

文章编号: 1007-9831 (2010) 04-0044-04

# 一种基于多级分类的西夏文字识别算法

门光福

(宁夏大学 数学计算机学院, 宁夏 银川 750021)

**摘要:** 字符识别是模式识别领域的一个传统课题, 汉字和古文字的识别是中文信息处理领域的一个重要研究课题. 采用了一种多级分类的方法对印刷体西夏文字进行识别, 在6 000余西夏文字中随机选取300字做测试, 平均识别速度达63字/s, 识别率可达到89%以上, 对西夏文献中的西夏文字的识别奠定了基础.

**关键词:** 西夏字; OCR; 多级分类

**中图分类号:** TP391      **文献标识码:** A      **doi:** 10.3969/j.issn.1007-9831.2010.04.015

随着西夏国(1038-1227年)政权的灭亡和党项族的消失, 西夏字已不再使用, 因此对于西夏国史料的记载很少. 近年来对西夏学的研究在国内外引起了广泛的关注, 西夏学作为一门新兴的学科, 其学科系统已经形成, 学科理论也正在产生和发展, 西夏学的研究已取得了不少成果. 西夏文字<sup>[1]</sup>是我国古代少数民族文字之一, 其笔划繁琐, 但结构严谨、合理, 字形优美, 其结构形态与汉字相仿, 现已挖掘整理出的西夏文字有6000余字, 笔划繁琐, 且绝大多数的笔划数均在14划以上, 因而西夏字是一种相似度远远高于汉字的象形字. 如何把西夏字记载的信息送入计算机, 将大量西夏文字资料方便、高效的以文本的形式进行存储, 从而便于编辑、整理和出版西夏文字相关文献, 是一个非常有现实意义和研究价值的课题.

文字识别输入, 是模式识别和人工智能领域的一个具体的研究方向, 是模式识别、图像处理与文字处理技术相结合的一种新技术. 在一定条件下, 文字识别输入比键盘输入更方便、快捷<sup>[2-3]</sup>. 目前西夏文字的录入一般都是通过专用软件, 如《夏汉字处理软件》<sup>[4]</sup>实现键盘录入. 这种录入方法, 通用性差, 录入速度慢, 如果要录入大量的西夏文字时, 将会耗费大量的人力, 如何快速高效录入西夏文字的问题亟待解决. 本文基于中文OCR(光学字符识别)的方法实现印刷体西夏文字的光学识别录入, 可以将印刷体西夏文字快速、准确地录入电子文档, 实现西夏文献的编辑、整理和出版, 为以后的手写体西夏文字和西夏文献中文字的识别研究奠定基础.

西夏文字识别的过程可以概括为从测量空间映射到特征空间, 再映射到模式空间. 其识别过程同其他字符识别过程类似, 从输入的西夏文字(待识别样本)提取描述该西夏文字的特征, 再要根据一定准则制定该样本所属的模式类别, 因此, 西夏文字描述、特征提取与选择及方案判决, 是西夏字识别过程的3个基本环节. 其中, 特征提取与选择是西夏文字识别的核心. 所以如何提取印刷体西夏文字图像中的西夏文字以及提取每个西夏文字的特征, 很大程度上影响到西夏文字的识别效率. 本文提出的基于多级分类的西夏文字识别算法是一种在保证一定的识别率的基础上, 能够快速并且以较低的硬件代价, 实现西夏文字的特征提取与选择的方法.

## 1 图像预处理

西夏文字的图像预处理是其识别过程中的第一步, 它的好坏直接影响西夏文字识别的效果. 当今的文

收稿日期: 2010-03-25

基金项目: 国家自然科学基金资助项目(60803104); 宁夏大学自然科学基金项目(ndrz09-34)

作者简介: 门光福(1978-), 男, 宁夏中卫人, 讲师, 硕士, 从事图像处理、模式识别研究. E-mail: mengf@nxu.edu.cn

字识别研究已趋向于多级、组合方法识别。无论是结构方法、统计方法，还是神经网络的应用，预处理和细化对信息提取都是有很大帮助的。它能起到“去粗取精，去伪存真”的作用，而且预处理和细化的结果对识别效果有直接的影响。西夏文字的预处理过程类似于其它图像识别预处理过程，其工作过程，见图1。

