北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

基于笔端形状相似性的汉字字体识别

王晓[1,2]　吕肖庆[1,2,†]　汤帜[1,2]

1. 数字出版技术国家重点实验室(筹), 北京 100080；2. 北京大学计算机科学技术研究所, 北京 100871；
† 通信作者, E-mail: lvxiaoqing@pku.edu.cn

摘要　提出一种基于笔端相似性的方法, 来解决在较大规模字体集上的单字符字体识别问题。该方法首先提取汉字笔画上的特定部位——笔端, 然后利用笔端形状作为汉字的字体特征来对其进行识别。实验证明, 该方法不但在常用字体集合上的识别效果优于同类方法, 而且在扩展后的大字体集合上也能达到较高的识别率。

关键词　字体识别; 形状相似性度量; 形状描述子; 笔端; 特征笔端

中图分类号　TP391

Optical Font Recognition of Chinese Based on the Stroke Tip Similarity

WANG Xiao[1,2] , LÜ Xiaoqing[1,2,†], TANG Zhi[1,2]

1. State Key Laboratory of Digital Publishing Technology(Peking University Founder Group Co., Ltd), Beijing 100080; 2. Institute of Computer Science and Technology of Peking University, Beijing 100871; † Corresponding author, E-mail: lvxiaoqing@pku.edu.cn

Abstract　This paper presents a novel method for the OFR of single Chinese character on a large font set. This method explores the specific parts of strokes of each character, called Stroke Tips, which is regarded as a good feature for font recognition. Experiments show that the recognition rate on common font sets is superior than that of other methods. Furthermore, experiments on an extended font set confirm the scalability of the proposed method, which means that this method is suitable for the OFR of single Chinese character on a large font set.

Key words　font recognition; similarity measure of shapes; shape descriptor; stroke tip; eigen stroke tip

以往, 字体识别的需求主要来源于文档电子化这一领域, 涉及的技术包括光学字符识别(OCR)、版面分析、理解与恢复等。近年来, 字体识别的需求有了新的变化。一方面, 由于 TrueType 等字体设计技术的广泛使用降低了字体设计人员创造新字体的难度, 各种各样的新字体陆续问世。而另一方面, 传统出版业在出版物中采用大量的新字体来增强读者的阅读体验。与此同时, 互联网的迅猛发展极大地推动了新字体的传播。报刊、杂志、电子邮件、网页、手机、电视等, 都可以很方便地借助于丰富多样的新字体来渲染内容, 从而吸引更多的读者。字体种类的迅速增长带来了字体识别的新需求。首先, 传统的文档电子化需要处理与识别的字体增多；其次, 字体所有者需要对拥有版权的字体进行保护；第三, 字体设计师以及普通用户需要根据少数字符来寻找相似字体。因此, 在较大规模字体数据集上的字体识别越来越受到重视。

根据文本相关性, 字体识别方法可以分为两类: 1) 文本无关的字体识别, 即在不知道待识别字符是哪个汉字的情况下判断其字体属性; 2) 文本相关的字体识别, 即借助待识别字符的编码缩小搜索空间进行识别。后者利用了字符先验知识, 往往具有更高的识别率, 可以扩展到更大的字体集合, 但是这种方法不能用于构建高精度的 OCR 系统。根据识别对象的不同, 字体识别又可分为基于文本块的字体识别和基于单字的识别。前者识别率高, 效果稳定, 更能容忍噪声和错误的分割, 但需要一块文本区域来提取字体纹理与笔画分布等全局特征, 不适用于

国家重点基础研究发展计划(2010CB735908)资助
收稿日期: 2012–06–03; 修回日期: 2012–08–23; 网络出版时间: 2012-10-26 17:55
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1755.023.html

文档版面的精确恢复等场景。基于单字的字体识别难度大, 但可用于更多的场景。对于英语等西文字体识别与检索的研究, 经常使用的字体集往往包含数百甚至数千字体。近几年, 较大规模的西文字体识别与检索[1−2]已有成效。然而, 对于中文, 在大字体集上的研究还较少。汉字字体识别的难度不仅在于汉字的结构复杂, 还在于每个汉字字符集包含了过多的字符。以往的实验大多只针对宋体、仿宋、黑体、楷体、隶书、魏碑和幼圆这 7 种字体的全部或部分集合进行。

