西夏学 第 6 辑 2010 年 9 月
Xixia Studies, Sep. 2010, Vol.6

西夏文古籍字库建立研究*

柳长青

摘 要: 本文指出以往的西夏文字库大多是通过人工隶定楷化后生成的，所造西夏字均带有明显的现代汉字风格，不能体现原始西夏古籍文献中西夏字所具有的力道与美感。为此文章就如何建立一套来源于西夏时期古籍文献的原始西夏文字库这一问题做了探讨，这对于推进西夏文字库的标准化及数字化有重要意义。

关键词: 西夏文 西夏文献 字库

一 引 言

西夏文是西夏时期（1038—1227）党项族的语言文字。[①]近年来，随着西夏考古成果及西夏历史文献的不断公布, 西夏文也再次被人们所关注。[②]如何利用信息技术处理西夏文已成为国内外西夏研究领域的热点。

当前，在西夏文字库研究领域主要有日本、中国和中国台湾地区的学者进行过研究工作。其中，1997 年中国学者李范文教授和日本学者中岛干起利用该所西夏排版计算机系统合著出版了《电脑处理西夏文〈杂字〉研究》一书。1999 年，由马希荣主编柳长青为主要完成人的国家自然科学基金项目“基于汉字字形的西夏文字研究”的成果“夏汉字处理及电子字典”软件由清华大学出版社正式出版，该成果建立了西夏文手写体和楷体共两套西夏 TrueType 字库, 是国内首个可以投入实际应用的基于 Windows 操作系统的西夏文录入编辑与排版软件[③]。台湾中研院语言学研究所的龚煌城和林英津于 2000 年合作开发了西夏文字形属性资料库，并建立了一套西夏文字库。2005 年景永时等利用方正典码系统制作了一套基于汉字库字形的西夏文字库，该字库包含了已有的 6000 余西夏字，并加入了 400 多个西夏字部首，同时还利用“万能五笔”输入法软件实现了西夏文四角号码输入。另外，以日本学者古家时雄为首的文字镜研究会也开发了汉字库, 该字库包含了九万个字符。共收录了汉字、甲骨文、金文、西夏文、梵文、越南喃字等多个矢量字体库。日本国立亚非语言文化研究所副教授荒川慎太郎与俄罗斯西夏学专家克恰诺夫合著的《西夏文字典》一书中的西夏文就采用了文字镜西夏文字库。北京中易中标电子信息股份公司也开发过“中华西夏文处理系统”，建立了包含 6000 多西夏文的西夏字库，并采用郑码输入法输入西夏字。

*本文所获基金资助项目：国家自然科学基金项目资助（批准号: 60803104）; 宁夏自然科学基金项目资助（NZ0836）; 2009 年宁夏高等学校科研项目后续资助。

① 陈育宁《宁夏通史》，宁夏人民出版社 2008 年。

② 史金波, 可恰诺夫《俄藏黑水城文献》，上海古籍文献出版社 1997 年；史金波，陈育宁《中国藏西夏文献》，敦煌文艺出版社 2005 年。

③ 马希荣《夏汉字处理及电子字典》，清华大学出版社 1999 年。

上述字库大多是通过人工隶定楷化后生成的，所造西夏字均带有明显的现代汉字风格，不能体现原始西夏古籍文献中西夏字所具有的力道与美感。而如何建立一种字形优美, 字体标准并能体现出西夏时期西夏文刻本特质的西夏文字库还属空白。因此，建立一套来源于西夏时期古籍文献的原始西夏文字库对于推进西夏文字库的标准化及数字化有重要意义。

二　西夏文字库编码

（一）国标 2312—80 码

西夏字属表意文字，若将已考证并解读的 6000 余西夏字存储于计算机中，一种简单的方法是将 GB2312—80 标准汉字库中的常用汉字用西夏字代替。

GB2312—80 标准中将常用汉字放在 94×94 的区域中，即总共可以放置 8836 个常用汉字。[①]而目前西夏字一共有 6073 个, 将 6073 个西夏字按汉字区位码排列在 16—94 区的 94 个码位中, 剩余的码位还可以放置 400 多西夏字部首。按照该方法建立的具有代表性的西夏字库是景永时、贾常业制作的“基于方正典码的西夏文字库”。该字库利用方正软件公司的“女娲补字”软件对楷体汉字库的汉字部首及结构进行拆解，通过拼接汉字部首及笔画部件造出西夏字库。由于利用汉字部首直接拼接西夏字，因此在显示空心字效果时产生了明显的轮廓线重叠与交叉现象，字例见图 1。正确的西夏字 Bezier 曲线空心效果见图 2。同时，由于占用汉字码位，因此在进行西夏文与汉字的混合排版时必须单独选择西夏字库才能显示和编辑西夏文字，而不能与汉字同时设置字体。用户在使用时需要在汉字库与西夏字库之间切换。此外，由于造字工具软件的限制采用该方法制作的西夏字处理软件不能很好的兼容 Windows Vista 和 Windows7 等新版操作系统。

