基于重叠动态网格和模糊隶属度的手写汉字特征抽取

吴天雷,马少平
(清华大学计算机科学与技术系;智能技术与系统国家重点实验室,北京 100084)

摘 要: 本文在基于动态网格的手写汉字特征抽取方法中引入重叠网格划分,定义了一种反映书写结构的加权点密度,并提出了一种根据密度投影计算模糊隶属度的方法,这些措施提高了特征的分类能力. 各种网格划分方法提取方向线素特征进行了试验比较,结果表明本文的特征抽取方法的在识别率上优于传统的动态网格方法和采用非线性归一化预处理的静态网格方法.

关键词: 手写汉字识别; 特征抽取; 模糊理论
中图分类号: TP39 **文献标识码**: A **文章编号**: 0372-2112 (2004) 02-0186-05

Feature Extraction for Handwritten Chinese Character by Overlapped Dynamic Meshing and Fuzzy Membership

WU Tian-lei, MA Shao-ping
(Dept. of Computer Science & Technology, Tsinghua University, State Key Laboratory of Intelligent Technology and System, Beijing 100084, China)

Abstract: Dynamic meshing integrates nonlinear shape normalization into feature extraction without introducing any new distortions. A feature extraction method using overlapped dynamic meshing is proposed for handwritten Chinese Character Recognition. a density image is defined for an input character image. Based on the density image, overlapped meshes and fuzzy membership are dynamically calculated. Fuzzy contour direction feature (FCDF) extracted with the proposed method is introduced. Experimental results show that significant improvement in recognition rate is yielded compared to conventional static or dynamic meshing methods.

Key words: handwritten Chinese character recognition; feature extraction; fuzzy theory

1 引言

特征抽取是汉字识别的一个研究热点[1~6],手写汉字识别系统的性能很大程度上依赖于所使用的特征. 在脱机手写汉字识别中广泛采用的是网格划分的特征抽取方法. 按照网格的构造方式的不同,可以分为静态网格[2~5]和动态网格[6]. 静态网格中网格是预先确定的,而动态网格中网格的大小和位置随输入图像的不同而动态改变的.

静态网格往往需要非线性归一化[7,8]预处理补偿笔画密度不均匀等手写变形,并将汉字图像规格化为固定大小的图像. 非线性归一化的基本思想是利用在水平和垂直坐标轴上的密度投影对原二值图像进行重采样得到笔画密度更均匀的二值图像. 文献[7]回顾和比较了基于点密度和基于由穿透数、笔画间隔或内切圆定义的线密度等非线性归一化方法. 如图 1 所示,(b)~(e)分别是这四种非线性归一化方法对同一汉字处理的结果,各种非线性归一化方法都不同程度地引入了畸变:汉字边缘的锯齿加大[2]、撇和捺的方向发生变化、笔画宽度变得不均匀等. 这些畸变影响着特征抽取乃至整个识别系统的性能.

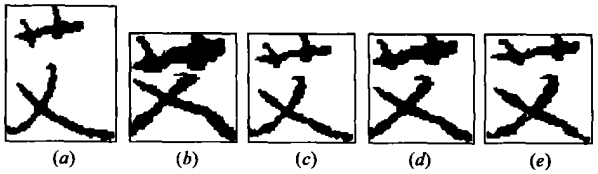

图 1 非线性归一化引入畸变的例子(a)原始汉字图像;(b)~(e)为非线性归一化后的图像

动态网格将非线性归一化中的密度均衡和网格划分的特征抽取方法结合起来,利用密度均衡得到非均匀网格以补偿笔画密度不均匀等手写变形. 动态网格在抽取特征时直接在原二值图像上进行,无需生成中间图像,不会引入任何新的畸变. 例如金连文的全局弹性网格[6]通过对汉字图像的水平和

收稿日期:2002-06-03;修回日期:2003-11-30
基金项目:国家重点基础研究(973) (No. G1998030509);自然科学基金项目(No. 60223004);国家 863 高科技项目(No. 2001AA114082)

垂直投影直方图分别进行均匀等分,形成划分汉字图像的非均匀网格.

静态网格普遍采用网格重叠[2~5]、对网格内的点进行加权[2,4]或引入模糊隶属度[5]等技术来提高特征的鲁棒性. 而传统的动态网格没有进行网格重叠,并且对网格中不同位置的点不加区别地对待,这在一定程度上影响了特征的分类性能. 依照动态网格的思想,我们根据密度投影进行重叠网格划分,并且提出了一种根据密度投影计算模糊隶属度的新方法.

