A CHARACTER RECOGNITION SCHEME BASED ON OBJECT ORIENTED DESIGN FOR TIBETAN BUDDHIST TEXTS

Chen-Yuan Liu [1]*, ***Huang-Cuang Lin*** [1], ***and Masami Kojima*** [2]

*[*1] Department of Information Technology and Communication, Tungnan University, No. 152, Sec. 3, PeiShen Rd., ShenKeng, Taipei, Taiwan*
Email: dori6803@ms37.hinet.net; cyliu@mail.tnit.edu.tw
[2] *Department of Information and Communication Engineering, Tohoku Institute of Technology, Sendai, Japan*

ABSTRACT

The purpose of this study is to develop a plausible method to code and compile Buddhist texts from original Tibetan scripts into Romanized form. Using GUI (Graphical User Interface) based on Object Oriented Design, a dictionary of Tibetan characters can be easily made for Buddhist literature researchers. It is hoped that a computer system capable of highly accurate character recognition will be actively used by all scholars engaged in Buddhist literature research. In the present study, an efficient automatic recognition method for Tibetan characters is established. The result of the experiments performed is that the recognition rate achieved is 99.4% for 28,954 characters.

Keywords: GUI, Tibetan, Buddhist texts, OOD, UML

1 INTRODUCTION

Buddhism is a religion that has been studied by Buddhist literature researchers all over the world from ancient times. Much of Buddhist literature was written using wooden blocked Tibetan language (Kojima et al., 1997). Some parts of this literature have already been printed for very important works.

As an example, we have used the “rGyal rabs gsal ba’i me long” published in 1993 in a volume of 250 pages. Computer recognition of these Tibetan printed texts would be eagerly welcomed by all scholars engaged in Buddhist literature studies because much printed Buddhist literature has recently been converted to this form. In this paper, we design a character recognition system for Tibetan characters by using UML (Unified Modeling Language) (Kojima et al., 2006), which is a newly developed method of OOD (Object Oriented Design) (Fujita et al., 2005; Kojima et al., 1995; Moriwaki et al., 1994). Using this GUI based on OOD, a Tibetan character dictionary can be easily made for Buddhist literature researchers (Choi et al., 2005; Mack et al., 2005).

2 EXPERIMENTS

A sample copy of the original Tibetan text is shown in Figure 1. The experimental system we used is schematically shown in Figure 2.

