国家社科基金后期资助项目

藏文识别原理与应用

The Principles and Application of Tibetan Character Recognition

江荻　编著

商務印書館
创于1897 The Commercial Press
2012年·北京

图书在版编目(CIP)数据

藏文识别原理与应用/江荻编著.—北京:商务印书馆,2012
ISBN 978-7-100-08724-7

Ⅰ.①藏… Ⅱ.①江… Ⅲ.①藏语—文字识别—研究 Ⅳ.①TP391.43

中国版本图书馆 CIP 数据核字(2011)第 232378 号

ZÀNGWÉN SHÍBIÉ YUÁNLǏ YǓ YÌNGYÒNG
藏文识别原理与应用
江荻 编著

商 务 印 书 馆 出 版
(北京王府井大街 36 号 邮政编码 100710)
商 务 印 书 馆 发 行
印刷厂印刷
ISBN 978-7-100-08724-7

20 年 月第 版 开本 × 1/
20 年 月北京第 次印刷 印张
定价: 元

项目主编：
江荻

编　委：
周学文　龙从军
康才畯　严海林

谨以此书纪念伟大的藏文创始人，藏族文化的先行者
土弥·桑布扎(thu mi sambhotta)
To the memory of the great creator of Tibetan writing system
and the forerunner of Tibetan cultures
ཐུ་མི་སམ་བྷོ་ཊ,617—700, A.D.

国家社科基金后期资助项目
出版说明

后期资助项目是国家社科基金设立的一类重要项目，旨在鼓励广大社科研究者潜心治学，扶持基础研究的优秀成果。它是经过严格评审，从接近完成的科研成果中遴选立项的。为扩大后期资助项目的影响，更好地推动学术发展，促进成果转化，全国哲学社会科学规划办公室按照“统一标识、统一版式、符合主题、封面各异”的总体要求，组织出版国家社科基金后期资助项目成果。

全国哲学社会科学规划办公室

序　一

文字识别是模式识别的一个分支，也是计算机视觉智能的重要应用。文字识别是计算机通过光学影像自动识别印刷在纸面上的文字，实现大量文字信息的高速输入，在信息处理中可广泛地应用。

在中国，文字识别不仅需要识别汉字，还需要识别各种少数民族语言文字，由于这些语言文字各有特点，解决它们的识别问题除了需要应用一般的识别技术外，还需要针对它们的特点，发展许多专门技术，才能达到高效、高精度的识别效果。这样，民族文字识别技术就成为中国中文信息处理领域中的一门独立性很强的技术之葩。

人们曾经担心汉字难以进入计算机会成为中国信息化的“瓶颈”，但今天，汉字进入计算机已不是难题，各种键盘输入、语音识别输入、文字识别输入的技术和方法呈现百花齐放、百家争艳的可喜局面。就本书主题的文字识别而言，印刷体汉字识别、手写体汉字识别、联机手写汉字输入都已不同程度地逐步走向高效和精确，应用面已扩展至国民经济和人们生活的各个领域。在汉字识别技术的带动下，近十余年来，我国少数民族文字识别技术同样取得了蓬勃的发展，各有关单位吸收汉字处理先进理念和技术方法，开发出藏语、维语、蒙古语等印刷字体的识别技术和产品，把少数民族语言文字识别理论和技术推到一个新的高度，做出了许多重要创新。

此处我想强调几个技术侧面。首先，这本书讨论的藏文印刷字体识别跟汉字相似，属于大字符集识别系统，仅在字符的数量上就较之西方字母小字符集识别艰难得多。据近年发布的国家标准，按照预组合形成的藏文字符超过了 7200 多个；其次，藏文识别同样存在复杂的内包结构识别和多层次叠置结构识别等高难现象，有些转写梵文的藏文字符甚至达到 5 层之多；再次，藏文相似字符数量不小，也是需要采用高精度处理的对象，增加了处理的复杂性。如果与汉字识别相比，藏文自身结构特点也要求识别技术有更多创新。例如，藏文字符不等高，不仅单字符字母不等高，大量组合型字符叠置层数不同高度也不同；藏文文本包含大量非字母型字符，形成字符不