2　重叠动态网格

从二值图像可以得到一个反映笔画疏密程度的密度图像,将密度图像投影到水平和垂直坐标轴上形成密度投影直方图,对投影直方图按照一定比例重叠划分就得到重叠动态网格. 图 2 的左下部分就是对一个输入图像进行 4 ×4 重叠动态网格划分的例子.

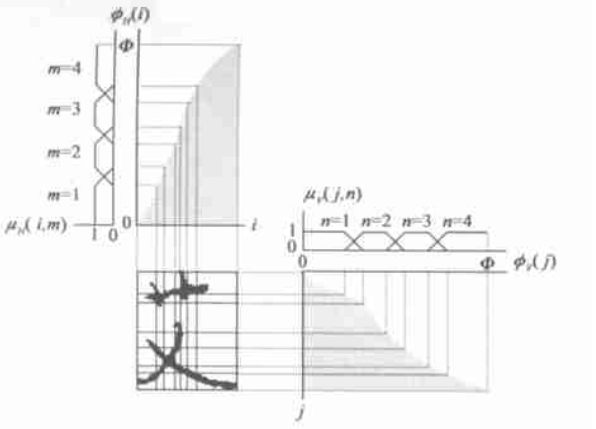

图 2　根据累积密度投影计算重叠网格和模糊隶属度的示意图

2.1　加权点密度

假设输入图像为 $f(i,j)$, $i=1,2,\cdots,I$; $j=1,2,\cdots,J$,其中 J 和 I 分别表示图像的行和列的数目. 输入图像中黑象素的值为 1,而白象素的值为 0. 我们可以得到 $f(i,j)$ 对应的密度图像为 $d(i,j)$,其中 $d(i,j)\geq 0$, $i=1,2,\cdots,I$; $j=1,2,\cdots,J$. 密度的定义是进行重叠动态网格划分的基础,可以采用非线性归一化方法中的各种密度定义[7,8],如最简单的点密度为: $d(i,j)=f(i,j)$.

我们定义了一种新的密度:加权点密度. 该密度可以看成是对点密度的一种改进:白象素点的密度不为 0,而是根据它到笔画边缘的距离进行加权,越靠近笔画边缘的白象素点的密度越大. 这样,一个点的密度的大小大致反映了该点附近的笔画疏密程度. 加权点密度的定义如下:

$$d(i,j)=\frac{1}{c(i,j)+1},\ i=1,2,\cdots,I;\ j=1,2,\cdots,J \qquad (1)$$

其中 $c(i,j)$ 为输入图像上的点 (i,j) 到最近的黑象素点的城市距离,可以通过距离变换得到. 对于黑象素点, $c(i,j)=0$.

2.2　重叠网格划分

在 $N\times N$ 重叠动态网格中,$2N$ 条水平(垂直)网格线将图像分为 N 个水平(垂直)区域. 在理想情况下,相邻区域互相重叠并且各重叠部分的密度和相等,单个区域的密度和也相等. 假设相邻区域重叠部分的密度和为 $\theta\times\Phi$,其中 $\theta(0\leq\theta\leq 1/(N+1))$ 是一个预先设定的表示重叠程度的常数,而 $\Phi=\sum_{i=1}^{I}\sum_{j=1}^{J}d(i,j)$ 为整个图像的密度和.

重叠动态网格划分是通过对累积密度投影进行均匀重叠划分来实现的. 设 $H(i)$ 和 $V(j)$ 分别表示密度图像在水平和垂直坐标轴上的密度投影函数,定义如下:

$$H(i)=\sum_{j=1}^{J}d(i,j),\ i=1,2,\cdots,I$$
$$V(j)=\sum_{i=1}^{I}d(i,j),\ j=1,2,\cdots,J \qquad (2)$$

而 $\phi_H(i)$ 和 $\phi_V(j)$ 分别表示水平和垂直坐标轴上的累积密度投影:

$$\phi_H(i)=\sum_{k=1}^{i}H(k),\ i=1,2,\cdots,I$$
$$\phi_V(j)=\sum_{k=1}^{j}V(k),\ j=1,2,\cdots,J \qquad (3)$$

从而,网格 (m,n) 可以如下定义:

$$mesh(m,n)=\{(i,j)\mid x_1(m)\leq i\leq x_2(m),\ y_1(n)\leq j\leq y_2(n)\} \qquad (4)$$