在基于文本块的字体识别方面, Zhu 等[3]采用 Gabor 滤波器提取字体特征, 然后对其进行全局纹理分析, 取得了较高的识别率。其后学者们[4−9]对此类方法做出改进, 取得了新的进展。杨志华等[10−11]提出了一种基于经验模式分解(empirical mode decomposition, EMD)的中文字体识别方法, 他们选择 5 个基本笔画特征来描述中文字体, 具体方法是对每一个给定的文本块计算笔画特征序列并且使用 EMD 进行分析, 生成一个特征向量, 最后采用最小距离分类器识别字体。基于单字符字体识别的代表性方法是 Ding 等[12]和陈力等[13]提出的基于小波变换的方法。这种方法首先通过小波变换从字符图像中提取大量的小波特征, 然后使用线性鉴别分析技术(linear discriminant analysis, LDA) 选择与字体信息相关的特征, 最后使用一种改进的二次鉴别函数(modified quadratic discriminant function, MQDF)分类器进行字体识别, 其他使用小波方法进行字体识别的研究工作包括文献[14−15]。Sun[16]利用汉字的笔画结构进行字体识别。该方法自动提取单个字符的笔画部分, 称为笔画模板(stroke template), 相同字体的笔画模板被存储在字体数据库中。对于新输入的字符, 将其笔画模板与字体数据库中的笔画模板一一比对, 最后使用贝叶斯分类器决定最有可能的字体类别。王恺等[17]使用一种基于特征点的个体分析法来解决汉字字体识别问题。

本文提出了一种利用汉字局部笔画结构的形状信息进行单字符字体识别的方法。这种方法首先提取单个汉字笔画的特定部位, 称为笔端, 然后使用一种笔端形状描述子(shape descriptor of stroke tip, SDST)来描述笔端, 并用欧氏距离度量笔端之间的相似度。这样每个汉字字符可以表示为若干 SDST, 而每种字体由从若干训练字符集提取的笔端描述子的聚类中心——特征笔端表示。对于一个待识别的汉字字符, 首先将其表示为 SDST 的集合, 然后用每个 SDST 与特征笔端一一比对, 选择距离最小的笔端作为 SDST 的类别, 最后综合全部 SDST 分类结果判断该汉字字符的所属字体。

1 原理与流程

字体间的一个主要区别体现在笔画的起笔与收笔部分。图 1 是多种字体的“捺”笔画在末端部分的形状比较, 其中(a)~(g)分别展示了 7 种字体。第 2 行形状取自第 1 行“永”字“捺”笔画的末端部分, 而第 3 行形状取自第 4 行“天”字“捺”笔画的末端。比较 2、3 行形状不难发现, 同种字体的笔画末端在形状上十分相似, 而不同字体笔画末端的形状有所不同。

我们称笔画在起笔与收笔时形成的独特形状为笔端。笔端形状与字体的紧密关系启发我们利用这种特征进行汉字字体识别, 基本流程如图 2 所示, 包括两个部分: 字体训练(虚线箭头)与单字字体识别(实线箭头)。字体训练部分的输入是用于训练的文本块图像或单字图像, 之后从中提取笔端结构,

