INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Jennil Thiyam
Roll Number	:	176155101
Programme of Study	:	Ph.D.

Thesis Title: A Structure-preserving Document Conversion System for Manipuri Documents in Bengali Script to Meetei Script

Name of Thesis Supervisor(s)	:	Dr. Sasnasam Ranbir Singh and Prof. Prabin Kumar Bora
Thesis Submitted to the Department/ Center	:	CLST
Date of completion of Thesis Viva-Voce Exam	:	05-06-2023
Key words for description of Thesis Work	:	Document segmentation, Image classification, OCR, Document conversion, Document intelligence

SHORT ABSTRACT

Manipuri, or Meeteilon, is one of the resource-poor languages of India and the lingua franca of the Indian state of Manipur. Though the Meetei Mayek (Manipuri script) is known to use for writing Manipuri documents since the early \ nth{6} AD, it was banned and replaced with the Bengali script} by the then king of Manipur in the 18th century. Since the late 70s, the Government of Manipur has made an effort to reintroduce Meetei Mayek and included it in Unicode in the year 2009. Meetel Mayek is progressively replacing the Bengali script in schools, colleges, offices, and other places. During the era of using Bengali script as Manipuri writing script (more than 300 years), a huge volume of Manipuri documents has been created in Bengali script. Almost all of the Manipuri literary materials are in Bengali script, and the population in Manipur is broadly divided into - Bengali script literate and Meetei Mayek literate. After a few decades, the majority of the Manipuri population will not be able to read/write Bengali script, creating a huge gap in accessing literary materials. Therefore, there is an urgent need to develop an effective system to convert Manipuri documents written in the Bengali script to Meetei Mayek to bridge the script divide. Motivated by the above concern, this thesis focuses on the following three research problems associated with the development of an automatic document conversion system (DCS) for the Manipuri documents in the Bengali script to Meetei Mayek.

- Document segmentation and region classification system:} Document segmentation and region classification are the first and foremost tasks in developing a DCS. In document region segmentation and classification, one of the prominent challenges is effectively segmenting non-textual regions that contain sparsely clustered pixels. While previous studies have primarily concentrated on using a single model to segment regions of interest (textual or non-textual), this thesis proposes a novel 2-tier feedback-based end-to-end deep learning and rule-based integrated framework. The framework aims to address this challenge by enabling joint segmentation and identification of regions of interest in a more efficient and effective manner. In addition, a dataset (document images and their corresponding mask images) for future similar research activities has also been created.

- Chart type classification model: In the field of chart type classification, false classification poses a significant challenge due to two main factors. First, there are confusing chart type pairs where multiple chart

types exhibit very similar characteristics. Second, noisy samples significantly contribute to misclassification as charts often contain additional components such as textual elements, the information presented in shapes, and marking points. To tackle these challenges, this study proposes a novel approach based on an attention and triplet loss-based Convolutional Neural Network (CNN) framework which enhances the ability of the model to distinguish between similar chart types and mitigate the impact of noisy samples. In addition, a dataset of 28 chart types has been proposed for future similar research.

- Manipuri OCR system: The development of an Optical Character Recognition (OCR) system from scratch poses a challenge, particularly when it comes to low-resource languages. Obtaining large text corpora from diverse environments is crucial for the effective functioning of the OCR system. However, curating such extensive corpora proves to be a challenging task for low-resource languages, as the availability of digitized textual data is severely limited. To address this challenge, this thesis employs an adaptive approach by leveraging existing OCR systems designed for similar scripts. By adopting this adaptive approach, the thesis aims to overcome the scarcity of resources for the low-resource language, Manipuri.

A prototype of the proposed DCS is implemented, and details are presented in Appendix A. In summary, this thesis makes important research contributions in terms of datasets and models for document segmentation, chart classification, and optical character recognition and develops a prototype end-to-end DCS.