2007－CH－76　(10)
2007／9／27

A New Multipurpose Character Recognition System with Easy Operation

Masami KOJIMA[1], Chen-Yuan Llu[2], Yoshiyuki KAWAZOE[3]

Abstract

The first purpose of this study is to develop a plausible method to code and compile Buddhist texts from original Tibetan scripts into Romanized form. It is confirmed to be able to make a dictionary Tibetan characters easily for Buddhist literature researchers by using GUI (Graphical User Interface) based on Object Oriented Design[Ref. 1-4]. Second subject is to extend this new method to be applicable for Chinese character recognition with similar easy operation. Finally the system aims to establish multipurpose character recognition.

1. Introduction

Buddhism is a religion which has been studied by Buddhist literature researchers all over the world from ancient time. There are many Buddhist literatures which are written using wooden blocked Tibetan language. Some part of these literatures has already been printed for very important literatures. As an example, we have used the " rGyal rabs gsal ba' i me long " published in 1993, as a volume of 250 pages. Computer recognition of these Tibetan printed texts would be eagerly welcome by all scholars engaged in Buddhist literature studies because of many printed Buddhist literatures are recently being converted in this form. In this paper, we design a character recognition system for Tibetan characters by using UML (Unified Modeling Language), which is a newly developed method of OOD (Object Oriented Design) [Ref. 1-3]. It is confirmed to be able to make the Tibetan character dictionary easily for Buddhist literature researchers by using this GUI based on OOD [Ref. 4].

1: Professor, Depart of Information and Communication Engineering, Tohoku Institute of Technology

2: Associate Professor, Department of Information Technology and Communication, Tung-Nan University of Technology

3: Professor, Institute for Material Research, Tohoku University

2. Experiments

A sample copy of the original Tibetan texts is shown in Fig. 1. The presently used experimental system is schematically shown in Fig. 2.

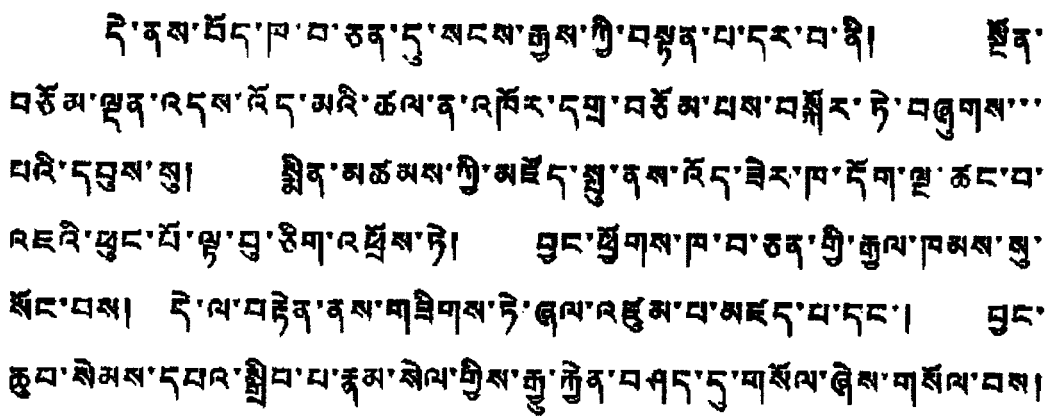

Fig. 1 Sample copy of the original Tibetan texts.

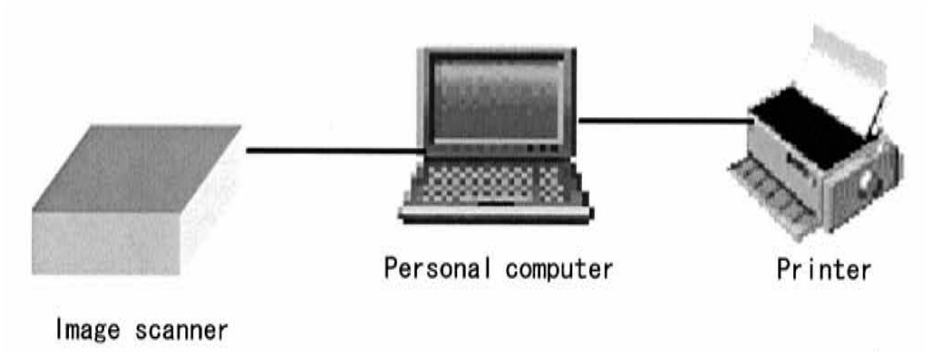

Fig. 2 Experimental system.

