Segmentation Method for Myanmar Character Recognition Using Block based Pixel Count and Aspect Ratio

Kyi Pyar Zaw
kyipyarzaw08@gmail.com
University of Computer Studies, Mandalay, UCSM

Zin Mar Kyu
zinmarkyu.pp@gmail.com
University of Computer Studies, Mandalay, UCSM

Character recognition is the process of converting a text image file into editable and searchable text file. Basically, there are three steps of character recognition such as character segmentation, feature extraction and classification. This paper mainly focus on the Myanmar character segmentation. Character segmentation is a vital area of research for optical character recognition. In this paper, the incoming text based images are segmented into lines, words and characters. Horizontal cropping is used for line segmentation and vertical cropping is used for vertically non-touching word and character segmentation. In a Myanmar compound word, there are one basic character and one or more extended characters. These basic character and extended characters may be connected or not according to the typing style or font style and Myanmar script nature. Therefore, it is difficult to segment these connected characters into individual characters. To solve this problem, we use block based pixel count and aspect ratio. This system can segment both touching characters and non-touching characters in text line image. Features are extracted from this segmenting characters. These individual characters are classified using eight directions chain code features and block based pixel count. Finally, the recognized text image is converted into editable text. In this paper, 92 characters in Myanmar script (34 consonants, 13 dependent vowels, 12 independent vowels, 1 punctuation mark, 10 digits, 8 medial, 5 compound medial, 3 tone characters and 6 compound words) are trained and different test line images that contain various words and characters are tested.