Research Report on Tibetan OCR (Printed Tibetan Character Recognition) part

Planned Completion Date : 12/30/08

Title: Research Report on Tibetan OCR (Printed Tibetan Character Recognition)

Researchers: Ngodrup, Putsering, Zhao dongcai, Bianbawabgdui,Liufang

Presentation: Software，Papers and Documentations

Abstraction: Information processing technology plays a more and more critical role in the development of information technology and modernization of Chinese society. As a key point of new high technology, China achieved great successes in the field of information processing, particularly in technology studies and products development. Currently, the technique of Chinese recognition system has already reached the international advanced level.

The development of Tibetan recognition system is still in a beginning stage. Along with the promotion of Tibetan information technology, there is an increasing demand on saving, processing large quantities of Tibetan Scripture books, texts and other materials on the computers. However, it consumes so much human power and time to input the texts and material into computer. In order to properly and efficiently process Tibetan related documents, it is quite urgent to develop a Tibetan character recognition system (Tibetan OCR) to process Tibetan texts and written materials into computer.

Character recognition is a field of research in pattern recognition and artificial intelligence. It is a new technology that is combined pattern recognition technique, image processing technique and word processing technique together, which usually is processed by the methods of feature discrimination and feature matching. Feature discrimination is a sort of classified discriminations according to common rules of the characters (such as English and Chinese). It is a unique method to recognize the character by analyzing the structure of the character in phases according to the degree of feature extraction of the character, but not by studying specific knowledge of the language. The key phase of character recognition is the feature extraction. The method of feature matching is a process of category comparison; the optimal result is also the final result of recognition.

The written Tibetan language is composed of a series of horizontally and/or vertically arranged letters which combine to form composite syllables and words. The thirty consonants and four vowels of the Tibetan alphabet are the most basic components of the written script. One or more letters are written to make a syllable segmentation unit (tsheg bar), the most basic meaningful element of the Tibetan language. Tibetan words consist of one or more syllables.

Each Tibetan syllable contains a root letter (ming gzhi) at its core. The root letter can be combined with one or more of the following: a prefix, a head letter, a subjoined letter, a vowel, a suffix, and/or a post suffix. Syllables are usually separated by a syllable marker (tsheg) or some other form of punctuation. Figure 1 shows the complete structure of a Tibetan syllable.

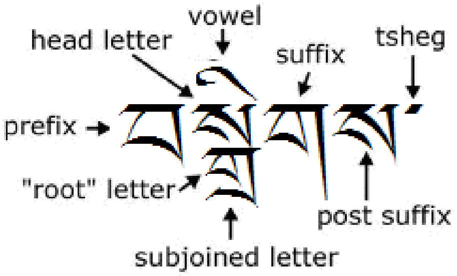


Figure1 The Complete Structure of a Tibetan Syllable

Although there are not many scripts in modern Tibetan, it is still quite difficult to extract one single script from one syllable, for Tibetan script is written both horizontally and vertically and there many distorted scripts and letters adhesion cases while writing Tibetan. Therefore, it usually chooses the character as the basic recognition unit in Tibetan OCR System. It is quite common in Tibetan to have similar characters, so it is the major issue on Tibetan Recognition if the system could effectively recognize the similar characters or not.

The recognition function of Tibetan OCR system are basically realized by following components, the basic structure of the system is shown in figure 2:

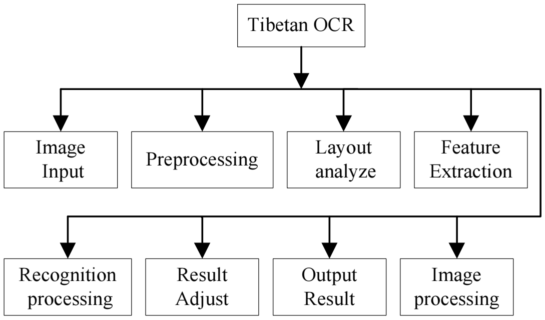


Figure 2 The Structure of Tibetan OCR system

The figure 3 shows the system operating procedures.

