脱机手写汉字识别的最优采样特征新方法

张 睿 丁晓青 方 驰
(清华大学电子工程系智能技术与系统国家重点实验室, 北京 100084)

摘 要 在脱机手写汉字识别中,笔画形变是造成识别率下降的主要原因,减少笔画形变的影响是提高脱机手写汉字识别率的关键.针对上述问题,提出了最优采样特征.该特征以目前被广泛应用的方向线素特征为基础,在一定的约束条件下,通过移动采样点的位置,可以适应笔画的形变.从而减少特征的类内方差,提高特征的可分性,改进了识别性能.通过在 THCHR 样本集上进行实验,并对最优采样特征和方向线素特征的实验结果进行比较,验证了最优采样特征的识别率优于方向线素特征.

关键词 脱机手写汉字识别 最优采样特征 统计模式识别方法
中图法分类号: TP391.43 **文献标识码**: A **文章编号**: 1006-8961(2002)02-0176-05

New Method of Optimal Sampling Features for Offline Handwritten Chinese Character Recognition

ZHANG Rui, DING Xiao-qing, FANG Chi
(*State Key Laboratory of Intelligent Technology and Systems, Department of Electronic Engineering, Tsinghua University, Beijing* 100084)

Abstract In offline handwritten Chinese character recognition, the high variability of the handwriting strokes is the main cause for lowering the recognition performance, thus decreasing the variability of the handwriting strokes is one effective and important way to improve the recognition accuracy. To solve this problem, we propose a new method of optimal sampling features, which are developed from the prevalently used directional features by following procedures. Firstly, four directional factor images are generated from an input binary character image. Next, these four images are transferred through a low-pass filter, and then these four low-passed images are sampled. The image values at these sampling positions produce a feature vector that is defined as sampling features. In the case of the sampling positions are uniform and fixed, the sampling features are subject to stroke variations, and these stroke variations will increase the within class pattern variability. In order to compensate for stroke variations, the sampling positions should be adaptable to these stroke variations. That is, the sampling positions should be displaced against reference patterns to decrease the within class variability, on the other hand the smoothness of the displacement should be preserved to keep the character's primary structure unchanged. The sampling features satisfying above conditions are defined as optimal sampling features. These two conditions could be expressed as a constrained minimization problem, thus optimal sampling features could be solved in an iteration procedure. For the sake of saving the time cost, a coarse-to-fine strategy is utilized. Finally, optimal sampling features are obtained, the discrimination of features is increased; and the recognition performance is improved. In order to demonstrate the effectiveness of optimal sampling features, we apply it to the THCHR database and compare it with directional features. The result shows that sampling features achieve higher recognition accuracy than directional features.

Keywords Offline handwritten Chinese character recognition, Optimal sampling features, Statistical pattern recognition

基金项目:国家 863 高技术计划(863-306-ZT03-03-1);国家自然科学基金(69972024)
收稿日期:2000-10-09;**改回日期**:2001-04-23

0 引　言

脱机手写汉字识别是光学字符识别(OCR)中的一个重要内容,是由计算机对输入的手写汉字图象进行识别,并输出相应的汉字内码.由于脱机手写汉字的笔画复杂、形变大、模式类别多,所以脱机手写汉字识别是字符识别领域中的难题.而笔画形变大是造成识别率下降的主要原因,因此减少笔画形变是脱机手写汉字识别中的重要内容.以前的研究人员提出了许多规一化方法[1],用以减少汉字图象上手写笔画的形变,其中,线性规一化方法能够减少汉字整体的大小和位置形变.非线性规一化方法基于密度均衡原则,能实现笔画均匀分布,减少汉字局部的形变.虽然线性和非线性规一化方法对于减少手写笔画形变都起到了一定的效果,但它们对汉字图象的变换都没有参考模板,与类别无关,所以限制了这些方法的效果.Wakahara 提出了自适应规一化方法[2],该方法将输入的汉字图象以各模板图象为参照进行动态变换[3],从而更大程度地减少了笔画形变.由于该方法对每个象素点都要做迭代运算,运算量巨大,为此采用骨骼化来减少象素点的数量,而在骨骼化过程中,又将引入了新的畸变,这也是该方法的局限之处.

