# A Study of Character Recognition for Wooden Blocked Tibetan Manuscript

**Masami Kojima**[*1]
**Yoshiyuki Kawazoe**[*2] **& Masayuki Kimura**[*3]

**Abstract**

The purpose of the present research is to develop a technique for reading the image data of wooden blacked Tibetan manuscripts by OCR, that is, for coding and compiling Buddhist texts automatically from manuscripts into a romanized form.

In this study, the segmentation of each syllable in Tibetan manuscripts is extremely difficult and then, we suggest that it is effectively performed by the way of recognizing the respective colored lines placed at each position to be segmented in the manuscripts. However, it needs Tibetan scholar's help to draw such a colored line, even if it is a very easy work for him. Any way it is the most important that one syllable which has a peculiar attribute is segmented efficiently with the present technique and becomes a capsular object to improve the rate of character recognition.

## Introduction

We designed a total system of character recognition for all procedures starting from reading the image data of wooden blocked Tibetan manuscript to print out the resultant Roman letters and compilling a syllable in the manuscripts by the concept of Object Oriented Design [Reference 1-5]. The most difficult problem in this study is the segmentation of one syllable in Tibetan manuscripts, [Reference 6]. Therefore, we would like to consider how to find the first part of the character of one syllable segmentation. We have started to examine some fundamental problems. Based on our experiment, it is clear that we have to consider character segmentation first. It is effectively performed by using the recognition of colored line to segment one syllable, and we now think that it is more practical than other methods considered up to the present. The most important result is that one syllable with peculiar attribute is segmented efficiently with the present method and we think that it become a capsular object to improve rate of character recognition.

[*1] Associate Professor, Tohoku Institute of Technology
[*2] Professor, Tohoku University
[*3] Vice President, Japan Advanced Institute of Science and Technology, Hokuriku

**Experiment**

A sample copy of wooden blocked Tibetan manuscripts is shown in Fig.1, and presently used experimental system is shown in Fig.2

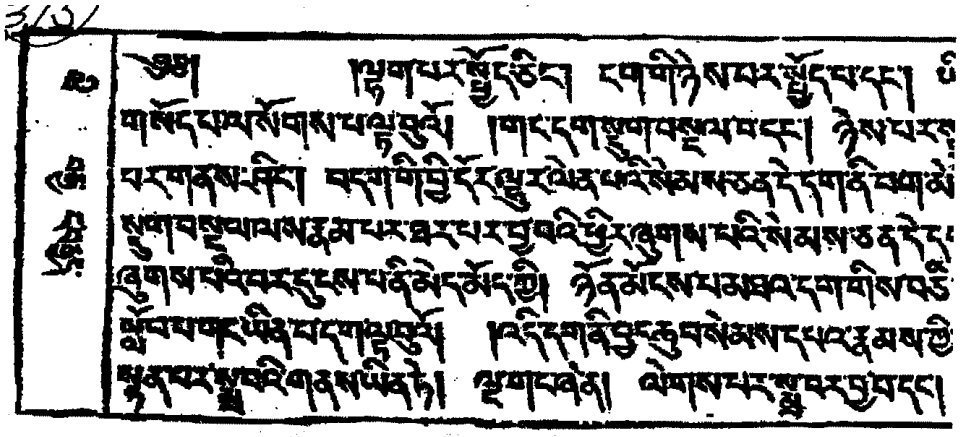

Fig.1 Part of wooden blocked Tibetan manuscripts
( Printed Derge )

