AN INVESTIGATION ON HANDWRITTEN CHARACTER RECOGNITION

B. Soujanya* and T. Sitamahalakshmi
Assistant Professor, Professor
Department of Computer Science and Engineering
GITAM (Deemed to be university), Visakhapatnam, Andhra Pradesh, India.

Abstract: The machine replication of human reading has been the subject of intensive research for more than three decades. A large number of research papers and reports have already been published on this topic. Many commercial establishments have manufactured recognizers for varying capabilities. The ultimate goal of developing a reading machine having the same reading capabilities of human still remains unachieved. So, there is still a great gap between human reading and machine reading capabilities and scope for future work. This paper discusses the different techniques, methodologies and research done in character recognition.

***Index Terms*–Character recognition, OCR, Pre-processing methods, Feature Extraction, Review**

I.INTRODUCTION:

Character recognition is one of the most challenging in the field of pattern recognition which presents very useful applications. It is widely used in areas like artificial intelligence, computer vision, pattern matching etc. Optical character recognition is one sub field that has wide range applications in converting handwritten text to digital format. The technique by which a computer system can recognize characters and other symbols written by hand in natural handwriting is called handwriting recognition system. The problem of handwriting recognition can be classified into off-line and on-line recognition. In offline recognition, only the image of the handwriting is available, while in the on-line case temporal information such as pen tip coordinates, as a function of time, is also available. Offline handwritten character recognition has another problem in knowing how the letter has been written by the writer. The complexity of the recognition is usually associated with the size of the language being considered. If the language contains more number of characters, the identification would be much more difficult than the case when the language contains lesser number of characters. Similarly we need to consider how the various characters are written and the differences between the various characters. They always have an effect on the performance of the handwriting recognition system.

Optical Character Recognition is based on mechanism consisting of a machine to recognize the scanned and digitized character images. Character recognition (CR) is one of the oldest applications of pattern recognition. The first efforts date from the early 1970s.Still, it remains an exceptionally difficult task to implement a character and numeral recognition framework that works under each possible condition and gives very precise outcomes.

In this paper a review of different methods adopted till to date in the field of OCR is presented.

II. REVIEW:

In early 1970s G. H. Granlund [1] has identified the requirements for pattern recognition as,

1. The pattern has to be treated by a feature extraction device which has to make measurements of parameters giving maximal information about the pattern and
2. This information has to be fed into a categorizer which makes a classification of the unknown samples.

Even though plenty of work has been done in understanding about classification of data, feature extraction is still a difficult task in developing a general approach. He used Fourier pre-processing techniques for classifying the characters by deriving information from ordinary Fourier coefficients. By identifying the relation between rotational symmetries of different angles and finding the relation between magnitude and these symmetries, he has been able to perform character recognition to a very good extent.

M Stoller [2] in 1971 invented Optical character recognition apparatus, which is program controlled image dissector tube that scans the printed information recorded on a storage medium to provide analog information signals. The analog signals are converted into digital data signals representative of the segmental brightness.

Hideo Ogawa and Keiji Taniguchi [3] in 1979 proposed techniques for calculating the stroke directions of thinned binary characters and for detecting the intersections and end points of strokes by means of pattern matching and weighting method as a pre-processing of handwritten Chinese character recognition. Global classification of handwritten Chinese characters by means of projection profiles of strokes also been proposed by them.

Fred W. M. Stentiford developed an automatic read and recognition system which reduces the manual involvement where feature selection may be chosen depending on the range of objective criteria and reduces training set error rate [4].L.A. Tamburino *et al.,* based on Fred WM Stentiford approach, developed automated feature detection using evolutionary learning processes [5]. This system contained a single feature detector and evolutionary learning system which evaluates this response and generates a modification to the detector. Based on the training set they identified that the software used is limited to few character types only and saw that it may serve as starting point for further investigation.

S. Impedovo *et. al.,* [6] in their survey on the methods of OCR from 1970s to 1999 highlighted various aspects and methodologies developed during that time.

The problems related to offline OCR identified by them are

1. Shape discrimination
2. Deformation of the image, which is caused by either noise like holes, disconnected line segments, isolated dots, translation and rotation.
3. Variations in size and pitch

They mentioned standards developed by CT Suen & S Mori and Canadian Standards Association, which are able to achieve less confusion and obtain higher recognition rates.

They highlighted that most of the OCR systems consists of parts as shown in figure 1.

