手写藏文字符识别研究

黄鹤鸣　马龙龙　赵维纳　著

科学出版社

手写藏文字符识别研究

黄鹤鸣　马龙龙　赵维纳　著

科学出版社
北京

内 容 简 介

本书主要介绍脱机手写藏文字符识别和联机手写藏文字符识别研究方面的最新进展，在简要介绍藏文信息处理、模式识别、数字图像处理等方面必要内容的基础上，重点介绍了手写藏文字符样本库的构建、脱机手写藏文字符识别方法、基于统计的联机手写藏文字符识别方法、融合统计和结构特征的联机手写藏文字符识别方法以及藏文字符的计算机排序等内容。在手写藏文字符样本数据库构建、预处理技术、特征提取、分类器设计以及后处理等方面进行了探索性研究，提出了一些符合藏文文字特点的新方法。

本书内容新颖，实用性强，理论与实际应用紧密结合，对从事文字识别研究的工作者特别是对从事少数民族文字识别研究的工作者提供参考，同时也为想了解手写藏文字符识别技术最新进展的科研工作者和工程技术人员提供较全面的参考。

图书在版编目（CIP）数据

手写藏文字符识别研究 / 黄鹤鸣，马龙龙，赵维纳著. —北京：科学出版社，2016.3

ISBN 978-7-03-047576-3

Ⅰ. ①手… Ⅱ. ①黄… ②马… ③赵… Ⅲ. ①藏语－手写字符识别－研究 Ⅳ. ①TP391.43

中国版本图书馆 CIP 数据核字(2016)第 046610 号

责任编辑：余 丁 赵艳春 / 责任校对：胡小洁

责任印制：徐晓晨 / 封面设计：迷底书装

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京京华虎彩印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 3 月第 一 版 开本：720×1 000 B5

2016 年 3 月第一次印刷 印张：12 1/4

字数：230 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

前　言

模式识别诞生于20世纪20年代，随着计算机的出现以及人工智能的兴起，模式识别在20世纪60年代迅速发展成为一门学科。模式识别的理论和方法在字符识别、生物身份认证、人脸识别、语音识别、说话人识别、信息检索、数据挖掘等方面得到了广泛应用。

文字识别是模式识别的一个重要研究内容，自20世纪60年代发表有关文字识别的第一篇研究论文以来，经过许多国家大量研究人员的不断努力，在汉字识别、英文字符识别以及数字识别等方面积累了丰富的研究成果。但有关手写藏文字符识别的报道较少，因此，本书重点介绍了脱机手写藏文字符识别和联机手写藏文字符识别方面的最新研究进展，希望能对从事文字识别研究的科研工作者特别是对从事少数民族语言文字识别研究的科研工作者提供较理想的参考，同时也为想了解手写藏文字符识别技术最新进展的研究者和工程技术人员提供较全面的参考资料。

本书在基本概念、基本理论和基本方法的论述上力求精练简洁、注重理论联系实际。本书共8章。第1章在简要介绍藏文文字特点以及藏文信息处理技术研究现状的基础上，较全面地回顾了藏文字符识别的研究现状。第2章和第3章分别简要介绍了数字图像处理和模式识别方面的基础知识。第4～7章是本书的重点，主要包括手写藏文字符样本库的构建、脱机手写藏文字符识别方法、基于统计的联机手写藏文字符识别方法以及融合统计和结构特征的联机手写藏文字符识别方法等方面的最新研究进展。后处理是文字识别系统的一个重要环节，而藏文字符的计算机排序是藏文字符识别后处理的基础，因此，第8章较详细地介绍了有关藏文字符计算机排序的研究成果。

本书在介绍手写藏文字符识别最新研究成果前补充了藏文信息处理技术研究现状、数字图像处理基础和模式识别基础等方面的内容，使本书在内容上更加完备，大部分读者在不必参考其他文献的情况下，能较顺利地阅读本书。

模式识别对数学有较高的要求，为了使读者更好地理解算法的本质，本书并没有刻意回避数学公式，要较好地理解本书读者需要在微积分、线性代数和概率论等方面具有较好的基础知识。

本书是由青海师范大学计算机学院黄鹤鸣、中国科学院软件研究所马龙龙和青海师范大学计算机学院赵维纳三位教师共同努力完成的，每位作者所撰写的内容基本上反映了该作者在近几年所取得的研究成果。其中，黄鹤鸣撰写了第1章的

