音節構造解析による活字チベット文字認識の高速化

山下康明[†1]　小島正美[†2]　木村正行[†3]

あらまし

本論文は、活字で出版されている重要なチベット仏典の自動認識を目的としている。実験に用いたのは中国で出版された「 西蔵王統記 」の30ページ分のデータである。これらのチベット文献は大変重要であり、多くのチベット学者からチベット仏典を認識して表音文字であるローマ字への高速変換処理が永年に亘って望まれてきた。本論文では、認識の高速化を進めるために、チベット文字の構造情報に着目してチベット文字をいくつかのクラスに分類し、その各クラス毎の辞書を用いた認識方法を考案した。その結果、認識速度が従来のパターンマッチングよりも２ないし３倍程度速く、かつ認識精度がほぼ同等な高速化が実現できた。なお、本論文で実験に用いた文献の18166文字のうち、１文字の切り出し率は99.0％で、正しく切り出しされた17985文字に対する認識率は99.0％と、実用レベルの精度を得ており、その活用が期待されている。

キーワード：高速文字認識、文字の構造解析、チベット活字文献

High-speed Recognition of Printed Tibetan Characters through Analysis of Tibetan Syllable Structures

Yasuaki Yamashita[†1], Masami Kojima[†2] and Masayuki Kimura[†3]

Abstract

Our research has originated from the desire to facilitate the work in coding/compiling Buddhist texts from their original Tibetan scripts into romanized form to encourage Buddhist literature studies by using the present-day computer assistance. As an example, we have used a 30 pages volume " rGyal rabs gsal ba'i me long " published 1993, to perform fast character recognition experiments based on the analysis of Tibetan syllable structures. It is hoped that a computer system capable of fast character recognition will be used actively by all scholars engaged in Buddhist literature studies.

The result of the experiments performed is that the segmentation rate achieved is about 99.0% for 18166 characters, and the recognition rate achieved is about 99.0 % for 17985 characters with such a speed 2 or 3 times faster than the ordinary pattern matching.

Keyword : High-speed Character Recognition, Analyzing the Structure of Character, Printed Tibetan Characters

---

†1 株式会社日立制作所計測器事業部
Instrument Div., Hitachi, Ltd.
†2 東北工業大学・通信工学科
Depart of Electrical Communication, Tohoku Institute of Technology
†3 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology, Hokuriku

1．はじめに

現在、チベット語で記載された重要な仏典文献の一部は既に活字で出版されており、これらの文献をコンピュータで自動認識して欲しいという要望がチベット学者から起きている[1)]。小島らは誤認識の多いチベット類似文字に対して、類似文字の特徴により４グループに分類し、それぞれのグループ毎に認識メソッドを固有に持たせることにより認識率の向上を目指した。その手法により単に重ね合わせによる認識率よりも改善され99％台の認識率が得られることを報告している[2)]。しかし、重ね合わせ法を主体とした文字認識法は、辞書文字の数が増えると、認識速度が遅くなるという欠点がある。チベット学者らが研究室でコンピュータを用いてチベット仏典文献から表音記号であるローマ字へ変換する場合、使用する多くのコンピュータはパーソナルコンピュータである。大量の文献をコンピュータで可読化しようとした場合、その処理速度は、研究者にとって無視できない問題となってきている。

そこで本論文では、チベット文字の音節構造（図４）に着目し、１音節中の文字数および文字の構造情報に基づいて、先ず１音節パターンの大まかな分類を行って認識対象となる候補文字のクラスを絞り、音節中の各文字の認識についてはそれが属するクラスの辞書を用いたパターンマッチングを行っている。本方法は、音節中の文字認識を先に行い、ついで音節パターンを識別する従来の方法[3)]に比して大幅な高速化が期待される。

