Turkish Journal of Computer and Mathematics Education Vol.12 No.6 (2021), 5320-5326

Research Article

CNN based Digital alphanumeric archaeolinguistics apprehension for ancient script detection

J.P.Premi [a] ,R.Madhumitha[b] , N.R.Raajan[c]

[a,b] A student of Communication Systems, School of EEE, SASTRA Deemed to be University, Thanjavur.
[c] Senior Associate Professor, School of EEE, SASTRA Deemed to be University, Thanjavur

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: OCR is a most prominently used system in computer vision space. In the era of computer vision, the recognitiontechnologies are indeed evolved but there are still some difficulties for computers when reading handwritten text which can be resolved only after the introduction of machine learning. Recognition and verification of handwritten information is still a challenging problem in machine learning. Optical Character Recognition is a process of recognizing text or information present inside the images and converting it into a digital formatted text. Text recognition has immense applications in the academics, research, commercial and industrial fields. This paper is about an Optical Character Recognition for text recognition from the images which could be in any of the forms of handwritten text files as well as from the ancient manuscript (Language-English). This paper presents a novel machine learning approach to recognize the characters using CNN and the accuracy is found to be 73% approximately within a fraction of second. Later the recognized images are converted into the text file and then get translated into the preferable languages.

Keywords: EMNIST datasets, Ancient manuscript, Neural network, Optical Character Recognition

1. Introduction

OCR stands for Optical Character Recognition and it refers to the technologies that help us to gather information from our environment whether it is in typed, digitally printed or in the form of handwritten and transform it into the digital text format. Once this was done then the digitally formatted text is easily search-able and editable into various documents like PDF files, Word files, plain text files, database and presentation files. Optical Character Recognition was originally developed to help visually impaired people and today it's usage was spread in various academic and industrial aspects. It was popularly used to gather data from ID's and passports, utility meters, promotional codes, serial number of any kind, automatic cartography, signature verification and identification and in automatic number-plate readers. Optical Character Recognition is employed when the information should be readable by both humans and to machines. As compared with other techniques of automatic detection and data capturing like Radio Frequency Identification, Bio-metrics, Magnetic stripes, voice recognition and smart cards, Optical Character Recognition is unique because it only takes fraction of second to get the required data, low system complexity and it eliminates the human error. OCR deals with the recognition of optically processed characters. OCR can be performed in two ways i.e., off-line and on-line. Off-line recognition was done when the writing or printing was completed whereas in on-line recognition the computer recognizes the characters when they are drawn. The paper deals with off-line handwritten text recognition as it is comparatively difficult, as each individual has a unique handwritten styles which is in different font, frequent change in style and sizes of the characters, overlapping between the letters and even due to incorrect meaning of the text, which brings the recognition system at risk. Character detection, recognition and feature extraction from the ancient manuscript was highly difficult task for the researchers due to various problems like quality degradation due to aging and climatic conditions, background images. So there is a need to preserve the valuable literatures in order to acquire knowledge for future generations. In this motive, the paper presents a machine recognition system that recognizes the characters from the text files and from the ancient manuscript images available in a palm leaf literature format. One of the major reason for poor character recognition was due to poor character segmentation. The proposed approach has better segmentation process by adaptive threshold value, noise level reduction, compute space vectors, binarization, bounding box analysis and perform pixel optimization. The proposed approach uses a Convolutional Neural Network trained with an E-MNIST dataset and character level recognition was done by Selection Auto Encoder Decoder (SAED) technique.

2. Significance of The Study

Before computers existed all the information was stored in written form, this is very inefficient form of storage as the paper information cannot be stored for very long time and can get lost or be destroyed. On the

contrary, information on computer is stored safely for long time and multiple copies of same information can be made easily. Thus after inventing the computers lot of money was wasted in manual labor for converting this

paper information into digital information. Instead machine learning can be used to identify and convert this paper information into digital information without human intervention or manual labor. This project is just an introduction to this approach.

3. Review of Related Studies

Enormous research work was done in this area and some of them are mentioned here.

