2007－CH－73 (4)
2007／1／27

オブジェクト指向設計によるチベット文字認識研究の発展

小島 正美[†1] 布宮 千夏子[†2] 川添 良幸[†3] 木村 正行[†4]

†1 東北工業大学・工学部 〒982-8577 仙台市太白区八木山香澄町 35-1

†2 山形県立山形職業能力開発専門校能力開発支援課 〒990－2473 山形県山形市松栄二丁目 2 番 1 号

†3 東北大学金属材料研究所 〒980 仙台市青葉区片平１－１

†4 北陸先端科学技術大学院大学 〒923-12 石川県能美郡辰口町旭台 15

あらまし 研究初期段階ではチベット文字認識における誤認識の多い類似文字にオブジェクト指向設計法を適用してきた．次に、チベット文字が固有に持つ有効な特徴情報を活用することにより文字認識率の向上に成功した．現在は、ユーザであるチベット学者側からの要求をより忠実に取り入れたオブジェクト指向設計によるチベット文字認識システムの構築を行なっている．

キーワード チベット文字認識手法、文字認識システム、ＵＭＬ、ＧＵＩ、オブジェクト指向設計

Progress in Tibetan Character Recognition by using Object Oriented Design

Masami KOJIMA[†1], Chikako NUNOMIYA[†2], Yoshiyuki KAWAZOE[†3] and Masayuki KIMURA[†4]

[†1] Department of Electrical Communication, Tohoku Institute of Technology

25-1,Kasumi-Cho, Yagiyama, Taihaku-Ku, Sendai, 982-8577, Japan E-mail: m.kojima@tohtech.ac.jp

[†2] Yamagata Vocational Skills Development Institute Skills Development Support Division

2-2-1,Matuei, Yamagata-shi, 990-2473, Japan

[†3] Institute for Material Research, Tohoku University, 1-1, Katahira, Aoba-Ku、Sendai, 980-8577, Japan

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

Abstract

In this paper, we design a desirable system of automatic character recognition for Tibetan characters by UML(Unified Modeling Language), which is a newly developed method of Object Oriented Design(OOD). The purpose of this study is to support the research on Tibetan literatures and establish character recognition method using characteristics of Tibetan characters.

Keyword Method of character recognition, Character recognition system, UML(Unified Modeling Language), GUI(Graphical User Interface),OOD(Object Oriented Design)

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大きな影響を及ぼしてきた．これまでに蓄積されてきたチベット文献の資料は膨大な量の遺産として今日我々に残されている．これらの重要な仏典文献をコンピュータで自動認識することができれば、それを活用することによりインド原典、チベット訳文献、漢訳文献などの研究者にとって、本来の文献学に専念できる点において大変有意義である[1)]．

チベット文字認識に関する研究の歩みを表１に示す。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年から重ね合わせ法と構造解析法を組み合わせた認識手法を主体に木版刷りチベット文献についての認識実験を行なってきた[2~4)]。１９９３年からはその改良版として重ね合わせ法で誤認識する類似文字に対してオブジェクト指向設計法[5~8)]による類似文字クラスを作成し、類似文字固有の認識メソッドを用いることにより文字認識率の向上を図ってきた[9~11)]．１９９８年からは重要な仏典文献の活字で出版された文献の認識も木版刷りチベット文献の認識と並行して行なってきた．さらにチベット文字１音節ごとでの文字認識を行なうことを取り入れてきた[12~17)]．２００４年からオブジェクト指向設計法の本来の目的であるユーザからの要求をいかに実現するかに重点を置いた文字認識法に着手している．文字認識を行なう場合、重ね合わせ法と構造解析法とがあるが、文字認識手法が簡便で認識精度の高い重ね合わせ法は文字の標準データとなる辞書文字作成を行う必要がある．しかし、従来は文字認識を設計する側で辞書文字作成を行ってきた．そのために、異なる文字種になった場合の対応が困難となっている．辞書文字作成をユーザであるチベット学者に行ってもらうことにより、より汎用性の高い文字認識システムを実現することが可能となった[18)]．