在模式识别中,特征是决定识别性能的一个关键.经验证表明,目前的脱机手写汉字识别中,方向线素特征是最有效的特征[4],并得到了广泛的应用[5].本文以方向线素特征为基础,在生成特征的过程中,借鉴自适应规一化方法中的动态变换思想,提出了最优采样特征.该特征减少了笔画形变引起的类内方差,提高了特征可分性.由于特征维数远小于象素点数,因此该方法相比于自适应规一化方法可以极大程度地减少运算时间.

1 采样特征

采样特征以方向线素特征为基础,特征的生成方法如图 1 所示,过程如下:

(1) 对输入的二值汉字图象进行预处理,即进行规一化和边缘平滑[1],得到 64×64 的二值汉字图象 $f(x,y),x,y=0,1,2,\cdots,64$.

(2) 采用与方向线素特征相同的方法[4],提取汉

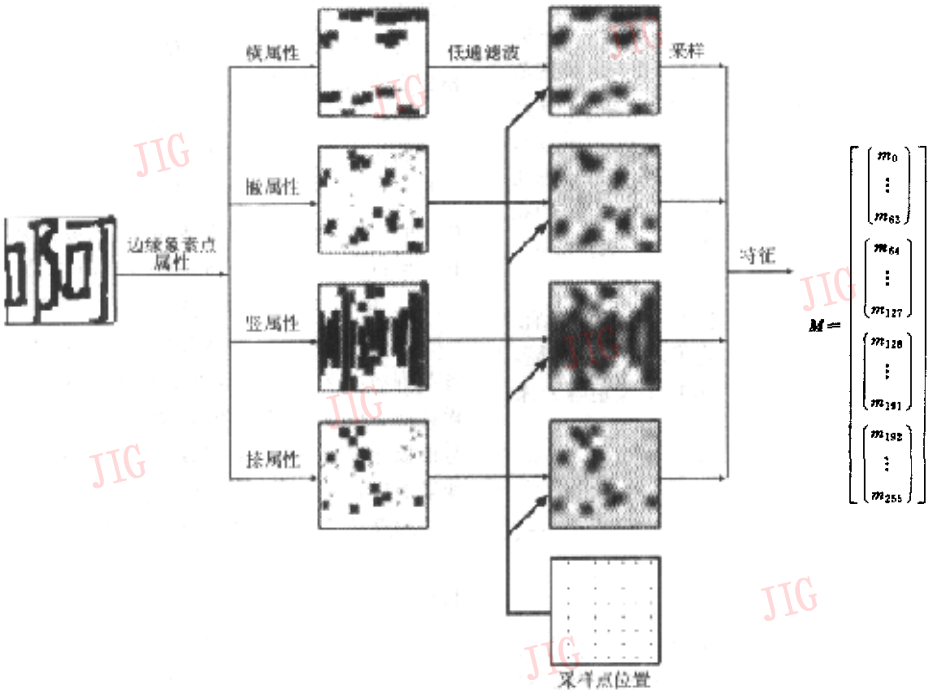

图 1　采样特征的生成过程

字的轮廓,并确定每个轮廓点的方向属性值:在 3×3 范围内根据汉字轮廓点与其相邻点连线的倾斜方向,将该方向在横、撇、竖、捺 4 种方向属性上进行量化.例如,当连接方向为 0°、45°、90°或 135°时,对应的方向属性值为 1,而其余 3 个方向属性值为 0.若连接方向为其他角度时,则与该角度相近的 2 个方向属性值分别为 0.5,而其余 2 个方向属性值为 0.

(3) 将倾斜方向的量化结果作为该轮廓点的方向属性值.再根据每个轮廓点的 4 个方向属性值,将图象 $f(x,y)$分成各代表不同方向属性的 4 幅图象 $f_k(x,y)(k=0,1,2,3)$,且每幅图象只包含该方向属性值非 0 的轮廓点.

