

オブジェクト指向設計を考慮した 木版刷チベット文献のイメージ文字認識

小島正美^{†1} 川添良幸^{†2} 木村正行^{†3}

^{†1} 東北工業大学通信工学科 〒982-8577 仙台市太白区八木山香澄町35-1

E-mail : mkojima@titan.tohtech.ac.jp

^{†2} 東北大学金属材料研究所

^{†3} 北陸先端科学技術大学院大学

要旨

本論文ではデルゲ版チベット文献をOCRにより自動認識することを目的としている。本研究の一番の問題点は、1音節区切り記号の識別が必要であり、この識別が大変困難なことである。そこで、この部分をカラーで線引きし、カラーの線引きを識別することにより1音節文字切り出しを行い、切り出し時に得られた文字特徴情報を切り出された文字に持たせるオブジェクト指向設計法の考え方をを用いて、認識を行うシステムを提案する。本方法の採用により、作業能率および1音節文字認識が従来の方法に比べて、飛躍的に向上する見込みである。

キーワード：木版刷、チベット文献、OCR、イメージ文字、1音節文字切り出し、オブジェクト指向

Automatic recognition of wooden blocked Tibetan image character by using object oriented design

Masami Kojima^{†1} Yoshiyuki Kawazoe^{†2} Masayuki Kimura^{†3}

^{†1} Tohoku Institute of Technology, E-mail : mkojima@titan.tohtech.ac.jp

^{†2} Institute for Materials Research, Tohoku University

^{†3} Japan Advanced Institute of Science and Technology, Hokuriku

Abstract

The purpose of this paper is to recognize automatically wooden blocked Tibetan manuscripts by OCR. The most difficult problem in this study is the segmentation of one syllable in Tibetan manuscripts. We suggest that it is effectively performed by using the recognition of colored line to segment the syllable, and object oriented method is meaningfully applied to this purpose. The result of our experiment shows clearly that the rate of segmentation is improved. Accordingly higher rate of character recognition is expected to be achieved.

Key Words : wooden blocked, Tibetan manuscripts, OCR, image characters, segmentation of one syllable, object oriented method

1. はじめに

本論文で認識対象としている文献は、図1に示すデルゲ版チベット文献である。この文献の大蔵経と蔵外だけでも表裏木版刷紙で18万枚ほどになる膨大なものであり、現在東北大学中央図書館に最重要資料として保管されている。コンピュータによりこれらの文献の自動認識を行うことができれば、インド原典、チベット訳文献、漢訳文献などの研究者が、単純な文字読取り作業から解放され本来の文献学に専念できる点で大変意義がある[文献1、2]。

著者らは東北大学文学部印度哲学研究室の協力を得ながら1989年から木版刷チベット文献の自動認識の研究に取組み、これまで正しく切り出しされた基本30子音については90%程度の認識率を得るに至っている[文献3]。しかし、チベット文字は1音節単位での表音認識を行わなければならない、そのためには1音節区切り記号の識別が必要である。また図1に示すように個々の文字が複雑に重なり合い、繋がっているため、1音節単位の文字切り出しは大変困難である。すなわち、最も困難なところは木版刷文献（イメージデータ）をスキャナで取り込み1文字切り出し（セグメンテーション）を行う部分であることがこれまでに分かっている。

そこで、1音節文字切り出しを可能とするアルゴリズムの検討を行ってきた。しかし、図1に示すような複雑な繋がり文字に対応する完全な切り出しプログラムを作成するには至らなかった。そこで、設計手法としてオブジェクト指向設計法[文献4～6]を取り入れて新たにこの問題に取り組んでみることにした。オブジェクト指向設計は次の5つのモデルから成っている[文献4]。要求モデル（Requirement Model）、分析モデル（Analysis Model）、設計モデル（Design Model）、実行モデル（Implementation Model）、試験モデル（Test Model）である。各モデル間を相互に往来しながら、目的とするシステムを実現させる。オブジェクト指向設計法においてもっとも重要とされているのは、「要求モデル」と「分析モデル」との往来である。著者らはその部分を特に重視して研究を行ってきた[文献7、8]。すなわち、実験システムのユーザであるチベット学者からの「要求モデル」に対する「分析モデル」を文字認識システムを構築する筆者らが検討し、設計することから始めた。この時、コンピュータで認識困難であった1音節区切り記号の識別は、チベット研究者にとって容易であることが分かった。また、チベット研究者がもっとも厄介だと思っていることは、1音節毎に切り出された文字のローマ字表記をキーボードから入力することであることも分かった。本論文ではチベット研究者がもっとも厄介と思っている部分は自動認識することとし、彼等にとって容易な1音節文字切り出しは、1音節区切り部にカラーペンで挿入してもらい、その線をコンピュータで自動識別するシステムとした。