第 1、3、4 节，第 2 章，第 3 章，第 4 章的第 1、2 节，第 5 章，第 8 章；马龙龙撰写了第 4 章的第 3 节，第 6 章，第 7 章；赵维纳撰写了第 1 章的第 2 节。黄鹤鸣负责全书的统稿工作。

本书的研究内容先后受到国家自然科学基金地区项目（61462072）、国家自然科学基金青年项目（61202220）以及藏文信息处理教育部重点实验室开放课题等项目的资助，特此感谢！同时，感谢蔡晓娟、格日扎西等同学所做的大量而细致的校对工作。

本书的出版恰遇青海师范大学华诞 60 周年，祝愿青海师范大学再谱华章！

由于时间仓促且水平所限，书中难免有不足之处，敬请广大同行和读者批评指正，相关内容请发电子邮件到 huanghm@qhnu.edu.cn，以便再版时补充和修改。

目　录

第1章 绪 论

文字识别是模式识别的一个重要研究内容，自20世纪60年代发表第一篇有关文字识别的研究论文以来，经过许多国家大量研究人员的不断探索和共同努力，在汉字识别、英文字符识别以及数字识别等方面积累了丰富的成果。但截至目前，除了在单体印刷体藏文字符识别方面的研究较成熟外，手写藏文字符识别方面的研究几乎空白，特别是脱机手写藏文字符识别和联机手写藏文字符识别更是几乎无人涉及。本章首先简要介绍了藏文文字和藏文信息处理技术，之后介绍了手写汉字的研究现状和手写藏文字符的研究现状，最后指出从事手写藏文字符识别研究的必要性以及可行性。

1.1 藏文文字简介

藏族是我国古老的民族之一，主要分布在西藏、青海、甘肃、四川、云南五省区。另外，尼泊尔、不丹、印度、巴基斯坦等国也有一部分藏族。藏语主要有卫藏、康、安多三大方言区。经过千多年的演变三大方言间的差异较大。尽管各地方言各异但文字仍然是统一的，书面语通用于整个藏族地区。

藏文是藏族人民的书面交流工具，藏文历史之悠久在国内仅次于汉文。史载藏文字的创造源远流长，距今有近1400年的历史。关于藏文的起源在当代学术界仍是一个有争议的问题。有些学者认为藏文源于印度，藏文源于印度说的主要观点是：公元629年，藏王松赞干布的内相吐弥·桑布扎在印度获得学业返藏后，奉命与桑阳顿涅等其他著名学者历时三年创制藏文。另有一些学者认为藏文源于象雄，藏文源于象雄说的主要观点是：在松赞干布之前就有藏文，早期的藏文由西藏本土教苯教的辛饶创制，吐弥·桑布扎习字创制藏文[1-3]。目前多数学者倾向于藏文源于印度的观点，这可能和松赞干布大力推广佛教有关。随着佛教在藏族地区的盛行和苯教的衰落，使藏文源于印度的说法占据主导地位。但无论是坚持藏文源于印度说的学者还是坚持藏文源于象雄说的学者，都承认吐弥·桑布扎在藏文的创制、规范以及发展方面所起的重要作用，藏文的历史也是从吐弥·桑布扎时代开始算起。总之，藏文是由古藏文演变为今用藏文，也是外来梵文影响的结果。

每种文字在创制之初都不可能十分完善，需要在实际运用中不断总结经验，逐步规范和完善。同样，藏文也有一个发展完善的过程。据史籍记载，藏文分别于公

元 8 世纪中叶至 9 世纪初叶、公元 9 世纪中叶和公元 11 世纪初叶进行过 3 次较大规模的厘定和规范。

和汉文、英文一样，藏文的书写走向是不同行从上到下书写而同一行从左向右书写。藏文字体大致可分为吾坚体和吾美体两大类，见图 1.1。图 1.1 中，第一行是手写吾坚体句子，第二行是手写吾美体句子。吾坚体书写规范相当于汉字的楷书，吾坚又称为有头字，常用于印刷、雕刻以及正规文书的书写。吾美又称为无头字，相当于汉字的行书，主要用于手抄记录。吾美又进一步可细分为笔画短促的“粗”，笔画细长的“粗仁”以及笔画转折处棱角突出的行书字体“珠”；还有一种书写迅速、笔画简化的草写体“丘”，适合速记，其形体和印刷体相比差别很大。一般而言，卫藏地区盛行吾美体，安多地区盛行吾坚体。