In the actual character recognition procedure, firstly the Tibetan texts are digitized using the image scanner with the precision of 300 dpi (digit per inch). The diagram of "use case" for Tibetan character recognition system is shown

in Fig. 3. It is very important that icon actor in this diagram is a Tibetan researcher who uses computer. It is possible to make the Tibetan character dictionary by using GUI based on OOD. An example of line segmentation is shown in Fig. 4. In this diagram, the image data digitized is shown in the left-hand inset, horizontal histograms are shown in upper part of right-hand inset. 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.

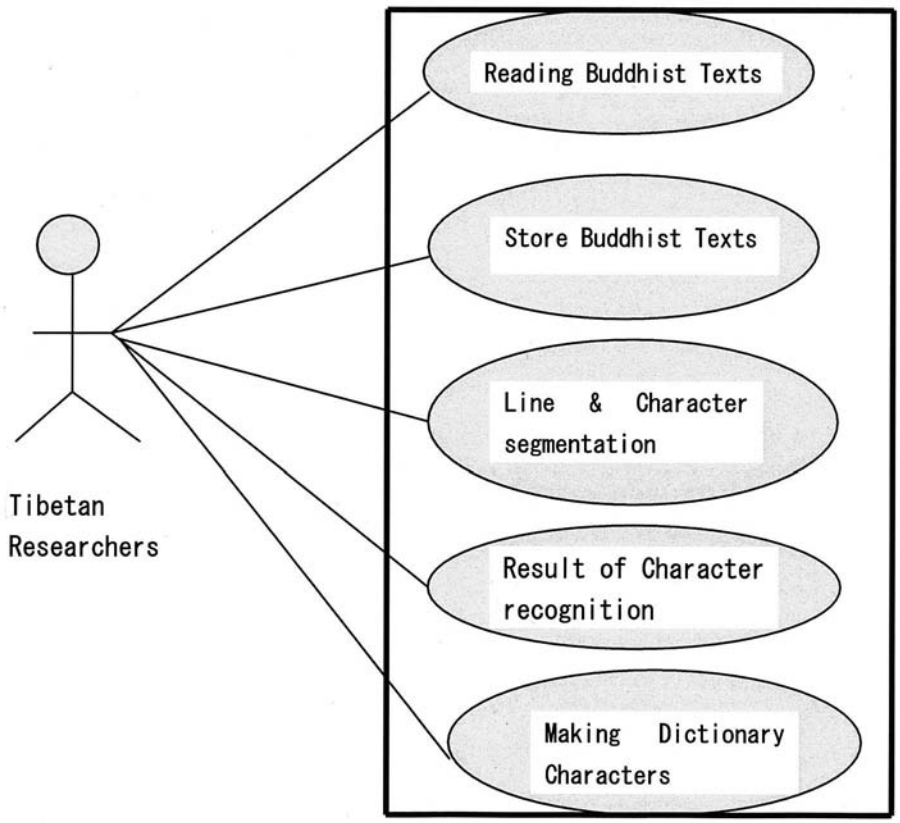

Fig. 3 "Use case" modeling for Tibetan character recognition system.

