ISRAR UDDIN
01-281121-005

Optical Character Recognition for Printed Urdu Nastaliq Font

A thesis submitted to the Department of Computer Engineering, Faculty of Engineering Sciences, Bahria University, Islamabad, in the partial fulfillment for the requirements of a Doctoral degree in Computer Engineering

Supervisor: Dr. Imran Siddiqi

Co-Supervisor: Dr. Shehzad Khalid

Department of Computer Engineering
Bahria University, Islamabad

March 2019

Abstract

Optical Character Recognition (OCR) is one of the most investigated pattern classification problems that has received remarkable research attention for more than half a century. From the simplest systems recognizing isolated digits to end-to-end recognition systems, applications of OCRs vary from postal mail sorting to reading systems in scene images facilitating autonomous navigation or assisting the visually impaired. Despite tremendous research endeavors and availability of commercial recognition engines for many scripts, recognition of cursive scripts still remains an open and challenging research problem mainly due to the complexity of script, segmentation issues and large number of classes to recognize. Among these, Urdu makes the subject of our study. More specifically, this study investigates the recognition of printed Urdu text in Nastaliq style, the most widely employed script for Urdu text that is more complex than the Naskh style of Arabic.

This work presents a holistic (segmentation-free) technique that exploits ligatures (partial words) as units of recognition. Urdu has a total of more than 26,000 unique ligatures, many of the ligatures, however, share the same main body (primary ligature) and differ only in the number and position of dots and diacritics (secondary ligatures). We exploit this idea to separately recognize the primary and secondary ligatures and later re-associate the two to recognize the complete ligature. Recognition is carried out using two techniques; the first of these is based on hand-crafted statistical features using hidden Markov models (HMMs). Features extracted using sliding windows are used to train a separate model for each ligature class. Feature sequences of the query ligature are fed to all the models and recognition is carried out through the model that reports the maximum probability. The second technique employs Convolutional Neural Networks (CNNs) to automatically extract useful feature representations from the classes and recognize the ligatures. We investigated the performance of a number of pre-trained networks using transfer learning techniques and trained our own set of networks from scratch as well.

Experimental study of the system is carried out on two benchmark datasets of Urdu text, the 'Urdu Printed Text Images' (UPTI) database and the 'Center of Language Engineering' (CLE) database. A number of experimental scenarios are considered for system evaluation and the realized recognition rates are compared with state-of-the-art recognition systems for printed Urdu text. An interesting aspect of experimental study is the combination of unique ligatures in the two datasets to generate a large set of around 2800 unique primary and secondary ligatures covering a major proportion of the Urdu corpus. The system reports high classification rates (88.10% and 94.78% on CLE and UPTI query ligatures respectively) demonstrating the effectiveness of the proposed recognition techniques which can be adapted for other cursive scripts as well. The findings of this study are expected to be useful for the document recognition community in general and researchers targeting cursive scripts in particular.

Contents

List of Tables