ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

A Survey on Historic Documents Preservation Using Multilingual Characters Recognition and Joint Words Segmentation in Devanagari

Sagar Thakare[1], Vina Lomte[2]

ME Student, Department of Computer Engineering, RMSSSOE, Pune, Maharashtra, India[1]

Assistant Professor and Head, Department of Computer Engineering, RMSSSOE, Pune, Maharashtra, India[2]

ABSTRACT: Handwritten character recognition of Devanagari script is an area of research in the field of pattern recognition. It is very important to preserve the historic documents. Historic documents can contain various characters, numbers and joint characters written in Devanagari language. The challenges in handwritten characters recognition is that, every person has his own style of writing the word, to recognize such variety of characters the optical character recognition system must be robust and Features extraction and classification process must be extensive. In this work, we propose a technique to recognize handwritten Devanagari characters using deep convolutional neural networks (DCNN) which are one of the recent techniques adopted from the deep learning community. We experimented the HPL Isolated Handwritten Devnagari Character Dataset and Devanagari Character Dataset provided by Kaggle. A layer-wise technique of DCNN has been employed that helped to achieve the highest recognition accuracy and also get a faster convergence rate.

KEYWORDS: Deep Convolution Neural Networks; Joint word Segmentation; Multilingual character recognition; Devanagari;

I. Introduction

Marathi is the Official language of Maharashtra state in India (written in Devanagari script) is most well-liked language after English and Hindi. Marathi handwritten character recognition has got lot of application in various areas like postal office, Bank sorting cheque electronically. The manually written Marathi characters acknowledgment by PC is a troublesome errand as the PC can without much of a stretch perceive contrast with composed characters, which can be. English Optical Character Recognition (OCR) has been widely studied in the last five decades and progressed to a level, sufficient to produce technology applications. But same is not the case for Indian languages which are difficult in terms of structure and computation. Digital document processing is acquisition recognition for application to office and library automation, bank, publishing houses communication technology, postal services and many other areas. With non-increasing requirement for office automation, it is necessary to provide practical and effective solutions. Marathi character recognition is becoming more and more important in the modern world. It helps human ease their jobs and solve more complex problems over the few past years, the numbers of companies involved in research on handwritten recognition are increasing continually. Being Devanagari is the prime language of India, spoken by more than 600 million people, should be given special attention so that document retrieval and analysis of rich ancient and modern Indian literature can be effectively done.

II. Related work

In [1] authors used Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to improve the performance of recognition system. In proposed system, first raw features are extracted using three different feature extraction methods: i) chain coding, ii) edge detection using gradient features and iii) direction feature techniques. In

DOI: 10.15680/IJIRCCE.2018.0606059

[2]the author gave given a detailed analysis of the process ofCNN algorithm both the forward process and back propagation.Then they applied the particular convolutional neural network to implement the typical face recognition problem by java. Then, a parallel strategy was proposed. In [3] the author has taken a different route and combine the representational power of large, multilayer neural networks together with recent developments in unsupervised feature learning, which allows them to use a common framework to train highly-accurate text detector and character recognizer modules. In [4] Author have proposed a technique to recognize handwritten Devanagari characters using deep convolutional neural networks (DCNN). In paper [5],the authors have presented a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. These method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Paper [6] is an attempt is made to recognize handwritten characters for English alphabets without feature extraction using multilayer Feed Forward neural network. Each character data set contains 26 alphabets. Fifty different character data sets are used for training the neural network. The trained network is used for classification and recognition. In the proposed system, each character is resized into 30x20 pixels, which is directly subjected to training. That is, each resized character has 600 pixels and these pixels are taken as features for training the neural network. In [7] the author has proposed a Convolutional Neural Network (CNN) based Optical Character Recognition system (OCR). In [8] proposed model, the handwriting images was first subjected to binarization process, the followed by the pixel matrix downsampling first using the column approach (C-DS), then combine raw and column approach (RC-DS). The compressed pixel (downsampled pixel matrix) then acted as an input vector for Artificial Neural Network (ANN).

III. Existing SYSTEM

In existing system, the authors have proposed a simple yet effective sequence-to-sequence neuralmodel for the joint task, based on a well-defined transition system,by using long short term memory (LSTM) neural networkstructures. They have conduct experiments on five different datasets.The results demonstrate that the proposed model is highly competitive. By using well-trained character-level embedding,the proposed neural joint model is able to obtain the best reportedperformances in the literature.