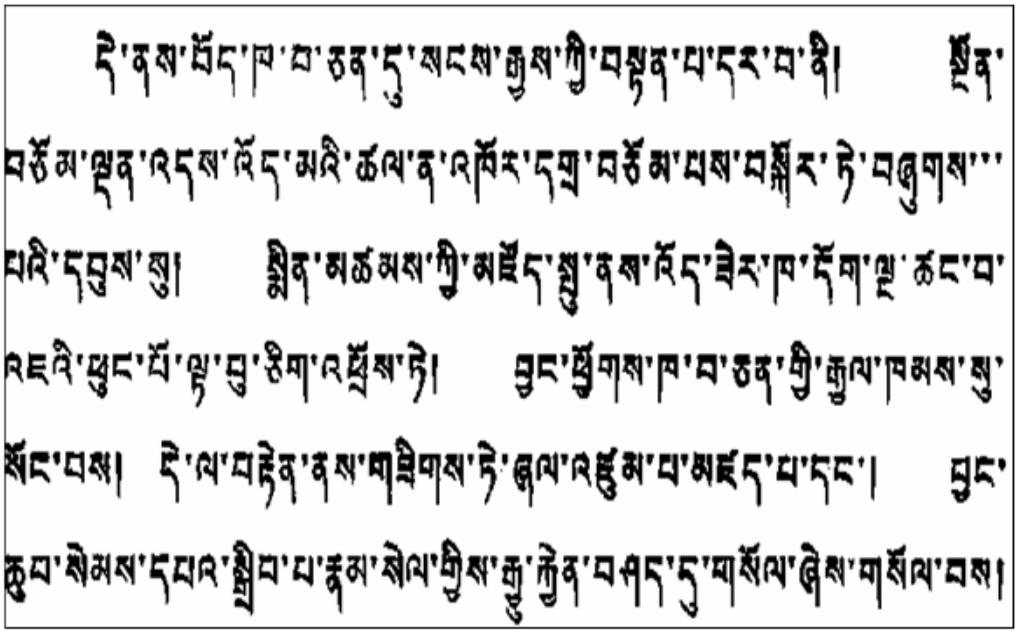

Figure 1. Sample copy of the original Tibetan text

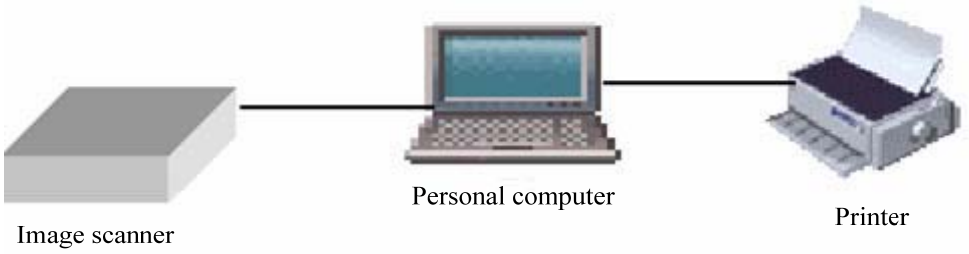

Figure 2. Experimental system

In the actual character recognition procedure, first the Tibetan texts are digitized using the image scanner with a precision of 300 dpi (digits per inch). The diagram of "use case" for a Tibetan character recognition system is shown in Figure 3. It is very important that the icon actor in this diagram is a Tibetan researcher who uses a computer. It is possible to make the Tibetan character dictionary by using GUI based on OOD. An example of line segmentation is shown in Figure 4. In this diagram, the image data digitized is shown in the left-hand insert and horizontal histograms are shown in the upper part of the right-hand insert. It is possible to read the Tibetan texts in sequence line by line, by touching the button for line segmentation shown in the bottom point of this diagram.