(4) 对各方向属性图象分别进行二维快速 DCT 变换,得到相应的频域图象

$$F_k(p,q)=\mathrm{DCT}[f_k(x,y)] \tag{1}$$

(5) 对各频域图象截取 8×8 的低频分量,去掉高频分量,得到

$$F_k'(p,q)=\begin{cases}F_k(p,q) & 0\le p,q\le 8\\ 0 & p,q>8\end{cases} \tag{2}$$

(6) 对截断后的频域图象进行二维快速 IDCT 变换得到原方向属性图象 $f_k(x,y)(k=0,1,2,3)$的低通图象

$$f_k'(x,y)=\mathrm{IDCT}[F_k'(p,q)] \tag{3}$$

(7) 在 4 幅低通图象上进行采样,并且 4 幅图象的采样位置保持一致,各有 64 个采样点.这些采样点排成 8 行 8 列,采样点记为$(i,j)(i,j=0,1,\cdots,7)$,表示该采样点在排列中的行列顺序.这些采样点处的图象值构成了特征矢量,其维数是 8×8×4=256 (8 行,8 列,4 幅图象).该矢量就是汉字的采样特征

$$N=(n_0,n_1,\cdots,n_{254},n_{255})^{\mathrm{T}} \tag{4}$$

在低通图象 $f_k'(x,y)$中,采样点位置是可变的.对于某幅图象,采样点的位置决定了所生成的特征,采样点的位置比较重要,这里用采样位置矩阵 P 来表示这些采样点的位置

$$P=\begin{pmatrix}p_{0,0} & p_{0,1} & \cdots & \cdots & p_{0,7}\\ p_{1,0} & & \vdots & & p_{1,7}\\ \vdots & \cdots & p_{i,j} & \cdots & \vdots\\ \vdots & & \vdots & & \vdots\\ p_{7,0} & p_{7,1} & \cdots & \cdots & p_{7,7}\end{pmatrix} \tag{5}$$

P 的大小是 8×8.每个元素 $p_{i,j}(x_{i,j},y_{i,j})^{\mathrm{T}}$ 是矢量,表示(i,j)采样点在图象 $f_k'(x,y)$中的坐标.对某幅低通图象来说,采样特征 N 依赖于采样位置矩阵 P,于是采样特征可以记为 $N(P)$.

2 最优采样特征

2.1 均匀采样特征

在生成采样特征的过程中,当采样位置为等间距均匀分布时,采样位置矩阵记为 $P^{(0)}$,得到的采样特征定义为均匀采样特征,记为 $N(P^{(0)})$.

由二维 DCT 变换的性质可知,在生成方向线素特征过程中均匀分块再求和的运算[4]等价于对图象低通滤波再进行均匀采样,因此均匀采样特征等价于方向线素特征.

由于均匀采样特征的采样位置固定,而手写汉字的笔画形变又较大,所以均匀采样特征容易受笔画形变的影响,笔画形变引起特征的类内方差增加,特征的可分性下降.

2.2 最优采样特征

为了克服手写汉字的笔画形变,特征的生成过程中采样位置应能适应笔画形变,即可根据模板对采样位置进行动态移动.此时采样位置矩阵 P 是一个变量,而在采样位置的变化过程中,既要减少笔画形变,又要保持汉字基本结构的稳定,则采样位置的移动应符合以下条件:

(1) 采样位置移动后,输入汉字的采样特征与模板之间的距离应减小.

(2) 采样位置的移动应满足平滑性.

满足上述条件的采样位置矩阵记为 $P^{(E)}$,生成的特征定义为最优采样特征,记为 $N(P^{(E)})$.上述的 $P^{(E)}$可以通过最优化问题来求解.该最优化问题由目标函数和约束条件组成,其中目标函数为

$$\min_P D(P,\overline{M})=\min_P\|N(P)-\overline{M}\|^2 \tag{6}$$

其中,$N(P)$表示输入汉字的采样特征;$\overline{M}$ 表示各类别的参考模板,其可以是该类训练样本均匀采样特征 $N(P^{(0)})$的平均值;$D(P,\overline{M})$是输入汉字特征与模板之间欧氏距离的平方(为了记述简单,省略了变量中代表不同模式类别的下标).

约束条件由下述迭代公式表示:

$$p_{i+m,j+n}^{(l+1)}-p_{i+m,j+n}^{(l)}=T_mT_n(p_{i,j}^{(l+1)}-p_{i,j}^{(l)}) \tag{7}$$

其中

$$T_m=\exp(-m^2/k) \tag{8}$$

$$T_n=\exp(-n^2/k) \tag{9}$$

$p_{i,j}^{(l)}$是第 l 次迭代中采样点(i,j)的位置$(x_{i,j},y_{i,j})^{\mathrm{T}}$;$p_{i,j}^{(l+1)}-p_{i,j}^{(l)}$是该采样点的移动;则

$p^{(l+1)}_{i+m,j+n}-p^{(l)}_{i+m,j+n}$表示为了保证平滑性,相邻采样点$(i+m,j+n)$相应的移动.