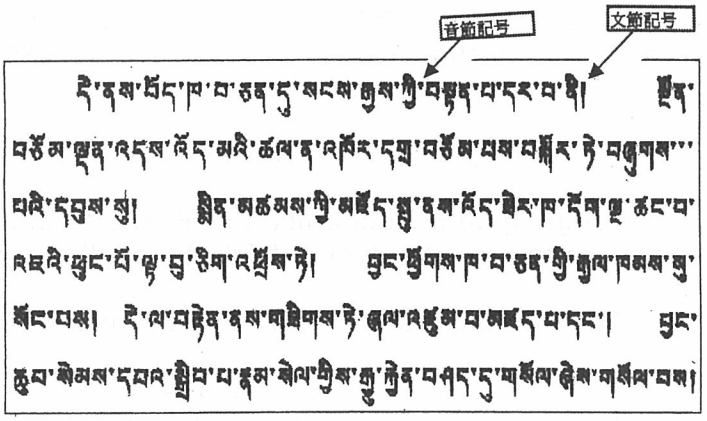


図１　認識実験に用いたチベット活字文献のパラグラフの１例

Fig. 1　An example of paragraphs in the Tibetan texts used in the recognition experiment.

認識実験では、チベット活字文献（「 西蔵王統記 」民族出版1993 ）の30頁分を認識対象データとして使用した。その一部を図１に示す。チベット文字は次章で述べるように、１音節単位で文字認識をしなければならない。１音節区切り記号および文単位となる文節記号を図１にそれぞれ矢印で示している。

2．チベット文字[4、5)]

チベット文字は図２に示すような１音節構造で示される。１音節文字の構成は基本30子音と４母音、93種の重層字からなる。基本30子音と４母音を図３に示す。基本子音を縦方向に重ね合わせた文字が重層字である。図２は、構成要素が最大な場合のチベット１音節文字の構成を示したもので、基字は基本30子音または重層字から構成され、基字以外の子音は基本30子音から構成される。前接字は「 ga 」、「 da 」、「 ba 」、「ma 」、「 'a 」の５文字、後接字は「 ga 」、「 da 」、「 ba 」、「 ma 」、「 'a 」、「nga 」、「 na 」、「 ra 」、「 la 」、「 sa 」の10文字、再後接字は「 da 」、「 sa 」の２文字である。接字は10種となる。母音記号は、基字の上部に母音記号「 i 」、「 e 」、「 o 」、が付くか、または基字の下部に母音記号「 u 」が付く。すなわち、上部かまたは

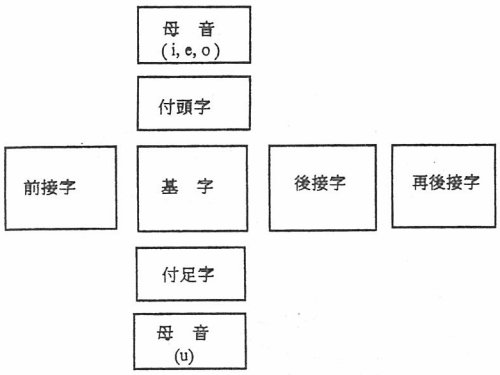


図２チベット文字で表わされたチベット語の音節構造

Fig.2 Structure of Tibetan syllables expressed with Tibetan script.

下部に母音記号が付いた文字は、基字の後に出現する「’ｉ」という文字を除いて総て基字となる。そのため、本論文では、便宜上、「’ｉ」という文字は母音が付いた文字であるが、基字にはならない文字なので文字認識する上で、接字グループに含めている。１音節中に存在する文字（最大４個）のどの文字にも母音がついていない場合、図４に示す７つの分節パターンを用いて、基字と接字を分類し、基字と判定された文字についてはチベット文字に内在している母音「ａ」を付けて読む。母音記号「ａ」は単体でしか存在しないため、チベット文法書では「ａ」は基本30子音としている。これらの音節単位を組み合わせて１文節となる。

３．チベット文字の分類

音節構造パターンに分類するためにチベット文字の特徴抽出を行い、チベット文字の分類を図５に示すように行う。図５の大きな楕円で囲んだ２つの部分の分類について述べる。

３-１）母音記号の有無による分類

図５の（ａ）の部分では、母音記号の有無の判定は、母音記号が存在する位置によって行う。チベット文献を行単位で切り出した時、水平方向のヒストグラムがピークとなる位置MHL（Main Horizontal Line と呼ぶ）より上部にある母音を識別し、分類する。著

