人文科学とコンピュータ　23－2
（１９９４．　９．　１６）

オブジェクト指向設計によるチベット文字認識について

小島正美　布宮千夏子　川村隆庸　秋山庸子　川添良幸　木村正行
（東北工大）　（山形県庁）　（日本ＩＢＭ）　（東　北　大）　（北陸先端大）

本研究では、１音節単位で切り出されたチベット文字に対してオブジェクト指向による認識実験を行なった。本研究で扱っているチベット活字文字の字種は、基本３０子音と重層字７６種および４母音と少ないが、類似した文字が大変多く、それに対応して種々の処理が必要となる。類似した文字に対して、文字の特徴に合わせた認識メソッドとその文字をカプセル化したオブジェクト指向設計による類似辞書を作成し、それを用いて文字認識を行うことにより本手法の有効性を確認することができた。

Recognition of Tibetan Characters by Object oriented designing

Masami Kojima† Chikako Nunomiya †† Takanobu Kawamura†††
Youko Akiyama†††† Yoshiyuki Kawazoe†††† Masayuki Kimura†††††

†Tohoku Institute of Technology,
35-1, Kasumi-Cho, Yagiyama, Taihaku-Ku, Sendai 982, Japan.
††Yamagata Prefectural Government,
8-1, 2 Chome, Matunami, Yamagata 990-70, Japan.
†††IBM Japan, Ltd.
19-21, Hakozaki-Cho, Nihonbashi, Chuou-Ku, Toukyou 103, Japan.
††††Institute for Materials Research, Tohoku University,
2-1-1, Katahira, Aoba-ku, Sendai 980-77, Japan.
††††† Japan Advanced Institute of Science and Technology, Hokuriku.
15, Asahidai, Tatunokuchi- Machi, Nomi- Gun, Ishikawa 923-12, Japan.

In this paper, we extract syllables from Tibetan manuscripts and try to recognize automatically the separated 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. Therefore, we apply an object oriented dictionary which is created by combining the categorization and the character identification procedures. From the experiment, it is confirmed to be possible to improve the rate of Tibetan character recognition dramatically by object oriented method.

1．はじめに

今回認識対象としたチベット活字文献[1]の冒頭部分を図１に示す。一般に文字認識を行なう場合、大きく分けて文字認識を行なう前までと後とに分けられ、前者は前処理部と言われる。前処理部においては、行切り出し、切り出した行の傾き補正、ノイズ除去、正規化、文字切り出しが行われる。チベット文字の場合は図２に示す様に、７つのパターンに分節される[2]。そのため、音節単位毎の文字切り出しを行なわなければならない。今回は、後者の文字認識の部分において、音節毎に切り出された文字のオブジェクト[3]を洗い出した。それを基にクラス設計[4]を行なった。これまでの認識実験で、誤認識するのは類似文字であることが分かっている[5]。類似文字に対しては文字の特徴に合わせた認識メソッドとその文字をカプセル化し、それを新たなオブジェクト類似辞書とし、これを用いて文字認識を行なうことにより本手法の有効性を確認した。

2．チベット文字

チベット文字の１音節構成の最大要素は、図２に示す７番目の分節パターンで、基字、付頭字、付足字、前接字、後接字、再後接字、母音記号の７種からなる。基字と付頭字、基字と付足字との組み合わせ文字は重層字と呼ばれ７６通りの種類がある。基字および重層字を子音と定義すれば、チベット文字の１音節構造は子音１ないし４個と母音１個との組み合わせからなる。母音記号「ｉ」、「ｅ」、「ｏ」は基字または重層字の上部に付き、母音記号「ｕ」は基字または重層字の下部に付く。基字または重層字の上部または下部に母音記号が存在しない場合は、通常のローマ字風の記述に従えば母音記号「ａ」を含んでいる。

チベット活字文字の基本３０子音及び４母音を図３に示す。表音記号「ｔｓａ」、「ｔｓｈａ」、「ｄｚａ」の文字は表音記号「ｃａ」、「ｃｈａ」、「ｊａ」の文字と上部のヒゲの部分を違い点「ｔｓ」とし（音素的には「ｔｓ」に対応する付加記号）、その組み合わせで表現することにした。

3．クラス設計

前処理部において３８９９音節文字中、正しく切り出された３８８５音節文字について、オブジェクト指向設計による認識実験を行なった。オブジェクト指向設計を行なうために、初めに図４に示す様にチベット文字を上部、基部、下部文字と分割し、認識する作業を記述した。その記述内容

