International Journal of Electronics and Communication Engineering (IJECE)
ISSN(P): 2278-9901; ISSN(E): 2278-991X
Vol. 8, Issue 3, Apr - May 2019; 15-20

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

SURVEY PAPER OF SCRIPT IDENTIFICATION OF TELUGU LANGUAGE USING OCR

B. Hari Kumar[1] & P. Chitra[2]

[1]Reserch Scholar, Department of ECE, Sathyabama Institute of Science and Technology, Chennai,India

[2]Professor, Department of ECE, Sathyabama Institute of Science and Technology, Chennai, India

ABSTRACT

This study provides a summary of the ongoing research and the development process of the optical character recognition (OCR) systems for Devanagari text. A file may contain words in more than a language in a multilingual country like India. Multilingual Optical Character Recognition (OCR) system is required for a multilingual environment to read the multilingual documents. The objective of OCR is an automatic reading of the optically sensed document to interpret human-readable fonts into machine-readable code. Reading of Devanagari script is still a challenging task as various approaches are available for Chinese, English, and Japanese for script acknowledgement to get 100 percent accuracy. A survey has been done on Telugu OCR System. Recognition involves character segmentation into the elements of the component and recognizing them. A heuristic method has been selected as the best classifier for the current work based on the identification precisions of multiple classifiers. In this article, a heuristic approach is developed for separation, feature extraction and recognition of Devanagari script. The subsequent portion of this paper would explain the development that has been made in OCR System application for Devanagari Script reorganization and for the future work's scope in Devanagari OCR systems.

KEYWORDS: *Handwriting Recognition, Optical Character Recognition(OCR), Character Recognition, Multi-Script Documents, Script Identification*

Article History

Received: *06 Apr 2019* **|** ***Revised:*** *10 Apr 2019* **|** ***Accepted:*** *15 May 2019*

INTRODUCTION

Optical character recognition is abbreviated as OCR. This is the translation of the typewritten or handwritten text into the machine-editable form. Optical Character Recognition (OCR) is the method of changing a text image to its relative text. The image in the document could be from newspaper clips, magazine papers, or from the textbooks and could be taken using a camera or scanner. India is a multi-lingual multi-script nation which has more than 18 regional languages derived from 12 various scripts. Example for such pages is question papers, money-order forms, bus reservation forms, language translation books that may contain text lines in more than one language/script forms. One important task of analyzing document image is automatic text information reading from the document image. Optical Character Recognition is a method that can convert text from digital image to editable text. Through optical mechanisms, it allows an appliance to recognize characters. The output of the OCR should be the same as input in formatting. The process includes some pre-processing of the image and then the attainment of important data about the text. That data or knowledge can be used to identify characters. OCR is becoming a significant portion of recent research-based computer applications.

About the advent of Unicode and support of complex scripts on personal computers, the significance of this application has improved. The present study is absorbed on examination of conceivable techniques when noise is present in the signal to develop an OCR system for Devanagari language. A thorough study of the Devanagari writing system has been performed to comprehend the core difficulties. Existing OCR systems are also researched to know the latest study ongoing in this industry. The importance was to find diacritic handling for Devanagari strings and workable segmentation technique and construct a recognition module for these strings. The whole methodology is planned to advance an OCR system for Devanagari and an application for testing is also made. The results from the tests are compared and reported with the earlier work done in this field.

Types of OCR

There are three types of OCR. They are:

Online Handwritten Text

Online handwritten text is directly written on an electronic medium using various digital devices. The output is a preparation of x-y coordinates that express the position of the pen as well as other data such as speed and pressure of writing.

Offline Handwritten Text

The manuscript created by an individual by writing with a pencil/pen on a paper, which is then scanned to the digitalized form is called Offline Handwritten Text.

Machine Printed Text

Machine printed texts are formed by offset processes and found commonly in printed forms.

Uses of OCR

To scan different document types such as images or PDF files and convert them into an editable file, Optical Character Recognition is used.