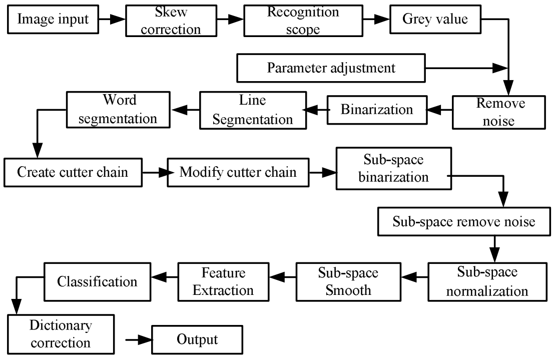


Figure 3 The operation procedure of the printed Tibetan OCR system

Status of Activity: Finished

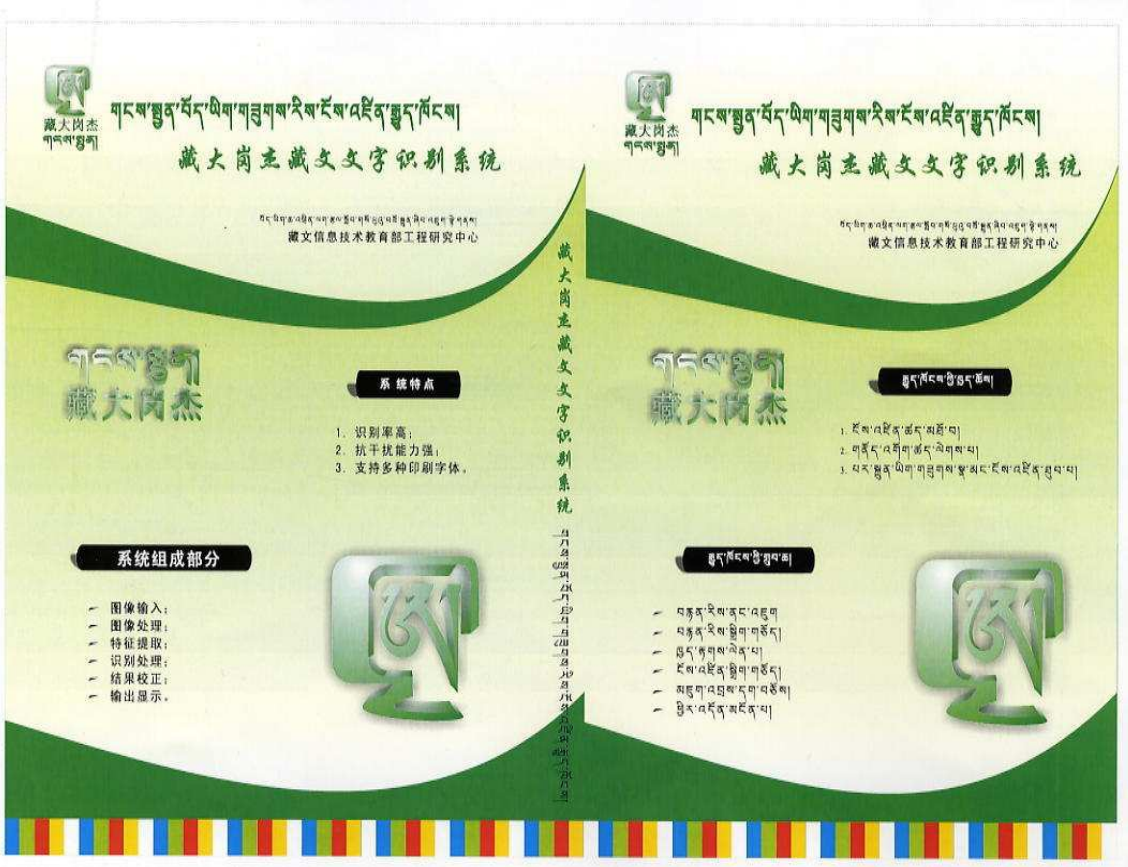
Papers:

(1) Ngodrup Ngodrup, Dongcai Zhao, Putseren , Daluosanglangjie, Fang Liu , BianBaWangDui, Study on Printed Tibetan Character Recognition, Computer Engineering and Application,2009.8
(2) Pu Tsering, Research on Extracting the Character of Tibetan Symbol at Different Printing Shape Letter, Journal of Tibet University. 2008.1

