REGION-BASED TOUCHED CHARACTER SEGMENTATION IN HANDWRITTEN WORDS

Fajri Kurniawan, Mohd Shafry Mohd Rahim, Daut Daman, Amjad Rehman
Dzulkifli Mohamad and Siti Mariyam Shamsuddin

Department of Computer Graphics and Multimedia
Universiti Teknologi Malaysia
Skudai 83110, Johor Bahru, Malaysia
kfajri2@live.utm.my

ABSTRACT. *This paper presents an intelligent technique for segmentation of off-line cursive handwritten words particularly on touching characters problem. In this study, Self Organizing Feature Maps (SOM) is implemented to identify the touching portion of the cursive words. The image of the connected characters is preprocessed and the core-zone is detected to overcome ascender and descender of the touched character. Prior to clustering, the pixels of the image were mapped into coordinate system as features vector. These features vector are clustered into three classes: left, right and middle region, and then vertical segmentation is performed using SOM to determine the winner node of middle region. The experiments are conducted using syntactic CCC database. The results show that the proposed algorithm yields promising segmentation output and feasible with other existing techniques.*
Keywords: Touching character segmentation, Self-organizing map, Cursive handwritten, Syntactic database

1. **Introduction.** Segmentation into isolated characters has played a crucial step for handwriting recognition systems [1,2,31,32]. One of the most challenging problems is segmentation on touching characters, because touching characters might lead to error-prone segmentation stage which also leads to recognition errors [3]. In addition, ambiguous touching character pairs such as "LI", "lo", "nn', "cl", "rn" tend to be misclassified by the/their recognizer/s. Therefore, one of the solutions of overcoming this problem is by implementing a lexicon-based OCR to obtain context information.

Many segmentation of touching string approaches have been reported. The segmentation strategies have been concluded as follows: based on "character like" properties method; recognition based method and holistic method [29]. Recently, some novel methods for touching string segmentation have been proposed, such as the Drop-fall based algorithms [25], the thin based algorithms [26], the water reservoir algorithms [27] and the stroke merging algorithms [28]. These methods depend on the features of the characters, the words or the string which contains the digits, characters, numeric symbols and so on. The selection of the features has been an art rather than a technique. In other words, features selection mainly depends on the experience of the author/s, and the priori of the string is ignored. In essence, the relationship between segmentation and recognition fall in a dilemmatic situation like the Catch-22 [29]. That is, each step depends on the other, and in complex cases it is paradoxical to seek a pattern that will match a member of the system's recognition alphabet of symbols without incorporating detailed knowledge of the structure of those symbols into the process. Consequently, the priori knowledge is