図３チベット活字基本30子音および4母音（ 表音はワイリー ）

Fig. 3 Basic 30 consonants and 4 vowels of Tibetan scripts. (Wylie type)

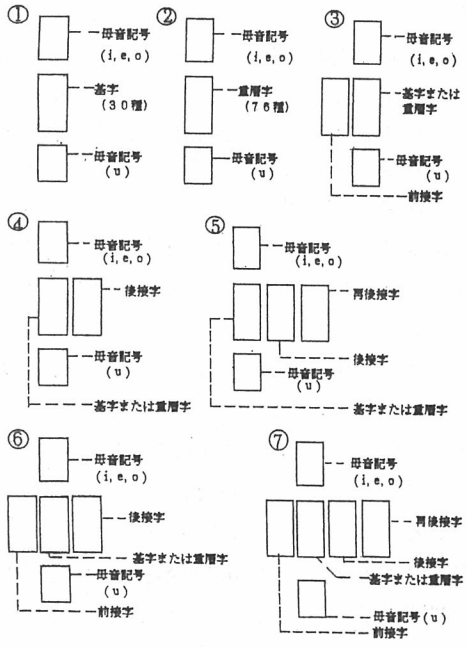


図４ チベット語の７つの分節構造パターン

Fig. 4 7 patterns in the structure of Tibetan syllables.

者らは、チベット文献を行単位で水平方向のヒストグラムをとり、MHLより上部にある母音を分割して認識する方法を用いる[6]。本論文では、統計的手法により、１音節単位にMHLの位置を決定し、母音の分割精度を向上させている。１音節単位での母音記号分割方法を図６に示す。１音節中の文字数を n、その各文字の上端位置を y_i とし、その平均値をμとすると、

$$\mu = \sum_{i=1}^{n} y_i / n \quad \cdots\cdots (1)$$

で表わされる。ここで、しきい値Θを次のように設定する。

母音有りクラス：if $y_i < \mu - \Theta$

母音なしクラス：otherwise

この方法によると、図６の音節は、文字数n=３、第２番目の文字が母音有り文字であることから、音節パターンは６となり、Char1は接字、Char2は母音有り文字、Char3は接字のそれぞれのクラスに分類される。基本30子音の中でMHL上部にヒゲがある文字（ tsa,tsha, dza ）の３字種は、母音有りクラスに含まれるに分類している。上の３字種は全て基字となる子音である。前章で述べたように、母音有り文字は基字となるので、識別の便利さからこれら３字種を母音有りクラスに入れている。

分類した結果を表１に示す。しきい値は実験的に求め、Θ＝4 dotで行い、その結果、全体の分類率は99.5％である。誤分類された文字はほとんど潰れや掠れ、または音節記号や文節記号から生じているノイズによるものである。

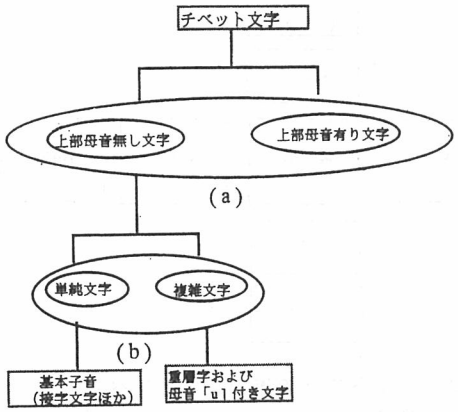


図５　チベット文字の分類図

Fig. 5　Classification chart for Tibetan characters.

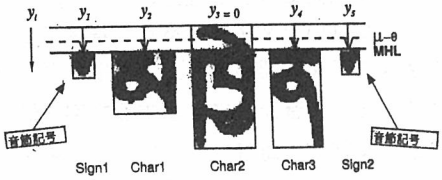


図６ MHL上部に位置する母音記号の識別

Fig. 6　Discrimination of vowels located in the upper part of MHL.