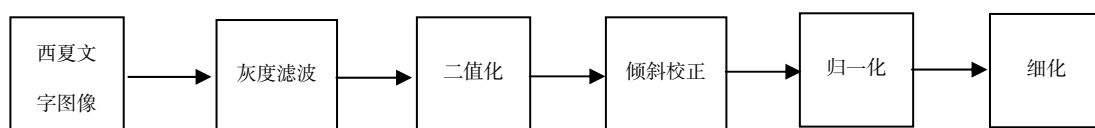


图1 西夏文字预处理过程

### 1.1 灰度滤波

灰度滤波的目的是对输入噪音较多的西夏文字灰度图像进行滤波，去除图像中的叉连、断点及模糊不清的部分，得到一幅较清晰的西夏文字灰度图像。本文选用灰度空间中值滤波法，具体算法是将图像中所有像素点的3\*3邻域内灰度值的中值代替该像素的值得到中值滤波后的图像。

### 1.2 二值化

将灰度滤波后的西夏文字图像二值化变为一幅二值的西夏文字图像。

设原始图像为  $f(x, y)$ ，阈值  $t=127$ ，则二值化后的图像  $g(x, y)$  为

$$g(x, y) = \begin{cases} 1 & f(x, y) > t \\ 0 & f(x, y) \leq t \end{cases}$$

### 1.3 倾斜校正

西夏文字图像可能存在倾斜，所以必须对它进行调整，使得字符都处于同一水平位置，那样既有利于字符的分割，也可以提高字符识别的准确率。调整的算法主要是根据图像上左右两边的黑色像素的平均高度来的。首先要分别计算图像左半边和右半边的像素的平均高度，然后求斜率，根据斜率重新组织图像。

### 1.4 归一化

经过处理的图像可能由于西夏文字大小存在较大的差异，而相对来说，统一尺寸的文字识别标准性更强，归一化的目的就是将所有的西夏文字图像统一到统一尺寸。具体算法为：先得到原来字符的高度，并与系统要求的高度做比较，得出要变换的系数，然后根据得到的系数求得变换后的应有的宽度，在得到宽度之后，把新图像里面的点按照插值的方法映射到原图像中。

### 1.5 细化

在对二值西夏文字图像的细化算法中，本文采用形态学细化算法<sup>[5]</sup>。与传统的 Hilditch 算法和 Pavlidis 算法相比，该算法是一种简单而有效的细化算法，可以很大程度上减少细化图像的毛刺现象，保持图像的连通性。传统的算法要根据目标点的8邻域的情况判断该点是否要去掉。本文所采用的算法增加了目标点邻域的范围，利用模板对要处理点的5\*5邻域  $S$ ，通过数学逻辑计算，综合判断该点是否可以删除。

设置一个5\*5的邻域  $S$  模板，见图2。 $S$  模板中各个位置上的取值取决于模板所对应图像中不同像素位置，如果  $S$  模板某一个位置上所对应的像素值为黑色，则模板上该位置赋为0，否则赋为1。可按4个条件来判断该像素点是否可以被删除。

S[0,0]	S[0,1]	S[0,2]	S[0,3]	S[0,4]
S[1,0]	S[1,1]	S[1,2]	S[1,3]	S[1,4]
S[2,0]	S[2,1]	S[2,2]	S[2,3]	S[2,4]
S[3,0]	S[3,1]	S[3,2]	S[3,3]	S[3,4]
S[4,0]	S[4,1]	S[4,2]	S[4,3]	S[4,4]

图2 5\*5的邻域  $S$  模板

条件1:  $2 \leq N(S[2, 2]) \leq 6$ ;

条件2:  $T(S[2, 2]) = 1$ ;

条件3:  $S[1, 2] * S[2, 1] * S[2, 3] = 0$  同时  $T(S[1, 2]) \neq 1$ ;

条件4:  $S[1, 2] * S[2, 1] * S[3, 2] = 0$  同时  $T(S[2, 1]) \neq 1$ 。

4个条件中:

$N(S[2, 2])$  表示以  $S[2, 2]$  为中心的3\*3邻域内目标像素的个数;

$T(S[2, 2])$  表示以  $S[2, 2]$  为中心的3\*3邻域内，序列  $S[1, 2] S[1, 1] S[2, 1] S[3, 1] S[3, 2] S[3, 3] S[2, 3] S[1, 3] S[1, 2]$  由0变1的变化次数;

$T(S[1, 2])$  表示以  $S[1, 2]$  为中心的3\*3邻域内，序列  $S[0, 2] S[0, 1] S[1, 1] S[2, 1] S[2, 2] S[2, 3] S[1, 3] S[0, 2] S[0, 3]$  由0变1的变化次数;

$T(S[2, 1])$  表示以  $S[2, 1]$  为中心的  $3 \times 3$  邻域内, 序列  $S[1, 1] S[1, 0] S[2, 0] S[3, 0] S[3, 1] S[3, 2] S[2, 2] S[1, 2] S[1, 1]$  由 0 变 1 的变化次数.

如果同时满足 4 个条件, 则删除该点, 否则保留该像素点; 重复判断像素点直到没有点可以删除.

原西夏文字及不同算法细化后的结果, 见图 3. 由图 3 可以看出, 与传统的 Hilditch 算法相比, 该算法有效地实现了细化, 得到图像的骨架, 保持了图像的连通性, 毛刺现象很少, 并且图像结果光滑, 更能体现文字的骨架特征.

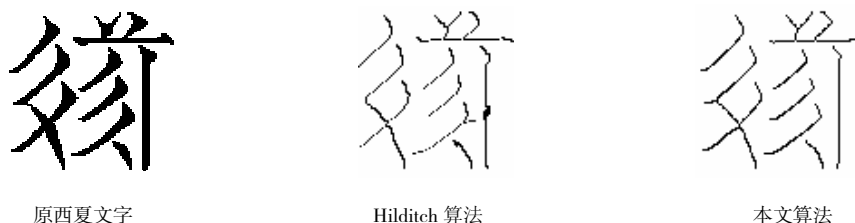


图 3 原西夏文字及不同算法细化后的结果

## 2 特征选择与提取

经过预处理后, 文字图像变成一幅点线图, 见图 4. 如果将文字的所有  $48 \times 48$  个点作为识别特征, 对于类别数达 6 000 之多的西夏文字并不合适, 并且西夏文字的点阵图像是有很大的冗余信息. 为此, 本文采用多级分类来对文字进行识别, 这样可以通过粗分类来缩小细分类的搜索范围, 即通过逐级缩小候选字集合的方法来进行识别, 这样使整体的识别速度不会因为字库的增大而成比例的增长<sup>[6]</sup>.

### 2.1 一级分类特征

将点线图像均匀划分为 36 个图像块, 其中每个图像块中包含 64 个点. 统计每个图像块中黑点的个数作为一维特征. 这样, 一级分类特征共有 36 维.

### 2.2 二级分类特征

由图 2 可以看出, 点线图像共有 48 行, 统计所有行中黑点的个数, 作为一维特征. 这样, 2 级分类特征共有 48 维.

### 2.3 三级分类特征

将点线图像均匀划分为  $12 \times 12$  个图像块, 其中每个图像块包含 16 个点. 统计每个图像块中黑点的个数作为一维特征. 这样, 3 级分类特征共有 114 维.

将每一级分类特征做归一化. 由于每一级特征的值不在同一个区间, 为此需对三级分类特征分别做归一化, 其目的是使所有的分级特征的每一维的值都在  $[0, 1]$  区间, 使比较样本与分类模式的数据统一起来.

## 3 识别算法

基于提取的 3 类特征, 采用三级分类对文字进行识别.

### 3.1 一级粗分类

求出所有候选字  $G_k = (g_{k1}, g_{k2}, \dots, g_{k36})$  分别与待识别文字  $X = (x_1, x_2, \dots, x_{36})$  一级分类特征的绝对值距离  $D_k = \sum_{i=1}^{36} |x_i - g_{ki}|$ , 检索出前 50 个最小距离的字集  $S_1$ , 并将它们作为下一级分类的候选字.