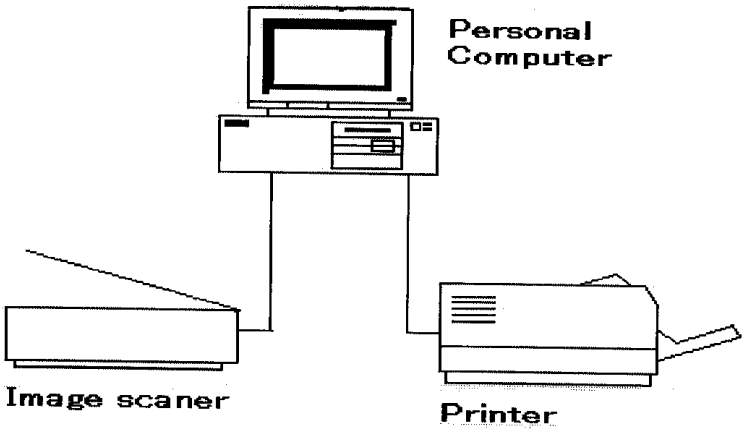


Fig.2 Experimental System

In the actual character recognition procedure, firstly Tibetan scholars draw colored lines on the copy of Tibetan manuscripts to segment syllables, and then the image data is digitized. The presently used wooden blocked Tibetan manuscripts with marked colored lines are shown in Fig.3. A typical example of horizontal histogram of Tibetan manuscripts is shown in Fig. 4. In this figure, we can observe the points of line segmentation which are consist of the start point given by (pl+p2)/2 and the end point given by 9(p2+p3)/10, where the constants are determined by experiment.

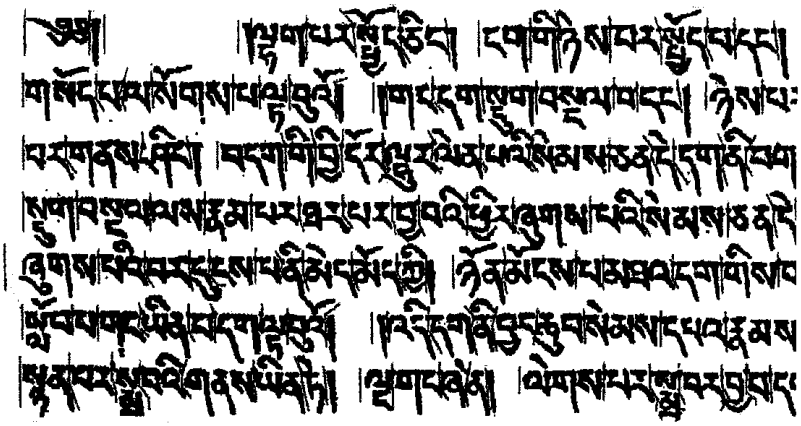

Fig. 3 Wooden blocked Tibetan manuscripts with colored lines( thin vertical lines ) to segment syllables

