Deep Digital oriental: A state of the art for Sanskrit oriental handwritten text recognition

Sree Ganesh Thottempudi

SRH University

Germany

Handwritten text recognition is an important research area that requires further analysis of the existing techniques. This article aims to demonstrate the state-of-the-art optical character recognition in Sanskrit (an oriental language from India) on three levels: 1. Comparison of various handwritten Sanskrit characters 2. Implementation of existing models on Sanskrit OCR 3. Development of hybrid deep models for less-resourced languages like Sanskrit. Sanskrit is technically a less-resourced language and holds plenty of manuscripts. It includes a comparative study of various Sanskrit handwritten character recognition techniques that use holistic, analytical, and segmentation-free approaches. The study begins by explaining the distinction between deep learning and machine learning approaches, followed by a description of the Sanskrit handwriting recognition process, which includes pre-processing, feature extraction, and segmentation. The main techniques used in handwriting recognition are illustrated, and a synthesis of these methods is presented. researchers explore these techniques and develop more advanced ones.

Many Indian manuscripts still contain unrevealed knowledge of ancient civilisation and the heritage of the Indian subcontinent. Digitising these ancient manuscripts and processing the language will help discover and preserve their knowledge. This paper aims to provide a detailed comparison of existing OCR tools that can be trained on Devanagari scripts. This paper's primary challenge is finding an optimal OCR tool for Devanagari scripts. Existing papers on this topic have focused on finding different machine-learning models to digitise Devanagari scripts. However, this paper focuses on finding state-of-the-art OCR tools that are easy for humanists and other professionals who work on preserving and researching Devanagari scripts. Hence, this paper compares Tesseract and Transcribes tools with a two-fold approach, i.e., using the pre-trained models and training the model from scratch with their dataset. For this comparison, the paper uses one of the two major epics of ancient Indian literature, "Ramayana." The evaluation of OCR tools is performed based on the character error rate, word error rate, and accuracy. According to the study, the Tesseract model provides the best results, with an accuracy of 90.2%, for digitising the chosen Devanagari scripts.

The ability to automatically recognize text on scanned handwritten Sanskrit images has led to the development of numerous applications, such as searching for words in large numbers of documents and editing previously printed documents. Recognising handwritten text in the Sanskrit script is a difficult task that has been addressed more recently compared to other domains. We also compared various methods proposed and applied to different types of images. After comparing existing processes, we presented our hybrid deep models to recognize handwritten Sanskrit documents. This paper offers a comprehensive review of these methods for Sanskrit OCR. It is the first survey to concentrate on Sanskrit handwriting recognition, including recognition rates and descriptions of test data for the approaches discussed in the context of deep learning techniques. The paper also overviews the field and discusses the methods and future research directions.

PROFILE

Lecture at SRH in the areas of Data Science, Python-NLP, provenance and scientific workflow systems, with multiple publications in top conferences and journals. Adjunct professor at IIT - Jodhpur In the area of Data Science and Semantic Web. Passionate about teaching, guiding, and motivating students. Full stack development for many Data Science applications.

EXPERIENCE

Professor in Computer Science and Data Science
SRH Berlin: March 2020 –

Teaching Modules

- Profiling Big Data
- Natural Language Processing
- Applied Research, Digital Humanities
- Case Studies (Industrial and Research)

Research Topics

- AI and Natural Language Processing
- Data Science, Semantic web and Digital technology
- Data Curation in distributed architectures
- Automated Feature Engineering

University of Heidelberg Research Scientist	**Heidelberg, Germany** Jan 2017 - Feb-2020
Georg-Eckert– Leibniz-Institut IT Consultant	**Braunschweig, Germany** Sep 2019 – Mar 2020
Success Metric PVT Consultant (R&D) (50%)	**Hyderabad, India** Jul 2017 – Jun 2019
DAASI International Solutions Engineer (R&D)	**Tüebingen, Germany** Aug 2015 – Dec 2016
GCDH Software Developer	**Göttingen, Germany** August 2012 – July 2015
CALTS & Adventum Solutions Pvt. Ltd. NLP-Software Developer	**Hyderabad, India** July 2006 – July 2012
Red Hat Inc. Language Maintainer - IT	**Pune, India** August 2007 – July 2008

I have been working for National and International IT companies as a R&D Consultant.