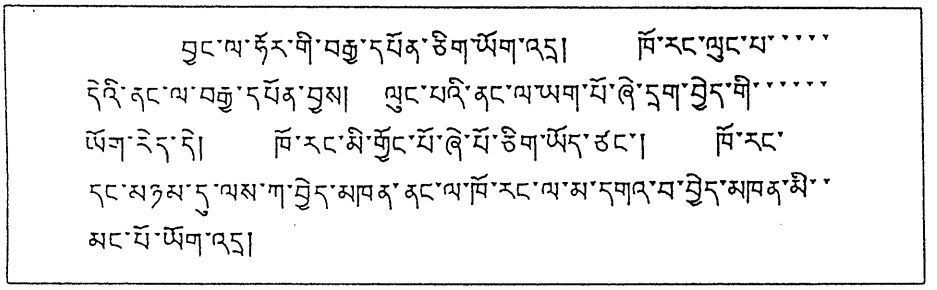


図１ 認識実験に用いたチベット活字文献

Fig. 1 Part of printed Tibetan texts used in the recognition experiments.

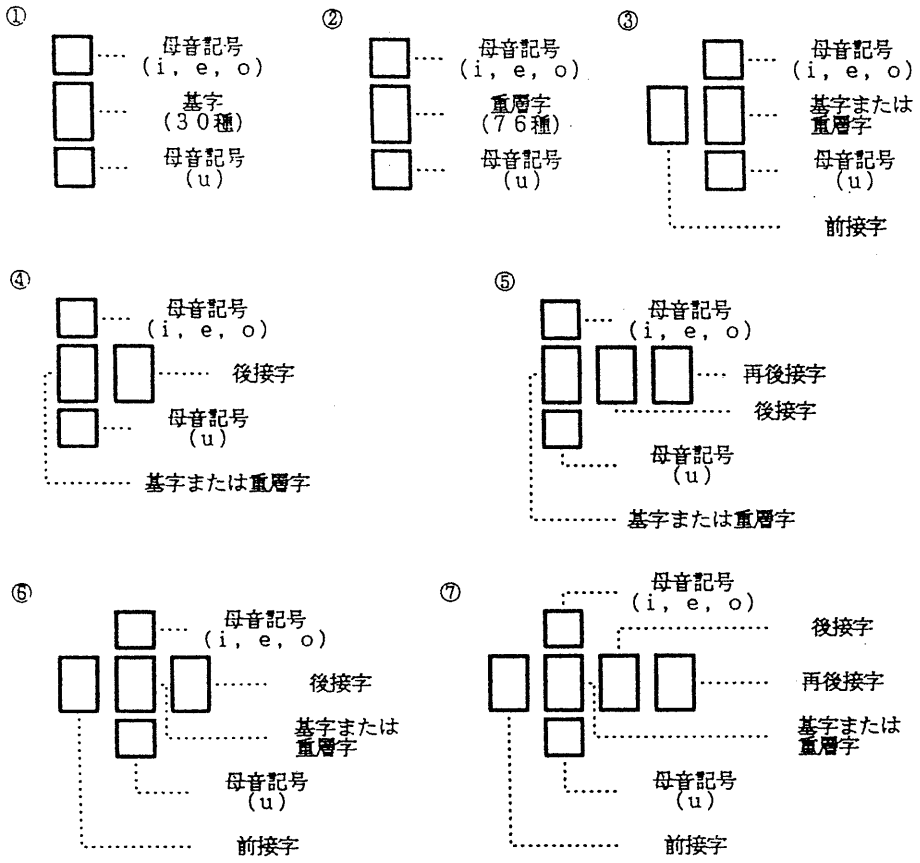
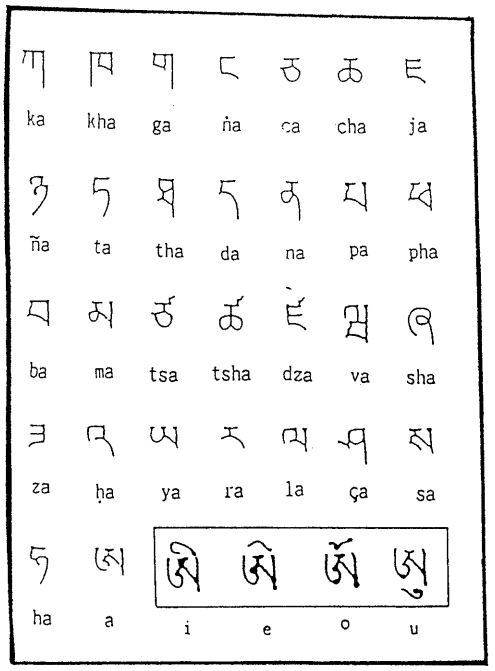


図2 チベット文字分節の7つのパターン

Fig. 2 7 patterns in syllabication of Tibetan characters.