其中 $m=1,2,\cdots,N$; $n=1,2,\cdots,N$,而四条网格线如下计算:

$$\begin{aligned}
x_1(m)&=\min\left\{i\,\middle|\,\frac{\phi_H(i)}{\Phi}\geq\frac{(m-1)(1-\theta)}{N}\right\}\\
x_2(m)&=\min\left\{i\,\middle|\,\frac{\phi_H(i)}{\Phi}\geq\frac{m(1-\theta)}{N}+\theta\right\}\\
y_1(n)&=\min\left\{j\,\middle|\,\frac{\phi_V(j)}{\Phi}\geq\frac{(n-1)(1-\theta)}{N}\right\}\\
y_2(n)&=\min\left\{j\,\middle|\,\frac{\phi_V(j)}{\Phi}\geq\frac{n(1-\theta)}{N}+\theta\right\}
\end{aligned} \qquad (5)$$

3　模糊隶属度

靠近网格边缘的点比较不稳定,图像加入轻微的扰动后,这些点可能就不再划分到该网格中. 考虑到这种非确定性,我们赋予网格中的每个点一个模糊隶属度:在重叠区域中,越靠近网格边缘的点的模糊隶属度越小,这些点在计算特征时所起的作用越小,从而增强特征的稳定性.

如图 2 所示,网格 (m,n) 对应于 ϕ_H 的一个区间,我们可以在该区间上定义具有如下性质的一维模糊隶属度 $\mu_H(i,m)$:非重叠部分的点的模糊隶属度为 1;重叠部分的点的模糊隶属度随着靠近区间的边缘而线性递减,到达区间边缘时模糊隶属度为 0. 在 ϕ_V 上也可以同样地定义一维模糊隶属度函数 $\mu_V(j,n)$. 对于网格 (m,n) 中任意一点 (i,j),根据 $\phi_H(i)$ 和 $\phi_V(j)$ 就可以分别计算出水平和垂直坐标轴上的一维模糊隶属度 $\mu_H(i,m)$ 和 $\mu_V(j,n)$,综合这两个一维模糊隶属度就得到点 (i,j) 在网格 (m,n) 中的二维模糊隶属度:

$$\mu(i,j,m,n)=\min(\mu_H(i,m),\mu_V(j,n)) \qquad (6)$$

从上述重叠网格的定义可以看出,当 $\theta=0$ 时,重叠动态网格退化为无重叠(相邻的网格有一条公共的网格线)和无模

糊隶属度的传统动态网格. 这是因为,对于每个网格,网格内的点的模糊隶属度等于 1,而网格外的点的模糊隶属度等于 0,即相当于没有模糊隶属度的情况.

4 特征抽取

下面我们以方向线素特征为例说明应用重叠动态网格进行汉字特征抽取的方法.

方向线素特征是手写体汉字识别中广泛采用的特征[2~6],一般说来,方向线素特征的抽取由网格划分、方向分解和特征计算等三个步骤构成. 其中,方向分解是将汉字的轮廓图像分解为各个方向的子图像,特征计算通常是在各子图像的每个网格中求加权累积直方图得到特征.

利用重叠动态网格,从汉字的二值图像中抽取模糊方向线素特征(FCDF)的步骤如下:

(1) 首先对二值图像进行边缘跟踪得到 8 方向 Freeman 链码表示的轮廓.

(2) 从轮廓链码得到与二值图像大小相同的四个特征图像 $f_d(d=0,1,2,3)$,分别对应于横、撇、竖和捺等四种轮廓方向,方法如下:首先初始化各特征图像的所有象素的值为 0. 然后遍历所有轮廓链码,对于链码的每个元素 $c(0 \le c \le 7)$,假设它表示从轮廓点 (i_1, j_1) 到相邻轮廓点 (i_2, j_2) 的方向,则我们在对应的特征图像中给相同位置的两个点的值加 1:

$$f_d(i_1,j_1) = f_d(i_1,j_1) + 1,\ f_d(i_2,j_2) = f_d(i_2,j_2) + 1$$

$$\text{其中 } d = \begin{cases} c, & c = 0,1,2,3 \\ c-4, & c = 4,5,6,7 \end{cases}.$$

(3) 按第 2 节的方法,得到二值图像对应的密度图像,而后根据累积密度投影计算 $N \times N$ 重叠动态网格及网格中各点的模糊隶属度.