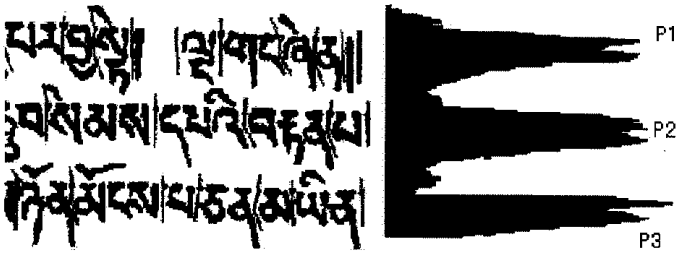


Fig. 4 Typical example of horizontal histograms of Tibetan manuscripts

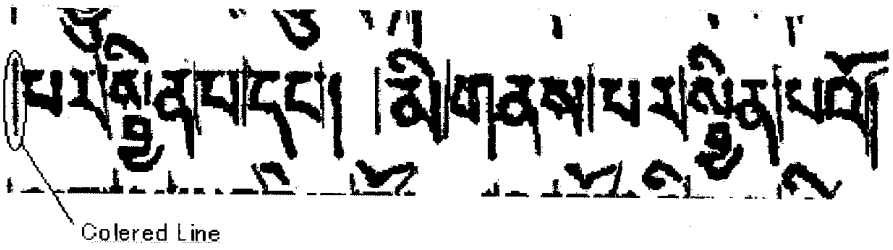


Fig. 5 Line segmentation

Figure 5 shows a sample of the line segmentation. In this case, it is necessary to take care the colored line drawn on the black area of character. It is not possible to recognize the colored line by computer, and we have to restore a break colored line, which is shown in Fig.6. Therefore, two colored lines are used for one syllable. In one of colored lines, it is drawn from the bottom of the colored line to the edge line of line segmentation as water falling. If it is not necessary to add the second colored line, it can be automatically drawn. Upper end of the colored line is the same except for the reverse of drawing direction. It is applied just like the same to another colored line. We treat image data outside of two colored lines as noise. We also treat image data separate from one syllable as noise.

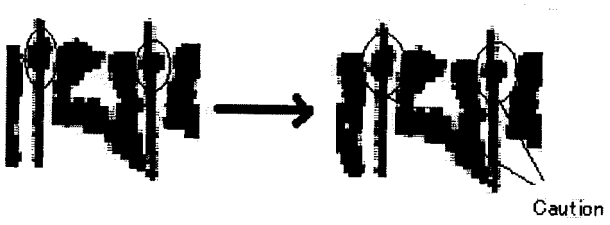


Fig. 6 Restoration of colored line

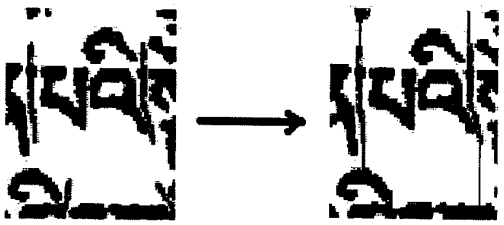

Fig.7 Water falling method （1）

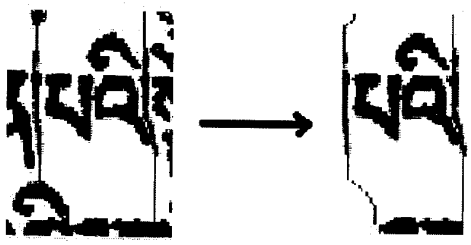

Fig.8 Water falling method （2）
(Delete the image out of the colored lines)

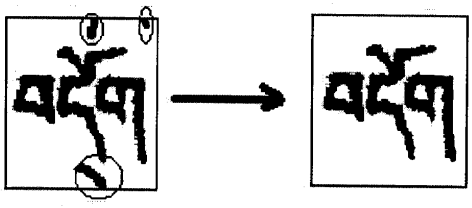

Fig. 9 Segmentation of one syllable
( Delete noise automatically )

These procedure are shown in Fig. 7,8,9. Now, successful segmentation rate of Tibetan manuscripts is achieved 95 % for 4,ooo syllables.

