ISSN1000-9825 Journal of Software 软件学报 2000,11(6):804~808

P形傅里叶变换在手写体汉字识别中的应用*

姚丹霖 陈火旺 殷建平

(国防科学技术大学计算机科学系 长沙 410073)

E-mail: ydl@chinabytemail.com

摘要 提出了曲线化汉字特征的概念,讨论了3类可曲线化的汉字统计特征.利用P形傅里叶算子,对这些曲线化特征作傅里叶变换,可提取到最终的识别特征.实验结果表明,该方法在手写体汉字识别领域具有良好的应用前景,尤其适用于细分类.

关键词 手写体,汉字识别,统计特征,细分类,傅里叶变换.

中图法分类号 TP391

用于手写体汉字识别(handwritten Chinese character recognition,简称HCCR)的汉字特征分为两大类:结构特征和统计特征.结构特征建立在笔画的基础之上,通常以笔画或笔画构成的部件之间的空间结构关系来描述汉字,能够体现汉字的本质特征.统计特征则建立在二值或灰度值点阵图像基础之上,通常是对汉字点阵信息进行非线形变换后提取,任何一种统计特征都难以体现汉字结构的本质.

对于手写体汉字,要准确提取到结构特征是极其困难的,因而难以采用基于结构特征的方法加以识别.尽管任何一种统计特征都与汉字的本质特征有较大区别,但由于其具有多样性、互补性、易于提取等优点,使其在手写体汉字识别领域占有十分突出的地位[1,2].不过,由于这些系统中的识别方法只对统计特征进行一些简单的变换,其性能不够理想.如果能够对统计特征进行某种复杂的变换,消除不具代表性的因素,则完全可以提高其性能.

频谱特征抽取法便是这样一种特征变换法.其基本思想是将平面上的曲线通过频带来表示,用傅里叶级数算子来描述.相似的原特征序列,在经过傅里叶变换后所得到的新特征序列中,其低频部分具有明显的差异.

文献[3]对传统的傅里叶算子进行了改进,使之适用于渐开曲线,称为P形傅里叶算子,简称P形算子.我们利用P形算子,使之作用于手写体汉字的多种统计特征,从中提取到对应的频谱特征,用于汉字识别,取得了良好的效果.

P形傅里叶算子最初应用于图像压缩和图像识别,本文表明,P形傅里叶算子同样也适用于文字识别.

1 P形傅里叶算子

设Z为二维平面上的一条曲线,将坐标系设为复平面坐标系,曲线Z的轨迹用该坐标系下的离散点阵坐标序列:$Z(0),\ldots,Z(k),\ldots,Z(m)$来表示.

首先,将长度$\delta_k=|Z(k+1)-Z(k)|(0\leqslant k<m)$不相等的上述轨迹坐标序列近似地变换成长度为定长$\delta$的轨迹坐标序列:$\tilde{Z}(0),\ldots,\tilde{Z}(k),\ldots,\tilde{Z}(n)$.

其次,记两个矢量$\tilde{Z}(k)-\tilde{Z}(k-1)$和$\tilde{Z}(k+1)-\tilde{Z}(k)$之间的夹角为$\alpha(k)(0<k<n)$,利用$\alpha(k)$定义曲线$Z$的角度变化的累积函数的全曲率函数如下:

---

* 作者姚丹霖,1963年生,工程师,主要研究领域为模式识别,软件工程.陈火旺,1935年生,教授,博士生导师,中国工程院院士,主要研究领域为软件工程,人工智能.殷建平,1963年生,博士后,教授,主要研究领域为软件工程,中文信息处理.

