

# 基于图形识别的汉字笔画分类方法

赵青,唐英敏

(北京大学 计算机科学与技术研究所,北京 100871)

**摘要:**提出了一种获得汉字笔画种类及其拓扑结构的新方法。通过分析 TrueType 字库存储的汉字字形信息,构造笔画图形并提取出笔画特征,采用统计分类和边界关键点定位相结合的方法,利用综合分类法实现了对二级 6763 个汉字的笔画种类和拓扑结构的基于 XML 的语义知识表达,可结合前端输入程序进行快速搜索和定位。为字体变形、字体自动生成、汉字知识挖掘等提供了必需的基本信息。实验表明这种方法能够准确、有效地识别汉字笔画及其拓扑结构。

**关键词:**汉字;笔画;形状识别;关键点;TrueType

**中图分类号:**TP391

**文献标识码:**A

**文章编号:**1673-629X(2009)10-0014-04

## Shape Recognition - Based Approach to Chinese Character Strokes' Classification

ZHAO Qing, TANG Ying-min

(Institute of Computer Science and Technology of Peking University, Beijing 100871, China)

**Abstract:** Propose an approach to classifying strokes and getting topology of Chinese characters. Shapes of strokes are constructed and characteristics of which are extracted through analysis of the TrueType font glyphs containing characters information. With composite classified algorithmic, XML-based semantic knowledge representation for stroke types and topology of 6763 characters is implemented by integrating four kinds of statistical approaches and layout of core points. This method supplies fast searching and locating by using foreground application, and supplies prerequisite basic information for glyph transmutation, automatic typeface creation and knowledge mining of Chinese characters. The experiments show the accuracy and efficiency in recognition and topology of strokes.

**Key words:** Chinese character; stroke; shape recognition; core point; TrueType

### 0 引言

信息时代的来临,让汉字的生存和汉字字体的创新面临前所未有的机遇和挑战。机遇在于高速发展的计算机行业和相关的文字、图形处理技术使得汉字字体信息化的过程可以耗费更小的人力,转而由计算机自动完成。挑战在于,这种自动完成对于研发人员来说仍是极其困难和艰巨的任务,因为虽然字库的格式已经有很多的国际规范,比如微软,Adobe 公司推出的 Type 1, TrueType, Open Type 等,但是其内的汉字存储信息仅仅是简单的二维空间点元素,无方向性、无结构、无语义,要从这样的数据中进行搜索和定位需要花费大量精力。因此,字体自动生成、字体变形、汉字知识挖掘、生僻字的互联网跨平台传输和表示等汉字处理领域迫切需要汉字的组成和结构信息。

结构化的、遵守一定语义规范的汉字构形信息能够为汉字的自动处理提供足够的信息量和检索方式。得到汉字构成和拓扑结构,并将其与字库软件结合起来,字库提供几何特征,汉字构形提供逻辑特征,是解决以上问题的重点。目前,多数研究人员使用人工方法开展研究工作,特别是对笔画和部件的识别、归类、构形式的生成等主要环节,皆依赖于研究者人脑对文字的认知和文字学专业知识,而计算机作为研究工具,仅执行统计等简单任务。没有以字库数据为基础的体现汉字所有特性的语义规则实现。

笔画是指书写时一次起落笔之间写下的部分,是构成汉字的基本元素。Windows 系统的 TrueType 字库中,宋体-方正超大字符集和仿宋-GB2312 两款字体提供了汉字笔画轮廓信息,即轮廓以笔画为基本单位存储,为从字库角度研究并实现汉字构形提供了基础,对于计算机实现前述汉字自动化工作具有十分迫切的需求和非常重要的现实意义。基于宋体-方正超大字符集字库提出一种获得汉字笔画种类及其拓扑结

收稿日期:2009-02-16;修回日期:2009-05-15

**作者简介:**赵青(1982-),女,山东青岛人,硕士研究生,研究方向为图形与文字信息处理;唐英敏,高级工程师,研究方向为图形与文字信息处理。

构的新方法。首先获笔画原始轮廓,并将其细化取得骨架图形,从两种图形提取笔画的特征,采用纵横比、四边码<sup>[1]</sup>、四角码和投影四种统计方法进行粗分类,然后跟踪笔画轮廓,提取边界关键点,利用关键点分布特征进行细分类,得到十九种笔画类型。最后计算笔画外包矩形的中心,确定笔画间的位置关系,建立其拓扑结构。计算结果全部以 XML 形式保存,便于理解和搜索。

## 1 笔画识别

TrueType 是 Apple 公司和 Microsoft 公司合作开发的页面描述语言<sup>[2]</sup>。TrueType 字库存储的是汉字轮廓信息,轮廓是一条封闭的曲线,由直线和 Bezier 曲线构成<sup>[3]</sup>,根据汉字构成的复杂程度,每个字由数量不等的轮廓组成。

