 95.35% accuracy and for small letters 92% accuracy is achieved.

The method proposed by Mohammad RezaSoheili, Mohammad Reza Yousefiet al., [20] used one dimensional LSTM network on grey scale images. The images used for OCR is not binarized. The network achieves less error rate low and high resolutions.

Yu Weng and Chunlei Xia [21] proposed method uses CNN to identify the handwritten Chinese offline characters using mobile devices.

NehaGautam and Soo See Chai [22] proposed method used to recognize the printed and handwritten brahmi characters. The technique uses thresholding,cropping and thinning for preprocessing the image. Character segmentation is performed using line and character detection methods. The geometric method used for feature extraction and zone and geometric method classification is used to recognize the characters in the final step. The method achieves the accuracy of 94.10% and 90.62% for printed and handwritten characters respectively.

Parulsahare& Sanjay B.Dhok [23] proposed method to recognize multilingual indian printed and handwritten characters. The characters in this method is segmented using structural property for unjointed characters and graph distance for joined characters. Three geometrical features are extracted and for identifying the characters k-nearest neighbor method is used. The proposed method has achieved good recognition rate for printed and handwritten characters.

Rio Anugrah, and KetutBayuYoghaBintoro [24] proposed method uses median filter and otsu's function for preprocessing and characters are segmented using connected components labeling. Finally a neural network called ANN is used for extraction of character features and character classification. This methods accuracy is based on the training of the data.

The method proposed by N Gomathi and A.K. Sampath[25] used for handwritten character recognition. First the image is preprocessed and then features are extracted using histogram of oriented

gradient. Finally character are classified using multi kernel function and it is based on fuzzy triangular membership function. The proposed methodology achieves higher accuracy for different document and document sets.

3. MAJOR PHASES OF OCR

The OCR is a complex process, which consists of following phases:

3.1 Image acquisition: It is the process of capturing or taking the scanned or camera or stored image for processing. Optical character recognition development and research can be traced back to early 1950s. The scanners were used to scan the images earlier was line scanners. After continuous research in this field full page scanner was introduced and later optical character recognition techniques were implemented to recognize the characters [26]. Now a days images are acquired using scanners cameras etc.

3.2 Preprocessing: After image is acquired different preprocessing steps such as binarization, Noise reduction, Skew Correction, Slant Removal are performed. The binarization or thresholding is process which converts the grey scale image to binary image. Unnecessary information can be removed using this method. Once the image is converted to binary image other methods are applied to highlight the text in the images and to improve the quality of the text.Filters are used such as averaging, max and min globally or locally to get quality image. After this differenttasks are performed such as dilation, erosion and thinning on the character document image.

3.3 Character segmentation: The individual characters from the document should be extracted to pass it to some model for recognition. So the character segmentation is a very important process. Once the preprocessing is performed the document is taken and different segmentation methods are applied such as line, word and finally character segmentation to get the characters for future processing.

3.4 Feature extraction*:* In this step various features of different characters are extracted which is used to identify the character. The selection of the right features and the total number of features to be used is an important research question.

3.5 Character classification: It is process of recognizing the objects or characters and recognizing to which category the character belongs. The image components relationships and feature extracted using different methods have been used for the grouping of the characters. Different set of characters are used for classification that is capital letters [A-Z], small letters [a-z] numbers [0-9] and special characters.

3.6 Post Processing: It includes the datadelivered by the image processing unit and contextual data to identify the recognition errors. It is the final step performed in OCR. The recognition errors may be used again to correct the classification errors which occurred during the initial process.

4. CONCLUSION

Although there are many algorithms available for character recognition, still further scope exists for the development of novel algorithms. This is due to the fact that there is possibility of generation of numerous images based the handwriting and printed document character styles and fonts. In this work, authors have discussed about the most recent methodologies, which work well in different scenarios and different types of text such as printed, online handwritten, offline handwritten and number plates etc. The important parameter has been considered in this study is accuracy of the output. Some of the methods used for identification are fuzzy logic, Support vector machines, neural networks, LSTM etc, and these methods achieve certainly better results.

REFERENCES:

[1]. Noman Islam, Zeeshan Islam, Nazia Noor “A Survey on Optical Character Recognition System” Journal of Information & Communication Technology-JICT Vol. 10 Issue. 2, December 2016.

[2]. Abdul Robby G, Antonia Tandra, Imelda Susanto, JeklinHarefa, AndryChowanda “Implementation of Optical Character Recognition using Tesseract with the Javanese Script Target in Android Application” 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12-13 September 2019.

[3]. NabeelOudah. M, Maher FaikEsmaile, EstabraqAbdulredaa “Optical Character Recognition Using Active Contour Segmentation” Number 1 Volume 24 January 2018 Journal of Engineering.