本文通讯联系人:姚丹霖,长沙 410073,国防科学技术大学计算机科学系

本文1998-09-29收到原稿,1999-06-11收到修改稿

$$\theta(0)=0,$$
$$\theta(k)=\theta(k-1)+\alpha(k),\quad k=1,\ldots,n-1. \tag{1}$$

利用该全曲率函数,定义复数函数 W 如下:

$$W(k)=\exp(\mathrm{i}\theta(k)),\quad k=0,\ldots,n-1. \tag{2}$$

对 P 表现 W,作离散傅里叶变换,得到

$$C(k)=\frac{1}{n}\sum_{j=0}^{n-1}W(j)\exp(-\frac{2jk\pi}{n}\mathrm{i}),\quad k=0,\ldots,n-1. \tag{3}$$

解式(3)可得

$$W(j)=\sum_{k=0}^{n-1}C(k)\exp(\frac{2jk\pi}{n}\mathrm{i}),\quad j=0,\ldots,n-1. \tag{4}$$

这表明,C 和 W 一一对应,函数 C 即为曲线 Z 的 P 形傅里叶算子(简称 P 形算子). 在 W 函数的定义中只涉及到曲线的曲率变化,因此,曲线的平移和缩放对函数 W 无影响,对函数 C 亦然.

令 N 为一个正整数,且 $N\leqslant n/2$,由 $C(k)$ 出发可定义 $\bar{C}(k)$:

$$\bar{C}(k)=\begin{cases}C(k) & k=0,1,\ldots,N-1,\\ C(n+k) & k=-N+1,\ldots,-2,-1,\\ 0 & \text{否则}.\end{cases} \tag{5}$$

称 $\bar{C}(k)(-N+1\leqslant k\leqslant N-1)$ 为曲线 Z 的 N 次 P 表现系数.

2 汉字曲线化特征

从前文所述可以看到,P 形算子只适合于二维平面中的连续曲线(不一定是光滑的),因此,为了对统计特征采用 P 形傅里叶算子进行变换,必须将此特征表示成二维平面中的一条连续曲线.

定义 1. 如果某类统计特征可用二维平面中的一条连续曲线来表示,则称该类统计特征是可曲线化的.

定义 2. 如果某类统计特征是可曲线化的,则将其曲线化后,得到其在二维平面中的曲线描述,称为曲线化特征.

定义 3. 所有能完整包围某汉字且其边与坐标轴平行的矩形中之面积最小者,称为该汉字的字形外框. 显然,字形外框具有 4 个不同的边缘,分别为左边缘、上边缘、右边缘和下边缘.

下面分别讨论 3 类可曲线化的统计特征:周边轮廓特征、笔画密度特征和区域投影轮廓特征. 为方便起见,在下文中如无特殊说明,均假定汉字已尺寸规范化,其外框大小为 $n\times n$.

2.1 周边轮廓(four-sides contour,简称 4SC)特征

定义 4. 对汉字二值化点阵图像中某一行,以汉字字形外框的左边缘为起点,水平向右扫描到第 1 个有效笔画或右边缘止,所经过的像素数,称为该行的左边缘轮廓特征值. n 个左边缘轮廓特征值构成的特征向量,称为左边缘轮廓特征.

与左边缘轮廓特征相类似,可定义汉字的右边缘、上边缘及下边缘轮廓特征.

定义 5. 由汉字的左、上、右、下 4 个边缘轮廓特征 F_{L4SC},F_{T4SC},F_{R4SC},F_{B4SC} 所组成的有序序列,称为该汉字的周边轮廓特征. 记为 $F_{4SC}=(F_{L4SC},F_{T4SC},F_{R4SC},F_{B4SC})$.