表１　年代ごとによるチベット文字認識手法

年代	認識手法
１９８９年～１９９２年	重ね合わせ法と構造解析法を併用した文字認識
１９９３年～１９９７年	オブジェクト指向設計法：誤認識の多い類似文字に適用
１９９８年～２００３年	１音節単位での文字の特徴情報としての取込み
２００４年～現在	ユーザからの要求を反映した文字認識システムの設計

2. チベット文字の特徴

チベット文字の１音節構成の最大要素は図１に示すように前接字、基字部、後接字、再後接字の４文字から構成される．基字部は付頭字、付足字、母音記号からなる．付頭字＋基字、基字＋付足字は重層字と呼ばれおよそ 80 通りの文字が存在する．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チベット文字に内在している母音記号「a」を付けて読む．チベット基本30子音と4母音を図2に示す．表音記号「ca」、「cha」、「ja」の文字は表音記号「tsa」、「tsha」、「dza」の文字と上部ヒゲの部分だけが異なる類似文字である．図2において、各文字に隣接している逆三角形の塊はチベット文字の1音節区切り記号でツエックと読む[19]．

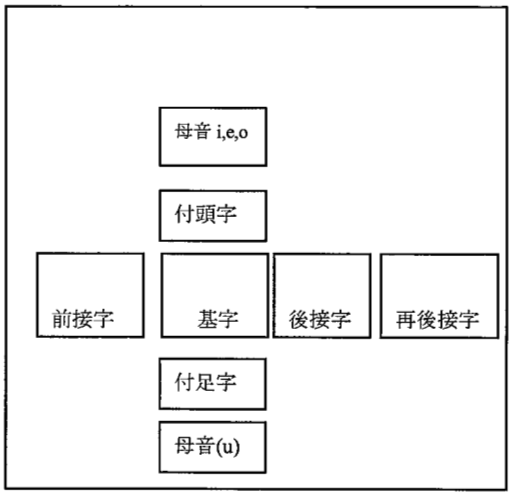


図1　チベット文字の1音節構成の最大要素　図2　チベット文字基本30子音と4母音

3．チベット文字実験

本論文において、チベット文字認識システムの設計にオブジェクト指向設計法を活用した推移を主に記述する．

以下、1993年から1997年代の実験、1998年から2003年代の実験、2004年から現在までの実験と分類して、文字認識に対してオブジェクト指向設計法を活用して高度化して来た経緯を記述する．

3－1．1993－1997年代の手法

一般に文字認識を行なおうとする場合、文字切り出しから認識まで多岐にわたる処理が必要となる．図3のように文字認識システムは大きく3段階に分けられ、段階ごとに設計開発をオブジェクト指向設計法で行ない、各段階間の連携を行なう統合的システムで行なう．本方法では、段階ごとに設計開発を他段階と独立して実現できる利点がある．第1段階は前処理部と呼ばれる部分で、イメージ文字入力から行、文節切り出し、文字切り出しを行なっている．第2段階は切り

出された文字を認識する部分であり、第３段階は認識した文字を１音節単位で表示する部分である．当初のオブジェクト指向設計法は主に第２段階に適用している．認識対象としているオブジェクトを洗い出し、オブジェクト間のメッセージ・インタラクト・ダイアグラムを作成し、それを基にクラス設計を行なう．オブジェクト指向設計法の重要な特徴であるオブジェクトとメソッドをカプセル化することを、類似文字認識に積極的に適用した．

図４に示すようにチベット文字特有の横棒（MHL：Mai Horizontal Line）文献番号）を境に上部と基部に分ける．下部に母音「u」が存在ときの切り出し位置（LCL : Low Cut Line）を境として基部と下部に分ける．「u」の文字の位置は固定されていないので、「u」の文字を認識することによりLCLを特定可能となる．そのため、下部文字については文字分割と認識を同時に行なう．このようにして、クラス「文字」はクラス「上部文字」、「下部文字」、「基部文字」に分割する．