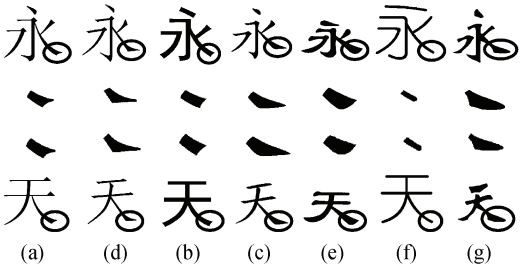


图 1 字体与笔端形状的关系

Fig. 1 Fonts and the shapes of stroke tips

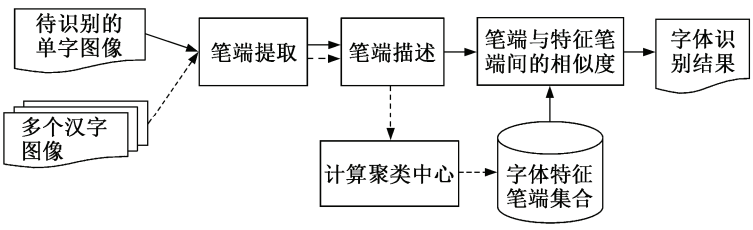


图 2 基于笔端相似度的汉字字体识别流程

Fig. 2 Flow chart of font recognition based on the similarity of stroke tips

并将其描述为特征向量, 最后计算出每种字体的聚类中心, 即特征笔端。单字字体识别部分的输入是单字的二值图像, 而提取描述笔端的方法与字体训练部分相同, 这样一个汉字表示为若干个特征向量的集合。之后, 利用特征向量间的欧氏距离度量待识别笔端与特征笔端间相似性。最后利用每个笔端最相似的特征笔端决定输入字符的字体。

2 笔端提取与描述

虽然人类很容易观察到笔端这种结构, 但是以往文献并未给出准确的定义, 因此, 我们首先要明确笔端这个概念, 即, 笔端是指笔画起始与收尾处占笔画一定比例的结构。另外, 在本文中有些字体的点笔画和转折部位也可以视作笔端。

从汉字中提取出用于区分字体的笔端并不容易。一方面, 将汉字分解为笔画已经是一项十分困难的工作, 例如在某些字体中, 如行书, 草书等, 并不存在明显可分割的笔画; 另一方面, 我们不仅需要从同一种笔画的不同变体提取出相似的笔端, 还需要从具有相似端部结构的不同笔画提取出相似的笔端。如图 3 所示, 同属于楷体的“仪”、“纹”、“女”和“业”4 个字中的圈出部分分别属于同种笔画的不同变体或者不同笔画, 这些部位具有很高的相似性。然而, 由于不同笔画之间以及同种笔画的变体之间在长度、宽度、方向等方面存在差异, 这给笔端提取造成了很大的困难。

2.1 笔端提取

Sun[16]提出的笔画模板(stroke template)与我们所定义的笔端有相似之处。这种算法首先将二值字符图像转化为骨架图像, 骨架的连通度与原始形状一致。然后以骨架上任意一个端点为起始点 p_e, 沿其所在的分支寻找第一个分叉点 p_b。设 $l(p_e, p_b)$为 p_e 和 p_d 之间骨架边的长度, 如果 $l(p_e, p_b)<\delta$, 则判断此骨架分支为毛刺并删除, 否则, 如果 $l>l_{min}$, 则截取这段分支对应的笔画段作为笔画模板。如果 p_b 不存在, 即沿着 p_e 所在的骨架边找到了另一个骨架端点 $p_{e'}$, 那么骨架边 skel(p_e, $p_{e'}$)对应的笔画段即作为笔画模板。这里δ和 l_{min} 是与字符高度相关的阈值。

图 3 同种字体在笔画变体间与不同笔画间的笔端形状相似性

Fig. 3 Stroke tip similarity of stroke variants from the same font

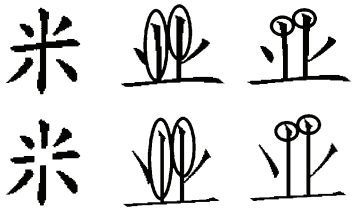
以“米”字为例, 这种算法提取到的笔画模板如图 4(a)所示, 与理想的笔端有相似之处, 但是存在着明显缺陷。其一, 从同种字体相同笔画提取的笔画模板存在较大差异。例如, 图 4(b)中分别是楷体与宋体的“业”字, 其笔画模板长短不一。其二, 对所有字体采用单一阈值去除毛刺并不合适。如果阈值过大, 容易去掉有意义的笔画模板; 如果阈值过小, 则提取的笔画模板容易在毛刺处提前截断。良好的笔端截取算法应当不受分叉点的位置影响, 只与笔端形状有关, 如图 4(c)所示。