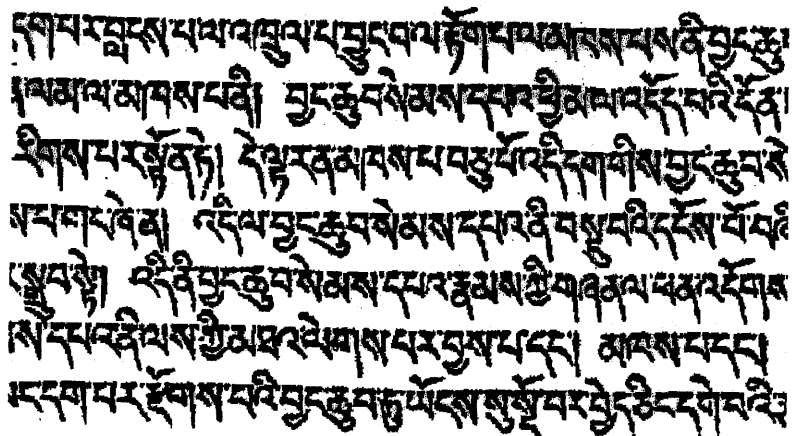


図1 認識対象としたデルゲ版チベット文献の一部

2. 実験

本論文で認識実験に用いた文献は図1に示すPechaと呼ばれる木版刷文献で、その大きさは幅10cm、長さが40cm程度である。そのため、0次処理としてA3版のスキャナ上におくことができる大きさまで木版文献を縮小コピーし、そのコピーした文献に1音節単位に図2に示すように赤ペンにより切り出し線を挿入し、文字部分（黒画素部）のみの水平方向射影をとる。図2に示す水平方向射影は対象画像を15%に圧縮したものを出力している。著者らの初めの予測では、チベット文字の特徴である音節区切り記号付近の主要水平線（これを我々はMHL: Main Horizontal Lineと呼んでいる）が行当たり1個の極大点となって現われるはずであった。しかしPechaのような横方向に長い木版刷チベット文献の場合、多くの文献はMHLが波状になっている。そのため図2に示すように二つの極大点が現われ、極大点の位置からだけでは正確な行切り出し位置を検出することは困難である。そこで、極大点の近傍を平滑化処理を行い、初めの極大点はp1（図2）、対象行の極大点はp2（図2）、次の行の極大点はp3（図2）とする。 $(p2 - p1) / 2 = L_{top}$ 、 $9 \cdot (p3 - p2) / 10 = L_{bot}$ とする。 L_{top} は上部の行切り出し箇所、 L_{bot} は下部の行切り出し箇所となる。ここで、切り出し位置を決定するために使用した定数値は実験値から求めた。このように、行切り出しを行うために、上下の切り出し位置に余裕を持たせている。

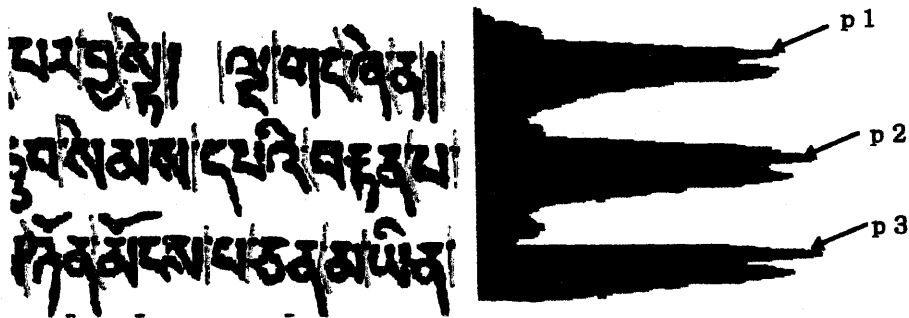


図2 チベット文献の水平方向射影図
(対象画像を15%に圧縮)

切り出し対象行を判別するために図3に示すように切り出し行の先頭に赤の標準線を記している。その標準線の幅内に存在する赤線は切り出し対象行の1音節文字切り出し線と認識し、その範囲外の赤線は消去する（図3参照）。次に、1音節文字切り出し線が図4に示すように音節区切り記号上に存在する場合、その部分の色識別が困難となる。そこで、1音節文字切り出し線の上部から下部への結合状態を調べ、切断されている箇所はその近傍の赤線と結合させることにより1音節文字切り出し線を修復する（図4参照）。このままでは、行切り出しにより上下行からの切り出しノイズが多くなる。そのため、図5に示すように、1音節文字切り出し線を上下行に向かい水を流すように延長していく。水が流れにくい場合、すなわち実験的に定められた範囲内で黒画素の断点が見つからない場合、強行突破してその部分を切断する。図5の下部で、左側は水が流れる方向へと切り出し線を引いた例で、右側は強行突破して切り出し線を引いた例である。上部については、画像を反転させて水が流れる方向と考えればよい。これを我々は「水流し切り出し法」と呼ぶことにする。