记 F_{L4SC},F_{T4SC},F_{R4SC},F_{B4SC} 中第 $i(0\leqslant i\leqslant n)$ 个分量分别为 F_{L4SC_i},F_{T4SC_i},F_{R4SC_i},F_{B4SC_i},将周边轮廓特征曲线化后,得到曲线化特征为

$$Z_{4SC}:Z_{4SC}(0),\ldots,Z_{4SC}(k),\ldots,Z_{4SC}(4n-1),$$

其中

$$Z_{4SC}(k)=a_k+b_k\mathrm{i},\quad 0\leqslant k<4n,$$

$$\begin{cases}a_k=k,\\ b_k=\begin{cases}F_{L4SC_k}, & \text{若 } 0\leqslant k<n,\\ F_{T4SC_{k-n}}, & \text{若 } 0\leqslant k<2n,\\ F_{R4SC_{k-2n}}, & \text{若 } 2n\leqslant k<3n,\\ F_{B4SC_{k-3n}}, & \text{若 } 3n\leqslant k<4n.\end{cases}\end{cases}$$

2.2 笔画密度(stroke density)特征

定义 6. 对汉字二值化点阵图像中的某一行,以汉字字形外框的左边缘为起点,水平向右扫描到右边缘为止,所经过的黑像素段数(有效笔画数),称为该行的纵向笔画密度特征值.n 个纵向笔画密度特征值构成的特征向量,称为纵向笔画密度特征.

与纵向笔画密度特征相类似,可定义汉字的横向、+45°对角线方向和−45°对角线方向的笔画密度特征.这 4 个方向的笔画密度特征统称为 VH2D(vertical-horizontal-diagonal-diagonal)笔画密度特征.

定义 7. 由汉字的纵向、横向、+45°对角线方向和−45°对角线方向 4 个笔画密度特征 F_{VSD},F_{HSD},F_{+45SD},F_{-45SD}所组成的有序序列,称为该汉字的 VH2D 笔画密度特征.记为

$$F_{SD}=\langle F_{VSD},F_{HSD},F_{+45SD},F_{-45SD}\rangle.$$

记 F_{VSD},F_{HSD}中第 $i(0\leqslant i\leqslant n)$个分量分别为 F_{VSD_i},F_{HSD_i},记 F_{+45SD},F_{-45SD}中第 $i(0\leqslant i\leqslant 2n-1)$个分量分别为 F_{+45SD_i},F_{-45SD_i},则将笔画密度特征曲线化后,得到曲线化特征为

$$Z_{SD}:\quad Z_{SD}(0),\ldots,Z_{SD}(k),\ldots,Z_{SD}(6n-3),$$

其中

$$Z_{SD}(k)=a_k+b_k\mathbf{i}\quad(0\leqslant k<6n-2),$$

$$\begin{cases}a_k=k,\\ b_k=\begin{cases}F_{VSD_k}, & \text{若 } 0\leqslant k<n,\\ F_{HSD_{k-n}}, & \text{若 } n\leqslant k<2n,\\ F_{+45SD_{k-2n}}, & \text{若 } 2n\leqslant k<4n-1,\\ F_{-45SD_{k-4n+1}}, & \text{若 } 4n-1\leqslant k<6n-2.\end{cases}\end{cases}$$

2.3 分区投影轮廓(regional projection contour,简称 RPC)特征

定义 8. 以汉字字形外框为基准,将汉字用纵向和横向二轴线分成 4 个区域,对每个区域内构成汉字笔画的黑像素,沿图 1(c)所示方向将其投影到轴线上,得到内部非空的投影实体,称为纵横分区投影.

(a) 纵向 (a) Vertical　(b) 横向 (b) Horizontal　(c) 纵横 (c) Vertical-Horizontal　(d) 双对角线 (d) Diagonal-Diagonal

Fig. 1 The regions and projection directions for regional projection contour features extraction

图 1 提取分区投影轮廓特征时的分区及投影方向

定义 9. 设 S 为汉字纵横分区投影实体,以汉字外框的左下角为坐标原点,水平向右为 X 轴,垂直向上为 Y 轴,从 45°直线与投影实体的第 1 个相交点出发,沿顺时针方向绕实体边缘一周,所经过的坐标轨迹,称为该汉字的纵横(vertical-horizontal)分区投影轮廓特征.