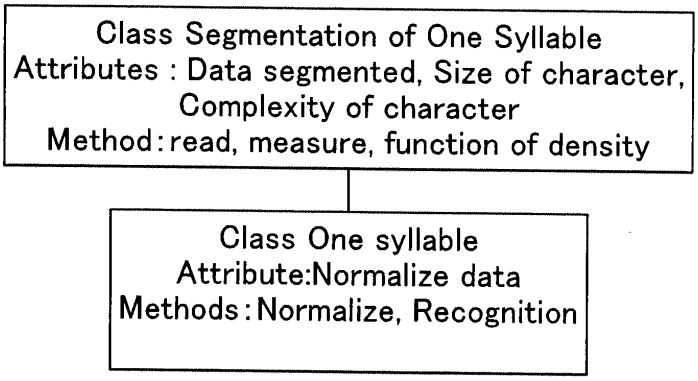


Fig. 1 0 Class chart of recognition Tibetan wooden blocked manuscripts

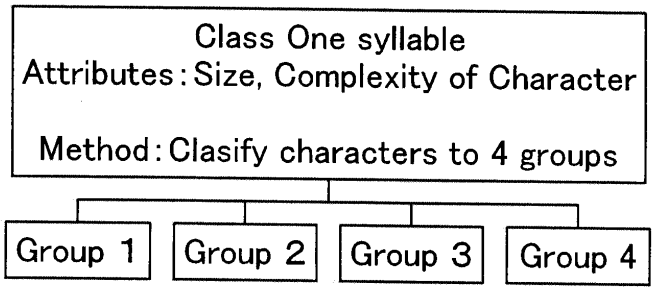


Fig.11 Class chart of classification of characters to 4 groups

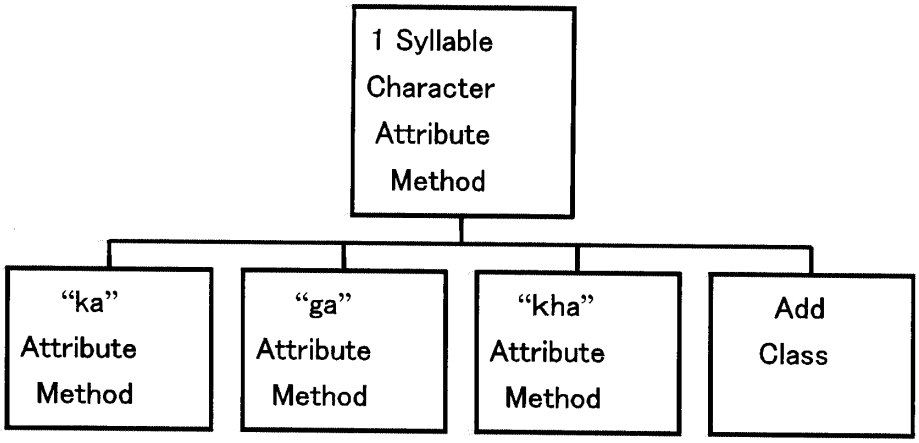


Fig.12 Class chart of one syllable

Next, we discuss the present recognition method of wooden blocked Tibetan manuscripts. We have the conclusion that it is necessary to keep should inherit the attribute of one syllable segmented until one syllable recognition is completed by using Object Oriented Design (OOD). The class chart is shown in Fig.10. There are two features for Tibetan characters in segmentation of one syllable. One is the size of one syllable, especially wide size. Because one syllable consists of four groups starting from one character to four characters. Another is the complexity of one syllable. The class chart of classifying one syllable to four groups are shown in Fig.11. The complexity of one syllable is calculated by using a function of horizontal density. Each syllable has a specific feature, which is named as attribute. It becomes a capsular object and the class chart is shown in Fig. 12.

**Conclusion**

For wooden blocked Tibetan manuscripts, we achieved a segmentation rate of 95 % for 4,000 syllables by using colored lines marked by human to segment syllable. The work of drawing colored lines is easier than to input data using keyboard. Therefore; this hybrid model is chosen. Of cause, it is hoped that recognition procedure is fully automatic. At least it is useful to perform character recognition using Object Oriented Design (OOD). We would like to emphasize that the recognition rate of wooden blocked Tibetan manuscripts improved by making a capsule object for one syllable.

**Acknowledgments**

Special thanks are expressed to Professor Lewis Lancaster of University of California at Berkeley for his continuous interest in our work. We are thankful to Dr. Simon Lin of Director of Computing Center, Academia Sinica for his kindness, and also thankful to President Keisyo Tsukamoto of Housen Gakuen Junior Collage and Professor Hirofumi Isoda of Tohoku University for their advice and presentation of Tibetan manuscripts.

**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. Principle of Object Oriented Analysis and Design, Englewood Cliffs, 1993.
4. Kojima, M. ct 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. 2611-2622, 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 I, pp. 527-533, 1997.
6. Kojima, M., Kawazoe, Y., and Kimura, M., Automatic Recognition of Tibetan Buddhist Text by Computer, 1999 EBTI, ECAI, SEER & PNC Joint Meeting, January 18-21, 1999 Taipei, pp. 387-393.