T_m,T_n定义为平滑函数.该函数决定了受采样点(i,j)移动的影响,相邻采样点$(i+m,j+n)$的移动.为了保证移动的平滑性,随着相邻采样点$(i+m,j+n)$与(i,j)距离的增大,采样点$(i+m,j+n)$受(i,j)影响而进行的相应移动应该逐渐减小,如图 2 所示.式中k决定了当采样点(i,j)移动时,对相邻采样点$(i+m,j+n)$位置移动的影响程度.通过实验比较,k值取为 3.

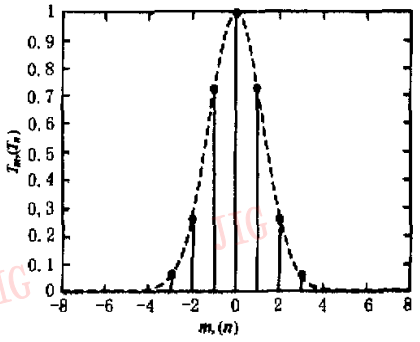

图 2　T_m,T_n函数的形式

在第l次迭代中,某个采样点(i,j)的位置在限定的范围内可以自由移动,同时按公式(7)的约束条件对其相邻的采样点$(i+m,j+n)$也进行相应的移动,选择使公式(6)中目标函数在上述移动中取值最小的采样位置$P^{(l)}$,该次迭代结束;然后在采样位置$P^{(l)}$的基础上,按顺序对下一个采样点的位置也在限定的范围内自由移动,而其他相邻点进行相应的移动,如此循环,直至目标函数不再减小,迭代过程结束.其中,限定采样点自由移动的范围是考虑到汉字笔画变形的实际程度,即可以减少运算时间,同时能更好地保证汉字基本结构的稳定.在 64×64 的图象中,采样点在x,y方向上的移动范围都指定为±8.

由式(6)、(7)定义的优化问题中,目标函数适应了笔画形变,同时约束条件保证了移动的平滑性.通过迭代运算求解该优化问题,就能得到最优采样位置矩阵$P^{(E)}$,然后在 4 幅低通图象$H_i(m,n)$$(i=1,2,3,4)$中,按矩阵$P^{(E)}$所指定的采样位置对低通图象(式(3)所描述的图象)进行采样,即得到最优采样特征$N(P^{(E)})$.

3　两级分类器

分类器采用最小距离分类器,即将输入汉字的最优采样特征与各模板进行匹配,将距离最小的模板对应的内码作为识别结果.由于汉字的类别多,因此为了减少运算时间消耗,可以采用两级分类器.

在粗分类中,采用均匀采样特征.因均匀采样特征没有求解最优采样位置的过程,故时间消耗较少.根据输入汉字的均匀采样特征与全体3 755个模板的欧氏距离,产生距离最小的 10 个候选字.

在细分类中,采用最优采样特征.参照 10 个候选模板分别得到对应的最优采样特征,同时得到与各候选模板的欧氏距离.将距离最小的模板对应的内码作为识别结果,输入汉字的识别过程结束.由于在生成最优采样特征的过程中,减少了模板的数量,因此速度可以得到极大的提高.

4　实验结果及讨论

实验中,所采用的脱机手写汉字样本库是 THCHR 样本集,共计1 500套样本,每套样本包含的字符集是国标一级3 755个汉字.根据汉字的识别率及书写工整程度将样本库分成训练样本集和测试样本集,两个样本集的识别率及书写工整程度基本相当,其中训练样本集包含1 400套样本,测试样本集包含 100 套样本.图 3 所示是其中的部分样本.

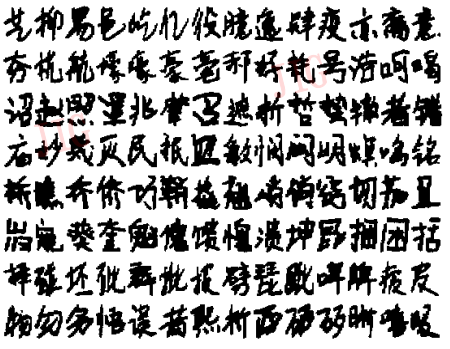

图 3　THCHR 样本集中的部分样本

如前所述,生成最优采样特征的过程是在约束条件下,最小化输入汉字采样特征与模板之间欧氏距离的迭代过程.可见,随着迭代次数的增加,输入汉字的采样特征与每个模板之间的距离都将逐渐减

小. 经过实验证明:由于约束条件满足移动的平滑性,保证了汉字的基本结构不变,因此在迭代过程中,随着采样位置的移动,输入汉字与正确模板之间距离减小的幅度比其他模板大,即提高了特征的可分性.