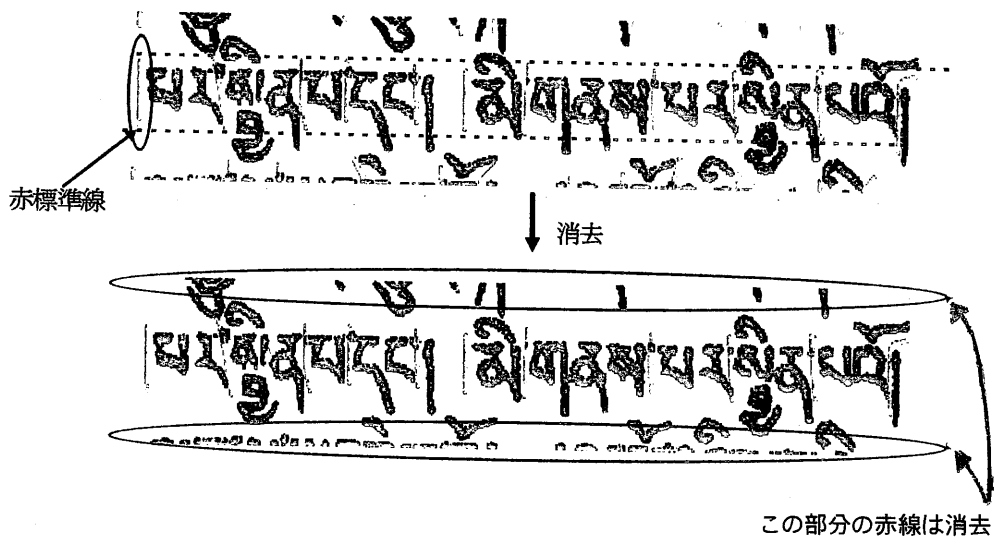


図3 赤の標準線をもとに1音節文字切出し線の識別

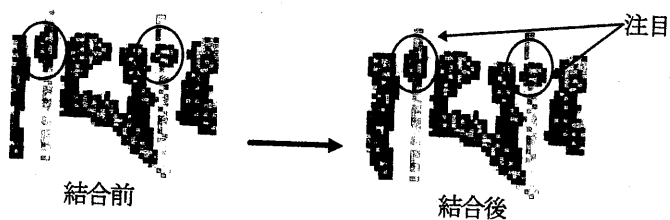


図4 掠れた1音節文字切出し線の修復

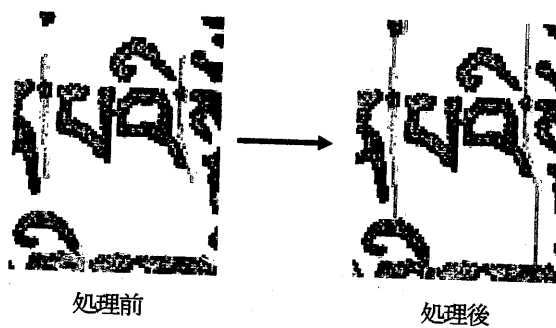


図5 1音節文字切出しを上下の行切出し線まで延長

このようにして、1音節文字切り出し線で囲まれた文字を切り出す。この状態で、上端および下端から連続し、中央の認識対象チベット文字から独立している黒画素はノイズとみなす。また、赤線の部分は消去し、そのために生じる孤立した黒画素は、その大きさ情報からノイズとみなして除去した例が図6である。その後、1音節文字の大きさ情報を測定し、非線形正規化処理を行い、文字の大きさを64ドット×64ドットにする。本論文では、1音節文字切り出し時の音節記号、文節記号、大きさ情報等を、非線形正規化したオブジェクト1音節文字に属性として持たせたファイル例を図7に示す。ここまでの処理が1音節文字切り出し処理となる。