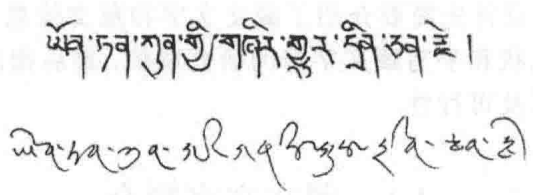

图 1.1　藏文的吾坚体和吾美体示例

1.1.1　本地藏文

藏文有两个字母表：本地藏文字母表(Local Tibetan Alphabet)和梵音藏文字母表(Transliterating Tibetan Alphabet)。本地藏文字母表有 30 个辅音字母和 4 个元音字母，其中，30 个辅音字母为：ཀ་ཁ་ག་ང་། ཅ་ཆ་ཇ་ཉ། ཏ་ཐ་ད་ན། པ་ཕ་བ་མ། ཙ་ཚ་ཛ་ཝ། ཞ་ཟ་འ་ཡ། ར་ལ་ཤ་ས། ཧ་ཨ། 。4 个元音字母为：ི་ུ་ེ་ོ། 。

通常意义上的藏文是指由本地藏文字母表中的 34 个字母通过纵向和横向拼合而成的文字。按照一套完整而严格的规则，藏文字母通过上下组合构成字符，字符通过左右组合进一步构成音节。为了简洁和美观，有些辅音字母在组合过程中会产生变形，例如：字母 ར 作为上加辅音时变形为 ྲ，而作为下加辅音时变形为 ྱ；字母 ཁ 作为下加辅音时变形为 ྐ，字母 ཡ 作为下加辅音时变形为 ྱ 等。另外，元音字母 ི、ུ、ེ 和 ོ在和辅音组合时只出现相应的元音符号 ི、ུ、ེ 和 ོ，而字母 ཨ 省略。

可以看出，藏文音节是字母的二维组合，例如：藏文音节 བསྒྲིགས 是由基本辅音、前加辅音、上加辅音、下加辅音、元音、后加辅音和又后加辅音七个字母构成，如图 1.2 所示。然而，并不是所有的藏文音节都有七个字母，对于多数音节而言，除基本辅音外，其他位置上的字母都可能空缺。特殊情况下，一个字母就是一个音节，这类似于有些汉字仅由一个笔画构成。

图 1.2 一个较典型的藏文音节

1.1.2 梵音藏文

本地藏文的创制能实现世俗的政治、文化、经济活动中信息交流的需求，但自佛教传入藏区后，随着势力的不断扩大，逐步在藏区形成了政教合一的政治制度，佛教影响着藏族地区的政治、文化、经济、社会等各个方面。为了能够准确地转写用梵文书写的佛教经典，专门创制了一个和本地藏文在字母表、拼写规则等方面差异较大的梵音藏文。在梵音藏文的辅助下，梵文和藏文之间能够实现一一对应：任何一个梵文句子都能唯一地转写成藏文句子；反之，任何一个藏文句子都能唯一地转写成相应的梵文句子。

梵音藏文有 34 个辅音字母和 16 个元音字母。其中，34 个辅音字母为：ཀ་ཁ་ག་ང་།ཅ་ཆ་ཇ་ཉ་།ཏ་ཐ་ད་ན་།པ་ཕ་བ་མ་།ཙ་ཚ་ཛ་ཝ་།ཞ་ཟ་འ་ཡ་།ར་ལ་ཤ་ས་ཧ་ཀྵ་།ཀྲ།。16 个元音为：ཨ་ཨཱ་ཨི་ཨཱི་ཨུ་ཨཱུ་ཨྲྀ་ཨཷ་ཨླྀ་ཨཹ་ཨེ་ཨཻ་ཨོ་ཨཽ་ཨཾ་ཨཿ。

梵音藏文和本地藏文之间的差异，主要体现在以下几个方面。

(1)字母表不同。如图 1.3 所示。虽然有些字母同时属于本地藏文字母表和梵音藏文字母表，但有些字母只出现在本地藏文字母表中，而有些字母只出现在梵音藏文字母表中。

(2)同时属于两个字母表中的字母在不同字母表中的顺序不同。例如：字母 ཡ、ར、ལ、ཤ 同时出现两个字母表中，这几个字母在梵音藏文字母表中的顺序为 ཡ、ར、ལ、ཤ，而在本地藏文字母表中的顺序为 ཤ、ཡ、ར、ལ。