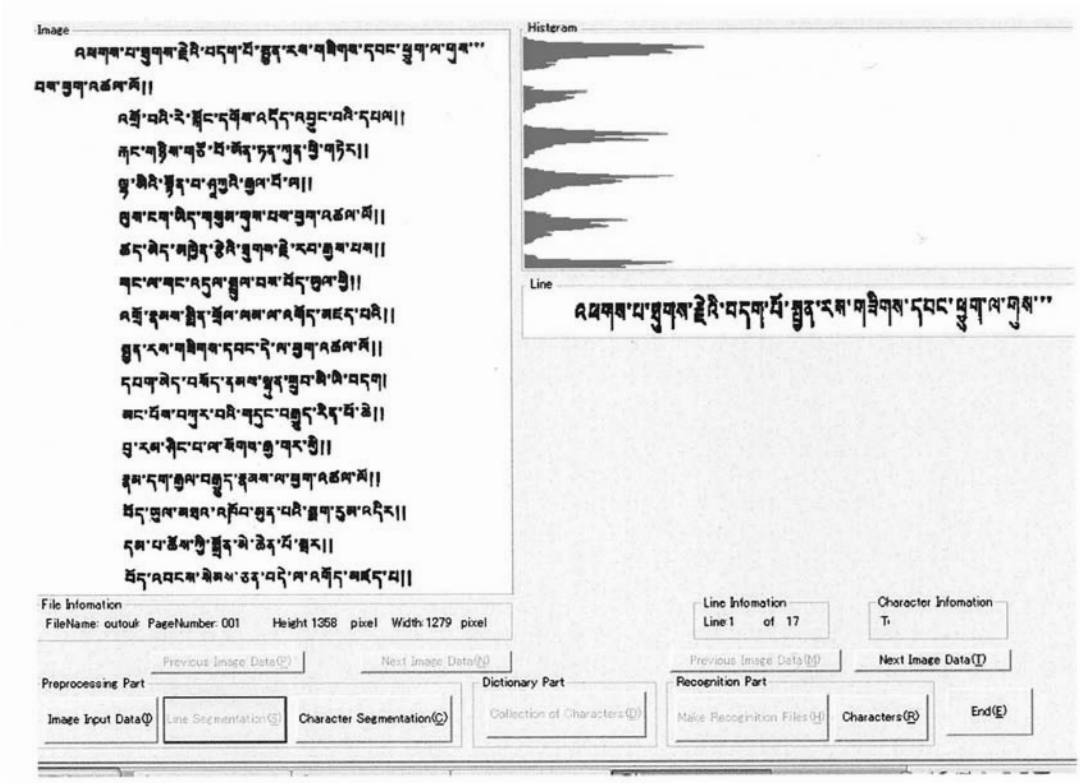

Fig. 4 Example of line segmentation.

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 Fig. 5. In the character segmentation, we have to segment one syllable and one character. When the first process of cutting out is done, two "tsheg" characters are not correctly extracted. The "tsheg" is shown by an arrow in Fig. 5. Now, it is very important in this paper to be able to make the Tibetan character dictionary easily for Buddhist literature researchers by using GUI based on OOD. The diagram of collecting dictionary characters is generated in Fig. 6, by touching the button for collecting dictionary characters in Fig. 5. When Tibetan researchers touch the start button in the upper part of right-hand inset of Fig. 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.