### 1.1 获取笔画图形

在宋体-方正超大字符集字库中,由于轮廓就是笔画,所以可将各个轮廓经过字库解析、数据提取、轮廓绘制和封闭图形填充三个步骤,直接存储为笔画图像,省略了基于图像的笔画抽取过程,且得到的笔画不会出现信息丢失或者变形。文中的笔画图形基于国标二级 6763 个汉字,共计 72717 个笔画,图 1 和图 2 是一些填充后的笔画图形示例。

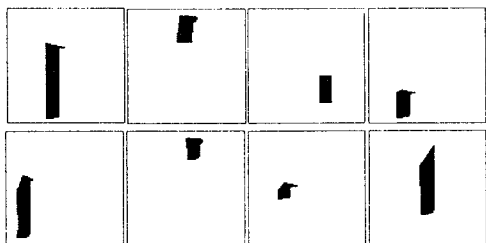


图 1 笔画“竖”

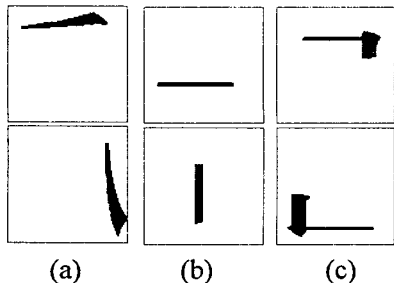


图 2 旋转造成的笔画种类变化

笔画的种类繁多,即使同一字体的同一笔画,在不同汉字或者同一汉字的不同位置都有可能发生局部变形甚至整体几何形状的改变,图 1 是笔画“竖”的八种不同形态。

有的笔画区别仅仅是旋转角度不同,如图 2 所示,2(a)为“撇”和“捺”,2(b)为“横”和“竖”,2(c)为“横折”

和“竖折”。笔画的特征易变、有交叠并且区分度不大,这使得识别笔画与识别一般的简单几何形状不同,单靠统计方法或者结构特征方法难以区分。由于笔画形状的多变性和旋转相关性,经过实验,许多形状识别的常规方法,如:模板匹配、不变矩<sup>[4]</sup>、傅里叶描述子<sup>[5,6]</sup>、自回归模型<sup>[7]</sup>等都不适用。因此,目前存在的比较成熟的字处理系统的汉字结构信息库大多采用人工分拣和识别,依靠长期积累和修正。以下三个小节是笔画识别方法的具体步骤。

### 1.2 细化笔画

为了减少需要处理的数据量,准确提取笔画特征,避免因为笔画粗细不均造成的统计偏差,需要使用细化算法,提取笔画骨架。由于笔画的形状相对简单、没有交叉和重叠,因此采用传统的基于数学形态学的并行细化方法<sup>[8]</sup>,具体算法如下:

1) 针对图像中的每个像素点,得到该点的八邻域,中心点编号为  $P_0$ ,其余各点按逆时针方向依次编号为  $P_i (i = 1, 2, \dots, 8)$ ,如图 3 所示。

$P_2$	$P_1$	$P_8$
$P_3$	$P_0$	$P_7$
$P_4$	$P_5$	$P_6$

图 3 八邻域编号

2) 进行两步迭代,删除冗余点。每步分别将满足条件的冗余点做标记,暂不删除,两步迭代结束后再一并删除。

第一步,对满足  $R1 \sim R4$  的点做标记:

$$R1: 2 \leq b(P_0) \leq 6$$

$$R2: X_R(P_0) = 1$$

$$R3: P_1 \times P_3 \times P_7 = 0$$

$$R4: P_1 \times P_3 \times P_5 = 0$$

第二步,对满足  $R5 \sim R8$  的点做标记:

$$R5: 2 \leq b(P_0) \leq 6$$

$$R6: X_R(P_0) = 1$$

$$R7: P_3 \times P_5 \times P_7 = 0$$

$$R8: P_1 \times P_4 \times P_7 = 0$$

对一幅笔画图像反复执行步骤 2),直至没有可删除的点为止,剩余部分就是骨架图形。从骨架图获得其外包矩形,计算统计特征时可以消除位移因素的影响,利于归一化计算,如图 4 所示,第一行分别是“竖”、“横折勾”、“撇”、“横”、“点”,第二行为其对应的骨架。

### 1.3 提取统计特征

特征的选取是笔画识别的关键,应尽量满足容易提取、分类能力强、较高稳定性和抗干扰性等条件<sup>[2]</sup>。文中的统计特征基于笔画骨架外包图形,粗分类阶段