与纵横分区投影轮廓特征相类似,可定义汉字的纵向(vertical)、横向(horizontal)、双对角线(diagonal-diagonal)分区投影轮廓特征.纵向、横向、纵横、双对角线 4 个分区投影轮廓特征分别记为 F_{VRPC},F_{HRPC},F_{VHRPC},F_{DDRPC}.图 1 给出了汉字的 4 种分区投影的轴线及其投影方向.

由定义可知,每个分区投影轮廓特征对应着二维平面中的一条曲线,因此,只需将定义 9 中的坐标系改为复平面坐标系,即可得到汉字分区投影轮廓之曲线化特征,分别记为:Z_{VRPC},Z_{HRPC},Z_{VHRPC},Z_{DDRPC}.

3 识别系统

基于本文已讨论的 P 形算子和 3 类可曲线化统计特征,我们建立了一个实验性识别系统.该系统由预处理器、特征提取与曲线化、P 形傅里叶变换和分类器等部分构成,如图 2 所示.

输入的原始二值化汉字点阵,由预处理器进行预处理.首先剔除飞斑、消除笔画内白点,然后进行梯度变换

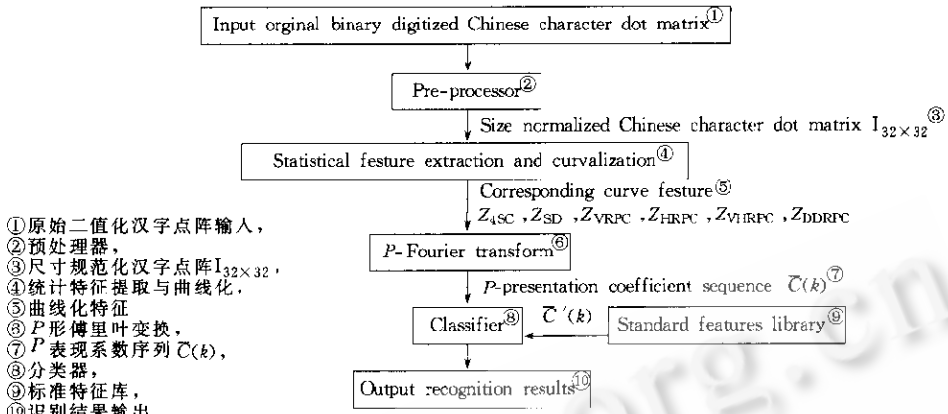

Fig. 2 The structure of recognition system

图2 识别系统结构

得到汉字笔画的轮廓[4],找出字形外框、计算文字重心,最后将其变换成 32×32 的统一尺寸.规范化后的汉字点阵送入特征提取与曲线化部分以提取汉字的 3 类共 6 个曲线化特征.

特征提取与曲线化部分依次提取汉字的周边轮廓特征、笔画密度特征、分区投影轮廓特征,并对提取到的特征进行曲线化,分别得到 6 个曲线化特征.

P 形傅里叶变换部分依据式(1)~(5)完成对曲线化特征的离散傅里叶变换,所得到的全体 P 表现系数序列即为识别特征.在将间距不等长的离散坐标点序列转换成间距等长的离散坐标点序列时,所取之长度为 $\delta=1$.

分类器以识别特征和标准特征之间的欧氏距离为判别准则,与识别特征距离最小的标准特征所对应的汉字为首候选字,次最小者为第 2 候选字,依此类推.最多保留 4 个候选字.

4 实验结果及分析

基于以上识别系统,我们对国标一级 3 755 个汉字进行了识别实验.共有 15 位志愿者,每个志愿者在带方框的稿纸上对国标一级 3 755 个汉字较为工整地书写 2 遍,共得到 30 份,总计 112 650 个汉字样本.从中选出 4 份共 15 020 个汉字样本,对每个样本,提取 6 个统计特征,曲线化后再变换成间距等长的离散坐标点序列.然后,分别取 $N-10$,$N-20$ 和 $N-30$ 对其进行 P 形傅里叶变换,得到对应样本的 10 次、20 次和 30 次 P 表现系数,构成 3 个标准特征库.对同一个汉字,在每个标准特征库中均保留了 4 个不同样本的全部特征.余下的 97 630 个样本,作为测试样本,进行识别实验.实验结果见表 1.