表１上部母音有り文字と上部母音無し文字の分類

Table 1　Classification between characters with　upper vowels and the others.

文字の種類 / 分類	上部母音有り文字	上部母音無し文字
正分類	4762	13223
誤分類	32	59
合　計	4794	13282
分類率	99.3%	99.6%

3-２）単純文字および複雑文字の分類

MHL上部に母音記号が存在しない文字のうち、図５に示すように基本子音のクラスを「単純文字」と定義し、重層字と母音記号「ｕ」の付随した文字からなるクラスを「複雑文字 」と定義する。重層字は基本子音を縦方向に組み合わせて構成される文字で、文字の大きさが基本子音に比較して縦長になる特徴がある。また、基本子音の下部に母音記号「ｕ」の付いた文字も同様な傾向となる。それ故、文字の縦横比（Ratio）と文字サイズ（Size）を、それぞれ、特徴量とする。 また、文字を縦方

向にスキャンしたときの文字（黒画素）領域との交差回数（Cym）が、「単純文字」は「複雑文字」より少ない傾向があるので、これを特徴量として用いる。これら３つの特徴量の定義式とその抽出方法を図７に示す。また１例として、単純文字「 nga 」、「 da 」と複雑文字「 khyu 」について、これら３つの特徴量の値を図８に示す。

さて、「 単純文字 」か「 複雑文字 」かの二つのクラスを判別する代表的な方法として、Fisherの線形判別関数[7]を用いる方法と、マハラノビスの距離[8]を用いる方法がある。この２つの方法を用いて、「単純文字」27字種10830字、「複雑文字 」127字種2586字の場合について判別（分類）実験を行った。その結果、線形判別関数と比較して、マハラノビス距離による方法は出現頻度の多い「単純文字クラス」の誤分類が多く、さらに計算量が大きくなる点で高速処理には不向きであることが判明した。それ故、以下では線形判別関数のみによる分類について述べることにする。

「 単純文字 」と「 複雑文字 」の判別する特徴量Cym、Ratio、Sizeを１個づつ個々に使用するよりは、２つ以上の特徴量を組み合わせて使用した方が判別の精度があがることが実験的に確認された。そこで２つ以上の特徴量で判別分析を行った。「 単純文字クラス 」と「 複雑文字クラス 」を判別するための関数としてFisherの線形判別関数を用いると、求める線形判別関数Zは

$$Z = a_1 X_1 + a_2 X_2 \quad \cdots\cdots (2)$$

となる。但し、２つの特徴量はそれぞれX_1、X_2で表す。「 単純文字クラス 」と「 複雑文字クラス 」のクラス間での分散と各々のクラス内での分散の F比（フィッシャー比 ）を最大となるように係数a_1、a_2を決定し、線形判別関数Zを求める。特徴量の組み合わせとして、CymとRatioおよびCym とSizeを用いた時の各文字の分布をそれぞれ図9及び図10示す。また、分類した結果を表２に示す。表２から、２つの特徴量の組み合わせでは、特徴量CymとSizeを用いたときが誤分類が小さく、

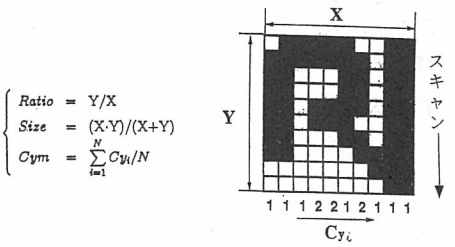
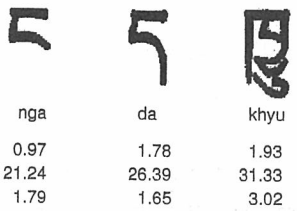


図7　チベット文字の３つの特徴量の定義式とその抽出法Cyiは横軸の座標i（ i = 1～N ）における交差回数

Fig. 7　Equations to define three features of Tibetan scripts and the ways to extract them. Cyi denotes the number of times to cross the elements of a character in i-th coordinate of the transversal axis.



	nga	da	khyu
Ratio:	0.97	1.78	1.93
Size:	21.24	26.39	31.33
Cym:	1.79	1.65	3.02

図8　基本子音（nga, da ）と重層字（khyu ）についての３つの特徴量(ratio, size and cym)の値の例

Fig. 8　Examples of values of three features about　two basic consonants (nga, da) and a combination　character (khyu).