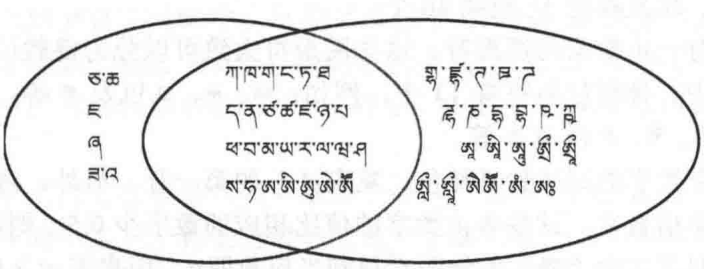

图 1.3 本地藏文字母表(左)和梵音藏文字母表(右)间的集合关系

(3)字母组合成字符或音节时的约束条件不同。一个藏文音节有一个基本辅音，

其他辅音根据它们相对于基本辅音的位置不同分别称为前加辅音、上加辅音、下加辅音、后加辅音以及又后加辅音等。基本辅音可以是 30 个辅音字母中的任何一个，一旦基本辅音确定后，其余位置上的辅音均会受它的制约。和本地藏文中字母间的这种很强的约束关系相比，梵音藏文没有前加辅音和后加辅音的概念，并且辅音之间可以随意组合。这种随意组合主要体现在以下三个方面。

①辅音字母间的纵向随意组合，例如：ཀྐ、ཀྵ、ཀྱ、དྲི、ཏྭ、ཏྣ、སྟྲ、ཐྱ、སྨྲ、ཙྭ、སྩྭ；

②辅音字母间的横向随意组合，例如：ཀཀཀཀ、ཀདཀ、ཀཀཀ、ཁངངཁ；

③字符间的横向随意组合，例如：ཀཀི、ཀཁ、ཀཆ、ཀཀ、ཀིཀི、ཀིཆི、ཀྐྱྀ。需要说明的是，此处的随意是相对于本地藏文音节中字母间规范且严谨的约束关系而言，并不是说任何梵音藏文字母间的随意组合都有意义。

本地藏文音节间有音节分隔符而梵音藏文中没有音节分隔符。英文用空格来分隔两个相邻的单词，类似地，本地藏文用音节分隔符“·”来分隔相邻的两个音节，见图 1.4 的第一行；而两个相邻的梵音藏文字符间没有任何分隔符号，见图 1.4 的第二行。

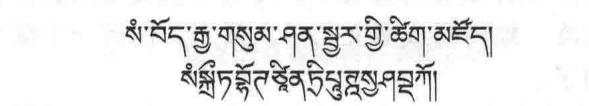

图 1.4　本地藏文和梵音藏文句子示例

1.1.3　藏文符号

除了字母外，构成藏文文字的元素还有 40 个标点符号、24 个图形符号以及 20 个藏文数字。

公元 7 世纪吞弥·桑布扎在创造藏文时，就开始应用音节分隔符“·”、句子符“།”、云头符“ ”等标点符号。后经历代添加，逐步形成了种类较多功能较齐全的藏文标点符号集。藏文中的标点符号主要有语音符、音节分隔符、字形区别符、首字符、注释符、层次符等 11 类共 40 个。

藏文中也有一定数量的图形符，这些图形符大致可以分为佛教仪轨符和天文历算符两类。其中，佛教仪轨符有 13 个，例如：卐、✣、☸ 以及 ❀ 等；天文历算符有 11 个，例如：☾、☽、☄ 以及 ☊ 等。

藏文中十个数字的写法独具特色，见表 1.1 的第一行。另外，藏文还有十个特殊的数字称为半值数字。这些半值数字的值比相应的数字少 0.5。例如：在数字四 ྔ 的中间加上类似汉文中“撇”的笔画 ྀ 得到半值数四 ྔྀ，用来表示 3.5。这些半值数字只能出现在个位上。例如：藏文数字 ྕྔྀ 表示 43.5 等。藏文中的十个半值数字见表 1.1 的第三行。

表 1.1 藏文数字和半值数字

藏文数字	༡	༢	༣	༤	༥	༦	༧	༨	༩	༠
阿拉伯数字	1	2	3	4	5	6	7	8	9	0
半值数字	༱	༲	༳	༴	༵	༶	༷	༸	༹	༺
阿拉伯数	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	−0.5