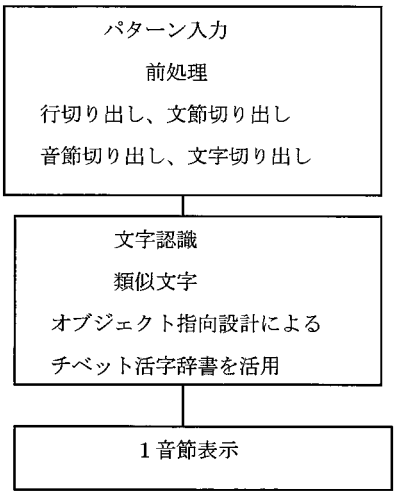


図３　文字認識システムのフローチャート

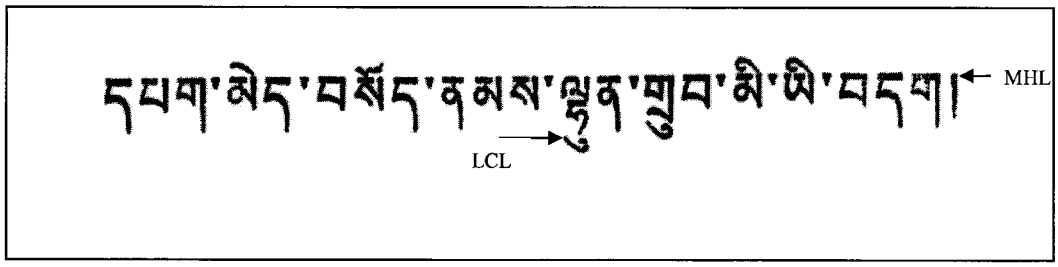


図４　チベット文字特有のMHLおよび「u」を識別するためのLCLの説明図

クラス「上部文字」はクラス「上部辞書」との重ね合わせ法により、第１位候補文字から第３位候補文字とユークリッド距離を持ったクラス「上部候補」を生成する．同様にクラス「下部文字」は、クラス「下部辞書」との重ね合わせ法により、母音「u」とユークリッド距離を持ったクラス「下部候補」を生成し、クラス「基部文字」はクラス「基部辞書」との重ね合わせ法により第

1位候補文字から第5位候補文字とユークリッド距離を持ったクラス「基部候補」を生成する．さらに、クラス「基部文字」は文字の構造解析のためにクラス「細線化データ」を生成する．クラス「基部辞書」の一部は、類似判定のためのクラス「類似文字群」と関係する．誤認識の多い類似文字を、文字の特徴により類似文字グループをカプセル化することにより、文字認識率の向上を逐次改善することを可能としている．「基部文字」をトリガーとしたメッセージ・インタラクト・ダイアグラムを図5に示す．