通过以上分析可知, 为了从同种字体获得相似的笔端形状, 需要更为精细的方法来确定截断位置。本文提出一种基于骨架的笔画形状描述子, 称为笔画长宽比(stroke aspect ratio, SAR), 用于计算笔端在骨架上的截断位置。形状 s 的笔画长宽比 sar(s)的计算公式如下:

$$\mathrm{sar}(s)=\frac{\text{骨架分支长度}}{\text{骨架分支点的最大半径}\times 2}。\tag{1}$$

图 5(b)中的形状 s'是对图 5(a)中形状 s 进行平移、旋转、缩放以及添加噪声后得到的结果。s 的骨架分支长度等于$|AB|+|BC|+|CD|$, 最大半径为 $r(B)=|AE|/2$, 则 $\mathrm{sar}(s)=(|AB|+|BC|+|CD|)/(2\times r(B))=(|AB|+|BC|+|CD|)/|AE|$, 而 $\mathrm{sar}(s')=(|A'B'|)+|B'C'|+|C'D'|)/(|A'E'|)$。由于 $A'B'$, $B'C'$, $C'D'$ 和 $A'E'$ 分别是 AB, BC, CD 和 AE 按照同比例缩放得到的, 所以 $\mathrm{sar}(s)=\mathrm{sar}(s')$。这说明笔画长宽比具有旋转、平移和缩放不变性, 并且具有一定的抗噪声能力。

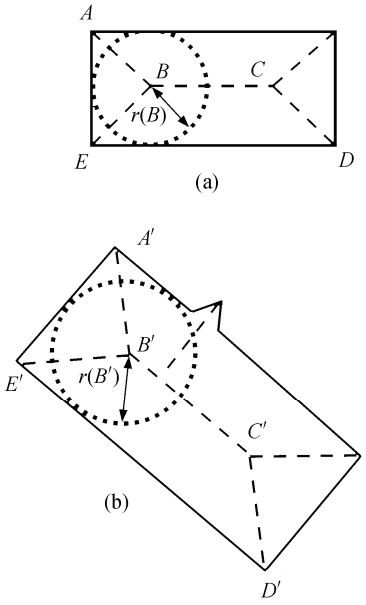
根据 SAR 的稳定性, 我们选取 SAR 为特定值的骨架点作为截断点, 从而保证从相同字体提取的笔端具有最大的相似性。



(a) (b) (c)

图 4 笔画模板与笔端的关系

Fig. 4 Stroke templates and stroke tips



(a) 实线是形状 s, 虚线部分是 s 的骨架; (b) 实线是形状 s', 虚线是 s'的骨架。s'是 s 平移、旋转、放缩并且添加边缘噪声的结果

图 5 笔画长宽比

Fig. 5 Stroke aspect ratio

从骨架上获得截断点之后, 还存在另外一个问题: 如何从骨架段恢复笔端结构。首先, 骨架点与形状上的点没有对应关系。虽然对于简单情况, 可以利用与骨架边垂直的切线切取笔端, 但是当骨架段的截取位置在分叉点附近时, 切线切出的笔端与理想的笔端有很大差异。其次, 笔端的截口形状也会对笔端形状产生影响。如果只用竖直或者水平的切线切取笔端, 那么笔画的角度有变化时, 切口形状会改变笔端的整体形状。根据以上分析, 我们利用中轴变换的逆变换来恢复笔端形状, 即以所截取的骨架段上的点为圆心, 用该点的内切圆覆盖周围区域, 这些圆的并集就可用来恢复笔端形状。由于提取骨架的算法是近似算法, 并且实现中轴变换的逆变换时存在误差, 从截得的骨架段恢复出来的形状不能完全与原始形状吻合, 因此在算法中, 我们将恢复形状时用的圆盘的半径增大几个像素, 再将结果与原始形状进行“与”运算, 将其结果作为最终的笔端形状。

2.2 笔端描述

笔端可以看成只包含单一轮廓、接近于凸多边形的特殊形状, 因此我们采用文献[18]提出的基于轮廓的形状描述子来描述并度量笔端间的相似性, 这种描述子称为基于多尺度曲率直方图的傅立叶描述子(multi-scale Fourier descriptor of curvature histogram, MFDCH)。另外, 由于笔端的提取依赖于汉字骨架, 因此骨架的信息也可以用来增强笔端描述。在本文中提出两种基于骨架的描述子, 与 MFDCH 相结合, 构成笔端形状描述子(shape descriptor of stroke tip, SDST)。笔画间的相似性由 SDST 间的距离来度量。