1.1.4 藏文字符集标准

要进行手写藏文字符识别研究，首先需要解决选择多少个字符进行识别以及选择哪些字符进行识别的问题，而这两个问题的答案应该以藏文字符集的国家/国际标准为依据。另外，建立和发布藏文字符集标准的曲折历史大致上反映了藏文信息处理技术的发展历史。因此，下文简要介绍藏文字符集的基本集、扩展集 A 以及扩展集 B。

1. 藏文字符集基本集

制定藏文字符集标准的目的是为了在计算机上实现藏文字符的正确表示、传输、交换、处理、存储、输入及显示。构成藏文的基本单位是字母，通过字母间纵向或横向组合形成藏文音节，音节之间组合形成词，从而进一步形成句子以及篇章。当然，在句子以及篇章中会用到标点符号、图形符号以及藏文数字。因此，如果能对藏文字母和基本符号进行编码就可以实现对所有藏文文档的编码，从而实现对藏文文档的计算机处理。

1997 年以前，藏文没有统一的字符集标准也没有统一的字符编码标准，研发藏文软件的各个公司采用各自的字符集标准和字符编码方案。各自为政的字符集标准和编码方案使得不同方案的用户间无法共享数据，并且由于国内各研发机构的编码方案大多借用汉字编码字符集 GB 2312 中的编码(Code point)为藏文字符编码，这使得藏、汉文混排时乱码现象不可避免[4]。

1997 年 7 月，ISO/IEC 通过了以我国提案为主的藏文字符集和编码方案，将藏文中的 193 个基本字母和符号包括辅音字母、元音字母、语音符、数字、标点符号、图形符号等收录到 ISO 10646 中。这些字符位于 Unicode 编码空间的基本多文种平面(Basic Multilingual Plane，BMP)，编码范围是 0F 行即 U+0F00-U+0FFF[5]。同年，我国发布了相应的国家标准 GB 16959-1997《信息技术 信息交换用藏文编码字符集基本集》[6]。

ISO 10646 采用动态组合方式为藏文字符进行编码。动态组合编码方式中，一个组合字符对应的不是一个独立的编码而是一个编码串，这个编码串由各个构成字母对应的编码组成。例如：组合字符 ཁྲ 对应的编码串是 0F66 0F90 0FB2 0F74，它们依次是字母 ཁ、ར 的不占位形式、ྲ 以及 ཱ 的编码。

简而言之，藏文字符集基本集是 ISO 10646 的一个子集，这个字符集的编码方

案是利用 195 个(包括后来增加的两个图形符号)基本字符的编码动态地给出所有组合字符的编码，从而实现对全部藏文的计算机处理。

ISO 10646 采用动态组合方式实现对全部藏文的计算机编码，具有以下几方面的优点：①动态组合编码方式体现了藏文字母组成字符以及音节的过程；②采用动态组合编码方式仅需对藏文中的基本字母和符号赋予适当的编码就能实现对全部藏文文档的计算机处理，因此，所需的 Unicode 编码很少，目前只使用了 195 个码值；③藏文字符的 Unicode 编码都在 BMP 平面，更容易得到主流厂商和技术的支持；④动态组合是藏文字符集国际标准推荐的编码方式，已得到 ISO 和 Unicode 技术委员会(Unicode Technical Committee)这两个信息技术领域国际标准权威发布机构的认可，将逐步取代其他编码方式。

2. 藏文字符集扩展集

藏文字符集国际标准的发布使藏文成为我国少数民族语言文字中第一个拥有字符集国际标准的文字，结束了藏文没有编码字符集标准的历史。但在计算机上实现藏文字符的动态组合需要OpenType字库和Unicode文字处理器(Unicode Script Processor，USP)的支持。这两个技术在一段时间内的相对滞后迫使藏文软件开发公司继续沿用各自此前的编码方案。

为了及早结束藏文字符编码各自为政的混乱局面，我国于 2006 年 10 月颁布了 GB/T 20542-2006《信息技术 藏文编码字符集 扩充集 A》[7]。扩充集 A 收录了 1536 个组合字符，位于 Unicode 编码空间 BMP 平面的用户专用区(Private User Area，PUA)，编码范围为 U+F300-U+F8FF。在此基础上，我国于 2008 年又颁布了 GB/T22238-2008《信息技术 藏文编码字符集 扩充集 B》[8]。扩充集 B 共收录了 5702 个组合字符，位于 Unicode 编码空间的 0F 辅助平面，编码范围为 U+0F0000-U+0F1625。