からオブジェクトとメソッドの洗い出しを行い、クラス設計を行なった。オブジェクト「上部文字」は属性としてオブジェクト「文字」の上部文字データ、母音記号「i」、「e」、「o」、違い点「ts」を持ち、上部を分割し、文字データに対して重心移動するというメソッドを持っている。この場合、オブジェクト「上部文字」はオブジェクト・クラスとなるので、この様な場合、この後単にクラスとだけ表現する。同様にして、クラス「基部文字」は属性として基部文字データ、基部表音記号「ka」、「kha」などの基本30子音と重層字76種を持ち、基部文字データを分割し、正規化するというメソッドを持っている。クラス「下部文字」は属性として下部文字データ、



←

図３ チベット活字基本３０子音及び４母音
（単独のａも子音に数える。ｉ，ｅ，ｏ，ｕは、実際は ａの文字に付いた上部、下部の部分のみ。）

Fig. 3 Basic 30 consonants and 4 vowels of Tibetan scripts. (**a** is included in consonants and only upper and lower parts of **i, e, o, u** are used.)

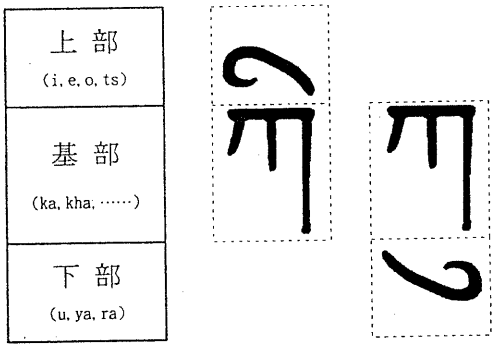


図４ 文字パターン分割

Fig. 4 Distinctions of character pattern.

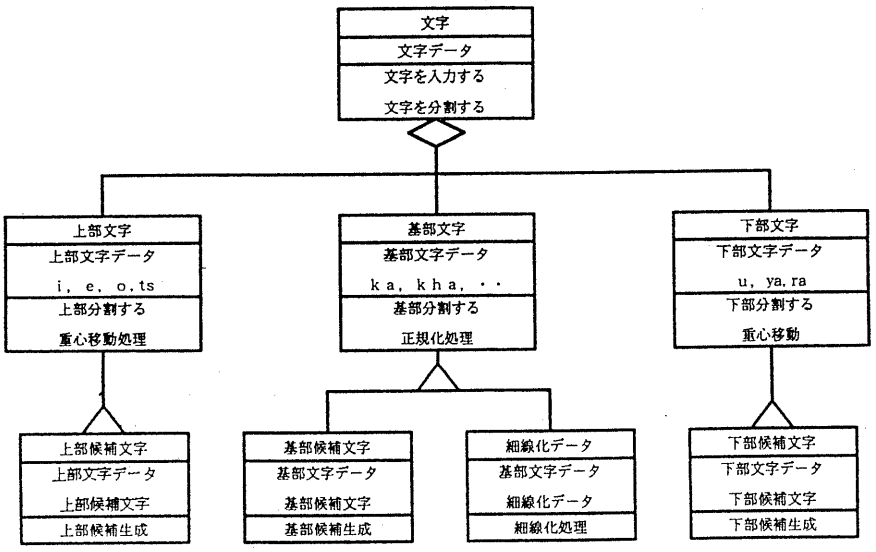


図５ 文字クラス図

Fig. 5 Class chart for Tibetan characters.