The proposed system model consists of 5 mains steps as i) The Transition System, ii) Seq2Seq Modeling, iii) Encoder, iv) Decoder, v)Training.

i) The Transition System:-

A transition system has two key components: (1)transition states and (2) a set of transition actions. A transitionstate defines representation of a partial result, while transitionactions are used to control how a transition state advance byone step. Initially, it is starting with an empty starting state, and then thestate advances gradually by a sequence of transition actions,until it reaches an end state representing a full result.

ii) Seq2Seq Modeling:-

A Seq2Seq model consists of two parts: (1) an encoderthat represents source input sequences, and (2) a decoder thatincrementally predicts the next coming symbol. The overallneural network structure of the joint model is depicted inbelow figure figure-1,inthe proposed Seq2Seq model, the encoder is usedto represent input Chinese character sequences, as shown bythe bottom region of the figure, and the decoder is used topredict the transition action sequences.

iii) Encoder:-

In this step the author has used a bi-directional LSTM to encode input Chinesecharacter sequences, following the majority Seq2Seq models. Give a sequence of Chinese charactersc1c2 _ _ _ cn, the bi-directional LSTM is built as follows. First,derive two sequences of input features x. i.e. forward directional and backward directional, then apply left-to-right LSTM on forward directional inputs and right-to-left LSTM on backward directional inputs.

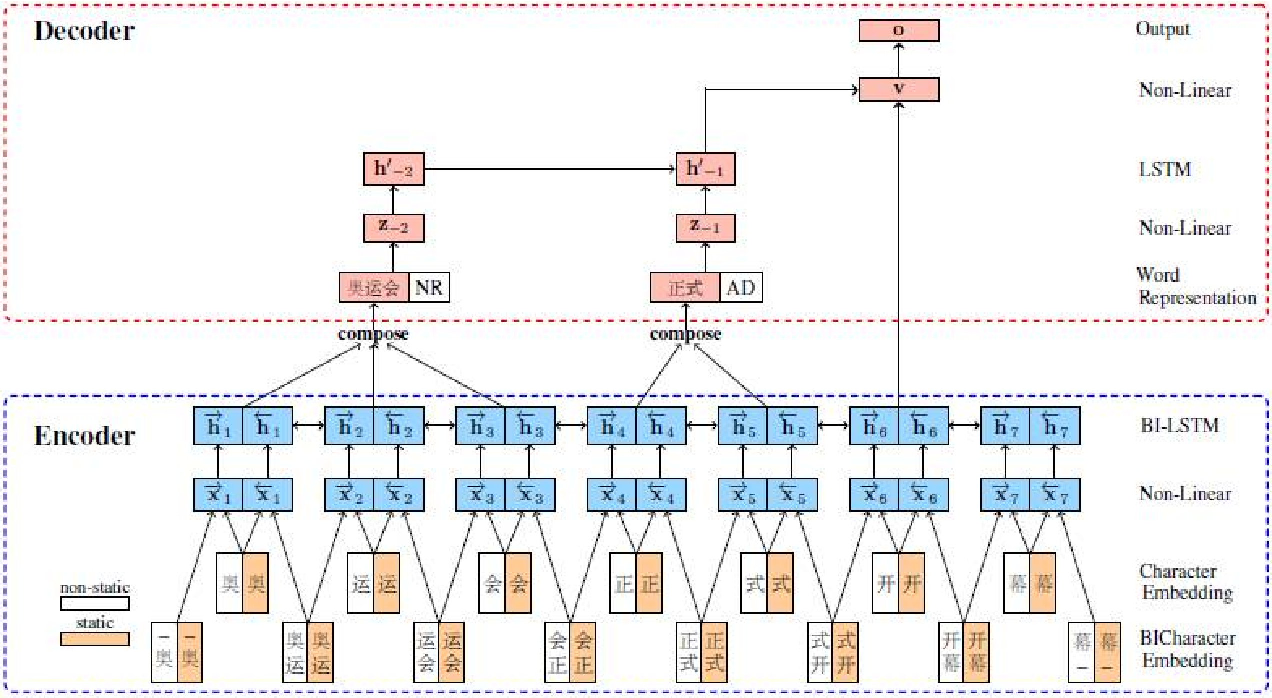

Figure-1: An example to illustrate the proposed model

iv) Decoder:-

The decoder aims to find a next-step action conditionedon historical actions. The overall decoding framework of our Seq2Seq model is shown by the upper part of Figure 1. On theone hand, we extract a source of features from encoder outputsfor prediction, and on the other hand, we build a left-to-rightLSTM over the generated output words incrementally, which isexploited as another source of features for action classification.
Decoder LSTM. For the left-to-right decoder LSTM, assumingthe word sequence is W-m…. W-2, W-1. First they have represented the discrete sequential words into dense hidden vectorsZ-m…Z-2, Z-1, and then they have calculate the LSTM outputs h-m…..h-2, h-1 incrementally.

v) Training:-

Proposed model by using a simple cross entropyloss, on the onehand, the training is highly efficient, and on the other hand,according to our preliminary experiments, the authors have found that thecross-entropy objective has been already to give strong performances.