等宽现象，例如出现频率最高的音节字之间的分音点和表示停顿的单垂符（标点符号）宽度仅为某些字符的 1/3 或 1/4。从技术的观点看，开展藏文识别丰富了中文信息处理领域的文字识别理论，开拓了文字识别技术新领域。这里要特别感谢从事这一领域攻关的科学家，正是他们多年如一日的辛勤工作，克服了重重困难，取得了可圈可点的创新成果，为中文信息处理做出了重要贡献。

我很高兴为江荻博士等人的这部新著写序。在当前 21 世纪，中国作为复苏的和崛起的大国，我国科技工作者肩负着光荣而艰巨的历史使命。如果我们的视野能跨越文化的藩篱，善于整合不同文化的知识资源，可以做出很有价值的创新，本书就是一个很好的范例。应当指出，中国是一个多民族和多元文化的国家，我国的语言文字丰富多彩，不仅有“蒙、藏、维”还有“朝、彝、傣”等 100 余种语言和 30 余种传统文字，包括各民族语言文字在内的中文信息处理领域还有许多问题有待于我们去探索和解决。

综上所述，本书全面深入地论述了藏文文字识别，既能紧密地联系实践又能提升到理论高度，是一部优秀的民族文字信息处理著作。我衷心希望，今后有更多像本书这样的著作面世，使中国各民族文字信息处理技术共同走向成熟，基于这些自主技术，中国各民族将共同实现可靠、低成本的信息化之路。

中国工程院院士
中国中文信息学会理事长
倪光南　教授

2011 年 3 月 10 日

序　二

我国是统一的多民族国家，各民族优秀文化是中华民族灿烂文化的重要组成部分。胡锦涛总书记在2005年的中央民族工作会议上明确指出："支持少数民族优秀文化的传承、发展和创新。"推动我国多种民族文字信息化的发展，关系到中华民族文化的发展，对于维护国家民族团结、和谐发展有重要的意义。

藏族是我国主要少数民族之一，藏文化有悠久灿烂的历史，保留有大量历史记载的藏文典籍，是中华文明极其珍贵的文化宝藏。为了保护和弘扬藏族历史文化，非常需要将其数字化和信息化，藏文文字识别是解决藏文信息化的计算机输入"瓶颈"问题的重要和有效手段，因此藏文识别问题受到各方人士的极大关注。

非常高兴地看到江荻博士、周学文工程师等完成的专著《藏文识别原理与应用》即将出版，我被邀为此书写序，感到十分荣幸。《藏文识别原理与应用》一书是作者在长期深入的研究基础上完成的，其中包括了对藏文文字特点的深入分析，对文字识别的理论和方法的详细介绍，对藏文识别预处理、印刷体藏文识别，以及藏文识别后处理进行了仔细的讨论。这是我国出版的第一本有关藏文识别的专著，对广大关心藏文识别或文字识别的读者将有重要的参考价值。该书的出版也反映了我国学术界对少数民族文字识别技术研究的热情，包含了该书作者江荻博士等在藏文识别原理和应用方面的研究成果，我相信它的出版必将推动我国民族文字分析、识别、处理研究的进一步深入发展。

清华大学电子工程系
丁晓青　教授
2009年9月26日

前　言

人类学家说：人类的历史可分为“史前”时期和“有史”时期，前者指文字发明和使用之前，后者当然是有了文字和有了书面语言的记录。这样的分期可是把文字放在了人类文明史上重中之重的地位。随着社会从采集、狩猎、游牧、农耕、工业化发展到现代的电子化、信息化，相信当今世界没有哪个主体国家或者地区不使用文字了。

最早的文字诞生于约公元前 3500 年的两河流域，即苏美尔人的楔形文字。此后的数千年中，世界各地陆续出现各种文字系统，汉字至迟在公元前 1300 年的商代就已存在，考古发现的甲骨文如此成熟，不能不令人把它的发明上推数千年之久；美洲玛雅文字最早的碑刻遗物约在公元 328 年；古埃及象形文字甚至在公元前 3100 年就已出现。想象一下，跨越了这么久远的历史，文字给人类带来的好处，文字对人类文明的贡献，一直都是那么辉煌，还有哪种文明创造能超越它的伟绩？它把语言的传播、传承、思维、交际等几大功能发挥得如此淋漓尽致，空间上的传播和时间上的传承，让远隔千山万水的人们能够沟通，让今人能够读懂远古的先哲，的确神奇。