Table 1 Four accumulated recognition rates of the experimental system

表 1 实验系统的前 4 位识别率

N	The first one recognition rate①(%)	The first two recognition rate②(%)	The first three recognition rate③(%)	The first four recognition rate④(%)
10	88.12	91.31	93.13	94.34
20	97.01	98.04	98.57	98.98
30	97.73	98.80	99.12	99.50

①首位识别率,②两位识别率,③3 位识别率,④4 位识别率.

从表 1 的数据可以看出,3 个 P 形算子中,30 次 P 形算子具有最好的识别结果,20 次 P 形算子次之,10 次 P 形算子最差.就 $N=20$ 和 $N=30$ 的识别结果进行对比,其差异并不显著.

系统提取统计特征的时间复杂性为 $O(n^2)$,P 形傅里叶变换的时间复杂性为 $O(n^3)$,总的时间复杂性为 $O(n^3)$.

5 结束语

本文提出了以间距等长的离散坐标点序列表示的曲线来描述汉字统计特征的方法,利用 P 形傅里叶算子

对曲线作离散傅里叶变换后提取频谱特征,用于识别手写体汉字.在此基础上实现了一个实验系统,对国标一级3 755个汉字进行识别实验,在最好的情况下取得了99.56%的良好结果,表明本方法是有效的.

由于对曲线化特征进行变换后所得到的频谱特征(P表现系数)数据量较大,因此,本方法适用于细分类,而不适用于粗分类.

考虑到汉字的复杂性和手写汉字的多样性,当字符集扩大或者样本差异性增大时,系统的识别率必然会下降.在此情况下,可以有两种解决途径:① 增加其他类型的统计特征以消除相似性;② 增加P表现的次数N以提高正确的识别率.另外,尚需考虑如何在不增加特征的情况下有效地实施粗分类以及如何有效地降低运算复杂性.这些都是有待今后进一步研究的课题.

参考文献

1 Cheng Heng-da, Xia D C. A novel parallel approach to character recognition and its VLSI implementation. Pattern Recognition, 1996,29(1):97～119

2 Park H S, Lee S W. Off-Line recognition of large-set handwritten characters with multiple hidden Markov models. Pattern Recognition, 1996,29(2):231～244

3 Uesaka Yosinori. A novel Fourier operator suitable for involute. Information Science Theory, 1984,J67-A(3):166～173
(上坂吉則.开曲线にも适用できる新しいフーリエ记述子.信学论,1984,J67-A(3),166～173)

4 Yao Dan lin, Yin Jian-ping. Gradient-Based feature extraction approach for Chinese character recognition. Journal of Yunnan University (Natural Sciences Edition), 1997,19(supplement):279～282
(姚丹霖,殷建平.基于梯度的汉字识别特征提取方法.云南大学学报(自然科学版),1997,19(增刊):279～282)

Application of *P*-Fourier Transform to Handwritten Chinese Character Recognition

YAO Dan-lin CHEN Huo-wang YIN Jian-ping

(*Department of Computer Science National University of Defense Technology Changsha* 410073)

Abstract In this paper, the conception of Chinese character's corresponding curve feature is proposed, and three kinds of statistical features of Chinese characters, which can be represented by curve, are discussed. Applying *P*-Fourier operator to them, the final features for character recognition are extracted. Experimental results indicate that the proposed method is very promising for the recognition of handwritten Chinese characters, and is especially suitable for the fine classification.

Key words Handwritten character, Chinese character recognition, statistical feature, fine classification, Fourier transform.