母音記号「u」、付足字「ｙａ」、「ｒａ」を持ち、下部を分割し、下部文字データに対して重心移動するというメソッドを持っている。クラス「文字」の設計図を図５に示す。

「文字」クラスに対応して設計した「辞書文字」クラスを図６に示す。チベット文字は図４に示す様に、上部、基部、下部と分割した場合、「辞書文字」クラスもそれに対応してクラス「上部辞書」、クラス「基部辞書」、クラス「下部辞書」とした方が分かり易くなる。クラス「上部辞書」と「下部辞書」は文字の大きさ情報も必要なため、単に重心移動による位置補正を行い、１０個程度のサンプルの平均値から１、０パターンとして作成した。クラス「基部辞書」については、基本３０子音と本論文で対象とした文献中に実際に出現した重層字５０種を対象に各々１０個程度のサンプルを正規化し、その平均値から１、０パターンとして作成した。クラス「基部辞書」において、基部文字データと基部辞書データとのハミング距離による重ね合わせ法を行い、その結果「基部候補文字」を生成する。クラス「基部候補文字」において、第１位候補文字が類似文字であると判定された場合は、クラス「類似文字群」で各類似文字群毎に識別を行なう。

類似文字群は図６に示す様に４つのグループに分類され、クラス「類似文字群１」は「ｂａ」、「ｐａ」、「ｐｈａ」の３文字からなり、誤認識の中で最も多かった文字群である。クラス「類似文字群２」は「ｂｙａ」、「ｐｙａ」、「ｐｈｙａ」、クラス「類似文字群３」は「ｍａ」、「ｓａ」、クラス「類似文字群４」は「ｄａ」、「ｎａ」の文字の組である。これらの文字は出現頻度も高く、誤認識の多くはこれらの文字間で起こっている。そのため、これらの文字のどれかが第１位候補文字となった場合、その文字が所属する類似文字群において、ホール数を調べる判定（ホール判定）を行なう。必要な場合は、文字上部の連続情報による判定（連続判定）を行なう。例えば、図７の類似文字群１において、ホール判定は黒画素で囲まれた部分をホールと判定し、その数とホールの面積を用いた。「ｂａ」と「ｐｈａ」はホール数１で、「ｐａ」はホール数０となる。また、「ｂａ」と「ｐｈａ」は同じホール数でもホールの面積により判定できる。ホール面積だけでは判定が困難な場合、上部の連続情報により判定する。「ｂａ」の文字は上部連続１で、「ｐｈａ」は２となる。

下部辞書においては、下部文字データが母音「u」であるか否かをハミング距離による重ね合わせ法で識別する。ハミング距離による重ね合わせ法だけでは、母音「u」と付足字「ｙａ」、「ｒａ」との識別が困難な時は、構造解析法により識別を行なった。

類似文字辞書を使わない場合、今回対象としたチベット文字３８８５音節文字中、正しく認識できたのは３５３５音節文字で認識率は９１％であるが、オブジェクト指向設計による類似辞書を活用することにより、９６．５％に向上した。実際の認識例を図８に示す。

４．まとめ

チベット文字の自動認識を行なおうとした場合、音節単位での文字切り出し及び認識を行なわなければならない。本研究では、オブジェクト指向設計による音節単位の文字切り出し及び認識を行なった。その結果、従来の重ね合わせ法に比べて、およそ６％程認識率が改善され、本手法の有効性が確認できた。

今後の課題として、誤認識の多くは文字の上部、基部、下部の切り出し時におきているので、これらの文字分割時における切り出し改善が挙げられる。また、誤認識する文字に対する認識メソッド