Software：GangJian (Snow eyes) Tibetan OCR is made in TU

Result:

1. Software for Tibetan OCR



གངས་སྤྱན་བོད་ཡིག་གཟུགས་རིས་ངོས་འཛིན་རྒྱུད་ཁོངས།

藏大岗杰藏文文字识别系统

བོད་ཡིག་ཆ་འཕྲིན་ལག་རྩལ་སློབ་གསོ་བུའི་བཟོ་སྐྲུན་ཞིབ་འཇུག་ལྟེ་གནས།

藏文信息技术教育部工程研究中心

系统特点

1. 识别率高；
2. 抗干扰能力强；
3. 支持多种印刷字体。

系统组成部分

- 图像输入；
- 图像处理；
- 特征提取；
- 识别处理；
- 结果校正；
- 输出显示。

藏大岗杰藏文文字识别系统

གངས་སྤྱན་བོད་ཡིག་གཟུགས་རིས་ངོས་འཛིན་རྒྱུད་ཁོངས།

藏大岗杰藏文文字识别系统

བོད་ཡིག་ཆ་འཕྲིན་ལག་རྩལ་སློབ་གསོ་བུའི་བཟོ་སྐྲུན་ཞིབ་འཇུག་ལྟེ་གནས།

藏文信息技术教育部工程研究中心

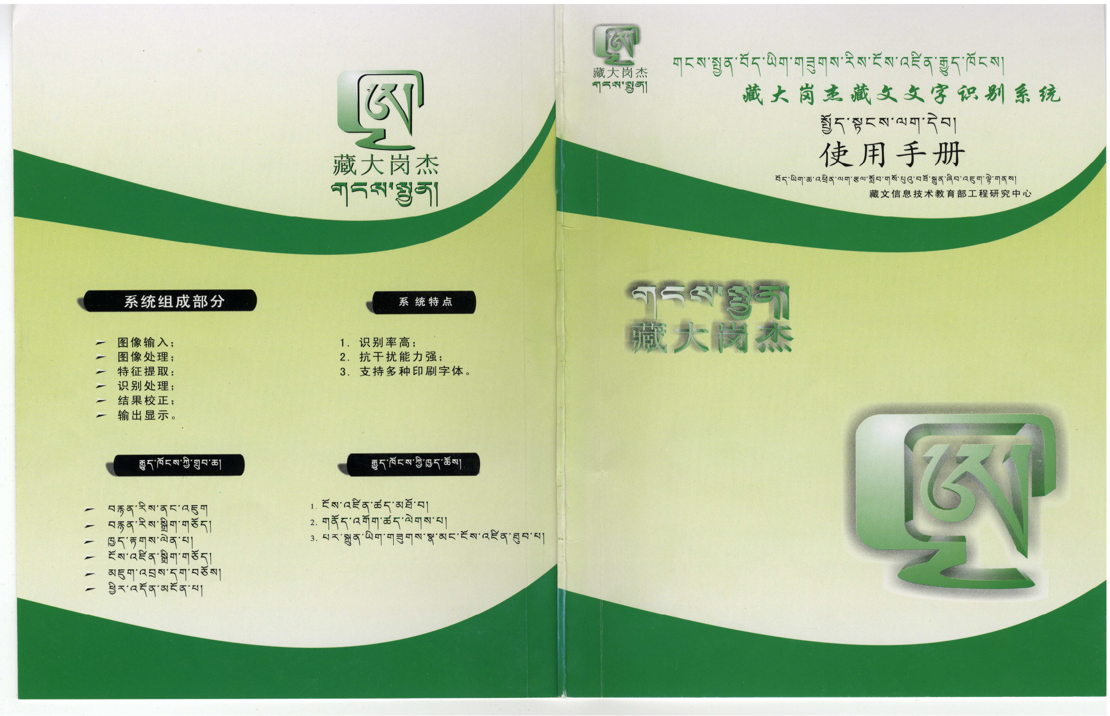
རྒྱུད་ཁོངས་ཀྱི་ཁྱད་ཆོས།

1. ངོས་འཛིན་ཚད་མཐོ་བ།
2. གནོད་འགོག་ཚད་ལེགས་པ།
3. པར་སྐྲུན་ཡིག་གཟུགས་སྣ་མང་ངོས་འཛིན་ཐུབ་པ།