(4) 在四个特征图像的每个网格中,以各点的模糊隶属度为权值对特征图像上的点进行加权累加,而后根据该网格的大小进行归一化,就得到了特征的一个元素. 这些特征串联后就得到一个维数为 $4N^2$ 的特征向量.

(5) 对特征向量的每一维元素进行开方变换使特征的分布更接近于高斯分布[9,10].

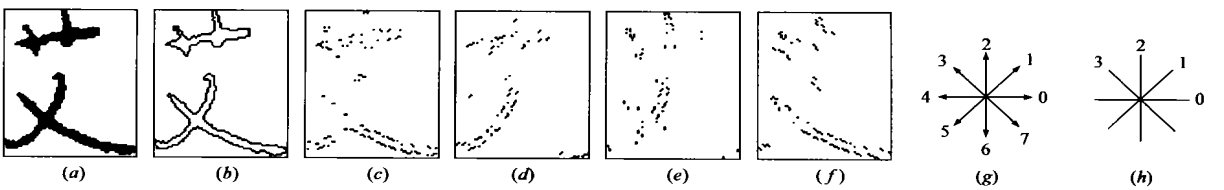

图 2 轮廓方向分解的示意图(a)输入图像;(b)轮廓;(c)(d)(e)(f)分别为横、撇、竖和捺等四个方向子图像;(g)8 方向 Freeman 链码;(h)链码分解对应的 4 个方向

上述算法中,步骤 1 和 2 是轮廓方向分解,如图 2 所示;步骤 3 进行的是重叠动态网格划分,获得的网格及模糊隶属度如式(4)~(6);步骤 4 进行的是特征计算,在特征图像 $f_d(d=0,1,2,3)$ 的某个网格 $mesh(m,n)$ 上提取的特征元素 $F_d(m,n)$ 如下所示:

$$F_d(m,n) = \frac{1}{g_d(m,n)} \sum_{(x,y) \in mesh(m,n)} f_d(x,y)\,\mu(x,y,m,n) \tag{7}$$

其中 $m = 1,2,\cdots,N$; $n = 1,2,\cdots,N$,而 $g_d(m,n)$ 是根据 $mesh(m,n)$ 的大小对该特征进行归一化的函数. 与静态网格相比,动态网格中没有非线性归一化预处理,而是在输入图像上直接抽取特征. 如果不对特征进行归一化处理,汉字图像进行缩放变换后,抽取的特征会有较大的变化. 由于方向线素特征是从单象素宽度的轮廓上抽取的,我们采用 0、45、90 和 135 度方向的直线在网格内的线段的最大象素数作为归一化函数来补偿图像大小的影响:

$$g_d(m,n) = \begin{cases} w, & d = 0 \\ \min(w,h), & d = 1,3 \\ h, & d = 2 \end{cases} \tag{8}$$

其中 $w = x_2(m) - x_1(m) + 1$ 和 $h = y_2(n) - y_1(n) + 1$ 分别为网格 $mesh(m,n)$ 的宽度和高度.

5 试验

试验用到了 1100 套脱机手写汉字样本,每套样本包含 3755 个国标一级汉字,各套样本分别由不同的人书写的. 其中 1000 套样本来自脱机手写汉字库 HCL2000[11]. 样本分为如下四个部分:

Data1:清华大学智能技术与系统实验室采集的 *smp*001~*smp*100 等 100 套样本.

Data2:HCL2000 中的 *hh*001~*hh*200 等 200 套样本.

Data3:HCL2000 中的 *hh*201~*hh*300 等 100 套样本.

Data4:HCL2000 中的 *xx*001~*xx*700 等 700 套样本.

Data1 是未经归一化的样本. 经统计,Data1 的样本的高度和宽度近似成正态分布,样本高度的均值和方差分别为 78.29 和 123.26,而样本宽度的均值和方差分别为 71.21 和 143.01. Data1 的部分样本如图 3 所示. HCL2000 的样本全部是归一化的样本,图像大小为 64 ×64,部分样本如图 4 所示. 其中 Data3 的部分样本书写较自由,识别上难于 Data2.

在试验 1 和试验 2 中用到的样本为 Data1,其中的前 50 套样本用于训练,而剩余的 50 套用于测试. 在试验 3 中,Data1、Data2 和 Data3 用于测试,而训练样本为 Data4.