扩展集 A 和 B 为每个藏文组合字符都赋予了独有的编码，例如，组合字符  的编码为 U+F367，这种编码方式被称为静态组合。虽然静态组合编码方式有效地回避了动态组合编码方式在字符显示方面遇到的困难，但同时存在一些不可避免的缺点：①扩展集 A 将 1536 个藏文字符放置在任何机构和个人都能随意使用的 PUA 区，为藏文字符留下了乱码的隐患；②Unicode 技术委员会曾明确指出，PUA 区中的字符不会得到 Unicode 标准的支持，因此，采用静态组合编码方式的扩展集 A 和 B 永远无法上升到国际标准的层面；③用静态组合方式为藏文编码需要用到基本集、扩展集 A 和扩展集 B 中的编码，但基本集和扩展集 A 中的编码是两字节的，而扩展集 B 中的编码是三字节的，编码长度的不统一会为软件开发带来许多不必要的麻烦。

令人欣慰的是，2003 年 4 月，微软推出第一个藏文 OpenType 字库 Microsoft Himalaya；同年 11 月，Windows 的 USP 开始全面支持藏文字符的动态组合，也就

是说从此 Windows 操作系统开始全面支持完全基于 ISO 10646 的藏文字符编码方式。对 Linux 操作系统来说，桌面环境 KDE 3.2 和 GNOME 2.8 分别于 2004 年 2 月和 2004 年 9 月开始全面支持藏文字符的动态组合，这意味着在 Linux 最普及的两个桌面环境上完全基于 ISO 10646 的藏文信息处理已成为可能[9]。

藏文字符集基本集、扩展集 A 以及扩展集 B 共收录藏文字符和图形符号 7433 个。由于本文讨论藏文字符的识别问题，不必考虑为了实现藏文字符的动态组合而收集的 74 个组合用字符。因此，藏文字符识别中，需要考虑的全部识别对象为 7359 个字符和图形符号。

1.2 藏文信息处理技术

自 20 世纪 80 年代初以来，一批学者和专家致力于藏文信息处理的研究，至今已有 30 多年的历史。

从文字和语言的角度，藏文信息处理技术可以分为字处理信息技术和语言信息处理技术。字处理信息技术主要研究字符编码、输入法、操作系统、排版系统、字库等文字层面的信息处理，而语言信息处理技术则主要指藏文文本的信息化处理，主要研究机器翻译、文本校对、信息检索、文本生成、文字识别等内容。相对而言，语言信息处理技术的研究范围更为广泛。

1.2.1 藏文操作系统

用计算机处理藏文，需要在计算机中实现藏文的输入、输出和显示，这需要对藏文字符集进行编码、制作藏文字库、研发藏文输入法软件。

20 世纪 80 年代初，一些专家和学者逐步开始了藏文信息处理的研究工作[10-12]。1981 年，张连生采用于道全提出的以数码代替藏文的编码方案，实现了首个藏文字符排序软件；之后，张连生采用李方桂提出的罗马转写方案为藏文编码方案，实现了一个集输入、显示和打印功能为一体的藏文字处理系统[13-14]；1984 年，俞乐等在 VICTOR 9000 微机上实现了一个具有输入、显示和打印功能的藏文字处理系统[15]；甘肃省计算中心和西北民族学院合作，在 WANGVS/80 机上也实现了一个藏文字处理系统 ZWCL[16]；航天部 710 所的罗圣仪在 PC-8001 和 IBM-PC 上实现了藏文字处理系统[17]。20 世纪 80 年代中期，以 CCDOS 为代表的汉字处理技术推动了藏文操作系统的发展。1986 年，青海省药品检验所俞汝龙、青海师范大学赵晨星、青海民族学院毛继祖、熊涛等与北京有线电厂合作，在 CCDOS2.13 下开发了与汉英文兼容的藏文操作系统 TCDOS[18]；后来在 TCDOS 的基础上，熊涛等与西北民族学院于洪志等合作开发了可挂接在 WPS 下的藏文轻印刷系统——兰海藏文系统[19]；1989 年，青海民族学院研发了 CTDOS 藏文操作系统；1992 年 10 月，西藏大学尼玛扎