རྒྱུད་ཁོངས་ཀྱི་གྲུབ་ཆ།

- བརྙན་རིས་ནང་འཇུག
- བརྙན་རིས་སྒྲིག་གཅོད།
- ཁྱད་རྟགས་ལེན་པ།
- ངོས་འཛིན་སྒྲིག་གཅོད།
- མཇུག་འབྲས་དག་བཅོས།
- ཕྱིར་འདོན་མངོན་པ།

2. Handbook for Tibetan OCR



藏大岗杰
གངས་སྐྱེད།

系统组成部分

- 图像输入；
- 图像处理；
- 特征提取；
- 识别处理；
- 结果校正；
- 输出显示。

系统特点

1. 识别率高；
2. 抗干扰能力强；
3. 支持多种印刷字体。

རྒྱུད་ཁོངས་ཀྱི་གྲུབ་ཆ།

- བརྙན་རིས་ནང་འཇུག
- བརྙན་རིས་སྒྲིག་གཅོད།
- ཁྱད་རྟགས་ལེན་པ།
- ངོས་འཛིན་སྒྲིག་གཅོད།
- མཇུག་འབྲས་དག་བཅོས།
- ཕྱིར་འདོན་མངོན་པ།

རྒྱུད་ཁོངས་ཀྱི་ཁྱད་ཆོས།

1. ངོས་འཛིན་ཚད་མཐོ་བ།
2. གནོད་འགོག་ཚད་ལེགས་པ།
3. པར་སྐྲུན་ཡིག་གཟུགས་སྣ་མང་ངོས་འཛིན་ཐུབ་པ།

藏大岗杰
གངས་སྐྱེད།

གངས་སྐྱེད་བོད་ཡིག་གཟུགས་རིས་ངོས་འཛིན་རྒྱུད་ཁོངས།
藏大岗杰藏文文字识别系统
སྤྱོད་སྟངས་ལག་དེབ།
使用手册
བོད་ཡིག་ཆ་འཕྲིན་ལག་རྩལ་སློབ་གསོ་པུའུ་བཟོ་སྐྲུན་ཞིབ་འཇུག་ལྟེ་གནས།
藏文信息技术教育部工程研究中心

གངས་སྐྱེད།
藏大岗杰

3. Test Results

Along with formulating Chinese national standard for Tibetan coded character set (including Basic Set, Extension Set A and Extension Set B), the standard for Tibetan coded character set are largely applied in all kinds of printed matters today. Tibetan OCR system that is developed and integrated based on all measures and algorithm listed above has run tests to 1,625 Tibetan scripts from Chinese national standard for Tibetan Coded Character Basic Set (partial) and Extension Set A (all). Under the condition of zero interference in scanning, the average recognition rate could reach to 98%, the average recognition speed could reach to 260 Tibetan scripts per second. Under the condition of existing common-mode interferences, the average recognition rate could reach to 92%, and the average recognition speed could reach to 260 Tibetan scripts per second. The detail test data is shown in the Table 1 and 2 below:

Table 1 Average Recognition Rates of Samples

	The Sample Quality		
Data	Zero Interference	Common Interference Partial Field Recognition	Scanning Entire Text Recognition
Training Samples	99.39%	95.76%	94.55%
Testing Samples	98.78%	93.93%	92.74%

Table 2 The Average Recognition Error of the Candidate Characters in samples

Data	Sample Quality					
	Zero Interference		Common Scanning Interference			
			Partial Field Recognition		Entire Text Recognition	
The Range of Candidate Character	5 letters	10 letters	5 letters	10 letters	5 letters	10 letters
Training Sample	99.52%	99.70%	96.97%	97.58%	95.61%	96.47%
Testing Sample	98.97%	99.21%	95.16%	95.76%	93.64%	94.98%