从粗分类简单有效的要求出发, 对于 36 维的一级特征矢量, 本文采用绝对值距离判别准则来进行识别. 虽然其缺点是对相似字的区分能力不如其它判别准则, 但作为粗分类, 只要求累积分类率高, 即只要能保证在前  $n$  (如  $n=50$ ) 位出现正选字即可, 再加上该准则具有计算量小的优点, 选用它作为粗分类的判

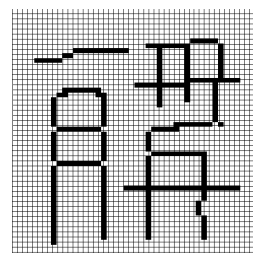


图 4 经过图像预处理后的西夏文字

别准则是合理的. 该粗分类特征比较稳定, 其前 50 个字的累积分类率可保证在 95% 以上, 因此可以说, 基本上达到了预期的要求.

### 3.2 二级细分类

令待识别字与前一步计算出的 50 个候选字集  $S_1$  的二级分类特征矢量分别为  $X = (x_1, x_2, \dots, x_{48})$  和  $G_k = (g_{k1}, g_{k2}, \dots, g_{k48})$ , ( $k=1, 2, \dots, 50$ ), 求出候选字与待识别文字的欧式距离为  $D_k = \sqrt{\sum_{i=1}^{48} (x_i - g_{ki})^2}$ , 检索出前 10 个最小距离的西夏文字集  $S_2$ , 并将其作为第 3 级分类的候选字集.

### 3.3 三级分类 (即识别)

在二级细分类计算出的 10 个候选字集  $S_2$  的基础上, 计算候选字与待识别文字三级分类特征的欧式距离  $D_k = \sqrt{\sum_{i=1}^{114} (x_i - g_{ki})^2}$ , 求出距离最小的候选字作为识别结果输出.

本文将《夏汉字典》中的西夏文字字库的 6021 个西夏文字作为训练样本, 随机选取训练样本中的 3 组 300 个西夏文字作为待识字, 采用多级分类识别算法对待识字进行识别的试验中, 平均识别速度为 63 字/s, 识别率分别达到 89.2%, 90.6%, 91.1%.

## 4 结束语

本文针对印刷体西夏文字的识别提出了一种多级分类识别算法, 实验结果表明, 对于西夏文字这样一种相似度远远高于汉字的象形文字, 本文算法能够有效地进行识别, 识别率可达到 89% 以上. 在识别速度和识别率等方面进行了合理的折中, 较好地弥补了西夏文字识别研究方面的不足, 为以后的手写体西夏文字、西夏文献图片资料中的西夏文字的自动识别算法等的深入研究奠定基础.

### 参考文献:

- [1] 李范文. 夏汉字典[M]. 北京: 中国社会科学出版社, 1998: 1-80.
- [2] 马希荣, 王行愚. 西夏文字特征提取的研究[J]. 计算机工程与应用, 2002 (13): 38-41.
- [3] 马希荣, 王行愚. 西夏文字识别中的图像预处理[J]. 计算机工程与应用, 2002 (2): 48-50.
- [4] 马希荣, 柳长青. 夏汉字处理系统及电子字典[M]. 北京: 清华大学出版社, 1999: 11.
- [5] 曹灿, 黄福莹. 数字图像细化算法研究与实现[J]. 广西物理, 2006, 27 (3): 40-42.
- [6] 任金昌, 赵荣椿, 张炜. 一种快速有效的印刷体文字识别算法[J]. 中国图像图形学报, 2001, 6 (10): 1 011-1 015.

## An algorithm for xixia characters recognition based on multi-stages classification

MEN Guang-fu

(School of Mathematics and Computer Science, Ningxia University, Yinchuan 750021, China)

**Abstract:** The characters recognition is a conventional project of model recognition domain. Both chinese characters and ancient characters recognition are an important project. A multi-stages classification for machine printed xixia characters is proposed, and the experiment is used to show the recognition speed and rate can reach 63 word per second, 89 percent. The investigation has built a solid theory foundation for the research and development of xixia characters recognition in xixia documents.

**Key words:** xixia characters; OCR; multi-stages classification