采用四个特征——纵横比、四边码、四角码和投影。下面给出这四个特征的定义：

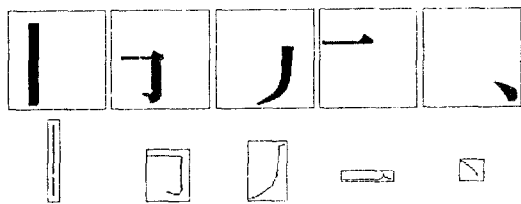


图 4 笔画及其骨架外包图

1) 纵横比:骨架外包图的长宽比值,如图 5(a)所示,计算方法见公式(1)。纵横比特征计算十分简单,但是在描述笔画大致形状方面非常有效。

$$\text{WHRate} = \frac{\text{Height}}{\text{Width}} \quad (1)$$

2) 四边码:在图形四周各划一条带,计算各带内的黑像素,分别除以 Height 或者 Width 做归一化处理,并把它量化成四个离散值。如图 5(b)所示,“撇”的四边码为(0,0,0,0)。

3) 四角码:根据笔画分布特点,提出的一种新的统计算法,截取图形四个角,计算各自内部黑像素,并分别除以四角面积做归一化处理,并量化成四个离散值。如图 5(c)所示,“撇”的四角码为(1,0,1,0)。

4) 投影:做两条对角线,分别在每个三角分割区域内垂直对角方向投影,将投影结果相加,除以对角线长度做归一化处理,并量化成两个离散值。如图 5(d)所示,“撇”的东北对角线的投影特征值为(0,1),其西南对角线投影特征值为(0,0)。

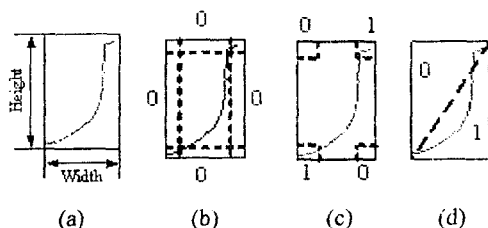


图 5 四个统计特征

在“横”、“竖”、“撇”、“捺”、“折”五种基本笔画基础上进行扩展,将笔画分为 19 类,每类各选择 100 个图形作为训练样本,提取出四个特征的分类域值。将每个图形获得的特征值与分类域值做比较,采用向量间的欧氏距离作为判别标准,划分到欧氏距离最小的相应的类别中。

#### 1.4 获取边界关键点

粗分类完成后,笔画图形被归为 19 类,此时,有相当数量的笔画被误分,例如“提”和“撇”等,在骨架图形上的统计特征十分接近,需要使用基于笔画轮廓的图形进一步细分类。在汉字单字识别中,特征点分类的方法是一种较好地利用汉字结构规律的方法。特征点

包括端点、折点、歧点、交点等。笔画是无交叉简单图形,所以采用曲线上局部区域的最大曲率点作为关键点,将关键点的空间位置分布和数量作为细分类的特征。其原理是跟踪轮廓曲线,用数字曲线的多边形近似方法去除边界冗余像素点,提取边界特征点作为原轮廓曲线的特征描述<sup>[3,9]</sup>,具体算法如下<sup>[3]</sup>:

第一步://初始化,从轮廓右上角的  $P_0$  出发沿笔画边界逆时针方向提取所有的关键点。

let  $i = 0$ ; let  $j = i + 2$ ; let  $P_0$  点作为第 1 个关键点;

第二步://在  $(P_i, P_j)$  区间的曲线上寻找到直线  $P_i P_j$  距离最大的点  $P_m$ 。

if  $j > N$  停止

else 找一个点  $P_m$ , 满足

$$d(i, j, m) = \max_{i < k < j} d(i, j, k)$$

第三步://判断  $P_m$  是否为关键点。

if  $d(i, j, m) > D$  then 提取  $P_m$  为新的关键点。 $i = m$ ;  $j = i + 2$ ;

goto 第二步;

else  $j = j + 1$ ; goto 第二步;

这里的  $d(i, j, m)$  表示点  $P_m$  到直线  $P_i P_j$  的距离,计算方法见公式(2)。将笔画的外包矩形平均分成四份,记录每个笔画在四个象限内的关键点个数,获得四维向量作为特征值。图 6(a) 是笔画“提”的关键点分布,其四维向量值是(1,1,1,0),图 6(b) 是笔画“撇”的关键点分布,其四维向量值是(3,0,1,0)。

$$d(i, j, m) = \left| \frac{(y_j - y_i)x_m - (x_j - x_i)y_m - (y_i x_j - x_i y_j)}{((x_j - x_i)^2 + (y_j - y_i)^2)^{1/2}} \right| \quad (2)$$