文字当然也随着时代发展，它的形式、载体、形态无论是勇往直前还是迂回曲折，都走到了新的时代、信息的时代、我们的时代。在这样一个追求效率的时代，聪明的人想着“偷懒”，敲键盘的活儿都要省掉，又发明了叫做 OCR（光学字符识别器）的玩意儿，这是一个奇特的装置。当人用智慧的双眼辨识手书铅印的文字时，他没有忘记让机器来代承他的辛劳。他努力锻造机器的“火眼金睛”，要它把千变万幻的图案转变为书面语言：这是“千、玖、恕”，不是“干、玫、怒”，那是“  ”，不是“  ”。上天保佑，人的历史又向前迈出了一步，“偷懒”大获成功。

回到本书的专题上来。史记藏文创制于公元 7 世纪，一个与中原内地鼎盛的唐朝同时代的吐蕃王朝时期。千余年来，藏文经历了多种字形字体的变化，而最典型、文献最丰富的要数历经千年逐渐规范的正楷书体（有头字），目前人们常用的类似雕版印刷而进一步规范的字体，也是本书讨论扫

描识别应用的字体。

这本书的具体内容留给读者阅评，略可概括的是，全书相对比较全面地介绍了藏文的字符分类和各类字形的特征，也详细叙述了藏文的识别模型和处理技术，对预处理、识别处理和后处理等标准识别程序均有讨论，还特别介绍了我们实验室开发的实验系统、清华大学开发的 TH-OCR 2007 多文种文字识别系统。相信本书在民族文字信息化处理蓬勃兴旺的今天，能对伟大祖国的民族文化和技术发展有所奉献。

本书的一个缺点是没有讨论雕版印刷藏文字体的识别，那是因为迄今还没有任何机构任何人开展过这方面的研究。鉴于此，设立在美国的西藏佛教资源中心（Tibetan Buddhist Resource Center）只好暂时采用扫描文档制作 PDF 文件的方式将大量雕版文献存储备用，拟待 OCR 技术开展之后再予识别。历史上的雕版印刷文献不知凡几，有如《布敦佛教史》、《雅隆教法史》、《王统世系明鉴》、《汉藏史集》、《贤者喜宴》、《米拉日巴传》、《萨迦班智达传》、《唐东杰布传》、《萨迦世系史》、《西藏王臣记》、《如意宝树史》、《土观宗派源流》、《青史》、《红史》、《新红史》等，还包括各个时代奉为瑰宝的《大藏经》，真希望这是学界和工程界下一个重要的攻关目标。

在本书出版之际，作为实验项目负责人，我要感谢项目团队的每个成员，他们兢兢业业、努力工作，都是有为的年轻人。还要感谢中国科学院倪光南教授、曹佑琦教授和清华大学丁晓青教授，倪教授是中文信息学会的理事长，曹教授是学会常务副理事长兼秘书长，他们热忱地鼎力支持少数民族语言文字信息处理事业，组织会议、举办讲座，对民族语文信息处理研究给予诸多支持。曹教授称赞几位作者首撰少数民族文字识别研究专著，提议请倪教授为本书写序，倪教授慨然惠允，使我们深感学会领导对藏语文信息处理研究的肯定和支持。丁教授是我国文字识别领域的权威专家，不仅开发了汉文识别，又组织开发了藏文、蒙古文和维吾尔文等文字识别系统。她应允为本书赐序，并欣然同意将她与王华博士合著的“多字体印刷藏文识别系统”附于本书，对培养新学，支持藏文和民族文字识别奉献之多，令人感慨。谢谢了！

江　荻

2009 年 9 月

于北京中关村

目　　录

第一章　绪　论

文字识别在现代学科分类中属于计算机模式识别与图像处理研究，以及各文字系统的应用领域。简单说，文字识别是利用光电原理将文字图形符号经过光电信号的处理转换为具有一定灰度值的数字信号，进而通过特征提取及匹配识别为计算机存储中可匹配的文字编码符号，这个过程人们通常称为光学字符识别 OCR(Optical Character Recognition)，所应用的装置称为光学字符识别发生器。

本书介绍藏文的识别研究和识别方法。由于文字识别方法需要建立确定的模式，在一定程度上依赖对识别对象的特征以及该文字的各种关联属性的认识，包括语言属性，所以本书还将讨论藏文文字的特点和语言的特点。

