人文科学とコンピュータ　30－3
（1996．5．17）

差分重み付ユークリッド距離法による木版刷チベット類似文字認識

小島正美　（東北工大）
川添良幸　（東北大）
木村正行　（北陸先端大）

本研究は、木版刷チベット文献の自動認識を行なうための基礎実験の一つとして、木版刷類似文字認識を対象に、差分重み付ユークリッド距離法を適用し、その有効性を確認しようというのものである。今回は木版刷文献に適用する前に、活字文字に適用しその有効性を確認し、木版刷文献に適用した場合の違い等を検討してみる。

Character Recognition of Wooden Blocked Tibetan Similar Manuscripts by Using Euclidean Distance with Deferential Weight

Masami Kojima†　Yoshiyuki Kawazoe††　Masayuki Kimura†††

† Tohoku Institute of Technology
35-1, Kasumi－Cho, Yagiyama, Taihaku－Ku, Sendai 982, Japan.

†† Institute for Material Research, Tohoku University
1-1, Katahira, Aoba－Ku, Sendai 980, Japan.

††† Japan Advanced Institute of Science and Technology, Hokuriku
15 Asahidai, Tatunokuchi－Machi, Nomi－Gun, Ishikawa 923-12, Japan.

The set of Tibetan characters consists of basic 30 consonants, 76 combination characters, and 4 vowels. Despite the limited number, there are many similar characters which are categorized into four groups. In this paper, we try to establish a new object oriented method that similar characters are recognized by themselves for wooden blocked Tibetan manuscripts by using Euclidean distance with deferential weight.

1. はじめに

インド仏教は、1200年近くチベット文化の主流を形成し、チベット人固有の文化に多大な影響を及ぼしてきた。この間蓄積されたチベット文献資料は、膨大な量の資産として今日我々に残されている。例えば、東北大学図書館所蔵のチベット文献は、多田師が将来されたデルゲ版大蔵経と蔵外だけでも表裏木版刷紙で18万枚に及んでいる。これらの文献をコンピュータで自動認識することができれば、インド原典、チベット訳文献、漢訳文献などの研究者が本来の文献学に専念できる点において大変意義がある［1、2］。

今回認識対象とした木版刷チベット文献のデルゲ版チベット文献［3］を図1に示す。一般に文字認識を行う場合、大きく分けて文字認識を行う前までと後とに分けられる。前者は前処理部と言われ、行切り出し、傾き補正、ノイズ除去、正規化、文字切り出しが行われる。後者は切り出された文字の認識を行なう．木版刷チベット文献の場合、前者の前処理部は大変重要であるが、今回は認識部を主に実験を行なった．この場合、誤認識は類似文字間で起こっていることがこれまでの実験で分かっている[4]。これらの誤認識する文字をオブジェクト指向設計によりあらかじめ設定し、類似文字認識の対策を行ない、その有効性を確認した[5]。しかし、この方法は認識対象文字に左右され、汎用性に欠けるということが後で分かった．そのため、認識対象文字がチベット文字以外の文字にでも応用できるように、類似文字自身が自ら類似文字であることを判定し、認識した候補文字を選定する「差分重み付ユークリッド距離法」を提案する。初めにチベット活字文字に適用し、続いて木版刷チベット文字に適用してみる。