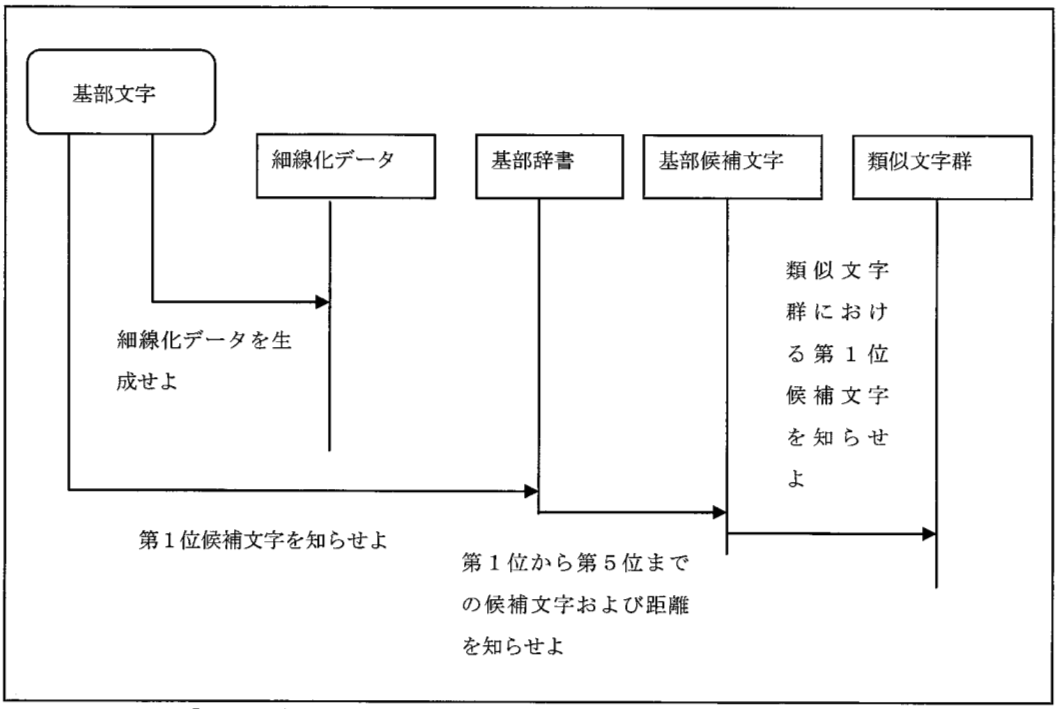
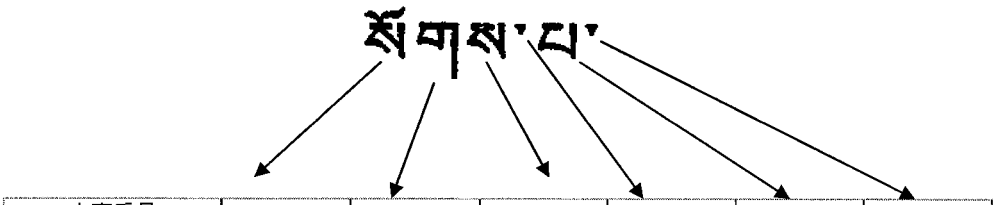


図5　「基部文字」をトリガーとしたメッセージ・インタラクト・ダイアグラム

3－2．1998－2003年代の手法

チベット文字は文字種により文字サイズが大きく異なる．また、字種により字体の複雑さ、構造が異なる．文字が固有に持っている特徴情報を文字切り出し時に取得し、切り出した文字の属性として持たせて、文字認識時まで継承させることにより文字認識率の向上が期待できる．チベット1音節文字は1ないし4個の文字から構成される．文字が2個以上から構成される場合は、どの文字が基字（母音をつけて読む文字）となるかを判定する必要がある．これらの判定を行うための重要な条件はMHL判定である．

文字の特徴情報を属性としたテーブルを図6に示す．これらの特徴情報から文字の大分類を行なうことが可能となり、高速・高精度な文字認識率を得るシステム設計を行なうことができる．現在、99%台の精度で大分類することが可能となっている．



文字番号	1	2	3	4	5	6
音節内文字数	1/3	2/3	3/3	・	1/1	・
縦サイズ(dot)	48	53	33	・	33	・
横サイズ(dot)	32	34	31	・	31	・
縦交差回数(回)	2.59	1.44	1.32	・	0.97	・
横交差回数(回)	1.79	1.58	2.06	・	1.70	・
上部付加母音	○	×	×	・	×	・
sa 判定	×	×	○	・	×	・
_a 判定	×	×	×	・	×	・
_i 判定	×	×	×	・	×	・
da判定	×	×	×	・	×	・
総合判定	基字	後接字	再後接字	ツェック	基字	ツェック
	(上部母音)		(sa)			