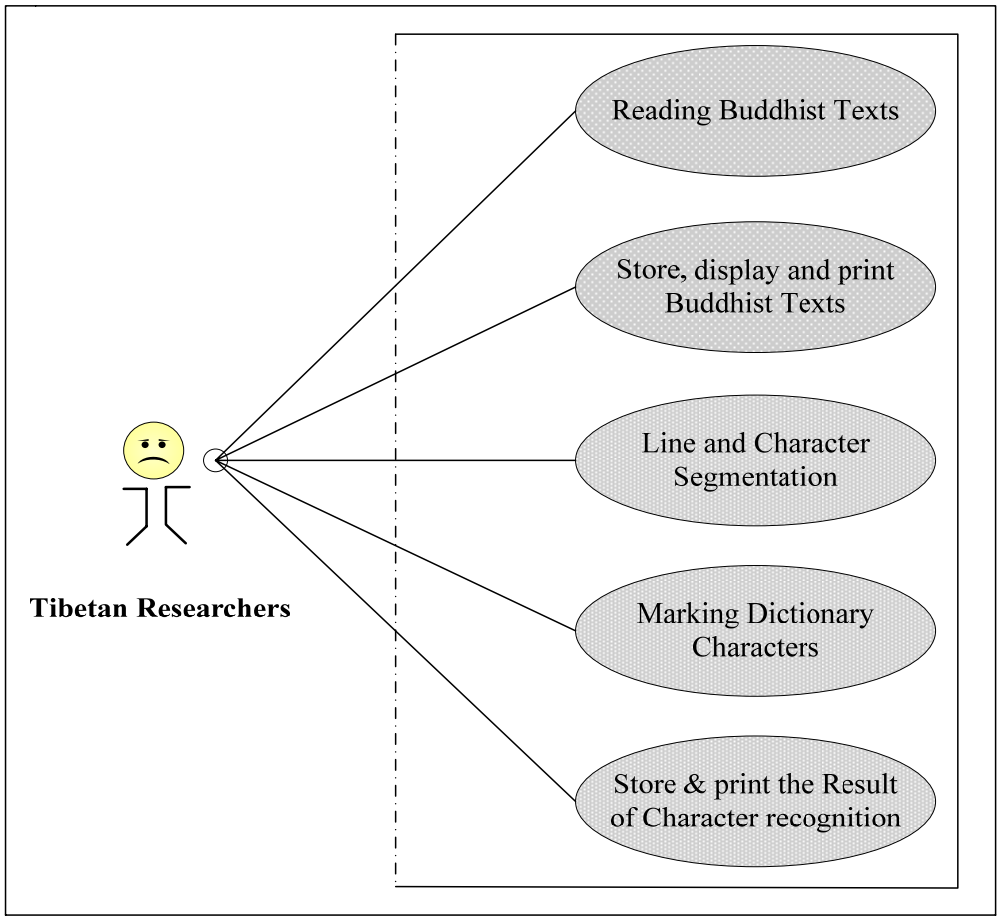

Figure 3. "Use case" modeling for Tibetan character recognition system

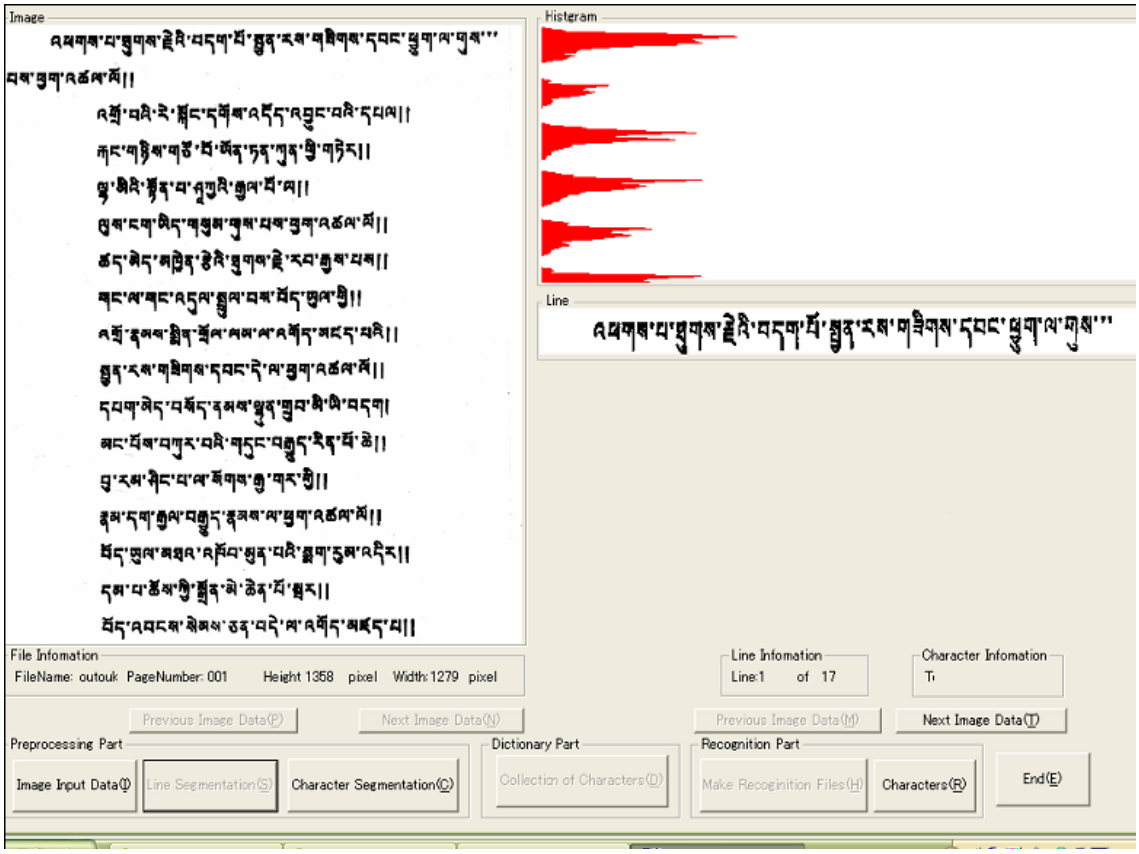

Figure 4. Example of line segmentation

Next, character segmentation is performed by touching the button for character segmentation shown also in the bottom of this diagram. An example of character segmentation is shown in Figure 5. In the character segmentation, we have segmented one syllable by extracting the character “tseg.” An arrow in Figure 5 shows the “tseg.” The diagram of collecting dictionary characters is generated in Figure 6, by touching the button for collecting dictionary characters in Figure 5. When Tibetan researchers touch the start button in the upper part of the right-hand insert of Figure 6, it is possible for them to collect dictionary characters automatically. Next, it is possible to make the dictionary characters by touching the button “making dictionary characters,” with the dictionary character name defined by Tibetan researchers. This operation is very easy for Tibetan researchers.