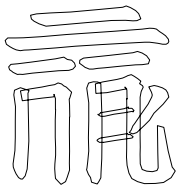

图 1 部首拼接西夏空心字

图 2 Bezier 曲线西夏古籍空心字

（二） GBK 大字符集[②]

GBK 全称“汉字内码扩展规范”，英文名称 Chinese Internal Code Specification，中华人民共和国全国信息技术标准化技术委员会于 1995 年 12 月 1 日制订。它延续了 GB2312－80 的编码体系结构，

① 张青、黄鹤鸣、章登义《基于 ISO/IEC 10646 标准的藏文编码转换的设计与实现》,《中文信息学报》2009 年第 4 期。

② 柳长青、马希荣《西夏字与汉字共存方案的实现》,《宁夏大学学报》（自然科学版）2001 年第 1 期。

采用双字节混合编码，与现有绝大多数操作系统、中文平台在一级内码兼容，支持现有的应用系统；在字汇上则与 GB 13000.1－93 兼容。同时还收录了藏文、蒙文、维吾尔文等主要的少数民族文字。GBK 码在一定程度上缓解了多语言字符同平面共存、汉字收字不足、繁简共存的问题。GBK 码的码位分配总体上采用 8140H－FEFEH 的矩形区域，并且剔除了 XX7FH 一条线，总共有 23940 个码位。这些码位被分配成汉字区、图形符号区和用户自定义区 3 个区域，见图 3。

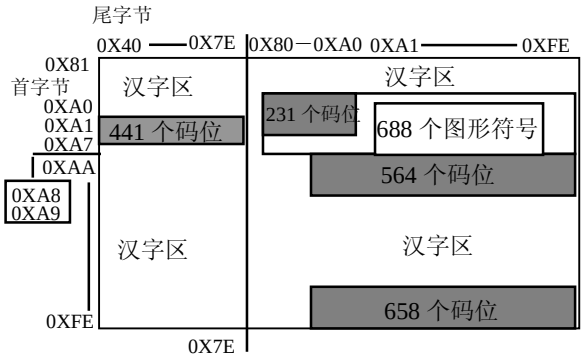

图 3　GBK 码位分配图

其中汉字区共有 21008 个码位，GB2312－80 的汉字被分配在 B0A1H－F7FEH 区，GB13000.1－93 的扩充汉字被分配在 8140H－A0FEH 及 AA40H－FEA0H 区，CJK 兼容汉字被分配在 FD9CH－FE4FH 区，80 个增补汉字部首及构件被安排在 FE50H－FEA0H 区。日本和中国台湾地区的学者所建立的西夏字库就在上述汉字的 CJK 兼容区内。中国学者马希荣等则使用 AAA1H－AFFEH、F8A1H－FEFEH 及 A140H－A7A0H 三个区域，即图 3 中的三个灰色矩形区域，该区域被称为用户自定义区。利用用户自定义区可以避免与汉字或其它字符的码位冲突问题，但需要采用字体位面技术，将 6073 个西夏字映射在 4 种字体库的用户自定义区中。每种字库提供 1894 个码位，4 个字库一共可放 1894×4=7576 个字符。这样就可以放下全部 6073 个西夏字且不占汉字码位。采用这种方法的技术被称为位面技术（见图 4）。这种方法创建的字库已不占用汉字码位，是对 GB2312—80 建库编码的改进，但该方法仍需在不同字体之间切换。