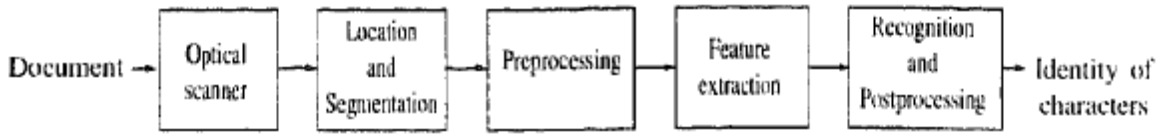

Figure 1. OCR system block diagram

Document is scanned by optical scanner and binary bitmapped image will be produced, which is mostly in black and white so as to reduce memory and this is known as thresholding. Software recognises the regions on this image and data will be written. It segments words of text into isolated characters. After segmentation, characters in the form of binary matrices are subjected to smoothing, elimination of noise, size normalization and other operations to facilitate the extraction of features in the subsequent stage. Identification of characters is achieved by comparing the extracted features with the statistics of features obtained from the set of samples used in the learning phase. Finally linguistic, contextual or statistical information can be used to resolve ambiguities of characters which have similar shapes or to correct words and phrases. He also compared different commercial CR systems available at that time for various application areas. By 1999 they observed that most of the systems available can recognise handwritten characters, however reading and recognizing cursive script is still a challenge.

Vedgupt Saraf and Rao [7] used genetic algorithm in character recognition of devnagari script. Through their work using genetic algorithm, they claim to have got an accuracy of around 97%-98%, although there are pairs that they found confusing. Pier Luca Lanzim *and* Politecnico di Milano [8], have used genetic algorithm for fast feature selection. The main advantage of their approach was that lesser processing time of CPU and the method was independent from a specific learning algorithm.

A neural network based classifier, called Multi-Layer perception (MLP), is used by S MShamim *et al.,* [9] to classify the handwritten digits. Multilayer perceptron consists of three different layers, input layer, hidden layer and output layer. Each of the layers can have certain number of nodes also called neurons and each node in a layer is connected to all other nodes to the next layer. They also used Support Vector Machine, Random Forest, Bayes Net, Naïve Bayes, J48 and Random Tree algorithms to test the accuracy in determining handwritten letters and word. To implement the above algorithms they use Waikato Environment for Knowledge Analysis(WEKA) which is a prominent suite of machine learning written in Java and developed at the University of Waikato. Finally it is found by them that the highest accuracy belongs to the Multilayer Perceptron classifier, followed by Support Vector Machine with a percentage of 87.97% and subsequently Random Forest Algorithm 85.75%, Bayes Net 84.35%,Naïve Bayes 81.85%, j48 79.51% and Random Tree75.06%. In this work they made an initial attempt that facilitates for recognition of handwritten numeral without using any standard classification techniques.

Vishwaraj *et al.,* [10]used Genetic Algorithm for Kannada Handwritten Character Recognition. They used kNN fitness function to determine the accuracy of recognising the characters. The whole system can be ratiocinated that the proposed recognised system has about 71.63% of recognised rate. Due to the application of Genetic algorithm, the feature selection had been done very easily and appropriately, and in total 490 images were used by them for training and for testing 490images (10images/character) were considered. Among 490 images, 351 images were recognised. Thus the recognised ratio is considered as 71.63%.

AyushPurohit and Shardul Singh Chauhan [11]did a survey on handwritten character recognition upto 2015. They identified the process of OCR as given in figure 2.

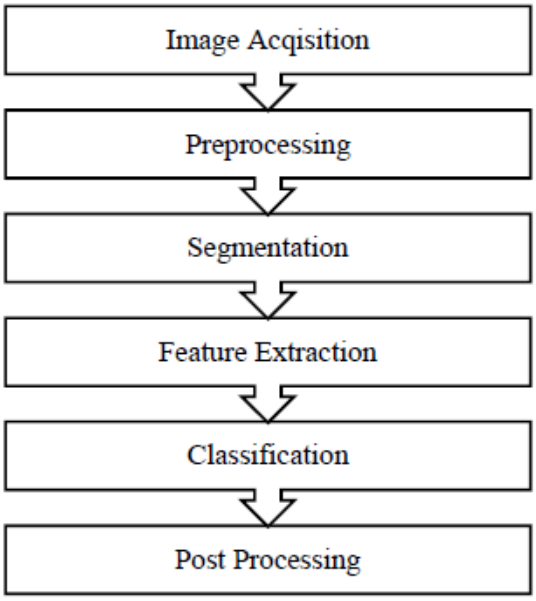

Figure 2 Block Diagram of Character Recognition

From the study done so far, it is analysed that the selection of the feature extraction techniques as well as the classification needs to be proper in order to attain good rate in recognizing the character. Studies in the paper reveals that there is still scope of enhancing the algorithms as well as enhancing the rate of recognition of characters.