5.1 不同密度的比较(试验 1)

在动态网格中可以采用不同的密度定义计算密度投影 $H(i)$ 和 $V(j)$. 除了本文的加权点密度,还有点密度[7]、笔画穿透数密度[7]、基于笔画间隔的线密度[7]、基于内切圆的线密度[8]等. 我们比较了这几种密度. 实验中采用 $N = 7$, $\theta = 0$ 的动态网格抽取方向线素特征,而后利用最小欧氏距离分类器进行识别. 从表 1 的结果可以看出,加权点密度获得的识别率稍低于基于笔画间隔的线密度,而明显高于其他的 3 种密度.

图 3　部分未经归一化的样本

图 4　HCL2000 的部分样本

在后面的实验中,我们采用加权点密度.

表 1　在动态网格($N=7$, $\theta=0$)中采用不同密度的比较

密度	笔画穿透数	基于内切圆的线密度	点密度	基于笔画间隔的线密度	加权点密度
识别率	84.01 %	90.62 %	91.07 %	91.97 %	91.72 %

5.2　网格数与重叠程度(试验 2)

试验中,我们采用不同网格数和重叠程度的动态网格抽取方向线素特征,而后利用最小欧氏距离分类器进行识别. 其中重叠参数 θ 从 0 到 0.12 按 0.02 的间隔递增(对于 $N=8$ 和 $N=9$, $\theta=0.12$ 超出了取值范围). 图 5 给出了识别率随网格数和重叠程度变化的情况. 对于每一种网格数,识别率达到最大值的重叠程度都是 $\theta=0.1$. 这表明,在相同的实验条件下,不同网格数下的最佳的重叠程度大致相同.

表 2 给出了 $\theta=0$ 和 $\theta=0.1$ 时不同网格数的重叠动态网格的比较. 从相同重叠程度下的识别率看, $N=8$ 和 $N=9$ 非常接近,稍高于 $N=7$,并且明显高于 $N=6$ 和 $N=5$ 的情况. 综合考虑识别率、存储与处理的代价,在实际应用中 $N=7$ 和 $N=8$ 是比较好的选择.

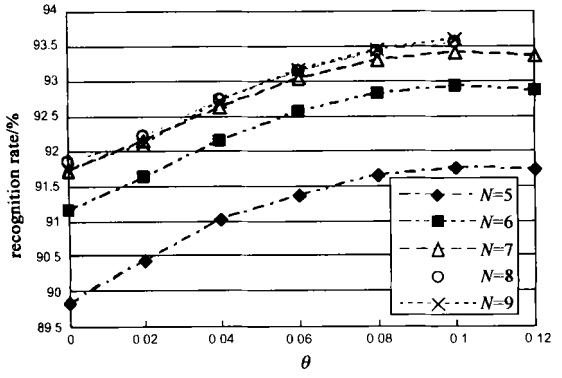

图 5　重叠动态网格的识别率(%)随网格数 N 和重叠参数变化的趋势图

表 2　在最小欧氏距离分类器下不同网格数的重叠动态网格的识别率(%)的比较

	$N=5$	$N=6$	$N=7$	$N=8$	$N=9$
$\theta=0$	89.83 %	91.16 %	91.72 %	91.84 %	91.75 %
$\theta=0.1$	91.76 %	92.94 %	93.41 %	93.57 %	93.60 %

5.3　各种方法的比较(试验 3)

为了检验各种网格划分方法在实际识别系统中的差距,我们采用了一个完整的手写汉字识别系统. 该系统包括原始特征抽取、特征变换、粗分类和细分类等四个部分. 其中,原始特征抽取是利用网格划分方法抽取方向线素特征作为原始特征. 为了减少存储和计算的代价并且提高特征的分类能力,维数较高的原始特征利用线性判决分析(LDA)[9]进行特征变换形成用于分类的 128 维特征. 粗分类采用最小欧氏距离分类器得到 30 个候选类,细分类利用 MQDF[12]从这些候选类中选择一个作为识别结果输出.

表 3　各种特征抽取方法在识别系统下的比较结果

	识别率(%)			Data1 上特征抽取的平均速度(毫秒/字)
	Data1	Data2	Data3	
静态网格	98.23	97.45	96.60	1.6
传统动态网格	97.96	97.53	96.67	0.76
重叠动态网格	98.55	98.10	97.39	0.88

我们实现了静态网格方法抽取的方向线素特征 DEF[4],该特征在日本的 ETL9B 数据库上获得了很高的识别率. 在 DEF 的抽取中,输入图像平滑后,利用基于加权点密度的非线性归一化方法归一化为 64×64 大小的点阵,归一化的图像平滑后利用预定义的 7×7 网格抽取方向线素特征. 参与比较的还有相同网格数的两种动态网格:传统动态网格($N=7$, $\theta=0$)和重叠动态网格($N=7$, $\theta=0.1$). 各种方法抽取的方向线素特征在我们的识别系统下进行了比较.