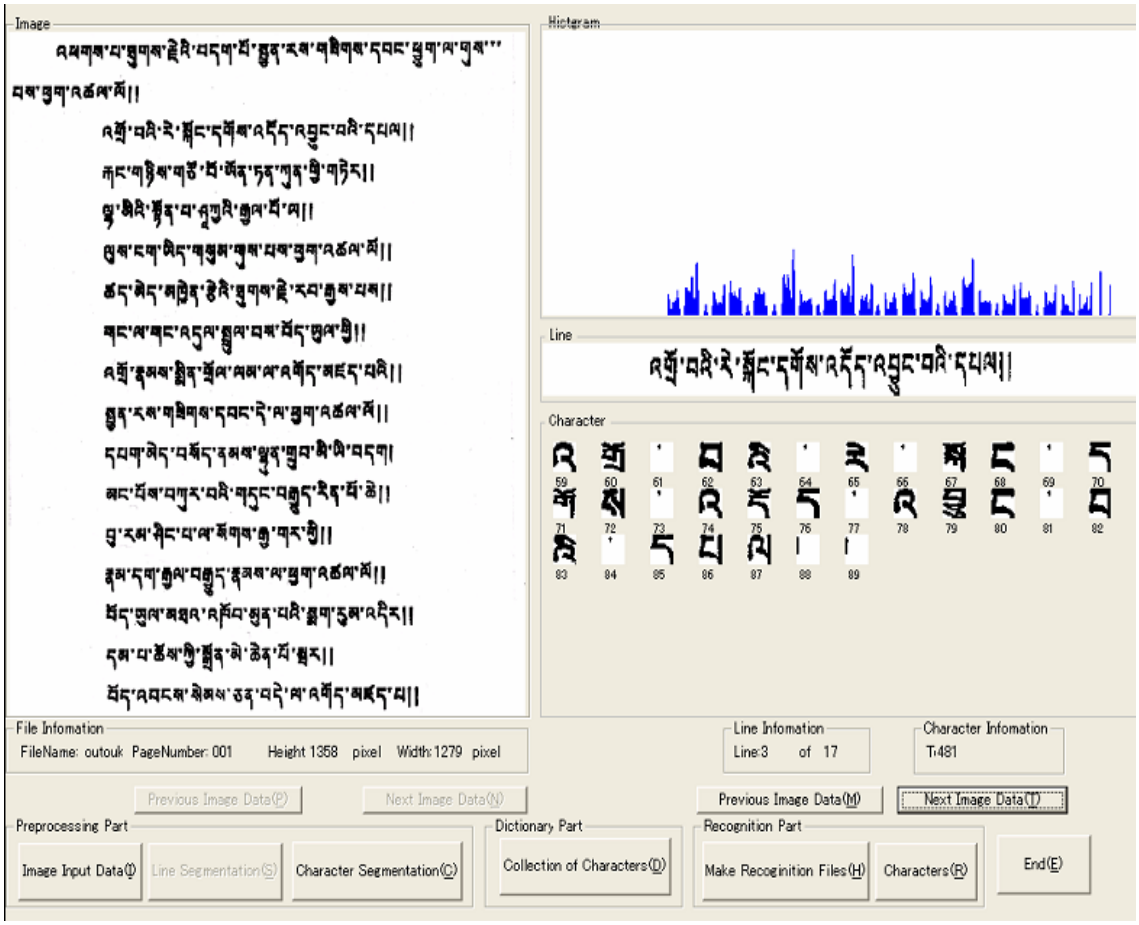

Figure 5. Example of character segmentation

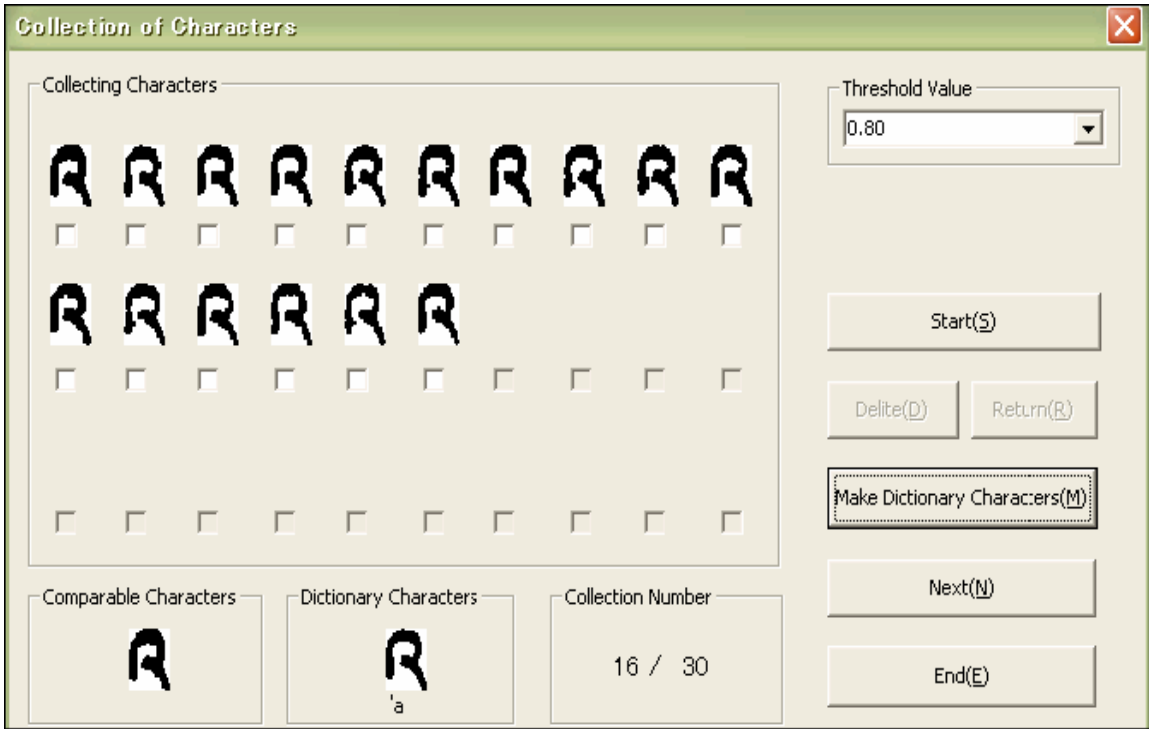

Figure 6. Example of collection of dictionary characters

Finally, it is possible automatically to recognize characters by touching the button for character recognition. These procedures are almost automatic using the GUI. A 99.9 % segmentation rate has been achieved for 141,988 characters in 250 pages of "rGyal rabs gsal ba'i me long." After obtaining the results of character recognition for 28,954 characters in 30 pages of "rGyal rabs gsal ba'i me long," we learned that mistakes mainly happen with similar characters. Group: "ba," "pa," and "pha" is shown in Figure 7 (Kojima et al., 1997; Kojima et al., 1995). OOD for these Tibetan characters is created by combining categorization and these characters, respectively. According to this additional procedure, 99.4 % recognition rate has been achieved.

The relationship between class for similar character and class for candidate characters is shown in Figure 8. Tibetan researchers without aid of computers performed all these operations.