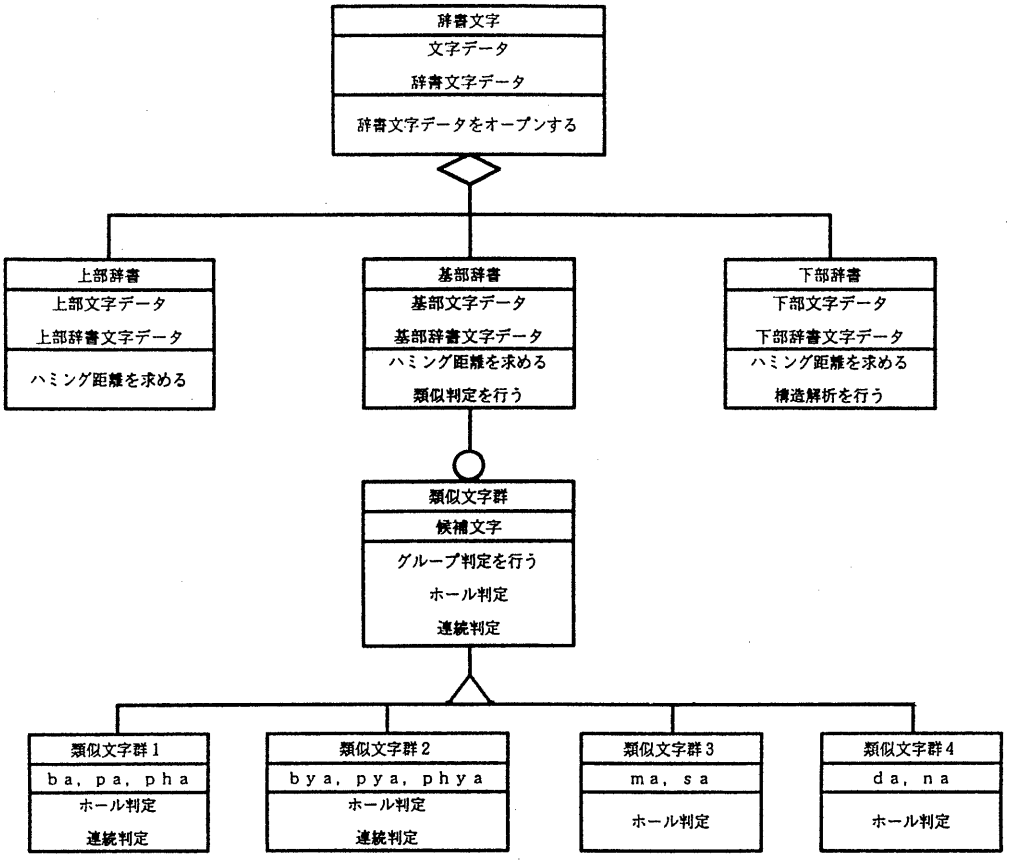


図6　辞書クラス図

Fig. 6　Class chart for Tibetan dictionary characters.

として、チベット文字の音節構造及び文法等から得られる知識を利用した認識方法を行っていきたい。さらに本研究で用いているオブジェクト指向設計による文字認識手法を、木版刷りのチベット文献の自動認識に適応することが考えられる。

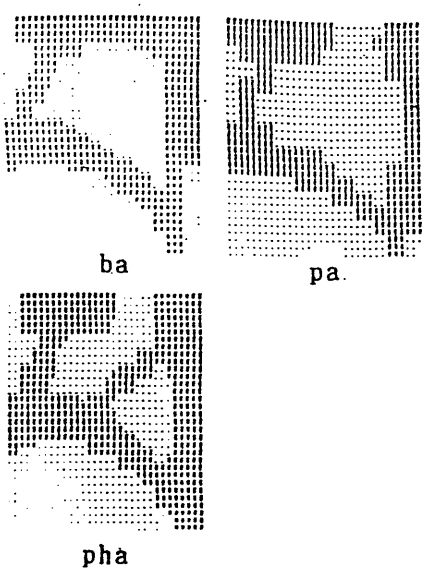


図７ 類似文字群１（ba, pa、pha）
Fig. 7 Group 1 of similar characters. (ba, pa, pha)

謝辞

本研究を進めるにあたり、貴重なアドバイスを頂いている宝仙学園短大塚本啓祥学長、東北大学文学部磯田熙文教授、医学部伊藤道哉助手、仙台電波高専山崎守一教授、実験をスムースにできるように心配りして頂きました東北大学情報処理教育センター並びに金属材料研究所材料情報科学室の皆様方に心から感謝致します。なお、本研究は文部省科研費一般研究（Ｃ）の補助を得て行っている。

参考文献

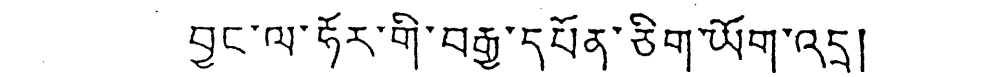
1）The seminar on Tibet : TEXTS OF TIBETAN FOLK-TALKS, IV, Tokyo, The Toyo Bunko, 1984, pp. 2-30.

2）稲葉：チベット語古典文法学、法蔵館、(1996).

3）J. Martin : Principles of Object Oriented Analysis and Design, Englewood Cliffs, (1993).

4）J．ランボー、M．ブラハ、W．プレメラニ、F．エディ、W．ローレンセン、羽生田訳：オブジェクト指向方法論ＯＭＴーモデル化と設計ー、トッパン、(1992. 7).

5）小島、川添、木村：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, (1991. 12), PP. 49-62.



' By n! ' l ' h(o) r ' g(i) ' b Rl(u) ' d p(o) n ' c(i) g ' y(o) g ' h! Dr |

図８ 認識結果出力例
Fig. 8 Results of Tibetan character recognition.