Chirag Patel proposed a case study for the OCR by Open Source OCR Tool [1], which gives an introductionto open source OCR tool Tesseract and provides a comparative experimental analysis2 of this OCR tool and Transym tool by taking car license as an input and various parameters are analysed. **Isaac Wu and Hsiao-ChenChang** presented a scheme that specifically used for signboard recognition by text recognition [2], this uses SIFTfeature matching method with traditional OCR to increase the success rate of recognition. **B.Gatos andD.Karras** proposed a rigorous non-linear method to solve large-scale OCR problems[3], by using typeset Greekcharacters and analysed the better performance as compared with other neural networks like Artificial Neural Networks. This non-linear method estimates the class discrimination ability of continuous valued features. It found to have high recognition performance in some real time applications. **Sang Sung Park, Won Gyo Jung** they constructed an OCR system using Artificial Neural Network and Back Propagation (BP) algorithm [4] to store the extracted characters automatically as a Database and can be easily used whenever necessary. The Method proposed in [5] was done in two stages, in which first phase neural network is used for recognition of isolated characters and in second phase for recognition of word and characters of variable length employing RNN and CNN integrated approach.

The remaining sections of the paper is structured as follows: Section II gives objectives of the study; Section

III proceed with the concepts of materials and the algorithm employed; Section IV shows the experimental results and detailed analysis; the final section provides conclusion.

4. Objectives of The Study

- To provide a higher accuracy model and faster computational method for recognition of the handwritten characters
- to evaluate the network performances like accuracy, threshold value
- to implement this model in real time applications like ancient script recognition (here in English language) from palm leaf image
- to expand this model for optical character recognition

5. Materials And Methodology

Neural computing is a new method in computer vision where the design of components is less well specified than other architecture. In this machine learning project, the convolutional neural network will recognize the characters, here the English alphabets from A-Z including both uppercase and lowercase and special characters like spaces even when found between the words of handwritten text and the character level segmentation and recognition was done for the ancient English literature manuscript written on palm leaf. This proposed system it will effectively recognize the characters using Convolutional Neural Network in Matlab software.