目前，藏文的识别虽处在起步阶段，国内外所开展的工作却已取得较好进展，除了理论和技术的探讨，也开发出可实用的产品。为了加强和推动这方面的研究和开发工作，本书有意尝试探索藏文光学字符识别的基础研究，并介绍目前藏文识别工作的进展。

1.1　藏文识别研究的背景

藏语是一种非常古老的语言。公元 7 世纪藏民族先民建立了吐蕃王朝，并创制了表征其独特文化的藏文。藏文典籍是一份恢宏的文化遗产，早期文献包括碑文、岩刻、敦煌石窟所藏藏文手卷、竹木简牍等。15 世纪永乐版《甘珠尔》是在南京刻印完成的，这项工作对后世影响极大，此后，木刻刊印的藏文文献在藏区空前兴盛起来，西藏、四川、甘肃、青海等各地较具规模的藏族寺院大量印刷藏文《大藏经》和其他佛教典籍，使得雕版印刷业全面发展起来，藏文文献也日益积累，数不胜数。

浩如烟海的藏文文献内容广泛，有为吐蕃赞普歌功颂德的传记，有记述

吐蕃君王与大臣的盟誓，还有吐蕃王国与中原唐朝的会盟祭祀，以及一千多年来的各类古代历史记载、佛教经典编译以及民间神话传说等等。藏文文献是我国除汉文之外，历史最悠久、文献最丰富的语言文化遗产。

正是由于这个原因，藏文古籍、文本的电子化和信息化处理成为当代社会所关注的课题。20 世纪 80 年代初期，汉字进入计算机已取得成功，开发出 CCDOS 这类汉文操作系统和相应的录入、编辑、排版、打印等应用软件。也就是在这个时期，包括少数民族语言文字处理在内的“计算机中文信息处理”概念也逐步形成。计算机学界和社会上兴起了一股中文计算机热潮，高校也开设了中文信息处理课程。随着 1981 年中国中文信息学会的成立以及学术活动的开展，1984 年从事少数民族语言文字计算机处理的专家学者也举办了第一届学术讨论会（呼和浩特，1984.10），并于 1985 年 10 月成立少数民族语言文字信息处理专业委员会，当时就制定了少数民族语言文字计算机处理系统、编码字符集、字模点阵及数据集和输入键盘布局等标准的研究和开发方向。此后的 20 年间，藏文信息处理的主要工作是建立藏文操作系统或与英文、汉文兼容的操作系统或系统工作平台，这项工作还包括了藏文字符编码标准、藏文编码输入、藏文编辑排版系统等。

具体来说，藏文计算机应用的开发经历了一些独特的过程。早在 20 世纪 80 年代初期，中国社会科学院民族研究所张连生（1983）尝试用计算机进行藏文词汇排序，他根据藏语专家于道泉先生（1982）提出的数码代替藏文字母的方法实验，设计出一种可行的排序方法，此后又在美国伊利诺伊大学利用 Plato 计算机实现了藏文字符输入、显示和输出的藏文字处理系统，开启了藏文文字计算机处理之先河。此后，航天部 710 所罗圣仪（1986）也在微机上实现了初级的藏文字处理系统。80 年代中后期，有关机构和人员开发出与汉英文兼容的藏文操作系统，例如俞乐（1984）等报道了利用 Basic 语言在 Victor 9000 微型计算机上开发了藏文字处理系统，俞汝龙、赵晨星、毛继祖等开发出藏文 TCDOS 系统，这个系统可算是最早投入使用的藏文操作系统，是在中文操作系统 CCDOS 基础上开发的，并发明了最早的藏文输入方法（俞汝龙，1987）。其后还有于江苏、葛小冲（1988）提出的藏文信息处理方案和 ZWDOS 系统，熊涛（1988）等开发的藏汉西文处理系统。到 90 年代，西藏大学尼玛扎西（1992）等开发出 TCE 藏汉英信息处理系统，彭寿全（1994）等开发出外挂式藏汉英混合处理系统及东洲藏卡系统，熊涛、于洪志等人合作开发了可挂接在金山 WPS 下的藏文轻印刷系统——兰海藏文系统（江嘎，1999）。随着计算机技术的发展，进一步开发的主要是基于 Windows 的藏文系统。西北民族学院于洪志、戴玉刚等 2000 年开发出基