The OCR system is used for the below reasons:

- Legal Billing (example, any government manuscript)
- Data Entry (example, passport, check, receipt)
- Editable text (example, contracts, resume)
- Save space (example, Free up storage space)

Motivation

Devanagari OCR system is required to alter many published Telugu books into computer text files in editable format. The present research is focused on the examination of an imaginable method to develop an OCR system for Devanagari script. The heuristic process is used to digitalize paper-based papers to reserve these papers and make them accessible fully searchable and processable in digital form.

The first step for altering the archives of hard copies into a digital archive is document scanning. The subsequent step is the application of Optical Character Recognition process, in the sense that the scanned image of each document will be

deciphered into machine processable text. Due to the documents' print quality and the error-prone design matching techniques of the OCR method, OCR mistakes occur.

Modern Optical Character Recognition processors have 99% of character recognition rates on the high-quality documents. Let's assume an average word length of 5 characters, which still means that there is a defect of one out of 20 words. Therefore, at least 5% of all managed words will comprise OCR mistakes. This error rate will even be more on the important documents because the quality of the print is to be of lower quality.

Once the OCR process is finished, many post-processing steps are essential dependent on the application, e.g. documents proof-reading for correcting spelling mistakes and OCR errors or tagging the documents with meta-data (author, year, etc.). Information that contains OCR errors or spelling mistakes is hard to process. For instance, a standard full-text search will not retrieve misspelled versions of a query string. To attain the demanding requirements of application toward nil errors, a post-processing phase to correct these mistakes is a very important part of the post-processing chain.

A post-processing mistake correction system can be semi-automatic, manual or fully automatic. A fully-automatic post-correction system does the correction of detecting errors by itself. Because manual or semi-automatic corrections need a lot of human time and effort, fully-automatic systems become essential to do a full correction. A semi-automatic post-correction system notices errors inevitably and suggests corrections to human correctors who then have to select the right proposal.

Related Works

Abdul Kawsar Tushar et. al. [8] proposed in his paper a model for knowledge transformation from one character acknowledgement task to another. An original form of convolutional neural network for transfer learning with competitive accuracy and curtailed time has conversed for this resolution.

Bindu Philipet al. (2009) [1] has suggested a technique for recognizing Malayalam characters in Malayalamtext documents by using an SVM (Support Vector Machine) method. In this writing, a cross-sectional analysis is achieved along each row of the regularized binary image medium leading to the formation of separate features. The planned algorithms have been verified on a diversity of printed Malayalam fonts and presently attain recognition rates between 95.31 % and 90.22%.

Chirag I Patel et al. (2011) [2] emphasize a technique to identify the characters in a given digitalized documents and read the changing effects of the Models using Artificial Neural Network by applying the back proliferating neural network to rise the script recognition's accuracy.

Gaurav Kumar et. al (2017) [7] used convolution neural network (CNN) for Devanagari character recognition. This study focused on handwritten images of Devanagari script digits. The comparison of the accuracy of existing classification models like KNN and SVM are also discussed. And, they constructed a CNN based learning model and tried to regulate the convolution impact, dropout, and connected layer fully on correctness. They got 99.07 % accuracy rate after training above model with our dataset having handwritten pictures.

Konkimalla Chandra Prakash et. al (2018) [6] has used CNN for the recognition of Telugu script. In this research, a record of Telugu fonts, a client-server solution for the algorithm's online deployment and deep learning based OCR algorithm are provided. The segmentation algorithm can be enhanced so that every character is segmented together with its gunintham and vattu.

Naveen Sankaranet.al (2012) [3] has proposed BLSTM (Bidirectional Long-Short Term Memory) based system that achieves credit at the word level. It outcomes in more than 20% development in the accuracy of word while comparing traditional OCR system. This method does not need a character to word segmentation, which is one of the most common cause for high word error rate.

For printed Hindi characters, Prasanta Pratim Bairagi et.al (2018) [4] described a system for OCR. The recognition accuracy of the prototype application is very challenging. In this article, the vertical and horizontal method is used for recognition of printed Hindi characters in Devanagari script, which provides more accuracy than other methods.