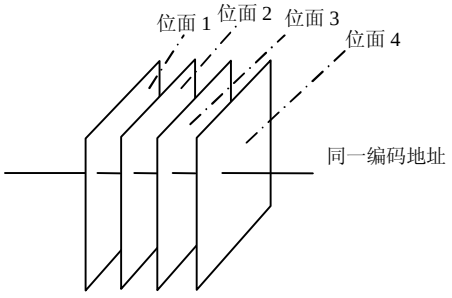

图 4 字体位面技术示意图

（三）Unicode 编码

Unicode 是国际组织制定的可以容纳世界上所有文字和符号的字符编码方案。Unicode 5.2 版本于 2009 年 10 月 1 日正式推出。目前实用的 Unicode 版本对应于 UCS—2, 使用 16 位的编码空间。这 16 位 Unicode 字符构成基本多文种平面（Basic Multilingual Plane, 简称 BMP）。BMP 字符的 Unicode 编码表示为 U+hhhh, 其中每个 h 代表一个十六进制数位。

美国加利福尼亚大学伯克利分校语言学系 Richard 博士于 2006 年向国际 Unicode 组织提交了西夏

文 Unicode 编码申请并被接受[①]。中国、日本和中国台湾的西夏研究人员也积极加入到这一工作当中。截止目前，Unicode 组织为西夏文分配的码位范围从 U+17000 至 U+18715 共 5910 个字符位，收录了 1986 年、1997 年李范文《夏汉字典》字体（1999 年马希荣制作 2 套字库），韩小忙字体（2004 年《西夏文正字研究》所用字体），荒川慎太郎字体（2006 年“文字镜研究会”），景永时字体（2008 年“西夏文字处理系统”所用字体）等作为西夏文字形数据库。西夏文 Unicode 编码的建立将解决占用汉字码位、夏汉同屏共存等问题，对于西夏文字库的国际标准化有非常重要的作用。目前，由于西夏文 Unicode 编码还未最终定稿，因此本文暂时将古籍字库放置在下列 Unicode 编码段：E000—E72C、E737—E76B、E866—F6FF、F71E—F8DA。

上述三种编码方案均有各自代表性的西夏文字库，在不同时期三种编码方案对西夏文字库的建立和发展起到了一定的推动作用。随着计算机技术的不断发展采用 Unicode 编码方案的字体库将是今后一段时期内的主流趋势。本文讨论的西夏文古籍字库将采用 Unicode 编码方案进行字库编码。

三、西夏古籍字模建立

西夏古籍字库中的每一个西夏字模均来自于原始文献扫描图，且最大程度的保留古籍文献中西夏字的原始风貌，尽量不做人工的后期加工，这样建立的西夏字体库也将比传统方法建立的西夏字库更为准确一些。西夏文古籍字模制作主要分为字稿选择、图像预处理、轮廓提取 3 个部分（见图 5）。其中，图像预处理又分为高精度电子化和数字化拟合 2 个步骤[②]。最后，利用拟合图像生成字模信息建立 TrueType 字体库文件。

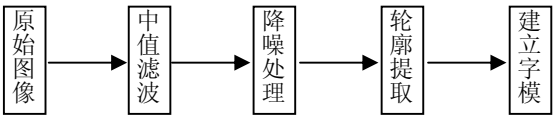

图 5 字模提取流程图

（一）西夏文稿图像二值化

古籍字库不同于一般字库，其字稿来源于西夏古籍文献辞书文献，辞书的选择以清晰可辨的刻本文献为主。本文主要选取了《同音》甲种、乙种本，《番汉合时掌中珠》甲种、乙种本，后期还将对西夏文献中的佛经刻本文献进行处理。确定字稿后，接下来将对选择的文献进行高精度电子化即高精度扫描及图像预处理。经过扫描后的图像需要进行二值化和去除噪声处理。全局阈值法是较为常用的一种图像二值化方法，首先利用公式（1）将真彩图像转换为灰度图像，其次利用中值滤波法对图像进行平滑处理。通过中值滤波能够有效消除文字外的噪点，并使得文字边缘清晰。最后，利用阈值变换，将灰度图像转换为二值图像（见图 6）。对转换为二值图像的西夏古籍文稿再做多次中值平滑处理。

$$Y=0.299*R+0.587*G+0.114*B \qquad (1)$$

① The Tangut UCS Encoding Project.西夏文和统一码[J/OL].http://unicode.org/—rscook/Xixia/,2006—7—12/2007—9—1.

