4

Multi-Font Printed Tibetan OCR

Xiaoqing Ding and Hua Wang

4.1 Introduction

Tibetan culture has a very long history and it is a splendid contribution both in China and over the world. As a significant representative of Tibetan culture, Tibetan language is still used by more than six million people in China at present. The Tibetan character set, which records Tibetan language and Tibetan culture, is very special and different in comparison with other character sets in the world, such as Chinese and so on. Therefore, research on Tibetan OCR, which will enable easier modernization of Tibetan culture and digitization of Tibetan document, is very important in theoretical value as well as in extensive application perspective. However, only few research works have been undertaken so far.

Masami et al. created an object-oriented dictionary [16], by combining categorization and character identification procedures to separately recognize basic consonants, combination characters and vowels. Furthermore, Euclidean distance with deferential weights [17] was designed to discriminate similar characters. Ma et al. established an experimental system [**?**] based on fuzzy line features and Euclidean distance classifier. Overall, the research of Tibetan OCR is still in its infancy and there remain several limitations in previous works. First, the importance of multi-font Tibetan character recognition problem has not been widely thought of. Reported researches are all focused on single font samples. Second, no effective and robust strategy to recognize actual Tibetan scripts has been proposed. Finally, huge dictionaries of Tibetan syllables are indispensable in achieving encouraging recognition results.

In this chapter, a novel and effective method based on statistical pattern recognition approach for multi-font printed Tibetan OCR is proposed. A robust Tibetan character recognition algorithm is designed whose destination

character set contains 584 modern Tibetan character categories used commonly in the Tibetan documents.

On the other hand, it is still challenging to develop a character-and-document recognition system, which can achieve extremely high recognition accuracy regardless of the quality of input scripts. Researchers have noted that most recognition errors that occur in an OCR system are due to character segmentation errors [15]. In this chapter, document segmentation, which is divided into two steps, is also discussed: line separation and character segmentation. To the authors' knowledge, no special technique aiming at printed Tibetan scripts' segmentation has been reported. However, many ideas can be borrowed from previous techniques that deal with other scripts [1, 5, 7, 11, 15]. Various algorithms [12] are available for skew detection and correction before text line separation. Histogram of horizontal projection is commonly used to separate text lines. Many segmentation strategies [1, 15] have been proposed to segment words into their character components. However, no exact previous technique can be directly implemented to solve Tibetan text segmentation. A comprehensive text segmentation method for multi-font Tibetan recognition is developed. Its validity is demonstrated by experimental results on a large-scale set.

The organization of this chapter is as follows. A brief introduction to the properties of Tibetan characters and scripts is given in Section 4.2. Section 4.3 describes the details of Tibetan recognition algorithm, including character normalization, statistical feature vector formulation and classifier design. The two-stage document segmentation strategy is discussed in Section 4.4. Experimental results are given in Section 4.5. The final section summarizes the chapter.

4.2 Properties of Tibetan Characters and Scripts

There are totally 34 basic elements (Figure 4.1), which consist of 30 consonants (3 of which have modified forms) and 4 vowels expanding the whole Tibetan character set in modern Tibetan language.

There are two kinds of characters used in written Tibetan language: (1) *consonants*, which serve as valid characters by themselves. We call them

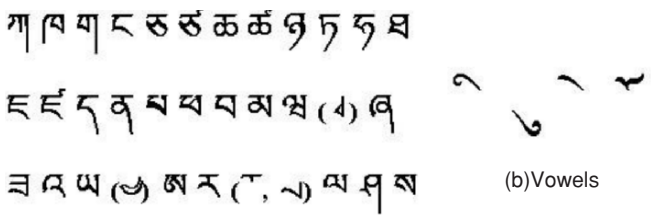

Fig. 4.1. Basic elements of Tibetan character set.