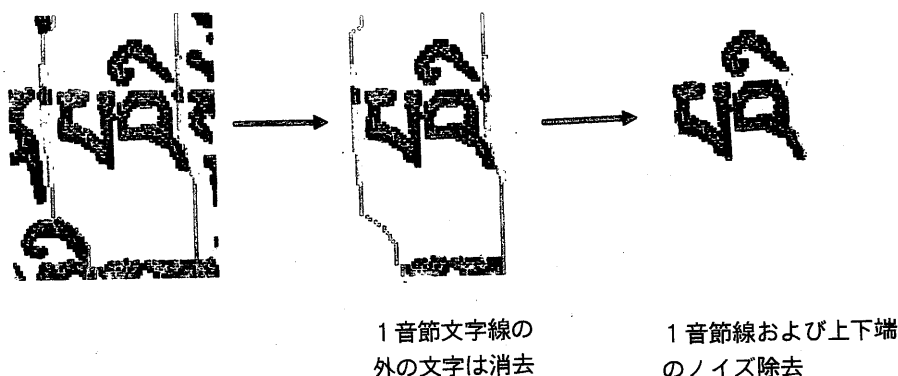


図6 1音節文字切出し例

ADDRESS	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00000000	63	75	74	5F	01	A7	55	16	1B	00	00	00	00	00	0E	00
00000010	00	3F	FF	FF	8F	FF	FE	18	00	3F	FF	FF	8F	FF	FE	18
00000020	00	3F	FF	FF	8F	FF	FE	18	00	3F	FF	FF	8F	FF	FE	18
00000030	00	FF	FF	F0	7F	FF	FF	FF	00	FF	FF	F0	7F	FF	FF	FF
00000040	00	FF	FF	F0	7F	FF	FF	FF	00	FF	FF	F0	7F	FF	FF	FF
00000050	00	FF	FF	FF	FF	FF	FF	FF	00	FF	FF	FF	FF	FF	FF	1F
00000060	00	FF	FF	FF	FF	FF	FE	1F	00	FF	FF	FF	FF	FF	FF	F8
...																
000031A0	00	00	00	00	A8	55	19	1C	00	00	00	00	00	00	0E	00

cut_

1ページの総音節数

音節コード、X軸、Y軸

画像データ

頭末文節記号コード

音節コード、X軸、Y軸

図7 1音節文字切出しファイル

3. まとめ

木版刷チベット文献8ページ分、認識対象の3、407音節文字中、1音節文字切り出しに失敗したのは175音節文字で、1音節文字切り出し率は95%程度となった。失敗した175音節文字の内訳は、切り出しノイズ:38、母音欠如:27、上下行からの繋がり文字:110となっている。上下行からの繋がり文字に関しては、繋がり情報を属性として持たせたオブジェクト1音節文字として切り出し、この部分は画面上でチェックする段階で入力することとする。この部分を失敗例から除いた場合の1音節文字切り出し率は98%程度となり、ほぼ実用の目処がたった。

今後の課題として、切り出し時に生じるノイズおよび母音欠如に対する対策、正しく切り出された1音節文字についてオブジェクト指向辞書文字を作成し、認識実験を行うことなどが挙げられる。

謝辞

日頃ご指導いただいております宝仙学園短期大学学長塚本啓祥教授並びに東北大学文学部磯田熙文教授に心から感謝いたします。著者らが試作している木版刷チベット文献認識システムのデモンストレーションを見ていただき、沢山のチベット学研究者の方々から貴重な意見をいただいております。これらの方々に感謝いたします。本研究の一部は文部省科学研究費平成7年度重点領域「人文科学とコンピュータ」課題番号「07207222」および平成9年度重点領域「人文科学とコンピュータ」課題番号「09204224」の補助を得て行いました。

<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総合研究」、東北大学特定領域研究組織TURNS017-報告書 (Feb. 1989); 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、Vol. 37, No.2 (Mar. 1989) .
- 3) 小島、川添、木村: 推論を用いたチベット文献中の文字自動認識; 日本印度学仏教学研究、第41巻、1号、pp.158-161, (Dec. 1992) .
- 4) Jacobson, I. : Object Oriented Software Engineering, Addison Wesley Publishing Company, (1992) .
- 5) Martin, J. : Principle of Object Oriented Analysis and Design, Englewood Cliffs (1993) .
- 6) J.ランボー, M.ブラハ, W.ブレメラニ, F.エディ, W.ローレン、羽生田訳: オブジェクト指向方法論OMT—モデル化と設計—、トッパン、(July 1992) .
- 7) 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, General Edited by Ernst Steinkellner, Volume I, PP. 527 - 533, (Apr. 1997) .
- 8) 小島他: オブジェクト指向設計による木版刷チベット文献の文字自動認識に関する研究、文部省科学研究費重点領域研究(課題番号07207200) 1995年度研究成果報告書、1996.