V. S. Dhaka *et al.,* [12] made a survey on the methods of offline handwritten English recognition. They pointed out that Anshul Gupta *et al.,[*13]focussed on different methods such as Segmentation using a heuristic algorithm, Manual marking of segmentation points and Training of the Artificial Neural Network (ANN) for feature based classification techniques for offline handwritten character recognition. They also used Convolutional Neural Networks(CNNs) for offline handwritten English character recognition. The Support Vector Machine (SVM) used by them predicted characters at 98.8% out of 74,000 data set. They also showed how *Aiquan Yuan et al.,*[14]used modified *LeNet-5* CNN model with special settings of the number of neurons in each layer and the connecting way between some layers for an error-samples-based reinforcement learning strategy. This has the ability to reject recognition results. They use 61,000 dataset and found the accuracy to be 90%.From the survey made by him ,has been noted that the errors in recognizing handwritten English characters are mainly due to incorrect character segmentation of touching or broken characters. Because of upper and lower modifiers of English text, many portions of two consecutive lines may also overlap and proper segmentation of such overlapped portions are needed to get higher accuracy. Many authors suggest that the post processing of classifier outputs by integrating a dictionary with the OCR system can significantly reduce the misclassifications in printed as well as handwritten word recognitions.

Vincent Christlein *et al.,* [15] proposed the use of activation features from CNNs as local descriptors for writer identification. A global descriptor is then formed by means of Gaussian Mixture model (GMM) supervector encoding, which has been further improved by normalization with the KL-Kernel. This method has been evaluated on two publicly available datasets: the ICDAR 2013 benchmark database and the CVL dataset. This work is shown in figures 3 and 4.

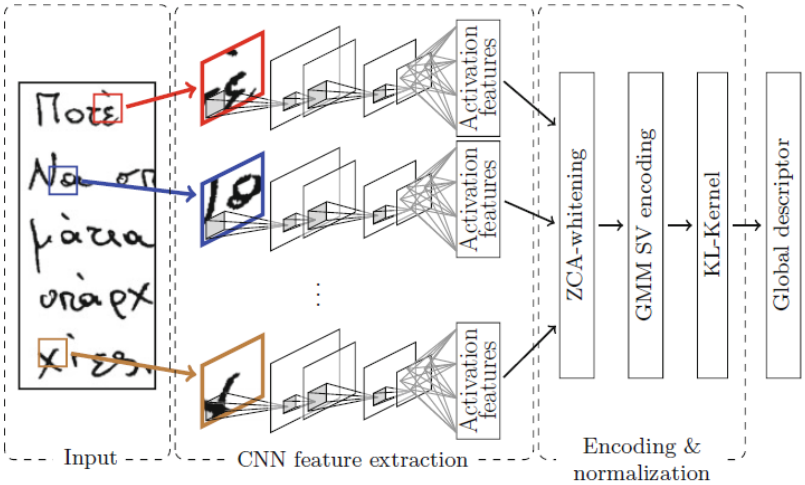

Figure 3 Overview of the encoding process.

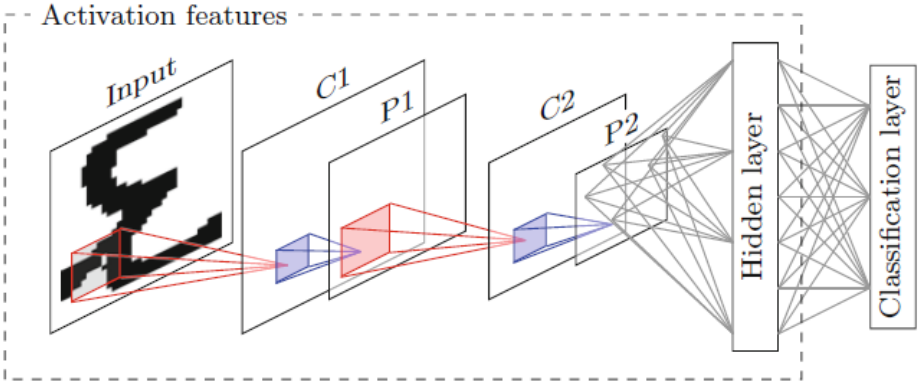

Figure 4 Schematic representation of the used CNN.