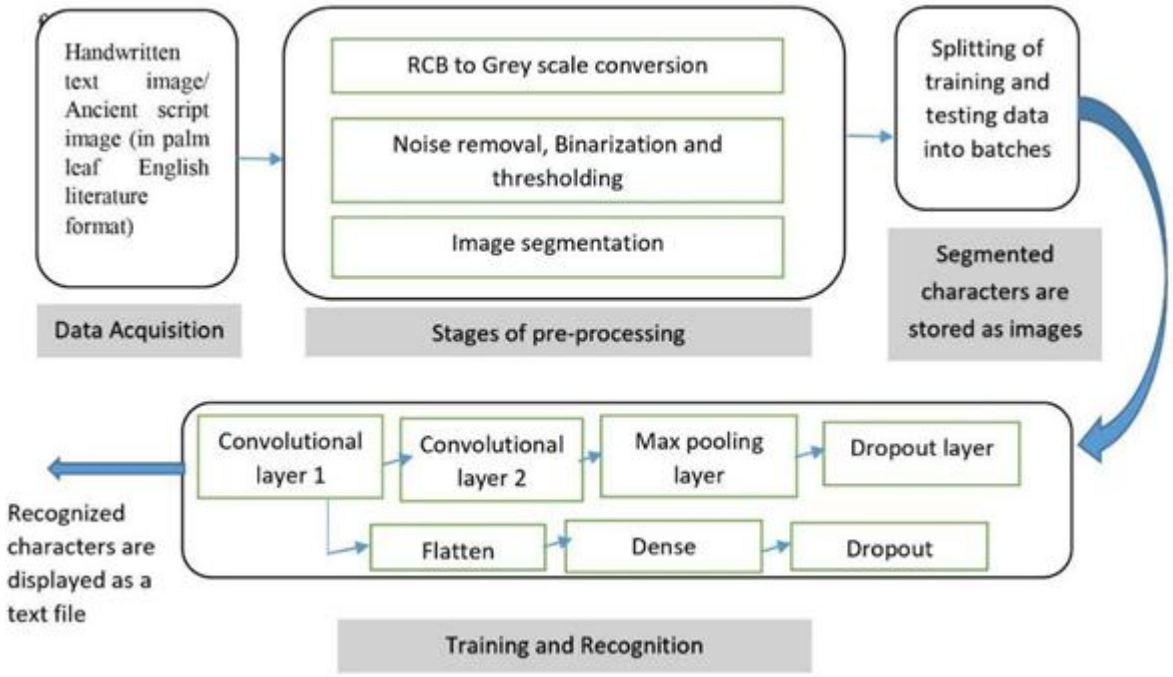

Fig no:1 Block diagram for Character recognition system

A. Data acquisition:

EMNIST stands for Extended Modified National Institute of Standards and Technology. The NIST Database_19 consists of handwritten digits and characters collected from around 500 writing sources. It was organized into six classes like Balanced Dataset, By-Class, By-Merge, Letter dataset, Digits Dataset and EMNIST MNIST Datasets. Here, only By-Class and By-Merge datasets are used. In the original NIST Special Database-19 consisting of a 731,668 training data and 82,587 testing data. The By-Class dataset holds 62 classes comprises of 10 digit classes, 26 lowercase alphabets classes, and 26 uppercase alphabets classes.

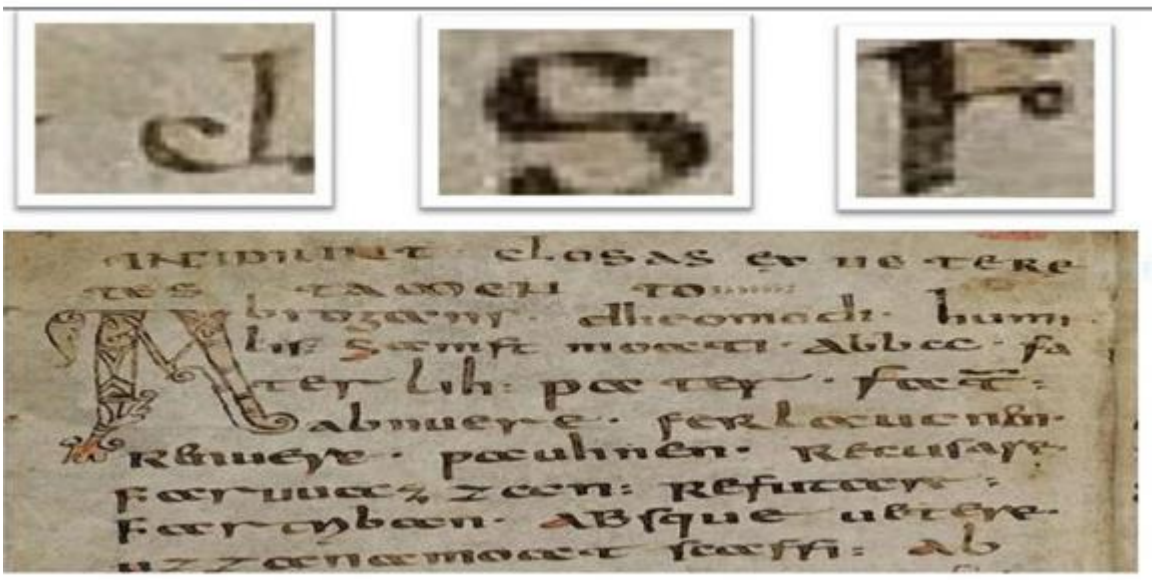

Fig no:2 Character level and line level text from an Ancient English language manuscript

Methodology:

Image Pre-processing:

Pre-processing was done to make the OCR system easily analyse and process the data. Here the input image which is RGB is converted into the binary scale (pixel value of 0 and 1) image followed by the noise removal. For the recognition of characters, the input image of size 28*28 is pre-processed using Convolutional Auto Encoder (CAE) to do binarization that encloses of an activation functions. In this

binary images, the characters are represented with the binary value 1 and the background has the binary value 0. Pixel normalization3 was done and the binary image get transformed into another binary image and store the detected word in a matrix form. Adaptive threshold was employed to map label to neighbouring pixel. A threshold can be calculated an algorithm as shown below.

a) Local extrema s(t) was calculated

b) Local maxima lmax(t) and Local minima lmin(t) was generated

c) Find the mean value (M) of lmax(t) and lmin(t)

d) Threshold function is given as:

T=M+K.sd

K is the tuning parameter and sd is a value standard deviation. Here the threshold value was found to be 0.44 in approx. Selection Auto Encoder Decoder(SAED) is used for binarization process. The experiment was done with 73% accuracy with a binarization time of 5 seconds.

Then the segmentation was done so that image becomes more meaningful and can be easily understood. In segmentation properties of image like position, edge colour, line width was detected. Then the space vectors are initialized to compute the total space between the adjacent characters. This method carves a path and the path with minimal energy is chosen.

Training and recognition:

After segmentation, the recognition was done by the three layers of CNN. This proposed approach employs convolutional layer, max pooling layer and dropout layer, flatten layer and dense layer. The image is firstly passed into a convolutional layer for two times which apply certain extracted features on the filter. The max pooling layer was used to reduce the spatial size of given image which further transferred into a dropout layer which prevents overfitting.

Table 1: summary of processing layers used:

Layer	Output shape	Parameters
Conv2d_1	(None.32,32,32)	895
Conv2d_2	(None.30,30,32)	9248
max_pooling2d	(None.15,15,32)	0
dropout_1	(None.15,15,32)	0
flatten_1(Flatten)	(None. 256)	0
dense_1(Dense)	(None. 512)	131584
dropout_4(Dropout)	(None. 512)	0

6. Result and conclusion

The CNN was used for the recognition of the characters and obtained the accuracy of 73% (approx.)