IV. Conclusion

An extensive survey has been carried out of the existing systems and we are come to this conclusion that, Large amount of work has been carried out in the areas of character recognition in various languages such as Japanese, Arabic, Chinese etc., but this has been observed that very minimal work has been carried out in the characters recognition in Devanagari language. Devanagari is primary language used for communication in Maharashtra state of India.
Over 600 million people speaks Marathi (which uses Devanagari script) in their day to day communication and Marathi ranks 19th in the list of most spoken languages in the world. Various documents are already written in Devanagari language and currently are being written also. Hence there is a need to store these historic and ancient documents digitally, so that which can be used in future for reference and existing of the language can be preserved digitally.

Based on this statistic, we come to conclusion that there is a need to develop a system which will recognize the characters written in Devanagari language. Like, i) recognize the numbers and letters written in Devanagari language. ii) Recognize the joint characters. I.e. the joint character means the character formed by joining two characters, one of which is a half and another is the full character. There is a really challenge in this, to read the character based on its various properties and train the system accordingly, iii) Recognize the characters written in different font sizes, iv) Recognise the character if it is written in slant pattern, because everyone has their own style of writing. v) Recognise the characters if they are written in multiple languages together for example, in the combination of English and Marathi

Please note: The paper which we have referred here is just for the reference purpose and also it is written for Chinese language, we have done the survey on this paper and have conclude our observation in this section.

REFERENCES

1. Sneha Shitole, Savitri Jadhav, ‘Recognition of Handwritten Devangari Characters using Linear Discrimination Analysis’, Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018) IEEE Xplore Compliant - Part Number:CFP18J06-ART, ISBN:978-1-5386-0807-4; DVD Part Number:CFP18J06DVD, ISBN:978-1-5386-0806-7, 2018.
2. Tianyi Liu, Shuangsang Fang, Yuehui Zhao, Peng Wang, Jun Zhang, ’Implementation on Training Convolution Neural Network’.
3. Tao Wang, David J. Wu, Adam Coates, Andrew Y. Ng, ‘End-to-End text Recognition with Convolution Neural Network’.
4. Mahesh Jangid, Sumit Srivastava, ‘Handwritten Devangari Characters Recognition Using Layer-wise Training of Deep Convolution Neural Networks and Adaptive Gradient Methods’, J. Imaging 2018, 4, 41; doi: 10.3390/jimaging4020041, 2018.
5. Tulshiram B. Pisal, Parshuram M. Kamble, ‘Marathi Handwritten Character Recognition by using
6. Probabilistic Neural Network Classification’, IJRCSIT I ISSN No. : 2319-5010 I Vol. 1 I Issue 1(A) I Feb. 2013
7. J. Pradeep, E. Srinivasan, S. Himavathi, ‘Neural network based handwritten character recognition system without feature extraction’. International Conference on Computer, Communication and Electrical Technology – ICCCET 2011, 18th & 19th March, 2011
8. Meduri Avadesh, Navneet Goyal, ‘Optical Character Recognition for Sanskrit Using Convolution Neural Networks’. 13th IAPR International Workshop on Document Analysis Systems, 2018.
9. Kani, Irman Harmadi and Agus Buono, ‘PIXEL DOWNSAMPLING FOR OPTIMIZATION OF ARTIFICIAL NEURAL, ARPN Journal of Engineering and Applied Sciences, VOL. 12, NO. 15, AUGUST 2017
10. Ammar Mohammed, Mohamed Karam, Hesham Hefny, ‘A hybrid approach for word segmentation’, SAI Intelligent Systems Conference 2015, November 10-11, 2015
11. Youlian Zhu, Cheng Huang, ‘An Adaptive Histogram Equalization Algorithm on the Image’, International Conference on Solid State Devices and Materials Science, 2012
12. Meishan Zhang, Nan Yu, Guohong Fu, ‘A Simple and Effective Neural Model for Joint Word Segmentation and POS Tagging’, IEEE Transaction paper, DOI 10.1109/TASLP.2018.2830117, IEEE/ACM
13. Rokus Arnold, Poth Milkos, ‘An Adaptive Histogram Equalization Algorithm on the Image’, 11th IEEE International Symposium on Computational Intelligence and Informatics • 18–20 November, 2010
14. Uday Modha, Preeti Dave, “Image Inpainting-Automatic Detection and Removal of Text From Images”, International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622 Vol. 2, Issue 2, 2012.
15. Muthukumar S, Dr.Krishnan .N, Pasupathi.P, Deepa. S, “Analysis of Image Inpainting Techniques with Exemplar, Poisson, Successive Elimination and 8 Pixel Neighborhood Methods”, International Journal of Computer Applications (0975 – 8887), Volume 9, No.11, 2010