图 4 所示是用易混淆字"哀、衰、衷"来说明生成最优采样特征过程中的目标函数,即特征与模板之间距离的变化.

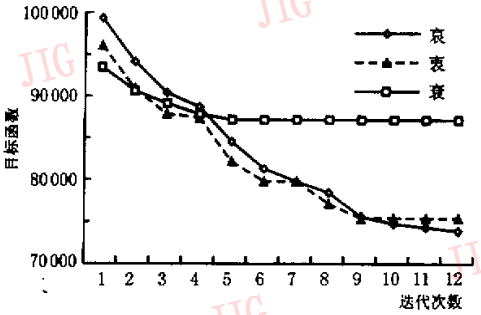

图 4 目标函数与迭代次数的关系

输入汉字是"哀". 在迭代次数为 0 时,输入与模板"哀"的距离比"衷"和"衰"的距离大,因此输入被误识为"衰". 随着迭代次数的增加,输入与各模板之间的距离都减小,但输入与模板"哀"之间距离减小的速度比"衷"和"衰"都快. 当目标函数达到最小时,迭代结束,此时输入与"哀"之间距离已经小于"衷"和"衰"的距离,则输入汉字被正确识别. 由此可见,最优采样特征可以减少类内方差,提高特征的可分性.

在相同的实验条件下,分别用最优采样特征和均匀采样特征,即方向线素特征对国标一级字符集中的 100 套测试样本进行了测试,结果表明,每套样本最优采样特征的识别率比均匀采样特征的识别率均有不同程度的提高. 表 1 给出了这两种特征在全部测试样本集上的平均识别率. 最优采样特征比均匀采样特征的识别率上升了 1.83%,错误率下降了 18.47%.

表 1 最优采样特征与均匀采样特征识别率比较

均匀采样特征	最优采样特征
90.09%	91.92%

5 结 论

综上所述,最优采样特征以方向线素为基础,采用了有参考模板的动态变换方法,提高了特征的可分性. 通过实验证明,该特征比方向线素特征提高了识别率. 总之,笔画形变是手写汉字识别的难点,基于动态变换的方法是解决笔画形变的一种有效途径.

参 考 文 献

1 Lee S W, Park J S. Nonlinear shape normalization methods for the recognition of large-set handwritten characters[J]. Pattern Recognition, 1994,27(7):895~902.

2 Wakahara T. Adaptive normalization of handwritten characters using global/local affine transformation[J]. IEEE PAMI, 1998, 20(12):1332~1341.

3 Burr D J. A dynamic model for image registration[J]. Computer Graphics And Image Processing, 1981,15(1):102~112.

4 Kimura F, Shridhar M. Handwritten numerical recognition based on multiple algorithms[J]. Pattern Recognition, 1991, 24(10):969~983.

5 Trier O D, Jain A K, Taxt T. Feature extraction methods for character recognition—a survey[J]. Pattern Recognition, 1996, 29(4):641~662.

6 Saito T, Yamada H. An analysis of hand-printed chinese characters by directional pattern matching approach[J]. Trans. IECE, 1982,65(5):550~557.

7 Wang P P, Shiau R C. Machine recognition of printed chinese character via transformation algorithms [J]. Pattern Recognition, 1973,3(5):303~321.

张 睿 1971 年生,现为清华大学电子工程系信号与信息处理专业博士研究生. 主要研究方向为模式识别、图象处理、文字识别.

丁晓青 1939 年生,现为清华大学电子工程系教授、博士生导师,智能图文信息处理责任教授. 主要研究方向为模式识别、图象处理、文字识别、神经网络应用、多媒体信息处理以及视频智能监测等. 发表论文 140 余篇,合作专著 2 本.

方 驰 1973 年生,2001 年获清华大学博士学位,现为清华大学电子工程系讲师. 主要研究方向为模式识别、图象处理、图象识别.