=(/)∗100

Where, Ar -recognition accuracy

Nr - recognized images

Nt - total number of images used. After the completion of recognition process

Fig no:3 Input image of handwritten text

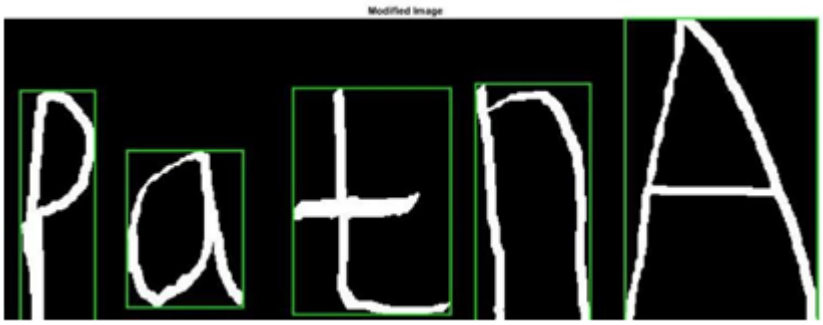

Fig no:4 Binarized image

```
Detected Text: PatnA
>>
```

Fig no:6 Detected handwritten text

Fig no:5 Segmented characters are stored as an separate image

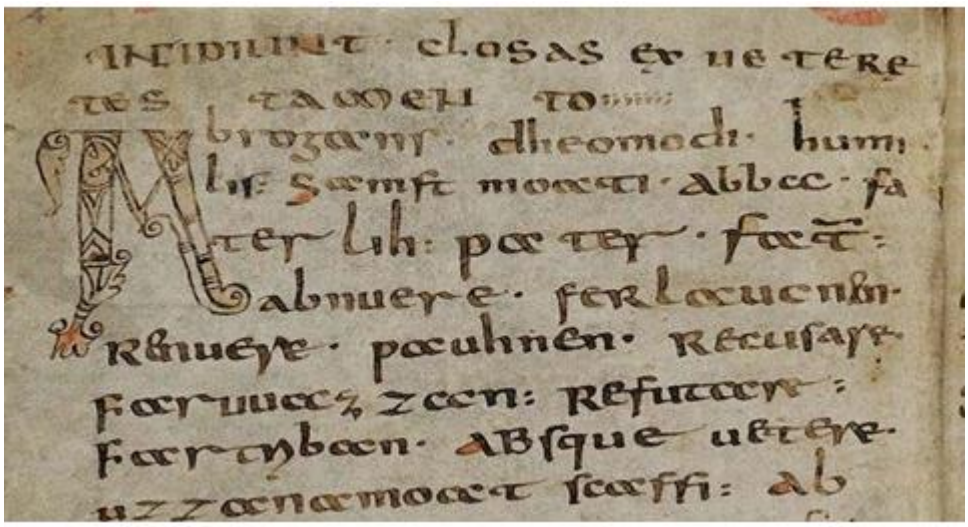

Fig no:7 Text from an Ancient English literature manuscript present in a palm leaf in degraded format

Fig no:8 Binarization done by Using Selection Encoder Decoder Technique and the Segmentation Using Seam Carbel Approach

Fig.8 Recognized characters are converted into text file

7. Conclusion

There are many languages found across the world each of them has different handwriting styles which can be recognized by this OCR system using CNN with higher recognition rate. Also here, character level recognition system is proposed for English text written on palm leaf. The ancient script shown here was in degraded format, which should be preserved and deciphered into text format for future generation. This proposed approach has very low wrong recognition accuracy but has high speed of performance. Performing with proper pre-processing tasks and valid database will leads to the successful detection of English alphabets and numerals from the images with better efficiency. This work can be further extended to multilingual ancient scripts recognition system and also with various form of documented data as front end. It can be used in Passport checking, Postal address verification, academic and industrial applications.

References

1. Chirag Patel, Atul Patel, Dharmendra Patel, "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study," International Journal of Computer Applications (0975 – 8887), Volume 55– No.10, October 2012.
2. Isaac Wu and Hsiao-Chen Chang, Signboard Optical Character Recognition.
3. B.Gatos, D.Karras and S.Perantonis, "Optical Character Recognition Using Novel Feature Extraction & Neural Network Classification Tech□niques," in IEEE, 1994.
4. Sang Sung Park, Won Gyo Jung, Young Geun Shin, Dong-Sik Jang, "Optical Character Recognition System Using BP Algorithm," IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.12, December 2008.
5. H. Meng and D. Morariu, "Khmer character recognition using artificial neural network," 2014 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2014, 2014, doi:10.1109/APSIPA.2014.7041824.
6. K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 1724–1734, 2014, doi: 10.3115/v1/d14-1179.

7. Batuhan Balci, Dan Saadati, Dan Shiferaw, "Handwritten Text Recognition using Deep Learning,", unpublished. Yi-Chao Wua, Fei Yina, Cheng-Lin Liu, "Improving handwritten Chinese text recognition using neural network language models and convolutional neural network shape models,",in Elsevier, 29 December 2016.
8. Raghuraj Singh, C.S.Yadav, Prabhat Verma, Vibhash Yadav, 'Optical Character Recognition Using Novel Feature Extraction & Neural Network Classification Techniques,", International Journal of Computer Science & CommunicationVol. 1, No. 1, January-June 2010, pp. 91-95.
9. F. Hussain and J. Je ong, Efficient deep neural network for digital image compression employing rectified linear neurons, Journal of Sensors, 2016, pp. 1-7.
10. IJ. Tsang, IR. Tsang and DV Dyck 1998. Handwritten character recognition based on moment features derived from image partition. In Int. Conf. image processing, vol. 2, 939– 942.
11. M. Shi, Y. Fujisawa, T. Wakabayashi and F. Kimura 2002. Handwritten numeral recognition using gradient and curvature of gray scale image. Pattern Recognition, vol. 35, no. 10, 2051–2059.