MFDCH 通过构造形状的曲率直方图来描述形状。首先计算笔端形状的质心, 并以其为原点, 以水平方向为起始线, 建立极坐标系。然后均匀选取极坐标下的 m 个方向, 以及 n 个距离, 由径向线与圆弧构成若干窗格, 使得笔端的每个轮廓点都位于一个窗格内。将 $m\times n$ 个窗格映射到 $m\times n$ 的二维直方图上, 窗格中的轮廓点的曲率按照式(2)累加到二维直方图:

$$X=\frac{\theta(s)+\frac{\pi}{2}}{2\pi}\times n,\quad Y=\frac{d(s)}{D}\times m, \tag{2}$$

其中 X, Y 是点 s 在直方图上的位置$\theta(s)$, 是点 s 的角度, $d(s)$是点 s 与质心间的距离。这种直方图称为曲率直方图。曲率直方图只具备平移不变性, 而形状描述子还需满足旋转不变性与尺度不变性。MFDCH 采用二维傅立叶变换来获得旋转不变性。二维傅立叶变换的公式如下:

$$\mathrm{FD}(u,v)=\frac{1}{mn}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}f(x,y)\mathrm{e}^{-\mathrm{j}2\pi\left(\frac{ux}{m}+\frac{vy}{n}\right)}, \tag{3}$$

$f(x, y)$是像素(x, y)的灰度值。忽略 FD 的相角, 直接使用其幅度$|FD(u, v)|$, 这样得到的向量具备了旋转不变性, 再将向量的每个元素除以第一个元素, 这样获得的新向量就具备了尺度不变性。称这个向量为曲率直方图傅立叶描述子, 用 FDCH 来表示:

$$\mathrm{FDCH}=\left\{\frac{|FD(0,0)|}{|FD(0,0)|},\frac{|FD(0,1)|}{|FD(0,0)|},\ldots,\frac{|FD(m,n)|}{|FD(0,0)|}\right\}。 \tag{4}$$

为了获得更全面的形状特征, MFDCH 引入多尺度平滑的方法, 依据是相似的形状在多尺度高斯平滑过程中产生的中间轮廓也应相似。设 k 次高斯平滑过程中使用的平滑参数依次为σ_1, σ_2,…, σ_k, 则 MFDCH 的表达式如下:

$$\mathrm{MFDCH}=\{\mathrm{FDCH}_{\sigma_1},\mathrm{FDCH}_{\sigma_2},\ldots,\mathrm{FDCH}_{\sigma_k}\}。 \tag{5}$$

由于笔端是根据汉字骨架提取出的形状, 因此骨架信息对于计算笔端相似性十分重要。本文增加了两个基于骨架的特征。第一个称为笔端横向伸展度, 记为 EP, 计算公式如下:

$$\mathrm{EP}=\frac{\text{笔端骨架半径的均值}}{\text{笔端骨架分支长度}}。 \tag{6}$$

EP 值越大, 笔端沿骨架分支伸展的程度越大, 笔端越宽, 越接近于圆形; 反之, 笔端越细长。第二个特征称为笔端平滑度, 记为 SM, 其计算公式如下:

$$\mathrm{SM}=\frac{\text{笔端区域面积}}{\sum(\text{笔端骨架点的半径})^2}。\tag{7}$$

SM 值越小, 说明笔端形状越不规则, 产生越多的骨架分叉。将这两个特征与 MFDCH 结合构成新的描述子, 称为笔端形状描述子(shape descriptor of stroke tip, SDST), 表示为

$$\mathrm{SDST}=(\mathrm{EP},\mathrm{SM},\mathrm{MFDCH})。\tag{8}$$

形状 s_i 和 s_j 间的相似性采用对应的 SDST 间的欧氏距离来度量, 形状间的欧式距离越小, 两个形状越相似; 反之, 两个形状越不相似:

$$D(s_i,s_j)=|\mathrm{SDST}_i-\mathrm{SDST}_j|。\tag{9}$$

3 特征笔端与字体识别

根据上述分析, 一个汉字可由其笔端表示, 进而被描述为若干个 SDST 向量, 不同汉字的 SDST 的数量也不同。对于一种字体, 它可以由多个汉字提取的 SDST 来表示。由于笔端间的相似性, SDST 所包含的信息可能会有冗余, 因此需要计算出尽可能少但又具有代表性的 SDST。具体方法是使用 *K*-均值聚类算法对 SDST 进行聚类, 这些聚类中心记为 $\{c_i\}$, i=1, 2,…k, 其中 $\{c_i\}$称为特征笔端。每种字体可以用这些特征笔端来表示。