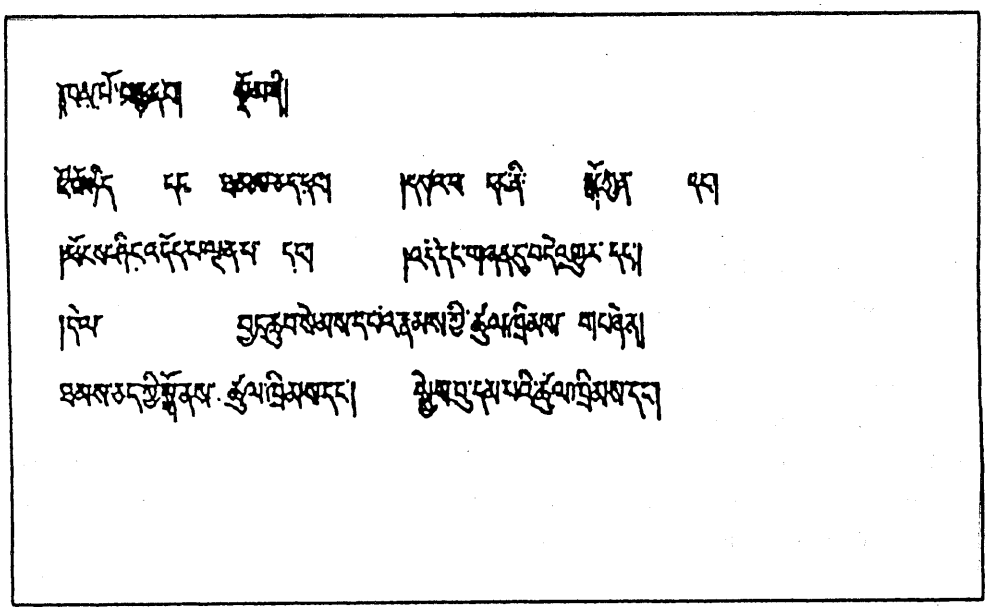

図1　認識対象とした木版刷チベット文献の一部

2．チベット文字

チベット文字［6］の１音節構成の最大要素は図２に示す音節構成であり、基字、付頭字、付足字、前接字、後接字、再後接字、母音の７種から構成される。なお、基字＋付頭字、基字＋付足字は重層字と呼ばれる。さらにサンスクリット文字からの転写文字などがある。母音記号のうち”ｉ”、”ｅ”、”ｏ”に相当する記号は基字または重層字の上部に付き、母音記号”ｕ”に相当する記号は下部に付く。チベット文字は単体で母音”ａ”を内在している。そのため、例えばチベット文字単体に母音記号”ｉ”が付加された場合、チベット文字に本来内在されている母音記号”ａ”が無視されることになる。母音記号を付加したこれらの文字を総計すると615文字となる。チベット文字の１音節構造は子音１乃至４個と母音の組み合わせからなる。子音の数が２個以上の場合、どの子音が基字となるかを判定しなければならず、分節パターン識別が必要となる。チベット文字の中で出現頻度が全体のおよそ80%を占める基本30子音および４母音を図３に示す（図は活字体を示し、表音はワイリ方式に従った）［7］。

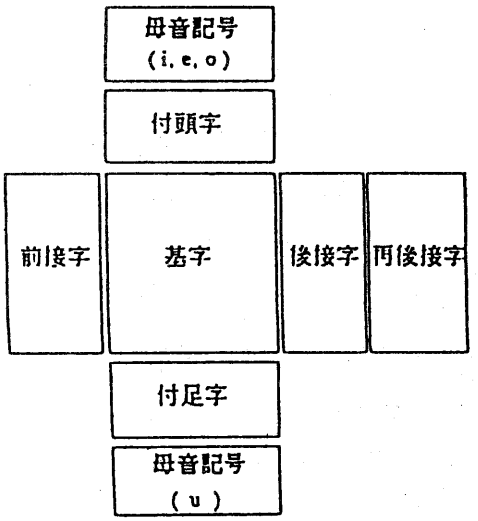

図2　チベット文字の音節構成

3．実験

3．1．文字切り出し

図１に示すデルグ版チベット文献は元の文献から予め切り張りして新たに作成されたデータを使用した。水平方向の射影により空白行を検出して、行データを取り出し、文節単位にデータが存在するので、１文節単位でデータをスキャナからコンピュータへ取り込む。取り込んだイメージデータに対してＬＰＰ（Local Projection Profile）法［8］により傾き補正を行った。