于 Windows 的同元藏文字处理软件(赵晓清,2008),青海师范大学开发出基于 Windows 的班智达藏文字处理软件(才藏太,2005),西藏大学尼玛扎西、洛藏等人 2001 年开发了火狐藏文字处理软件(陈玉忠,2003)。

在工业化实现方面,1989 年,中国藏学研究中心在整理出版"中华大藏经"(藏文版)工作的推动下,与华光集团合作研制藏文计算机排版系统,该系统 1993 年正式应用于藏文《大藏经》排版,取得巨大成功。藏文版华光藏文软件有 9 种字体:6 种正楷和 3 种草写体,有独立的编辑器,也可在 Word 里进行编辑,可编辑藏、梵、汉、英四种文字,可处理公文、报刊、藏文经卷等(赵晓清,2008)。随后,北大方正也研制出了藏文书报排版系统,在全国藏文书报印刷方面占据极大市场份额(赵晓清,2008)。到 21 世纪初,这些系统进一步从 DOS 操作系统迁移到基于 Windows 的系统,迄今一直是我国藏文印刷出版业的主要技术基础和应用产品。可以说,以上各类研究和应用项目铺垫了藏文信息处理的基础,具有筚路蓝缕以启山林之功。

1.2 藏文识别研究的技术基础

围绕藏文输入、输出和信息处理的实践,以及藏文操作系统和排版印刷系统的开发,藏文计算机处理技术获得飞跃发展,其中最突出的进展表现在藏文编码标准、藏文字体标准、键盘输入技术、藏文排版和印刷技术、藏文网络实现技术。这些技术的发展又在一定程度上推进了藏文文本信息处理的进展,也促动了藏文语音识别、文字识别的研究。目前来看,文本研究方面的进展主要有:藏文电子词典建设与索引排序方法、文本资源建设及文献分类、文本统计和熵值计算、分词方法和识别算法,以及藏文的拉丁转写方案等,这些研究都为古籍和文本的信息化奠定了基础。以下我们结合文字识别简略叙述作为藏文识别相关技术基础的研究,这些研究对于藏文文字识别来说都起着关键的作用。

藏文编码标准的建立是藏文信息处理和文字识别最重要的基础支撑技术。1997 年国际标准化组织(ISO/IEC)通过了中国提出的藏文编码字符集方案,中国国家标准委员会 1998 年 1 月也颁布了藏文编码国家标准,使藏文成为 ISO/IEC 10646 通用多八位编码字符集的重要组成部分(国家质量技术监督局标准,1998)。

建立在 ISO10646-1 基本平面 00 组 00 平面的藏文《基本集》(UCS 的基本多文种平面,机内码 0F00-0FBF,占用 192 个码位)提供了 168 个编码

字符。其中，辅音字符 41 个、组合用辅音字符 36 个、元音符号 15 个、变音符 13 个、数字符号 20 个，其他是篇章起始符、标点符号、装饰符号等等。后来在国际统一编码(Unicode3.0)之后，藏文编码空间进一步扩展，获得 256 个码位，编码空间是 0F00－0FFF，增添的各种编码达到 201 个(江荻，龙从军，2010)。藏文标准编码的实现为藏文识别构建了文本数据对应技术，通过扫描识别的图形符号可以顺利转化为编码显现形式，利于对扫描识别结果进行判断和取舍，也为进一步的后处理构建了字、词、句的语境分析提供了技术支撑。

近年值得一提的发展是，为了解决藏文字符组合中出现的多种技术难点，例如叠置引擎技术，国内藏文编码专家进一步提出藏文大字符集的编码观点，陆续提出建立预组合字符集方案，目前《信息技术 信息交换用藏文编码字符集 扩充集 A》已由国家标准技术委员会作为国家标准发布，共有 1536 个垂直预组合字符。例如：ཀྱི(kyi)、ཀྲི(kri)、ཀྱུ(rku)等。《扩充集 A》安置在 GB13000 的基本多文种平面专用用户区，其编码位置是 F300～F8FF，共占用 1536 个编码位置(国家质量技术监督局标准，2007)。另一项标准《信息技术信息交换用藏文编码字符集 扩充集 B》也于 2009 年发布，共收入 5669 个垂直预组合字符。在 GB13000 专用平面 0F 平面上的编码，共占用从 000F0000～000F1624 位置的 5702 个编码位置。