図６　文字の特徴情報を属性として持たせるテーブル

３－３．２００４年－現在の手法

オブジェクト指向設計法でもっとも重要なことはユーザであるチベット学者からの要求を取り入れることである．すなわち、チベット学者が文字認識システムを操作することを想定したシナリオを作成している．そのため、シナリオから詳細な分析を行うためにＵＭＬ（Unified Modeling Language）によるシステム設計を行っている．その結果、従来までの文字認識システムでは考慮されていなかった前処理部に図７に示すような辞書文字作成を含め、ユーザであるチベット学者に辞書文字作成を行ってもらう部分が重要であることが分かった．このことにより、種々のタイプの認識対象文献に対応することを可能としている．辞書文字作成を前処理部に組み込んでいるために、ユーザであるチベット学者の操作を考慮したGUI（Graphical User Interface）機能をシステムに取り入れている．GUI機能を用いて文字認識している例を図８に示す．

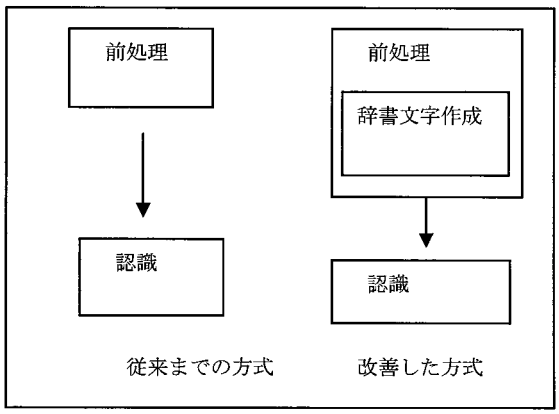


図７　従来までの文字認識と本方式の違い

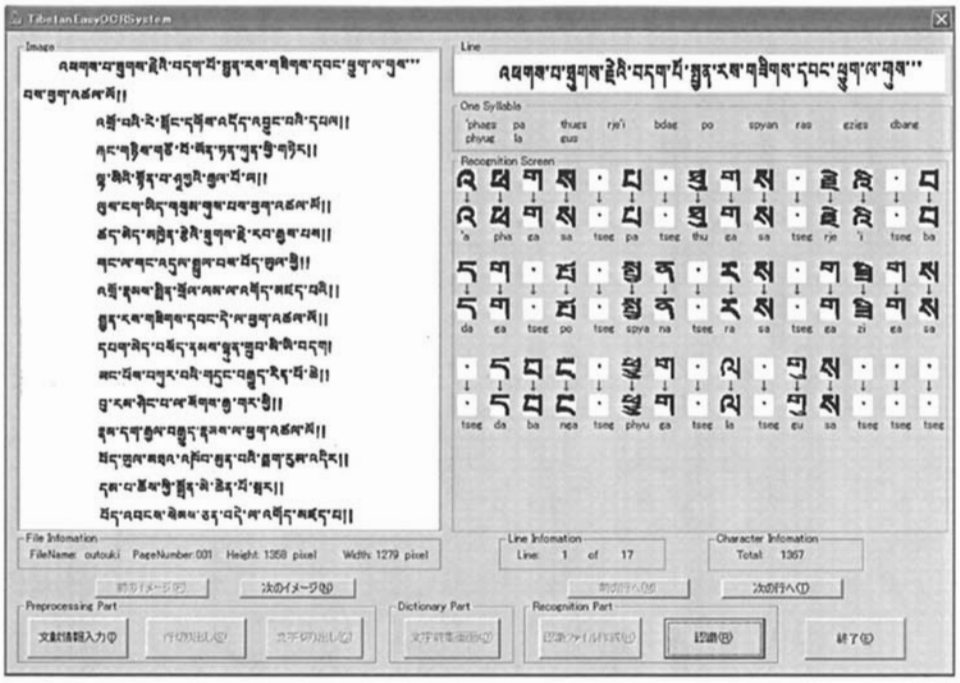


図８　GUI 機能を用いた認識画面
（上部のイメージデータの下に１音節表示もされている）

本方式において、認識対象とした西蔵文献１ページから３０ページ[20]までの認識対象文字28,954個に対して認識率 99.4%を実現することができた．オブジェクト指向設計法の考え方を積極的に取り入れることにより、ユーザであるチベット学者にとって実用可能な文字認識システムが実現可能となった．