Figure 7. Similar characters of "ba," "pa," and "pha"

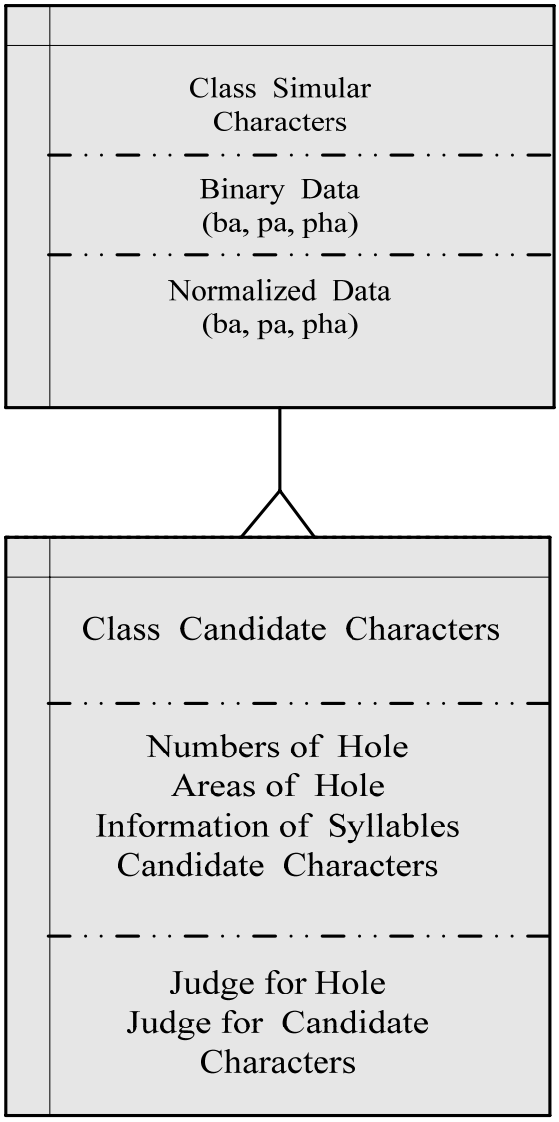

Figure 8. Relationship between class for similar characters and class for candidate characters

3 CONCLUSION

In the present study, an efficient recognition method for Tibetan characters is established. We achieved 99.4 % recognition rate for 28,954 characters in a test case. Tibetan character recognition equipment using GUI is easy to use by Tibetan researchers and has been systematized. We will next try to recognize wooden blocked Tibetan manuscripts.

4 ACKNOWLEDGEMENTS

We are thankful to Professor Kazuo Hyoudo of Otani University of Japan for his advice and the presentations of Tibetan scripts.

5 REFERENCES

Choi, Y. J., McCarthy, K. L., & McCarthy, M. J. (2005) A MATLAB graphical user interface program for tomographic viscometer data processing. *Computers and Electronics in Agriculture 47(1),* pp. 59-67.

Fujita, M., Sasaki, S., & Matsui, K. (2005) Object-oriented analysis and design of hardware/software co-designs with dependence analysis for design reuse. *Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration, IRI - 2005*, pp. 318-325.

Kojima, M., Takagi, H., Kawazoe, Y., & Kimura, M. (2006) A Convenient Recognition System of Tibetan Characters by UML. *IPSJ SIG Technical Report, 2006-CH-71,* pp. 9-14.

Kojima, M., Nunomiya, C., Kawamura, T., Akiyama, Y., & Kawazoe, Y. (1995) Recognition of Similar Characters by using Object Oriented Design Printed Tibetan Dictionary. *Transaction of Information Processing Society of Japan 36(11),* pp. 2611-2621.

Kojima, M., Kawazoe, Y., & Kimura, M. (1997) Automatic Tibetan Script Recognition by Computer. Steinkellner, E. (ed) *Proceeding of the 7th Seminar of the International Association for Tibetan Studies, Graz, 1995, Vol. 1*, pp. 527-533.

Mack, A., Kloos, U., Weltz, D., Scheib, S.G., Bottcher, H., & Seifert, V. (2005) Development of an interactive graphical user interface for therapy simulation. *Australasian Physical and Engineering Sciences in Medicine 28(4),* pp. 223-231.

Moriwaki, T. & Nunobiki, M. (1994) Object-oriented design support system for machine tools. *Journal of Intelligent Manufacturing 5(1),* pp. 47-54.