図3　チベット文字基本30子音と4母音

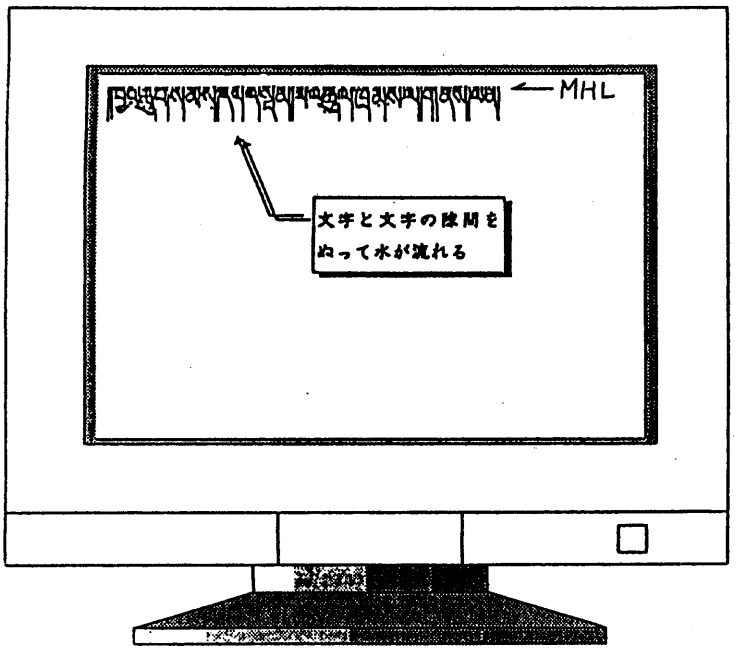

図4　「水流し法」による自動1文字切り出し実行画面

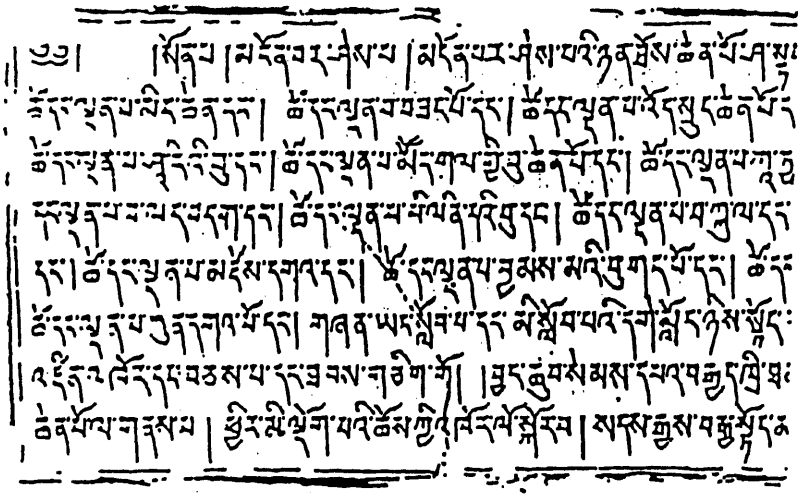

図5　北京版チベット大蔵経の中の正法白蓮華経の冒頭部分

傾き補正を行ったイメージデータから図４に示すようにMHL（Main Horizontal Line）（チベット文字の水平方向射影で特徴的なピークを現す場所；これを我々は”MHL”と呼んでいる）を基準にスペースを検出した場合、水を流すようにして切り出す方法（これを我々は”水流し法”と命名している）により１文字切り出しを行った。本手法によりデルグ版5,000文字に対して1,175文字の文字が切り出し可能となり、この段階における1文字切り出し率はおよそ24%である。図５に示す同じような木版刷文献である北京版チベット文献の場合は、3,844文字中2,768文字が正しく切り出しできて、その切り出し率はおよそ72%である［4］。このように一見同じように見える木版刷文献でも、文字間のスペースが微妙に異なり、それが文字切り出し率に大きく影響していることが分かる。現在、１文字幅情報などを取り入れて、実験を行っている。

３．２ 高度な認識システムの開発

従来の認識プログラム開発においては、処理機能を表すプログラムとデータを蓄えるデータベースの２つを別々に設計してきた。この事がプログラム開発をより複雑で分りにくいものにしてきた。オブジェクト指向開発は、この機能とデータをオブジェクトとして１つにまとめ、設計を一元化することにより、プログラムの品質・生産性・拡張性を高めようとするものである[9]。

今回の認識実験はアナログ辞書文字とサンプル(認識対象)文字とのユークリッド距離を算出して第１位候補文字と第２位以降の候補文字との距離が実験で定めたある値以上ある場合は、第１位候補文字を認識文字とした。第１位候補文字と第２位以降の候補文字との距離が実験で定めた値以内に接近している場合は、これらの文字群は総て類似文字であると判定し、候補文字のアナログ辞書同士の差分を取り、新たに差分の部分に重みをかけてユークリッド距離を求め、距離がもっとも近いアナログ辞書文字を候補文字とする「差分重み付ユークリッド距離法」を適用する。辞書文字を作成した530文字に対するユークリッド距離による認識実験（クローズ実験）を行った。その認識率はおよそ96%である。

４．まとめ

木版刷チベット文献の場合、文字同士が複雑につながっている。そのため、水流し法だけではデルグ版チベット5,000文字に対しておよそ24%程度の１文字切り出し率しか得られていない。認識精度の向上には、高精度な１文字切り出しが要求されている。今後、チベット文字の特徴抽出をしながらさらに有効な文字切り出し法を検討していきたい。例えばチベット活字文字の場合、１文字切り出しが容易であるため10,000文字を越える文字数に対して、オープン実験でおよそ９９%の認識率を得ている［7］。

オブジェクト指向設計により、基本子音の辞書文字作成に用いた文字データ530文字に対して、およそ96%の認識率を得ている。しかし、実用を考えた場合、オープン実験の認識率が重要となる。今後はオープン実験において90%を越える認識率を得られるように、認識ロジックの検討をしていきたい。

謝辞

本研究を進めるにあたり、チベット文字に関して大変貴重なアドバイスをいただいております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に心より感謝いたします。また、チベット活字文字フォントを使用させていただいております大谷大学真宗総合研究所兵藤一夫助教授に感謝いたします。本研究の実験は、文部省科学研究費「重点」の補助を得て購入しましたＩＢＭパーソナルコンピュータＴｈｉｎｋＰａｄを用い行いました。

文献

1）塚本：インド文学の形成と展開、”サンスクリット・チベット語のコンピュータによる統合研究、東北大学特定領域研究組織 TURNS017-報告書（1992. 2）；磯田：チベット文字の特色とコンピュータ利用について、ibid.

2）川添：コンピュータによる仏教混淆梵語の研究（2）、印度学仏教学研究37巻第2号（1989. 3）.

3）羽田野他：瑜伽論菩薩地、チベット佛典研究会、pp. 901-961、（1976）.

4）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, pp. 49-62,（1991. 12）.

5)小島、布宮、川村、秋山、川添：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).

6）稲葉：チベット語古典文法学、法藏館、（1966）.

)Masami Kojima, Yoshiyuki Kawazoe and Masayuki Kimura : Automatic Tibatan Scripts Recognition by Computer, 7th Seminar of the International Association for Tibetan Studies, (to be published in 1997).

8)秋山、増田：書式指定情報によらない紙面構成要素抽出法、電子通信学会論文誌（D）, J66-D, No. 1,（1983. 1）.

9)I. Jacobson : Object Oriented Software Engineering, Addison Wesely Publishing Company,（1992).