假设目前已用 SDST 表示的字体为 M 种, 用 F_k 来表示第 k 种字体。用 $c_{k,j}$ 表示第 k 种字体的第 j 个特征笔端, 用 s_k 表示 F_k 中特征笔端的个数。那么全部字体可以表示为

$$F_k=\{C_{k,j}\,|\,1\leqslant j\leqslant s_k\},\ \forall 1\leqslant k\leqslant M。\tag{10}$$

当需要对一个汉字字符进行识别的时候, 首先将这个字符表示为 $\{\mathrm{SDST}_j\}$, 其中 i=1, 2,…, t, t 为从汉字中提取到的笔端数量。这里将 $\{\mathrm{SDST}_j\}$简记为 $\{S_j\}$, 每个 S_j 所属类别的判断公式如下:

$$S_j\text{属于}F_k,\ \text{如果}P(F_k\,|\,S_j)>P(F_i\,|\,S_j),\ \forall i\neq k。\tag{11}$$

利用 S_j 与 F_i 全部特征笔端的最近距离来判断 S_j 属于 F_i 概率的大小, 即

$$P(F_k\,|\,S_j)>P(F_i\,|\,S_j),$$

$$\text{如果}\min(D(S_j,C_{k,t}))<\min(D(S_j,C_{i,t})),\tag{12}$$

这样, S_j 的类别由与 S_j 距离最小的特征笔端的类别决定。

对全部 r 个 SDST 分类之后, 待识别的汉字字符利用这 r 个 SDST 投票来决定汉字的字体, 得票最多的字体就是待识别汉字字符的字体。

4 实验设计与结果分析

4.1 常用字体集上的字体识别

第一组实验用来验证基于笔端形状形似性的字体识别方法对于常用字体的识别效果。这些字体包括宋体、仿宋、黑体、楷体以及隶书共 5 种字体, 每种字体包含正规、加粗、斜体和粗斜 4 种样式。我们选用国标一级汉字字库中前 798 个汉字用于实验, 汉字图像数据集的获取方法是将所有汉字以 18 号字打印在 A4 纸上, 然后以 1200 dpi 扫描为黑白图像, 再将扫描后的图像分割为单字图像。

在这个数据集合上, 我们随机选取 598 个汉字用于计算特征笔端, 其余 200 个汉字用于测试识别率。我们采用 5 种方式划分测试集, 即将测试集随机划分为 200、100、50、40 和 20 组, 并将每一组字符集视为从同一个文本块所得, 这样, 这 5 种方式对应的文本块大小依次为 1、2、4、5 和 10 个字符。测试一用来测试本文方法对于单字符的字体识别效果, 测试五用来与文献[16]对比。文献[16] 对每种字型使用 20 个文本块来测试, 每个文本块包含 20~30 个同种字型的汉字。测试方法是对每个文本块的全部汉字提取笔画模板, 然后利用提取到的全部笔端识别文本块包含的字型。

提取笔端时 sar 值为 1, *K*-均值聚类算法中的 *K* 设为 40。整组实验被重复执行 3 次, 每次按照上述的 5 种方式测试 20 种字型的识别率, 最后取 3 次实验结果的平均值作为最终的识别率。平均识别率的结果如表 1 所示。其中对比方法是 Sun[16]提出的基于笔画模板的方法。通过表 1 可以看出, 随着用于提取笔端的字符数量增加, 文本块的识别率也随之上升。测试五采用了与对比方法相同数量的文本块

表 1 常用字体集上的平均识别率

Tabel 1 Recognition rate on common font set

测试	识别方法	测试文本块数	文本块含字符数	平均识别率/%
测试一	基于笔端	200	1	74.96
测试二	基于笔端	100	2	87.23
测试三	基于笔端	50	4	95.05
测试四	基于笔端	40	5	95.75
测试五	基于笔端	20	10	98.88
对比方法	基于笔画模版	20	20~30	98.75