４．まとめ

チベット文字認識を行なうとき、誤認識率の多い文字に対してオブジェクト指向類似辞書文字を用いることにより、誤認識した文字の改善した分、着実な認識率改善がなされ、有効性を確認してきた．本研究により、本来のオブジェクト指向設計法の目的である文字認識システムを活用するユーザであるチベット学者らに有効な文字認識システムが実現した．

謝辞

本研究を進めるにあたり、大変貴重な文献の提供ならびに種々アドバイスいただきました東北大学塚本啓祥名誉教授、磯田文名誉教授、大谷大学兵藤一夫教授に深謝いたします．

参考文献

１）塚本啓祥：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);磯田熙文：チベット文字の特色とコンピュータ利用について、ibid..

2）小島正美、川添良幸、木村正行：チベット文献の自動認識について、印度学仏教学研究、39巻2号、pp.207-211、(1991).
3）小島正美、川添良幸、木村正行：木版刷りチベット文献の文字自動認識の試み、情報知識学会誌、2巻1号、pp.49-62、(1991).
4）小島正美、川添良幸、木村正行：推論を用いたチベット文献中の文字自動認識、印度学仏教学研究、41巻1号, pp.158-161、(1992).
5）j. ランボー、M. ブラハ、M. プレメラニ、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).
6）Jacobson, I. : Object Oriented Software Engineering, Addison Wesley Publishing Company, (1992).
7）Martin, J.: Principle of Object Oriented Analysisi and Design, Englewood Cliffs(1993).
8）小島正美、川添良幸、木村正行：木版刷りチベット文献の自動認識、塚本啓祥教授還暦記念論文集、佼成出版社、「知の邂逅・仏教と科学」、pp.563-571,(1993).
9）小島正美、川添良幸、木村正行：木版刷りチベット文献中の文字特徴抽出、印度学仏教学研究、42巻、2号、pp.214-217, (1994).
10）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法によるチベット活字辞書を用いた類似文字認識、情報処理学会誌、36巻、 11号、pp.2611-2621,(1995).
11)Masami KOJIMA, Yoshiyuki KAWAZOE and Masayuki KIMURA : Automatic Tibetan Scripts Recognition by Computer, 7th Seminar of the Association for Tibatan Studies, Vol. 1, (1997).
12）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、日本西蔵学会会報, 43号, pp.31-38, (1998).
13) Masami KOJIMA, Yoshiyuki KAWAZOE and Masayuki KIMURA: Automatic Recognition of Tibetan Buddhist Texts by Computer, 1999 EBTI, ECAI, SEER & PNC Joint Meeting, pp.387-393, (1999).
14)山下康明、小島正美、木村正行：音節構造解析による活字チベット文字認識の高速化、情報処理学会「人文科学とコンピュータシンポジウム」、pp. 53-60, (1999).
15) 小島正美、川添良幸、木村正行：自己検証可能なチベット活字文献自動認識、情報処理学会「人文科学とコンピュータシンポジウム」、pp. 271-278, (2000).
16) 小島正美、川添良幸、木村正行：木版刷チベット文献1音節切り出し法について、印度学仏教学研究、51巻、 1 、pp. 342-346, (2002).
17)布宮千夏子、小島正美、川添良幸：オブジェクト指向設計によるチベット文字認識システム―認識前処理部について―、情報処理学会、CH-60, pp.25-32, (2003).
18) 小島正美、高木恒男、川添良幸、木村正行：効果的なチベット文字認識システム、情報処理学会「人文科学とコンピュータシンポジウム」、pp.9-16 (2006).
19)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).
20)王統記（蔵文）（rgyal rabs gsal ba'i me long）薩迦索南堅賛著　嘉賽阿旺洛桑，工布吉村編　民族出版社 (1981).