Jorge Sueiras *et al.,* [16]proposed a new deep neural architecture to offline handwriting recognition which combines a deep convolutional neural network with an encoder-decoder, called sequence to sequence, to solve the problem of recognizing isolated handwritten words. They tested the proposed model on two handwritten databases (IAM and RIMES) under several experiments to determine the optimal parametrization of the model. Without using any language model and with closed dictionary, they obtained a word error rate 12.7% in the test set in IAM and 6.6% in RIMES.

Bai et al. has shown that fully convolutional methods can outperform recurrent networks on sequence modeling problems [17] . By using an initial CNN to calculate the number of symbols in a word block, word blocks can be resized to a canonical representation tuned to a Fully Convolutional neural Network (FCN) architecture. Knowing the average symbol width, this FCN can then perform accurate symbol prediction without post processing. And this has been used by *R. Ptucha et al.,*[18] to obtain character based classification without relying on predefined dictionaries or contextual information.

Hung Tuan Nguyen *et al.,* [19]proposed an end-to-end deep-learning method for text-independent writer identification that does not require prior identification of features. A Convolutional Neural Network (CNN) has been trained initially to extract local features, which represent characteristics of individual handwriting in the whole character images and their sub-regions is given in figure 5.

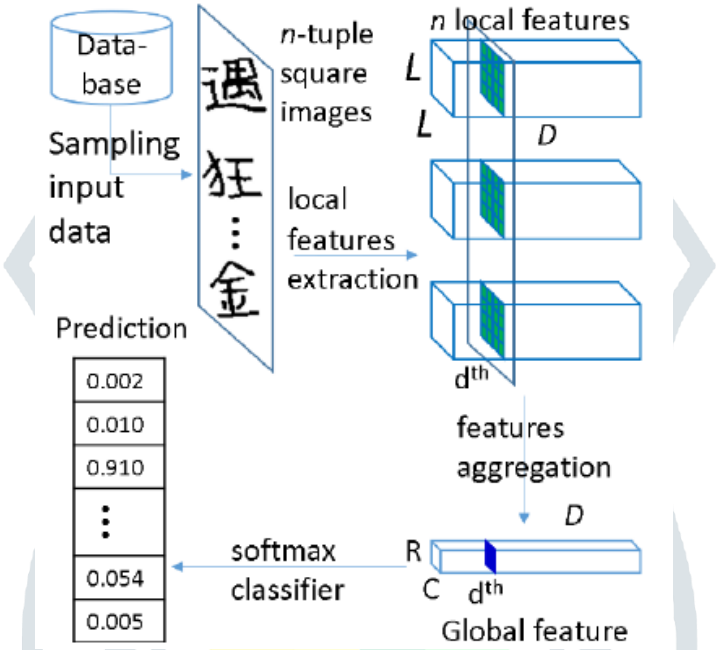

Figure 5 Overview of the method developed.

They conducted experiments on the JEITA-HP database of offline handwritten Japanese character patterns. With 200 characters, their method achieved an accuracy of 99.97% to classify 100 writers. Even when using 50 characters for 100 writers or 100 characters for 400writers, this method achieved accuracy levels of 92.80% or 93.82%, respectively. This approach overcomes the difficulties of gathering handwritten character patterns in the same category for writer identification.

DhanyaSudarsan and Dr.Shelbi Joseph [20] used contour detection for segmentation and CNN for classification of Malayalam handwritten text from Malayalam manuscripts like palm leafs, official documents in Government offices etc. Euler number based feature extraction and fuzzy membership function is based on a Master-Pattern and a pattern score for recognition of handwritten characters. The system was tested and proved to obtain an accuracy of 96.46%.

Lyzandra D'souza and Maruska Mascarenhas [21] have developed Handwritten Mathematical Expression recognising method using CNN as classifier. In preprocessing stage Noise Reduction, Binarization and Thinning have been done. The segmentation of the Handwritten Mathematical Expression (HME) into individual math symbols is done using Vertical Projection Profile Cutting (VPPC), horizontal projection called Horizontal Projection Profile Cutting (HPPC) and Connected component (CC) is applied on HME binarized image. SpNet-CNN with 83 different classes has been used by them as a classifier. It comprises of the following layers input, convolutional, max pooling, fully connected and softmax. This system will work best for isolated symbols. But for merged or connected or joined symbols will give better recognition results when they are segmented properly.

