2018 13th IAPR International Workshop on Document Analysis Systems

Optical Character Recognition for Sanskrit using Convolution Neural Networks

Meduri Avadesh* and Navneet Goyal**

Department of Mechanical Engneering,*
*Department of Computer Science**,*
Birla Institute of Technology and Science Pilani, Pilani Campus, India
f2015641@pilani.bits-pilani.ac.in, goel@pilani.bits-pilani.ac.in***

Abstract—**Ancient Sanskrit manuscripts are a rich source of knowledge about Science, Mathematics, Hindu mythology, Indian civilization, and culture. It therefore becomes critical that access to these manuscripts is made easy, to share this knowledge with the world and to facilitate further research on this Ancient literature. In this paper, we propose a Convolutional Neural Network (CNN) based Optical Character Recognition system (OCR) which accurately digitizes Ancient Sanskrit manuscripts (Devanagari Script) that are not necessarily in good condition. We use an image segmentation algorithm for calculating pixel intensities to identify letters in the image. The OCR considers typical compound characters (half letter combinations) as separate classes in order to improve the segmentation accuracy. The novelty of the OCR is its robustness to image quality, image contrast, font style and font size, which makes it an ideal choice for digitizing soiled and poorly maintained Sanskrit manuscripts.**

Index Terms—**Devanagari Script, Sanskrit, Hindi, Deep Learning, OCR, digitization, Optical character recognition, CNN**

I. Introduction

Sanskrit is gaining importance in various academic communities due to the presence of ancient scientific and mathematical research work written in this language. Scientists all over the world, are spending increasing amount of time trying to understand these ancient research manuscripts. However, the lack of accurately digitized and tagged versions of Sanskrit manuscripts is a major bottleneck. In addition to this, the poor maintenance and text quality adds to the problem. Hence, it becomes essential to digitize such ancient manuscripts which are not only important for research but, are also an important part of the culture and heritage of India. In order to facilitate digitization of ancient Sanskrit material, we build an Indic Optical Character Recognition System (OCR), specifically for Sanskrit.

In the recent years, several OCRs have been developed for various Indian languages such as Hindi, Bangla, Telugu etc. [10,11,12,13]. However, very little work has been done to develop good OCRs for Sanskrit. Even though both Hindi and Sanskrit are written in the Devanagari script, it is important to use a Sanskrit OCR instead of a Hindi OCR to digitize Sanskrit text due to the significant difference in complexity between the two languages. Sanskrit text consists of several compound characters which are formed by different combinations of half letter and full letter consonants. Some examples of compound characters are shown in Fig 3 and Fig 4. Since such compound characters are either less frequent or completely absent in Hindi text, Hindi OCRs would not be trained to segment and classify such characters correctly. Subsequently, the Hindi OCRs would display poor results in Sanskrit text.

Most of the recent Indic OCR systems make use of machine learning algorithms such as support vector machines (SVMs) [12] and artificial neural networks (ANNs) [11,16] to classify letters in the image. These classifier models used in the OCRs are trained with input images that are often downsampled by applying PCA [15,16], Gabour Filters [15,27], Geometric Feature Graphs [27] etc., in order to reduce the complexity of the data. However, this results in a loss of important information necessary to make the classifier robust. For example, the SVM classifier [12] displays different classification accuracy for different font styles, showing that it does not generalize to different font styles. In addition to this, existing Indic OCRs display poor results on degraded or poorly maintained documents or materials and their digitizing capability is limited to good quality text documents [27].

In order to develop a robust OCR system which can digitize soiled and noisy documents with high accuracy, we propose the use of Convolution neural networks (convnets) as opposed to the popular use of SVMs and ANNs, as convnets possess very high learning capacity and the capability to handle high dimensional data such as images [2]. Convnets have displayed these characteristics consistently in various large-scale image classification and video recognition tasks. Popular Convnet architectures such as the GoogLeNet [5], ResNet [24], VGG Net [6] have achieved state of the art results in popular image classification challenges like the ILSVRC challenge or the ImageNet Challenge. In addition to this, researchers make use of convnets for various other tasks such as human pose estimation, dense semantic segmentation [26] etc.

The main contributions of the paper are 1) Developing an OCR framework for Sanskrit which can digitize soiled and poorly maintained documents 2) The use of CNNs as classifiers for Sanskrit OCRs 3) A Sanskrit letter dataset consisting of 11,230 images belonging 602 classes [1].

The rest of the paper is organized as follows. We first review the related work in section 2 and describe the features of Devanagari script in section 3. In section 4, we discuss the approach used to segment letters in the image and the proce-

[1] Link for dataset: https://github.com/avadesh02/Sanskrit-letter-dataset

978-1-5386-3346-5/18 $31.00 © 2018 IEEE
DOI 10.1109/DAS.2018.50