从表 3 的结果可以看出:在所有样本上,我们的方法的识别率均高于其他两种方法. 在未经归一化的 Data1 上的比较结果应该最能够反映各种方法的水平:传统动态网格得到的识别率低于静态网格方法,而重叠动态网格的识别率超过了静态网格方法;从特征抽取速度上看,重叠动态网格也比静态网格方法快. 在 Data1、Data2 和 Data3 上,重叠动态网格的识别率分别比传统动态网格方法高 0.59 %、0.57 %和 0.72 %,而错误率相对下降了 28.9 %、23.1 %和 21.6 %,这表明了我们方法的有效性.

值得一提的是,在 Data1 上,不管是样本的高度和宽度的分布,还是在是否经过归一化等方面,都与训练样本存在较大的差异. 如果直接使用未经归一化的样本来训练,识别率可能会更高.

6　结束语

文中提出了一种基于重叠动态网格划分和模糊隶属度的特征抽取方法. 与采用非线性归一化预处理的静态网格方法相比,该方法在汉字图像上直接进行特征抽取,无需生成新的图像,因此不会引入新的噪声和畸变;此外,由于动态网格不需要生成新的图像和进行额外的图像平滑,特征提取速度也

高于静态网格方法.与传统的动态网格方法相比,网格的重叠和模糊隶属度的加入,使网格对于笔划位移和局部变形不敏感,具有较强的鲁棒性.试验结果也表明该方法抽取的方向线素特征的分类能力明显提高.

只要定义合适的密度函数,本文的方法可以直接应用到基于灰度图像的汉字特征抽取中.进一步的研究工作包括将该方法应用到除方向线素特征之外的其他特征的抽取中.

参考文献:

[1] Q T Trier, A K Jain, T Taxt. Feature extraction methods for character recognition [J]. Pattern Recognition, 1996, 29: 641 - 662.

[2] F Kimura, T Wakabayashi, S Tsuruoka. Improvement of handwritten japanese character recognition using weighted direction code histogram [J]. Pattern Recognition, 1997, 30: 1329 - 1337.

[3] Jia-yong ZHANG, Xiao-qing DING, Chang-song LIU. Multi-scale feature extraction and nested-subset classifier design for high accuracy handwritten character recognition [A]. 15th International Conference on Pattern Recognition [C]. Piscataway, NJ: IEEE Press, 2000. 2. 581 - 584.

[4] N Kato, M Suzuki, et al. A handwritten character recognition using directional element feature and asymmetric mahalanobis distance [J]. IEEE Trans on PAMI, 1999, 21: 258 - 262.

[5] 马少平,夏莹,朱小燕.基于模糊方向线素特征的手写体汉字识别 [J].清华大学学报(自然科学版),1997,37: 42 - 45.

[6] 金连文,徐秉铮.手写体汉字识别中的一种新的特征提取方法 - 弹性网格方向分解特征 [J].电路与系统学报,1997,2: 7 - 12.

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

[8] H Yamada, K Yamamoto, T Saito. A nonlinear normalization method for hand printed kanji character recognition-line density equatization [J]. Pattern Recognition, 1990, 23: 1023 - 1029.

[9] K Fukunaga. Introduction to Statistical Pattern Recognition (2nd Edition) [M]. Boston: Academic Press, 1990.

[10] A J Richard, W W Dean. Applied Multivariate Statistical Analysis [M]. Englewood Cliffs, N J: Prentice Hall Press, 1982.

[11] 郭军,蔺志青,张洪刚.一个新的脱机手写汉字数据库模型及其应用 [J].电子学报,2000,28,(5): 115 - 116.

[12] F Kimura, K Takashina, et al. Modified quadratic discriminant functions and the application to Chinese character recognition [J]. IEEE Trans on PAMI, 1987, PAMI-9: 149 - 153.

作者简介:

吴天雷 男,1979 年 7 月出生于福建长汀,清华大学计算机系硕士研究生,主要研究领域:汉字识别和自然语言处理.

马少平 男,1961 年 2 月出生于河北省唐山市,清华大学计算机系教授,主要研究领域:汉字识别和信息检索.