Konkimalla Chandra Prakash*et al.*,[22] proposed a 2-CNN architecture for OCR of Telugu script. The pipeline followed by them is, skew correction– word segmentation – character segmentation – recognition. They found that a single CNN would be futile because of the huge number of classes arising from various permutations of the main character, vattu and gunintham. Therefore, they have used two CNN architectures for classifying the character. The first CNN is used for identifying the main character and the second CNN for identifying the vattuand/or gunintam present along with the main character. This work has spanned the entire Telugu language while creating the dataset, without leaving any further possibility of increase in data. The segmentation algorithm can be improved so that every character is segmented together with its vattu and gunintham. Network accuracy can be further improved to make the classifier better. To facilitate usability, they have developed an Android app that deploys the proposed OCR solution.

Harathi Surya Patchipala [23]developed a method for handwritten telugu character recognition using CNN. The classifier used to train the data is Deep Convolutional Neural Network and the input to the convolutional neural network is a 4D Tensor. The 80 x 80 image is reformatted into a 4D tensor with shape (samples, channels, rows, columns) which is then passed through a stack of different kinds of layers as shown in figure 6.

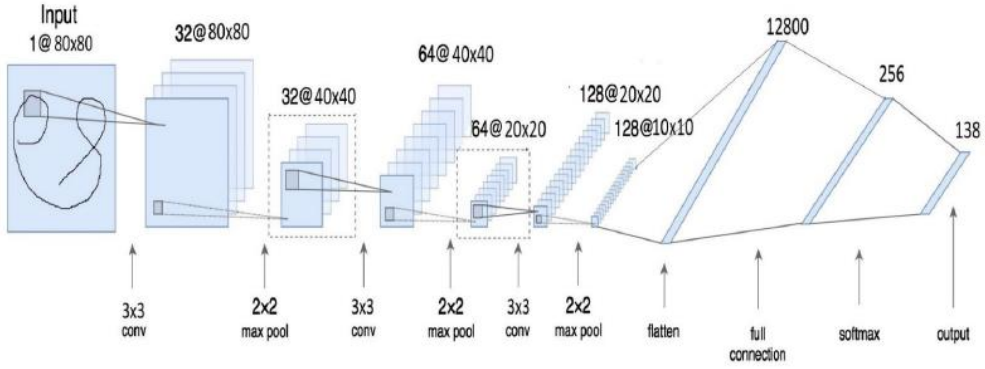

Figure 6. CNN Architecture for Handwritten Telugu Character Recognition

This algorithm used only single characters of Telugu language. But there are many characters with '*vattu*' and '*gunintham*'. As there are very less contributions on Telugu language characters, they couldn't find the dataset consisting of all these characters. But the characters used in this project are like vowels and consonants in English.

A 5-layer CNN based handwritten character recognition of multiple scripts like English, Devanagari, Bangla, Telugu and Oriya has been developed by Durjoy Sen Maitra *et al.,* [24] and they were able to achieve very high accuracy in determining the characters as well as numbers as given in table 1.

Table-1: Accuracy figures (%)

Bangla Basic Characters	Bangla Numeral	Devanagari Numeral	Oriya Numeral	Telugu Numeral	English Numeral
95.6	98.375	98.54	97.2	96.5	99.10

III. CONCLUSIONS

This paper discusses in detail all advances in area of handwritten Character recognition(CR). The most accurate solution provided in this area directly or indirectly depends upon the quality as well as the nature of the material to be read. Various techniques have been described in this paper for character recognition in handwritten recognition system. From the study done so far, it is analysed that the selection of feature extraction as well as the classification techniques needs to be proper in order to attain good rate in recognizing the character. There is still scope for enhancement of algorithms in recognition of characters.

REFERENCES

[1] G. H. Granlund, "Fourier Preprocessing for Hand Print Character Recognition," in IEEE Transactions on Computers, vol. 21, no. , pp. 195-201, 1972. doi:10.1109/TC.1972.5008926

[2]United States Patent 1 91 Stoller 1 1 Mar. 27, 1973 [54] optical character recognition Reversal Dissector," Vol. 12, No. 9, Feb. 1970. P

[3]Hideo Ogawa, Keiji Taniguchi, "Preprocessing for Chinese character recognition and global classification of handwritten Chinese characters" Pattern Recognition,11(1979) Pages 1-7

[4]Fred W. M. Stentiford "Automatic Feature Design for Optical Character Recognition Using an Evolutionary Search Procedure" IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: PAMI-7 , Issue: 3 , May 1985) 349 - 355

[5]L. A. Tamburino, M. M. Rizki and W. Van Valkenburgh, "Automated feature detection using evolutionary learning processes," Proceedings of the IEEE National Aerospace and Electronics Conference 1989, pp. 1080-1087 vol.3.