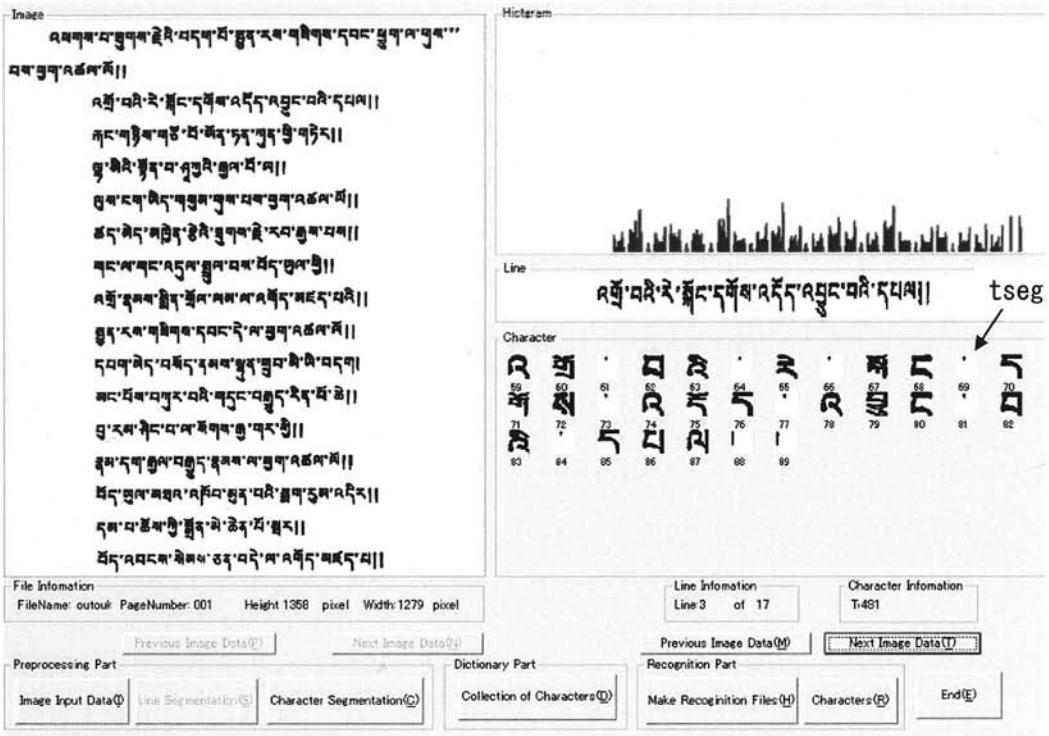

Fig. 5 Example of character segmentation.

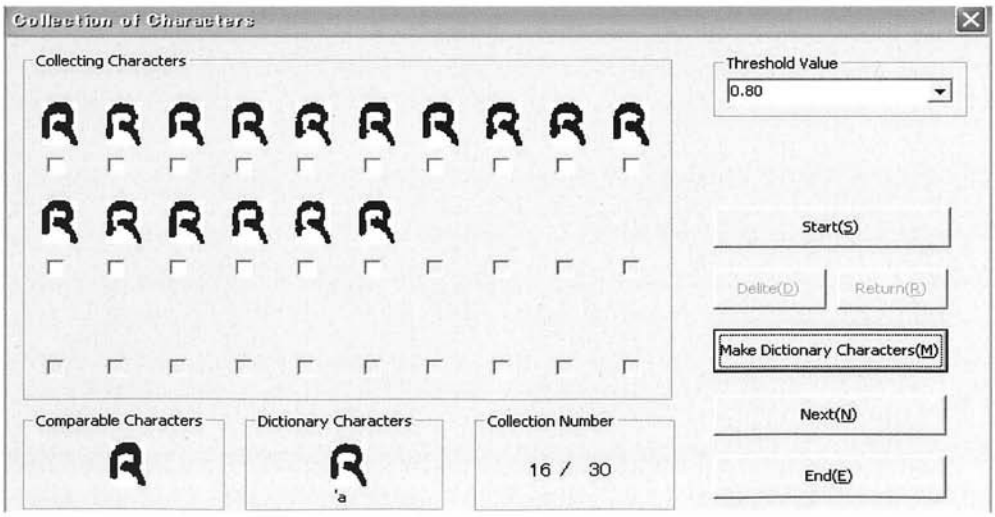

Fig. 6 Example of collection of dictionary characters.

Finally, it is possible automatically to recognize characters by touching the button for character recognition. These procedures are almost automated by using GUI based on OOD. 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, According to this additional procedure, 99.4 % recognition rate has been achieved.

The above introduced character recognition system for Tibetan characters is now extended to be applicable to recognize Chinese characters by rotating the image data of Chinese manuscript by 90 degree as shown in Fig.8.

Fig. 8 Example of the image data rotated by 90 degree.

Fig. 9 Example of Chinese character segmentation.

After line segmentation, character segmentation is performed by touching the button for character segmentation shown also in the bottom of this diagram in Fig.9. After character segmentation, the character is again rotated by 90 degree that is shown in Fig.10. A sample result of Chinese characters recognition by using this method is shown in Fig. 11.

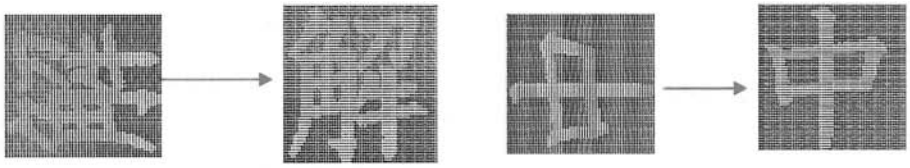

Fig.10 Example of Chinese characters rotated by 90 degree.

Fig.11 Example of Chinese character recognition.

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. The Tibetan character recognition equipment using GUI for easy to use by Tibetan researchers is systematized. We will try to recognize wooden blocked Tibetan manuscripts and publish the results. Moreover, this system is shown to be able to be applied for the recognition of Chinese characters by rotating 90 degrees the image data automatically.

Acknowledgments

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

References

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