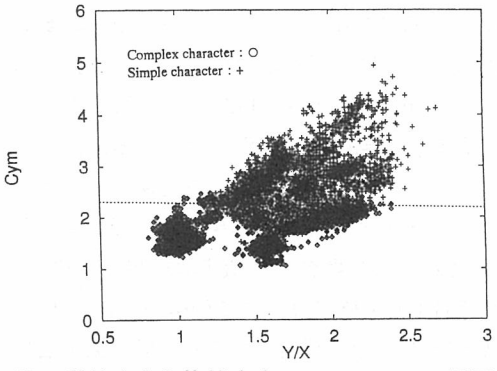


図9 単純文字と複雑文字のY/X - Cym 平面上の分布図

Fig. 9　Distribution map of simple characters and complex characters　in the Y/X - Cym plane.

3つの特徴量を用いたときと同じ程度の分類率であることが分かる。

線形判別関数を用いた「単純文字クラス」と「複雑文字クラス」との分類のクローズおよびオープン実験の結果を表3に示す。表4には線形判別関数による誤分類文字を示す。括弧内の数値は誤分類された文字数である。

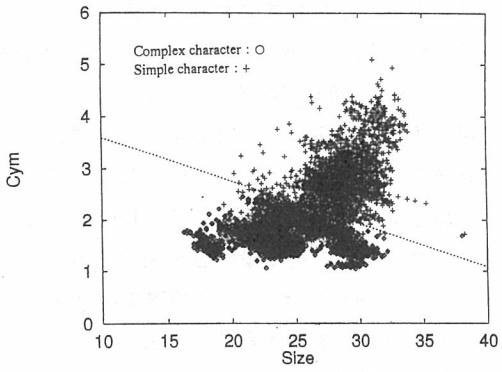


図10　単純文字と複雑文字のSize - Cym 平面上の分布図

Fig. 10 Distribution map of simple characters and complex characters in the Size - Cym plane.

4．認識実験

認識対象となる文献データをイメージスキャナで取り込み、文字切り出しを行った。文字切り出し率（精度）は99.0%で、正しく切り出された332字種、17985文字について認識実験を行った。

入力文字パターン$\mathbf{x}$と標準パターン$\mathbf{x}^{(j)}$間の距離計算は式（3）の単純類似度$s^{(j)}$ 9）を用いた。

$$s^{(j)}=\cos\Theta^{(j)}=(\mathbf{x},\mathbf{x}^{(j)})/(\|\mathbf{x}\|\cdot\|\mathbf{x}^{(j)}\|)\cdots(3)$$

$$(0\leqq s^{(j)}\leqq 1)$$

ここで、$s^{(j)}$はカテゴリjの$\mathbf{x}^{(j)}$と入力$\mathbf{x}$との間の単純類似度を表わす。入力文字パターン$\mathbf{x}$は、32×32次元の2階調のドットパターンとし、同じ次元数で作成されている標準パターン$\mathbf{x}^{(j)}$（辞書）との類似度計算を行う。辞書の作成は、実験データ30頁分の中に出現した332字種に対して行った。辞書作成に用いた文字数は17895字、字種毎に文字の大きさにバラツキがあるため、32×32次元に正規化処理を行った。文字認識のフローチャート図を図11に示す。図11において、(1)母音の有無のみの判別を用いた場合、（2）複雑文字と単純文字の判別のみを用いた場合について判別の特徴量を2個と3個とした各々の場合を考えて、それぞれの判別結果に応じてクラス別けされた各クラスの辞書を用いて単純類似度を計算する。これらの各場合の認識実験結果と、従来の全数整合による認識実験結果を表5に示す。表6には、図11の最終ブロックでの処理、即ち入力パターンの第1候補の類似度が実験的に決定されたしきい値を超えない場合、つまり実験結果の信頼性に問題があると見なされた場合には、他のクラスの辞書と再び距離計算を行う処理を表5の全数整合以外の各処理に追加した場合の認識結果を示す。その結果、計算量は、他のクラスとの認識率を追加しない場合より、母音の有無の分類法では、1．1倍、単純文字と複雑文字との分類法では1．2倍増加しているが、認識率はいづれも向上し、母音の