[6]S. Impedovo , L. Ottaviano and S. Occhinegro "Optical Character Recognition — A Survey" International Journal of Pattern Recognition and Artificial Intelligence 05 (1991)1-24

[7]VedguptSaraf, D.S. and Rao, 2013. "Devnagari script character recognition using genetic algorithm for better efficiency", IJSCE, ISSN: 2231-2307, Volume-2, Issue-4, April 2013

[8]Pier Luca Lanzi and Politecnico di Milano, 1997. "Fast feature selection with genetic algorithm: A filterapproach", IEEE, 0-7803-3949-5/97, 1997.

[9]S M Shamim, Mohammad BadrulAlam Miah, AngonaSarker, Masud Rana & Abdullah Al Jobair , "Handwritten Digit Recognition using Machine Learning Algorithms" Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence Volume 18 Issue 1 Version 1.0 Year 2018, 8pages

[10]Vishwaraj, Karthik S P, Sreedharamurthy S K, " Kannada Handwritten Character Recognition Using Genetic Algorithm " International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 02, Issue 08, 2015, 369-374

[11]AyushPurohit and Shardul Singh Chauhan" A Literature Survey on Handwritten Character Recognition " International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, 1-5

[12]V. S. Dhaka et al., " Offline Handwritten English Script Recognition: A Survey" International Journal of Advanced Networking and Applications (IJANA), Special Conference Issue: National Conference on Cloud Computing & Big Data (2014)114-124

[13]Anshul Gupta, Manisha Srivastava and ChitralekhaMahanta"Offline Handwritten Character Recognition Using Neural Network" 2011 International Conference on Computer Applications and Industrial Electronics (ICCAIE 2011).

[14]Aiquan Yuan, Gang Bai, Lijing Jiao, YajieLiu "Offline Handwritten English CharacterRecognition based on Convolutional NeuralNetwork" 10th IAPR International Workshop onDocument Analysis Systems 2012.

[15]Christlein V., Bernecker D., Maier A., Angelopoulou E. (2015) Offline Writer Identification Using Convolutional Neural Network Activation Features. In: Gall J., Gehler P., Leibe B. (eds) Pattern Recognition. DAGM 2015. Lecture Notes in Computer Science, vol 9358. Springer, Cham

[16]Velez, Oßine Continuous Handwriting Recognition Using Sequence to Sequence Neural Networks, Neurocomputing (2018), doi: 10.1016/j.neucom.2018.02.008

[17]S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv: 1803.01271 (2018).

[18]Raymond Ptucha et al. "Intelligent character recognition using fully convolutional neural networks" Pattern Recognition, 88 (2019) 604-613

[19]Hung Tuan Nguyen ,Cuong Tuan Nguyen , Takeya Ino , BipinIndurkhya , Masaki Nakagawa , Text-Independent Writer IdentiÞcation using Convolutional Neural Network, Pattern Recognition Letters (2018), doi: 10.1016/j.patrec.2018.07.022

[20]D. Sudarsan and S. Joseph, "A Novel Approach for Handwriting Recognition in Malayalam Manuscripts using Contour Detection and Convolutional Neural Nets," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1818-1824. doi: 10.1109/ICACCI.2018.8554592

[21]L. D'souza and M. Mascarenhas, "Offline Handwritten Mathematical Expression Recognition using Convolutional Neural Network," 2018 International Conference on Information , Communication, Engineering and Technology (ICICET), Pune, 2018, pp.1-3. doi: 10.1109/ICICET.2018.8533789

[22]K. Chandra Prakash, Y. M. Srikar, G. Trishal, S. Mandai and S. S. Channappayya, "Optical Character Recognition (OCR) for Telugu: Database, Algorithm and Application," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 3963-3967. doi: 10.1109/ICIP.2018.8451438

[23]https://github.com/Harathi123/Telugu-Character-Recognition-using-CNN/blob/master/project_report.pdf

[24]D. S.Maitra, U. Bhattacharya and S. K. Parui, "CNN based common approach to handwritten character recognition of multiple scripts," 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, 2015, pp. 1021-1025.doi: 10.1109/ICDAR.2015.7333916