② 刘瀚猛、芮建武、白真龙等《藏文字库标准符合性自动检测方案设计与实现》，《中文信息学报》2008 年第 3 期。

图 6 彩色、灰度及二值图

（二）Level Set 轮廓提取[①]

Level Set 方法是一种较有特色的边缘检测方法。该方法将两维（三维）的闭合曲线（曲面）演化问题转化为三维（四维）空间水平集函数曲面的隐含方式来求解，避免了拓扑结构变化的处理，计算稳定。Level Set 方法把随时间运动的物质界面看作某个函数 $\phi(\vec{x},t)$ 的零等值面，$\phi(\vec{x},t)$ 满足一定的方程。在每个时刻 t, 我们只要求出函数 $\phi(\vec{x},t)$ 的值，就可以知道其等值面的位置，也就是运动界面的位置。构造函数 $\phi(\vec{x},t)$，使得在任意时刻，运动界面 $\Gamma(t)$ 恰是 $\phi(\vec{x},t)$ 的零等值面，即

$$\Gamma(t)=\{\vec{x}\in\Omega:\phi(\vec{x},t)=0\} \qquad (2)$$

$\phi(\vec{x},t)$ 的初值应满足在 $\Gamma(t)$ 附近为法向单调，在 $\Gamma(t)$ 上为零。一般可取 $\phi(\vec{x},0)$ 为 $\vec{x}$ 点到界面 $\Gamma(0)$ 的符号距离，用 $d(\vec{x},\Gamma(0))$ 表示。

$$\phi(\vec{x},0)=\begin{cases} d(\vec{x},\Gamma(0)) & x\in\Omega^1 \\ 0 & x\in\Gamma(0) \\ -d(\vec{x},\Gamma(0)) & x\in\Omega^2 \end{cases} \qquad (3)$$

为了保证在任意时刻函数 ϕ 的零等值面就是活动界面，ϕ 要满足一定的控制方程。由于首先需要求得在全图像范围内的各个网格点到当前轮廓曲线距离为零的点，然后依次连接所有得到的点来获取新的初始轮廓曲线。如此反复迭代直到图象分割完成。通过采用文献①中的窄带方法将计算区域局限在曲线周围一个较小的区域里，当曲线演化到区域的边界时，再重新以当前新得到的曲线为中心建立区域。本文采用上述方法得到了较精确的西夏字轮廓线（见图 7）。

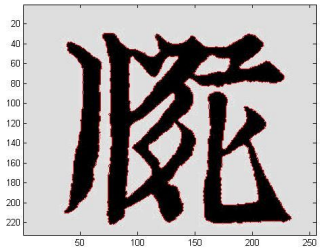

图 7 西夏字轮廓提取图

四、字库建立

True Type 是由 Apple Computer 公司和 Microsoft 公司联合提出的一种数学字形描述技术。它用采用几何学中二次 B 样条曲线及直线来描述字体的外形轮廓，并含有字形构造、颜色填充、数字描述函数、流程条件控制、栅格处理控制、附加提示控制等指令。True Type Font 特点是既可以作打印字体，

① 柳长青《基于 LevelSet 方法的西夏字轮廓提取》,《中文信息学报》2009 年第 4 期。

又可以用作屏幕显示字体。由于它是用指令对字形进行描述，因此与分辨率无关，输出时总能按照打印机的分辨率输出。无论放大或缩小，字符总是光滑的，不会出现锯齿现象。True Type 字体技术具有：1. 真正的所见即所得字体；2. 支持字体嵌入技术；3. 兼容性较好等优点。

（一）True Type 文件结构及其操作

True Type 文件是树形结构, 文件首部共 12 个字节, 记录了 True Type 文件的版本号(定点小数，占用四字节)、描述表数目(无符号短整型，占用二字节), 以及描述表快速查找范围、入口选择、范围调整(均为无符号短整型，占用二字节)的信息。接着是描述表目录, 每个目录项 16 个字节, 包括描述表名称(4 个字节，每个字节都记录表名称的 ASCII 码)、描述表内容的校验和描述表的位置偏移值、长度等。最后是各个具体的描述表，其中“GLYF”表存放的是 True Type 字模数据，存放所有的轮廓描述信息，包括数据信息和指令信息。而简单字模和复合字模的描述方法也不同。简单字模在此表中存放的是一系列的在线和轮廓点的坐标以及作用于此简单字模的指令信息；复合字模在此表存放构成复合字模的简单字模序号和作用于复合字模的指令信息。每个字模的轮廓描述信息都有一个头结构:

USHORT　number Of Contours　//轮廓数目

Word xMin　//轮廓点的最小 x 坐标

Word yMin　//轮廓点的最小 y 坐标

Word xMax　//轮廓点的最大 x 坐标

Word yMax　//轮廓点的最大 y 坐标

如果 number Of Contours 大于等于 0, 则此字模为简单字模, number Of Contours 代表它包含的封闭轮廓个数；如果 number Of Contours 小于 0，则此字模为复合字模。读取操作主要通过调用 Windows API 来完成，以下是读取字体轮廓线信息的 API 函数：

DWORD Get Glyph Outline(HDC hdc, UINT uChar, UINT uFormat, LPGLYPHMETRICS lpgm, DWORD cbBuffer, LPVOID lpvBuffer, CONST MAT2 *lpmat2)；

该函数功能是获取进指定设备的 True Type 字体的字符轮廓或位图信息。通过 GetGlyphOutline 函数可以从字库中提取字体轮廓信息，利用这些轮廓信息可以还原显示 TrueType 字体。

利用上述结构生成了 5652 个古籍西夏字模。其中，有 421 个西夏字由于字稿图像过于模糊及残缺无法提取完整的轮廓线信息，故拟采用部件法构造，这将在今后的研究中进一步开展工作。表 1 是从已提取的 5652 个字模中随机选取的 5 个字例。

原始图像	古籍字库	空心效果	字稿来源
			《文海宝韵》甲种本
			《文海宝韵》乙种本
			《同音》甲种本
			《同音》乙种本
			《同音》丁种本

表 1　TrueType 西夏古籍字库字例表

（二）字体库链接技术

建立了西夏古籍字库后，为了避免用户在汉字库与西夏字库之间切换所引起的使用不便，则需将西夏字库与汉字字库进行字体链接。所谓“字体链接技术”是指在不改变系统字库的基础上生成若干个与系统字库相链接的用户自定义字库。由于本文所建立西夏古籍字库已采用了 Unicode 字符编码方案，有独立的码位空间，因此，其本身不占用汉字码位，不与汉字码位冲突。将 Unicode 编码与字体链接技术结合使用就可以实现西夏字库和汉字库的无切换编辑录入[①]。这种方法不需要对汉字库做任何修改，被链接的西夏字库物理上仍然是一个完整独立的西夏文字库。采用链接技术后将生成一个键值属性，该属性的作用是将汉字库与西夏文字库做一个逻辑链接。如果将西夏古籍字库与操作系统字体链接，那么无论当前编辑的字体是那一种字体，用户只需要调用西夏文输入法就可以实现西夏文字的录入而无需再选择“西夏古籍字体”。

上述基于简体中文版 Windows 操作系统的西夏古籍字库及配套四角号码输入法研究成果已开发完成，目前正在实际测试、修正阶段。该软件可运行于简体中文 Windows XP、vista 及 Windows7 等操作系统下，能够完成汉字和西夏字的无切换录入、编辑等功能。同时，为方便非中文操作系统用户使用，软件同时提供独立的西夏古籍字体库，供用户自定义安装使用。

五、结束语

西夏文古籍字库的建立研究，近年来已受到了西夏研究学者及计算机学者的密切关注。本文的西夏古籍字库是基于西夏古籍文献原始图像之上的字库建立研究。该字库能够较好地保持西夏文原有的风格与特质，并能真实的还原西夏文字的形态。限于原始资料有限，有约 400 余西夏字原始图像无法较好地提取轮廓，因此，今后还将进一步研究利用古籍文字部首部件拼接组合古籍西夏文字的处理方法，并借此再进一步研究西夏文字构造原理及西夏字的结构规律。西夏古籍字库的建立，对于西夏文信息处理及西夏古籍文献的数字化有一定的促进作用，同时为西夏文国家标准字库的建立提供了有益的参考。对于其他少数民族语言文字的古籍字体字库建立也有一定的借鉴与参考价值。[②]

（作者通讯地址：宁夏大学西夏学研究院　银川　750021；宁夏大学数学计算机学院　银川　750021）

① 顾绍通、马小虎、杨亦鸣《基于字形拓扑结构的甲骨文输入编码研究》，《中文信息学报》2008 年第 4 期。

② 柳长青《网络下的西夏文及西夏文献处理研究》，《宁夏社会科学》2008 年第 5 期；高定国、欧珠《藏文编码字符集的优化研究》，《中文信息学报》2008 年第 4 期。