Automatic Recognition of Tibetan Buddhist Text by Computer

Masami Kojima*1, Yoshiyuki Kawazoe*2 and Masayuki Kimura*3

***1 Dept. of Electrical Communication, Tohoku Institute of Technology**
(E-mail : mkojima@tohtech.ac.jp)
***2 Institute for Materials Research, Tohoku University**
***3 Japan Advanced Institute of Science and Technology, Hokuriku**

Abstract

The purpose of this study is to develop a plausible method to code and compile Buddhist texts automatically from original Tibetan scripts into the Romanized form. We extract syllable from Tibetan texts and recognize automatically the Tibetan characters. The set of Tibetan characters consists of basic 30 consonants, 76 combination characters, and 4 vowels. Despite of the limited number of Tibetan characters, there are many similar characters in shape. Therefore, to separately recognize them we apply an Object Oriented Dictionary (OOD) which is created combining the categorization and character identification procedures. From our experiment, it is confirmed that it is possible to improve the rate of Tibetan character recognition dramatically by Object Oriented Method [Ref. 1,3]. We would like to express our opinion on automatic character recognition for wooden blocked Tibetan manuscripts.

1. Introduction

Most of the computer systems have been developed to solve numerical problems. Therefore, they do not have satisfactory pattern recognition function specifically found in human intelligence. This research has originated from the desire to facilitate the work in coding and compiling Buddhist texts written in Tibetan scripts into the Romanized form to encourage Buddhist literature studies by using the present-day computer assistance. As an example, we have used the " rGyal rabs gsal ba'i me long " published by " Mi rigs dpe skrun khang ", published in February 1993, as a volume of 250 pages to perform the present experiment by using OOD characters. It is hoped that a computer system automatic recognition of these Tibetan printed texts would be eagerly welcome by all scholars engaged in Buddhist literature studies

because of many printed Buddhist literature's are recently being converted in this form.

We are also studying automatic recognition of wooden blocked Tibetan manuscripts. The most difficult problem in this case is the segmentation of one syllable in Tibetan manuscripts. We suggest that it is effectively performed by using the recognition of colored line to segment the syllable by Tibetan researchers, and object oriented method is meaningfully applied to this purpose. We design dictionary characters of syllables by using segmented data beforehand. Accordingly higher rate of character recognition would be expected to be achieved for wooden blocked Tibetan manuscripts which has already been achieved for printed Tibetan texts. We are hardly studying to recognize wooden blocked manuscripts that we think the most important in Tibetan text recognition.

2. Experiments

A sample copy of the original Tibetan text is shown in Fig. 1. The presently used experimental system is shown in Fig. 2. In the actual character recognition procedure, firstly image data is digitized, where the segment is expand by a factor of about 1.4, and read in sequence, line by line using the image scanner, Epson GT-6000 with the precision of 100 dpi (digit per inch). The obtain image data are sent to a personal computer, NEC PC9800, and the data are further preprocessed (for example alignment, segmentation, rejection of noise, and normalization), and recognized. An example of character segmentation is shown in Fig. 3. Results of recognition are printed by using a laser printer, CANON LBP-B406E. The time needed for all procedures starting from reading one character to printed the recognized character is about 5 seconds.

These procedures are automated by using a personal computer. A 99.9 % segmentation rate has been achieved for 141,988 characters in 250 pages of " rGyal rabs gsal ba'i me long ". After the result of character recognition for 17,753 characters within 30 pages of " rGyal rabs gsal ba'i me long ", we know that mistakes mainly happen specifically between the two groups of similar characters ; group 1 : " ba ", " pa " and " pha " and group 2 : " ma " and " sa " that is shown in Fig. 4 [Ref. 4,5]. Therefore, Object Oriented Dictionary (OOD) about these Tibetan characters is created by combining categorization and these characters, respectively.

According to this experiment, a 98.8 % recognition rate has been achieved. Moreover, a 99.0 % recognition rate is realized by combining OOD with differential weight Euclidean distance. An example of the results of Tibetan character recognition is shown in Fig. 5.

A sample copy of the original Tibetan manuscripts is shown in Fig. 6. The colored line to segment the syllable by Tibetan researchers is shown in Fig. 7. By all of these trials, we achieved a segmentation rate of more than 95 % for about 4,000 sylables.

3. Conclusions

In the present study, an efficient automatic recognition method for Tibetan characters is established. We achieved a segmentation rate of more than 99.9 % for 141,988 characters. We achieved a recognition rate of more than 99.0 % for 17,753 characters. The equipment for easy to use by Tibetan researchers is systematized. We are now trying to recognize wooden blocked Tibetan manuscripts.

Acknowledgment

Special thanks are expressed to Professor Lewis Lancaster of University of California at Berkeley for his kindness to help publishing this paper. We are also thankful to President Keisyo Tsukamoto of Housen Gakuen Junior Collage and Professor Hirofumi Isoda of Tohoku University and Professor Kazuo Hyoudo of Otani University for their advice and the presentations of Tibetan scripts.

Reference

1) Rumbaugh, J. : Object Oriented Modeling and Design, Englewood Cliffs, 1991.
2) Jacobson, I. : Object Oriented Software Engineering, Addison Wesley Publishing Company, 1992.
3) Martin, J. : Principles of Object Oriented Analysis and Design, Englewood Cliffs, 1993.
4) Kojima, M. et al. : Recognition of Similar Characters by Using Object Oriented Design Printed Tibetan Dictionary, Transaction of Information Processing Society

of Japan, Vol. 36, No. 11, pp. 2611-2621, 1995.

5) Kojima, M., Kawazoe, Y., and kimura, M. : Automatic Tibetan Script Recognition by Computer, Proceeding of the 7th Seminar of the International Association for Tibetan Studies, Graz, 1995, edited by Ernst Steinkellner, Volume 1, pp. 527-533, 1997.

<<<<< Captions >>>>>

Fig. 1 Original Printed Tibetan Text

Fig. 2 Experimental System

Fig. 3 Example of Character Segmentation

Fig. 4 Two Groups of Similar Characters

Fig. 5 Sample Screen Display of Automatic Recognition of Tibetan Characters

Fig. 6 Wooden Blocked Tibetan Manuscripts

Fig. 7 Wooden Blocked Tibetan Manuscripts with Colored Lines to segment Syllables