[4]. Muhammad NaeemAyyaz , Imran Javed and Waqar Mahmood “Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction” Pak. J. Engg. & Appl. Sci. Vol. 10, Jan., 2012

[5].Xu-Yao Zhang ,YoshuaBengio , Cheng-Lin Liu “Online and Offline Handwritten Chinese Character Recognition: A Comprehensive Study and New Benchmark” Computer Vision and Pattern Recognition 2016.

[6]. N. Venkata Rao, DR. A.S.C.S.Sastry, A.S.N.Chakravarthy, Kalyanchakravarthi P “Optical character recognition technique algorithms” Journal of Theoretical and Applied Information Technology January 2016.

[7]. Christian BartzHaojin Yang ChristophMeinel “STN-OCR: A single Neural Network for Text Detection and Text Recognition” Computer Vision and Pattern Recognition 2017.

[8]. ZbigniewWojna, Alex Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, Julian Ibarz “Attention-based Extraction of Structured Information from Street View Imagery” 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 2017.

[9]. Vladimir Rybalkin, Norbert Wehn, Mohammad Reza Yousefi and Didier Stricker “Hardware Architecture of Bidirectional Long Short-Term Memory Neural Network for Optical Character Recognition” Design, Automation and Test in Europe (DATE) 2017.

[10].BingcongLi , Xin Tang , Xianbiao Qi ,Yihao Chen, Rong Xiao “Hamming OCR: A Locality Sensitive Hashing Neural Network for Scene Text Recognition”

[11]. GauriKatiyar and ShabanaMehfuz “MLPNN Based Handwritten Character Recognition Using Combined Feature Extraction” International Conference on Computing, Communication and Automation (ICCCA2015)

[12]. Durjoy Sen Maitra, Ujjwal Bhattacharya and Swapan K. Parui “CNN Based Common Approach to Handwritten Character Recognition of Multiple Scripts” International Conference on Document Analysis and Recognition (ICDAR) 2015.

[13]. Zheheng Rao, Chunyan Zeng, Minghu Wu, Zhifeng Wang, Nan Zhao, Min Liu,Xiangkui Wan “Research on a handwritten character recognition algorithm based on an extended nonlinear kernel residual network”

[14]. FeiXie , Ming Zhang , Jing Zhao, Jiquan Yang, Yijian Liu and Xinyue Yuan “A Robust License Plate Detection and Character Recognition Algorithm Based on a Combined Feature Extraction Model and BPNN” Journal of Advanced Transportation, Volume 2018.

[15]. Shrinivas R. Zanwar , Sandipann P. Narote , Abhilasha S. Narote , Ulhas B. Shinde “An Effectual Optical Character Recognition Using Efficient Learning System” International Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019)

[16]. NaumanSaleem, HassamMuazzam, H.M.Tahir, Umar Farooq “Automatic License Plate Recognition Using Extracted Features” International Symposium on Computational and Business Intelligence 2016.

[17]. Shrawan Ram, Shloak Gupta &Basant Agarwal “Devanagri character recognition model using deep convolution neural network” Journal of Statistics and Management Systems 2018.

[18]. S´ergioMontazzolli Silva, and Cl´audioRosito Jung “License Plate Detection and Recognition in Unconstrained Scenarios” Conference: European Conference on Computer Vision (ECCV 2018).

[19]. NishaSharma ,Bhupendra Kumar and Vandita Singh “Recognition of Off-line Hand printed English Characters, Numerals and Special Symbols” 5th International Conference- Confluence The Next Generation Information Technology Summit (Confluence) 2014.

[20]. Mohammad Reza Yousefi, Mohammad Reza Soheili, Thomas M. Breuelt, EhsanollahKabir and Didier Stricker “Binarization-free OCR for Historical Documents Using LSTM Networks” 13th International Conference on Document Analysis and Recognition (ICDAR) 2015.

[21]. Yu Weng and Chunlei Xia “A New Deep Learning-Based Handwritten Character Recognition System on Mobile Computing Devices” Mobile Networks and Applications 2019

[22].Neha Gautam and Soo See Chai “Optical Character Recognition for Brahmi Script Using Geometric Method” Journal of Telecommunication, Electronic and Computer Engineering e-ISSN: 2289-8131 Vol. 9 2017.

[23]. PARUL SAHARE AND SANJAY B. DHOK “Multilingual Character Segmentation and Recognition Schemes for Indian Document Images” IEEE Access 2016.

[24]. Rio Anugrah, and KetutBayuYoghaBintoro “Latin Letters Recognition Using Optical Character Recognition to Convert Printed Media in to Digital Format” JurnalElektronikadan Telekomunikasi (JET), Volume. 17, December 2017

[25]. A.K. Sampath and N Gomathi “Fuzzy-based multi-kernel spherical support vector machine for effective handwritten character recognition” Sādhanā42, 2017.

[26]. Mohammed cheriet, Nawwaf karma, Ching Y Suen, Cheng Lin Liu “Character recognition systems – A guide for students and practitioners” A John wiley and sons inc, Publication.