表2 線形識別関数による単純文字と複雑文字の誤分類数（括弧内は判別誤り率）

Table 2 Number of misclassified characters between simple characters and complex characters by using a linear discriminant function, and a number in a parenthesis denotes an error rate.

特徴量＼文字の種類	単純文字	複雑文字
Cym, Ratio	99(0.9%)	207(8.0%)
Cym, Size	36(0.3%)	163(6.3%)
Cym, Ratio, Size	34(0.3%)	157(6.1%)

有無の分類では、全数整合とほぼ同一となった。

5．まとめ

実験の結果、正しく切り出しされた17985文字中、チベット文字の特徴である図5に示した分類法により、全数マッチングとほぼ同じ99%の文字認識率を保ちながら、認識速度を全数整合法に比べて2～3倍にすることができた。また、チベット文字の構造上の特徴により「単純文字」、「複雑文字」の識別は、文字の縦方向交差回数を特徴量とすることが有効であ

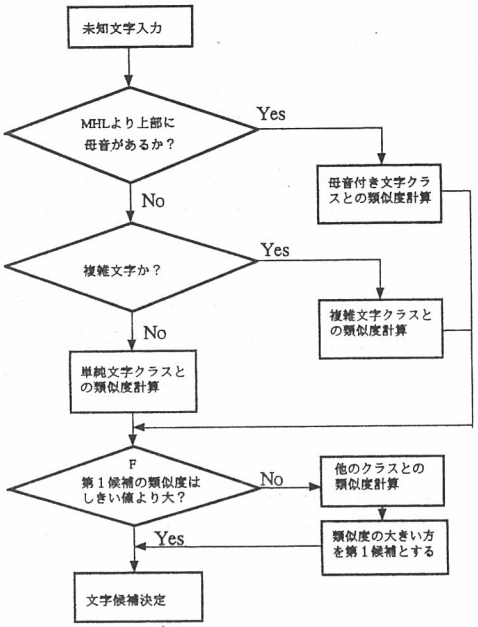


図11　チベット文字認識のフローチャート

Fig. 11　Flowchart of the process to recognize Tibetan scripts.

表3 線形識別関数による単純文字と複雑文字の判別誤り率（クローズ、オープン実験）

特徴量＼文字の種類	単純文字		複雑文字	
	Close	Open	Close	Open
Cym, Ratio	0.9%	0.9%	8.0%	8.0%
Cym, Size	0.3%	0.5%	6.3%	6.6%
Cym, Ratio, Size	0.3%	0.5%	6.1%	6.2%

表4 チベット文字の2及び3特徴量を用いた線形判別関数の各々による誤分類文字（但し、括弧内の数字は対応する 文字の誤り率を表わす。）

Table 4 Characters misclassified with those two linear discriminant functions which use two and three features, respectively and a number in a parenthesis denotes an error rate of the corresponding character.

特徴量数	文字の種類	誤分類
2特徴量 Cym, Size	単純文字	全27字種中3字種 zha (32/37), ha (1/7), na (3/1074)
	複雑文字	全127字種中36字種 shwa (9/9), rka (4/4), dwa (2/2), rnga (2/2), gra (20/23), rma (10/14), rtsa (6/13), zla (7/16), kya (13/37), sra (21/69), etc.
3特徴量 Cym, Ratio, Size	単純文字	全27字種中3字種 zha (30/37), ja (1/4), na (1/1074)
	複雑文字	全127字種中37字種 shwa (9/9), rka (4/4), dwa (2/2), rnga (2/2), gra (17/23), rma (8/14), rtsa (4/13), zla (7/16), kya (11/37), sra (21/69), etc.