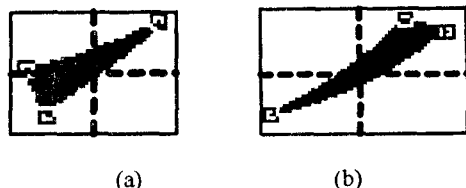


图 6 “提”和“撇”的关键点分布

#### 1.5 分类结果

细分类完成后,笔画识别过程结束,经过人工筛选,72717 个笔画误分类的个数为 4626 个,正确率达到了 93.64%。笔画自动识别过程极大地简化了笔画分类工作。表 1 列出了十九种笔画的类型及其数量。

## 2 拓扑结构的建立

笔画间的位置关系可用八方向描述,分别为:北(N)、西北(WN)、西(W)、西南(WS)、南(S)、东南(ES)、

表 1 笔画分类结果

笔画	数量	笔画	数量
㇏	9662	㇏	142
㇏	434	㇏	3259
㇏	157	㇏	1645
㇏	20149	㇏	222
㇏	8695	㇏	89
㇏	62	㇏	84
㇏	1172	㇏	143
㇏	1996	㇏	137
㇏	11370	㇏	261
丨	13038		

东(E)、东北(EN)。在 1.2 小节获得的笔画外包图形是包括笔画的最小矩形,其中心点坐标可作为笔画的位置描述。因此,将每个笔画外包框的中心点坐标记录在 XML 文件中,笔画间的位置便可直接比较坐标得到。

下面的 XML 子节点描述的是“丈”字,记录了 Unicode 码、笔画个数、每个笔画的中心点坐标和类型编码:

```
<Character char="丈" code="4e08">
<StrokeCount>3</StrokeCount>
<Strokes>
  <Stroke pointX="504" pointY="221">
    heng
  </Stroke>
  <Stroke pointX="392" pointY="499">
    pie
  </Stroke>
  <Stroke pointX="593" pointY="637">
    na
  </Stroke>
</Strokes>
</Character>
```

(上接第 13 页)

and mapping (SLAM): Toward exact localization without explicit localization[J]. IEEE Transactions on Robotics & Automation,2001,17(2):125-136.

[6] Kwon T B, Song J B. Real-time building of a thinning-based topological map[J]. Intelligent Service Robotics,2008,1

### 3 结束语

二级汉字的笔画数量庞大,形状极其多变,人类视觉对部分独立笔画都难以分辨,让计算机快速准确地识别物体的形状更加困难。该文利用综合分类法,先预处理汉字轮廓图形得到汉字笔画;然后基于笔画骨架采用统计方法获得纵横比、四边码、四角码、投影四个特征向量,对笔画进行粗分类;最后跟踪笔画轮廓,获取关键点特征,进行细分类,对粗分类过程中的误分笔画进一步处理,获得了较高的分类正确率。笔画类型和拓扑关系是下一步进行汉字语义规则描述的基础,具有十分重要的意义。值得一提的是,笔画细化过程会产生畸变和噪音,影响统计的正确性,因此可在细化之前进行低通滤波以平滑图像和弱化细节,使细化之后的骨架具有更强的分类特性。

#### 参考文献:

[1] 丁晓青. 汉字识别——原理方法与实现[M]. 北京:高等教育出版社,1992.

[2] Microsoft Corporation. TrueType 字型核心技术[M]. 北京:学苑出版社,1993.

[3] Chen H H, Su J S. A syntactic approach to shape recognition [C]//In: Proc Int'l Computer Symp. Tainan, Taiwan: [s. n.],1986:103-122.

[4] Belkasim S O, Shridhar M, Ahmadi M. Pattern recognition with moment invariants: A comparative study and new results [J]. Pattern Recognition,1991,24(12):1117-1138.

[5] Kauppinen H, Seppanen T, Pietikainen M. An Experimental Comparison of Autoregressive and Fourier-Based Descriptors in 2D Shape Classification[J]. IEEE Trans on PAMI,1995,17(2):201-207.

[6] Persoon E, Fu L S. Shape discrimination using Fourier descriptors[J]. IEEE Trans on PAMI,1986,8(3):388-397.

[7] Dubois S R, Glanz F H. An autoregressive model approach to two-dimensional shape classification[J]. IEEE Trans on PAMI,1986,8(1):55-66.

[8] Lam L, Lee S W, Suen C Y. Thinning methodologies - A comprehensive survey [J]. IEEE Trans on PAMI,1992,14(9):869-885.

[9] Sklansky J, Gonzalez V. Fast polygonal approximation of digitized curves[J]. Pattern Recognition,1980,13(5):327-331.

(3):211-220.

[7] 史绍强. 一种改进型的汉字字符图像细化算法[J]. 计算机技术与发展,2007,17(9):88-91.

[8] 龙占超,蔡超. 一种新的指纹细化算法[J]. 计算机技术与发展,2007,17(3):147-149.