Fast Recognition of Asian Characters Based on Database Methodologies

Woong-Kee Loh[1], Young-Ho Park[2], and Yong-Ik Yoon[2]

[1] Department of Computer Science & Engineering, University of Minnesota
200 Union Street SE, Minneapolis, MN 55455, USA
lohw@cs.umn.edu
[2] Department of Multimedia Science, Sookmyung Women's University
53-12 Chungpa-Dong, Yongsan-gu, Seoul 140-742, Korea
{yhpark,yiyoon}@sm.ac.kr

Abstract. Character recognition has been an active research area in the field of pattern recognition. The existing character recognition algorithms are focused mainly on increasing the recognition rate. However, as in the recent Google Library Project, the requirement for speeding up recognition of enormous amount of documents is growing. Moreover, the existing algorithms do not pay enough attention to Asian characters. In this paper, we propose an algorithm for fast recognition of Asian characters based on the database methodologies. Since the number of Asian characters is very large and their shapes are complicated, Asian characters require much more recognition time than numeric and Roman characters. The proposed algorithm extracts the feature from each of Asian characters through the Discrete Fourier Transform (DFT) and optimizes the recognition speed by storing and retrieving the features using a multidimensional index. We improve the recognition speed of the proposed algorithm using the association rule technique, which is a widely adopted data mining technique. The proposed algorithm has the advantage that it can be applied regardless of the language, size, and font of the characters to be recognized.

Keywords: character recognition, Discrete Fourier Transform, multidimensional index, association rule.

1 Introduction

Character recognition has been an active research area since 1980s in the field of pattern recognition. Many character recognition algorithms have been implemented for various applications including postal services, and many commercial software packages have been released until recently.

The characters to be recognized can be categorized into two groups: printed and handwritten characters [7,13]. The printed characters are those in printed materials such as textbooks, magazines, and newspapers. A lot of research on recognition of printed characters has been performed historically, and the recognition rate on printed numbers and Roman characters reaches almost up to 100%.

R. Cooper and J. Kennedy (Eds.): BNCOD 2007, LNCS 4587, pp. 37–48, 2007.

The handwritten characters are those written by human hands. Since the variation in handwritten characters is more salient than the printed characters, the recognition rate on the former is generally much lower than the latter. Although the recognition rate on correctly handwritten characters reaches up to 90%, the recognition on cursive and unsegmented characters is still a tough research issue. The recognition process of printed characters consists of two phases [13]. In the first phase, a template is generated for each of characters to be recognized by extracting abstract feature from the shape of the character, and in the second phase, given a scanned character as an input, a template with the closest feature to the character is returned.

The existing character recognition algorithms in the pattern recognition field are focused mainly on the correctness of recognition, but not on recognition speed. The requirement on recognition speed was originated by the applications that recognize enormous amount of documents. An example is Google Library Project, which digitizes and recognizes the contents of books stocked in big libraries and provides the service of searching on the book contents [9]. Moreover, the existing algorithms do not pay enough attention to Asian characters. In general, different character recognition algorithms should be used depending on the character set and the number, size, and font of the characters to be recognized. Especially, the number of Asian characters is much more than Roman characters and their shapes are much more complicated. For example, the number of all Korean characters is as many as 11,172, and even the number of frequently used ones goes up to 2,350 [12].

In this paper, we propose an algorithm for fast recognition of printed Asian characters based on the database methodologies. The proposed algorithm extracts the feature from each Asian character through the Discrete Fourier Transform (DFT) [1,15], and optimizes recognition speed by storing and retrieving the features in a multidimensional index [3,5]. We improve the recognition speed of the proposed algorithm using the association rule technique [2], which is a widely adopted data mining technique. In the proposed algorithm, an association rule is a pattern of character sequence that frequently appears in a document, and is used to improve the recognition speed by reducing the number of unnecessary feature comparisons. The proposed algorithm has the advantage that it can be applied regardless of the language, size, and font of the characters to be recognized.

This paper is organized as follows. In Section 2, we briefly describe previous related work. In Section 3, we explain in detail about the proposed algorithm. In Section 4, we perform experiments to evaluate the performance of the proposed algorithm. Finally, we conclude this paper in Section 5.

2 Related Work

The algorithms for printed character recognition generate a template for each character to be recognized by extracting abstract feature from the shape of the character [13]. Since the performance of the algorithms is highly dependent on