Survey and bibliography of Arabic optical text recognition

Badr Al-Badr[a],*, Sabri A. Mahmoud[b]

Department of Computer Science and Engineering, Mail Stop FR-35, University of Washington, Seattle, WA 98195, USA
Computer Engineering Department, College of Computers and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia

Abstract

Research work on Arabic optical text recognition (AOTR), although lagging that of other languages, is becoming more intensive than before and commercial systems for AOTR are becoming available. This paper presents a comprehensive survey and bibliography of research on AOTR, by covering all the research publications on AOTR to which the authors had access. This paper introduces the general topic of optical character recognition (OCR), and highlights the characteristics of Arabic text. It also presents an historical review of the Arabic text recognition systems. Further, this paper reports on the state of the art in AOTR research, and lists the specifications of commercially available systems for AOTR. In this paper, we first underline the capabilities of different AOTR systems, and then introduce a five stage model for AOTR systems and classify research work according to this model. We devote a section to each of the stages of this model: preprocessing, segmentation, feature extraction, classification, and post-processing. In the preprocessing section, we emphasize handling degraded documents, and thinning of Arabic text. In the segmentation section, we discuss methods of segmenting Arabic text and categorize the methods into five general approaches. In the feature extraction and classification sections, we highlight the main techniques and analyze AOTR research works based on those techniques. We then discuss approaches for post-processing and show their relation to the Arabic language. We conclude by pointing problems and directions for future research on AOTR.