用于测试，识别率与对比方法相当，但是每个文本块只包含了10个字符，远远小于对比方法中采用的文本块大小。因此，基于笔端相似性的方法要优于基于笔画模板的方法。

然而，通过表 1 中测试一的结果可见，本文的方法对于包含不同样式的字体集上的识别效果并不理想。这是因为，同一种字体的 4 种样式在笔端形状上的区别很小，容易造成混淆。按照测试一的方式进行实验，但在计算识别率时采用只考虑字体名称，例如，字型为宋体–正规的汉字只要被识别为宋体，不再区分样式之间的差异，都认为识别结果正确，那么 5 种字体的平均识别率为 92.45%。

4.2 大字体集上的字体识别

第二组实验中采用比较大的数据集，这个数据集包括 23 字体。对于每种字体，共收集 396 个字符的图像，每个字符采用初号字，利用计算机以 600 dpi 保存为图像。我们随机选取了 296 个字符来计算聚类中心，再用剩余的 100 个字符做测试。

SDST 的尺度选择为 1, 2, 4, 8, 16 共 5 个尺度，每个尺度上采用 100 维的向量。计算特征笔端的方法使用了 *K*-均值聚类算法。聚类中心数量为 50。表 2 给出了 23 种字体及其识别率。23 种字体的平均识别率为 88.49%，证明基于笔端相似性进行字体识别方法的有效性。从表中还可以看出，个别字体的识别率并不理想，原因之一是有些字体笔端本身与其他字体的笔端相近，仅靠笔端很难区分；另一个原因是从某些汉字中提取的笔端，其截断面占笔端的比例较大，从这部分提取的信息淹没了笔端形状信息，进而导致误识别。

在基于笔端相似性的字体识别方法中，特征笔端用于表示字体信息。特征笔端的个数对本方法能否扩展到更大的数据集上有着至关重要的影响。在保证识别率的前提下，特征笔端个数越少意味着每个特征笔端包含的字体信息越丰富。数量较少的特征笔端一方面可以降低识别时需要比对的笔端描述子的个数，从而减少识别时间，另一方面也可以避免过多的笔端描述子对相似度计算的干扰，从而能够向规模更大的数据集上扩展。图 5 是多组实验的识别率变化线。每组实验除了聚类中心设定的个数外，都采用与 23 种字体实验相同的参数设置。横坐标是设定的聚类中心数，即特征笔端的数量。从图中可以看出，当聚类中心数量超过 40 时，识别率的提高并不明显。这说明对于 23 种字体的识别来说，每种字体的字体信息采用 40 个特征笔端较为合理。

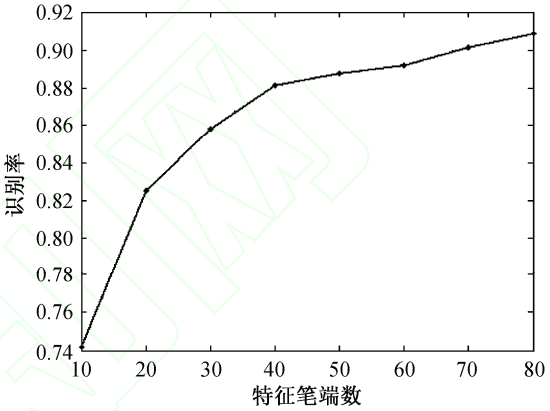


图 6 特征笔端的数量对字体识别的影响

Fig. 6 Relationship of number of eigen stroke tips and the recognition rate

5 总结与展望

本文提出一种基于笔端相似性的汉字字体识别方法，用于解决在较大规模字体数据集上进行字体识别的问题。本文定义了“笔端”这一在人们视觉感知上存在的汉字部件，并且提出一种基于骨架的汉字笔端提取算法，这种结构比现有文献中的同类方法能更精练地表示字体信息。本文还提出一种基于形状的笔端描述子 SDST，并且利用 SDST 间的距离来定义笔端相似度，最后使用 *K*-均值算法将每种字