另一项作为藏文识别技术基础的研究是关于藏文文本字符和结构的统计研究。迄今，藏文专家已开展了部分藏文结构的统计分析，提出藏文 25 种字结构形态(参见表 1.1)及其出现比率(江荻，1998)。根据这项研究，藏文“基字＋后加字”结构占全部结构出现比率的 31％，单纯“基字”结构的藏文字占 25％，“前加字＋基字＋后加字”结构占 12％，仅此三项共计占全部结构的近 70％以上。而出现最少的结构与出现最多的结构之间相差了 1000 倍以上。表 1.1 列出了有关数据。

表 1.1　藏文结构的动态比率数据

	结构类型	比率％		结构类型	比率％
1	基＋后	30.914	9	前＋基	1.777
2	基	25.235	10	上＋基	1.693
3	前＋基＋后	12.109	11	前＋基＋下＋后	1.333
4	基＋下＋后	7.134	12	前＋基＋后＋重	1.283
5	基＋后＋重	4.549	13	基＋下＋后＋重	1.106
6	上＋基＋后	4.501	14	上＋基＋后＋重	1.083
7	基＋下	2.222	15	上＋基＋下	0.682
8	上＋基＋下＋后	1.963	16	前＋基＋下	0.657

续表

	结构类型	比率%		结构类型	比率%
17	前＋上＋基＋后	0.647	22	前＋上＋基	0.110
18	前＋上＋基＋下＋后	0.328	23	前＋上＋基＋下＋后＋重	0.066
19	前＋基＋下＋后＋重	0.225	24	前＋上＋基＋下	0.060
20	上＋基＋下＋后＋重	0.183	25	基＋下＋下	0.024
21	前＋上＋基＋后＋重	0.115			

目前，藏文识别领域有两种识别策略，一种认为应采用预组合整字识别方法，即基字与上、下加字及元音叠置的字符可作为整体模式识别对象(Ding，Wang，2006)，另一种认为可按多字母或切分出预组合字符中的字母构件来识别(王维兰，1999)。根据以上结构数据统计，后一策略要解决约30%预组合字符的构件切分问题，考虑到组合种类多，叠置层次复杂，例如上＋基、基＋下、上＋基＋下等，这并不是一项易行的方案。前一策略把切分处理的难点转化为前期词典数据，提高识别处理的效率。所以，了解藏文结构及其统计数据对制定识别方案有重要指导作用。

除了藏文结构的分析与统计外，相关的统计还包括藏文字符和藏字(音节字)出现频率的统计，每个结构位置上字符出现概率统计，例如ག(g-)、ད(d-)等5个前加字数据统计和བ(-b)、མ(-m)等后加字数据统计(江荻，1998)。关于藏文音节字的长度统计，也有不同统计方法和不同的结论。例如以现代藏文词典静态数据统计获得藏文音节字平均长度为3.678字符(江荻，1995)。另一项统计计算了藏文音节字占据的编码位置长度，以部分《丹珠尔》文本动态统计的结果是，藏文音节字平均编码位置长度(除去音节点)为2.0(严海林，2005)。

中国国家标准藏文字体(GB16960－1997)和藏文电子词典的开发也是藏文识别重要的技术支撑及应用平台。藏文字体繁多、形体各异，因此设计多字形计算机用藏文字体也是开发藏文OCR系统的重要内容。目前各种藏文计算机系统汇集的字体不一，多寡不一，据估计约有近20种不同藏文计算机用字形，例如，北大方正藏文系统中已有七种字体：白体、黑体、标题黑、新白体、新黑体、竹体、美术体。相关内容可参见北大方正藏文排版系统的产品附录“藏梵文输入法使用手册”。

关于藏文电子词典的开发，目前已有多项报道。词典的规模也已达到相当规模，最早的藏文电子词典由中国社会科学院民族学与人类学研究所建立，首先建立的词典是以口语为主的词典，收词3万余条(江荻，1995)，其后又建立了以藏语分词以及句法分析为目的的语法信息词典，每个词条附加了多项词法和句法属性信息，并且添加了词法和句法实例。例如藏语词法的特征