Abstract

Optical character recognition (OCR) has been an active area of research over last few decades. Several OCR systems are available today in the market; however, their performance degrades significantly with different fonts, orientation, quality, and presence of compound characters in the script. Our research goal is to achieve higher recognition accuracy for existing OCR systems for Bangla, which is the second-most popular language/script in India and among the top ten popular languages/scripts all over the world. Researchers of this domain aim in finding out new methods that can enhance the performance of existing systems. At the same time, use of existing schemes up to its fullest potential for solving application-specific problems is also important. Several significant contributions have been made in this dissertation toward solving some of the well-known problems in the domain of document image processing in general, and Bangla OCR in particular. To start with, we propose an *adaptive-cum-interpolative binarization* method in a *multi-scale framework* to handle degraded documents, a character segmentation method based on *vertex characterization* of *isothetic covers*, and a *medial-axis-based thinning* strategy tailored for thinning printed and handwritten characters in Bangla and few other Indian scripts. We also present certain novel features based on the *topology* of a character, captured by structural features of the strokes and their spatial relation. We use these topological features for recognition of printed and handwritten Bangla characters. Finally, we focus on the challenges to recognize compound characters in Bangla script in order to further improve the OCR performance. We examine the multiple *segmentation hypotheses* to segment the compound characters into prominent shape components. *Grouping principles* specific to language and structural characteristics are used to identify these shape components in a compound character. Identification of the compound character is done by taking note of the identity of the shape components using *topological features* and *template matching* technique. Exhaustive experimentation has been done in each stage to evaluate the performance of our proposed techniques.