Convolution Neural Network

Convolutional Neural Network (CNN) is an inspired trainable architecture of machine learning that can study from skills like standard multilayer neural networks. CNN involve numerous layers of overlain slating groups of small neurons to attain improved depiction of the original image. CNN is used widely for video and image recognition. There are three major types of layers used to shape the architecture of CNN.

Convolution Layer

The convolution layer is the essential part of a CNN. It convolves the input image with a set of learnable weights or filters, each making one aspect map in the output picture.

Fully-Connected Layer

In the neural network, the fully-connected layer is used for the reasoning at high-level. It takes all neurons in the earlier layer and joins it to every single neuron. Their stimulations can be calculated with a matrix development followed by a bias offset like a standard neural network.

Pooling Layer

The pooling layer is used to decrease the spatial size of the representation progressively to decrease the number of computation and parameters in the network. The pooling layer takes minor rectangular blocks from the subsamples and convolution layer and it produces a single output from that block. There are numerous ways to do this pooling, such as taking the maximum or average, or a learned linear grouping of the neurons in the block.

RESULTS AND DISCUSSIONS

The dataset was verified on diverse cataloguing existing algorithms like KNN, SVM, and Gradient Descent. The below table shows the accuracy result.

Table 1

Classifier	Accuracy(%)
SVM	81.29
Gradient Descent	86.82
KNN	97.10
CNN	99.07

We tested the dataset of transferring the Devanagari script to Convolution Neural network from Novel method.

Table 2

Transfer Learning Comparison. Source Task	Destination Task	Best Accuracy Achieved	Best Accuracy in Epoch Number	Accuracy After 10 Epochs
Urdu	Bangla	96.99%	42	93.90%
Bangla	Urdu	97.79%	48	97.12%
Hindi	Bangla	98.66%	38	95.45%
Urdu	Hindi	95.88%	77	92.11%
Bangla	Hindi	98.57%	52	93.67%
Hindi	Urdu	98.57%	64	92.44%

REFERENCES

1. *B. Philip and R. D. Sudhaker Samuel. "An Efficient OCR for Printed Malayalam Text using Novel Segmentation Algorithm and SVM Classifiers" International Journal of Recent Trends in Engineering, Issue. 1, Vol. 1, May 2009*
2. *Chirag I Patel, Ripal Patel, Palak Patel Handwritten Character Recognition using Neural Network International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011*
3. *Sankaran, Naveen, and C. V. Jawahar. "Recognition of printed Devanagari text using BLSTM Neural Network." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.*
4. *Darshan Kawade1, Neha Kaunds1, Vedang Date 1, Rajendra Pawar, "OCR for Devanagari Script Using Tensorflow Technology" International Journal of Innovative Research in Computer and Communication EngineeringVol. 6, Issue 3, March 2018.*
5. *Prashant S. Kolhe & S. G. Shinde, Devanagari OCR Using KNN and Moment, International Journal of Computer Science and Engineering(IJCSE), Volume 2, Issue 2, April-May 2013, Pp 93-100.*
6. *Prasanta Pratim Bairagi, "Optical Character Recognition for Hindiusing Eight neighbourhood of a pixel networks," International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 05 | May-2018.*
7. *Konkimalla Chandra Prakash, Y. M. Srikar, Gayam Trishal, Souraj Mandal, Sumohana S. Channappayya optical character recognition (ocr) for telugu Indian Institute of Technology Hyderabad 25 Dec 2018.*
8. *Gaurav Kumar, Sachin Kumar, "CNN Based Handwritten Devanagari Digits Recognition", International Journal of Computer Sciences and Engineering ,Volume-5, Issue-7, june 2017.*
9. *Abdul Kawsar Tushar, Akm Ashiquzzaman, Afia Afrin, and Md. Rashedul Islam, "A Novel Transfer Learning Approach upon Hindi, Arabic, and Bangla Numerals using Convolutional Neural Networks".*