表 2 包含 23 种字体数据集上的字体识别率

Table 2 Recognition rate on 23-font set

字体	识别率/%	字体	识别率/%	字体	识别率/%	字体	识别率/%
宋体	99.67	仿宋	98.67	黑体	95.33	楷体	98.00
幼圆	98.67	隶书	73.67	华文新魏	64.00	方正胖娃	81.00
方正剪纸	71.33	方正少儿	97.33	方正水黑	80.33	方正北魏楷书	98.00
方正超粗黑简	83.33	方正古隶	79.00	方正琥珀	87.33	方正华隶	90.00
方正铁筋隶书	95.67	方正雅艺	98.67	方正姚体	92.67	方正毡笔黑	89.33
方正大标宋	88.00	方正粗倩	88.33	方正美黑	87.00		

体表示为少量的聚类中心, 即特征笔端。使用本文的方法, 在 5 种常用字体数据集合上进行字体识别实验, 识别率达到 99.6%, 高于同类方法的实验效果, 与以往的字体识别研究最大的区别是, 本文在较大规模的字体集合上取得了较好的识别效果。

不过, 这种方法对文本图像的清晰度要求较高。下一步的工作包括在小字号数据集合上的实验, 以及如何从小字号汉字获得清晰轮廓。另外, 对于少数字体的个别汉字, 笔端不易提取, 因此还可将笔端方法与其他方法相结合, 进一步提高识别率。

参考文献

[1] Solli M, Lenz R. FyFont: find-your-font in large font databases // Proceedings of the 15th Scandinavian conference on Image analysis. Aalborg, Denmark: Springer-Verlag, 2007: 432–441

[2] Lidke J, Thurau C, Bauckhage C, The Snippet Statistics of Font Recognition // 20th International Conference on Pattern Recognition. Istanbul, 2010: 1868–1871

[3] Zhu Yong, Tan Tieniu, Wang Yunhong. Font recognition based on global texture analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(10): 1192–1200

[4] 杨芳, 田学东, 郭宝兰. 用于字体识别的 Gabor 滤波角度 GA 优化方法. 计算机工程与应用, 2003(25): 83–85

[5] 许春晔, 郭宝兰. 基于 Gabor 函数的汉字字体识别. 河北大学学报: 自然科学版, 2001(2): 167–170, 190

[6] 田学东, 郭宝兰. 基于 Gabor 变换的汉字字体识别研究. 计算机工程与应用, 2002(20): 89–91

[7] 田学东, 郭宝兰. 基于纹理特征的汉字字体识别研究. 计算机工程, 2002(6): 156–157

[8] 朱学芳, 邹文豪, 朱鹏. 基于 Gabor 函数的字体识别实验研究. 第十五届全国图象图形学学术会议论文集, 广州: 清华大学出版社, 2010: 204–209

[9] 朱学芳, 邹文豪, 王栾井. 对字体识别中 Gabor 滤波器参数的实验研究 // 第六届全国信息获取与处理学术会议论文集(1). 焦作: 仪器仪表学报杂志社, 2008: 424–428

[10] Yang Zhihua, Yang Lihua, Qi Dongxu, et al. An EMD-based recognition method for Chinese fonts and styles. Pattern Recognition Letters, 2006, 27(14): 1692–1701

[11] 杨志华, 齐东旭, 杨力华, 等. 基于经验模式分解的汉字字体识别方法. 软件学报, 2005(8): 1438–1444

[12] Ding Xiaoqing, Chen Li, Wu Tao. Character independent font recognition on a single chinese character. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(2): 195–204

[13] 陈力, 丁晓青. 基于小波特征的单字符汉字字体识别. 电子学报, 2004(2): 177–180

[14] 王连银. 基于小波变换的印刷体汉字字体识别研究. 科技资讯, 2007(14): 61–62

[15] 王洪, 汪同庆, 刘建胜, 等. 基于小波包纹理分析的字体识别方法. 光电工程, 2002(增刊 1): 62–65

[16] Sun Hungming. Multi-linguistic optical font recognition using stroke templates // 18th International Conference on Pattern Recognition. Hong Kong, 2006: 889–892

[17] 王恺, 靳简明, 史广顺, 等. 基于特征点的汉字字体识别研究. 电子与信息学报, 2008(2): 272–276

[18] 吕肖庆, 李沫楠, 蔡凯伟, 等. 一种基于图形识别的甲骨文分类方法. 北京信息科技大学学报: 自然科学版, 2010(增刊 2): 92–96