表5　4つの各認識プロセスで得られた認識結果の比較

Table 5　Comparison of recognition results obtained by the 4 types of recognition processes, respectively.

実験項目＼認識方法	全数整合	母音の有無	2特徴量判別分析	3特徴量判別分析
誤認識文字数	141	310	405	398
認識率	99.2%	98.3%	97.7%	97.7%
計算回数／1入力文字	332.0	137.3	94.2	94.9
認識速度改善率	1.00	2.39	3.34	3.29

表6　従来の全数パターンマッチングプロセス及び表５の従来のタイプ以外の３つのタイプの各々に図11のフローチャートの最終段階Fを追加した新しい３つのタイプの認識プロセスによって得られた認識結果の比較

Table 6　Comparison of recognition results obtained by the ordinary pattern matching process and the 3 types of new process formed by adding the final stage F of the flowchart in Fig. 11 to each one of the 3 types of recognition processes except the ordinary one in Table 5.

認識方法 / 実験項目	全数整合	母音の有無	2特徴量判別分析	3特徴量判別分析
誤認識文字数	141	159	213	214
認識率	99.2%	99.1%	98.8%	98.8%
他クラスとの認識文字数	-	792	922	933
計算回数／１入力文字	332.0	155.2	116.6	115.9
認識速度改善率	1.00	2.33	2.97	3.02

り、文字サイズの特徴量と組み合わせて用いることにより、認識率は全数整合法に比して0.4％減少するが、認識速度は３倍程度まで向上させることができ、チベット学者らの要望に答えることができた。

本論文では、チベット文字の特徴である構造情報を用いて、文字の分類を行ったが、今後の課題として、チベット文字の文法情報を取り入れることにより、全数整合法と同等以上の認識率を確保しながら、認識処理速度をより一層向上させ、チベット学者にとって、さらに利用しやすいシステム構築を目指していきたい。

謝辞

大変貴重な文献資料など、ご協力いただいた大谷大学兵藤一夫助教授に心から感謝いたします。また、御助言・御討論を頂いた北陸先端科学技術大学院大学下平博助教授らに心から感謝いたします。

参考文献

1)小島正美、川添良幸、木村正行：コンピュータによるチベット文献の自動認識、日本西蔵学会々報、第43号、pp. 31-38 、(March 1998).

2)小島正美、布宮千夏子、川村隆庸、秋山庸子、川添良幸：オブジェクト指向設計によるチベット活字辞書を用いた類似文字認識、情報処理学会論文誌、Vol. 36, No. 11, pp. 2611-2621, (Nov. 1995).

3） Kojima, M., Kawazoe, Y., and Kimura, M., : Automatic Tibetan Scripts Recognition by Computer, 7th Seminar of the International Association for Tibetan Studies, Vol.1, pp. 527 -533, (April 1997).

4)ロサン・トンデン著、石濱裕美子・ケルサン・タウ訳：現代チベット語会話 、（株）世界聖典刊行協会、（July 1997).

5）兵藤一夫：チベット語文献におけるコンピュータ利用、勉誠出版、人文学と情報処理No.18、特集「挑戦古文書ＯＣＲ」pp.43-49, (Nov. 1998).

6)小島正美、川添良幸、木村正行：木版刷チベット文献の文字自動認識の試み、情報知識学会誌、Vol. 2, No. 1, pp. 49-62 (1991).

7)舟久保登：パターン認識、共立出版、情報電子入門シリーズ、(Dec. 1991).

8)有馬哲、石村貞夫：多変量解析のはなし、東京図書、(June 1989).

9)飯島泰蔵：パターン認識理論、森北出版、基礎情報工学シリーズ、(Dec. 1991).