文章编号: 1003-0077(2007)03-0117-05

基于弹性网格模糊特征的手写体汉字识别方法

刘 伟,朱宁波,何浩智,李德鑫,孙发军

(湖南大学 计算机与通信学院,湖南 长沙 410082)

摘 要: 网格方向特征在手写体汉字识别系统中得到广泛应用,被认为是目前较成熟的手写体汉字特征之一。网格技术是网格方向特征的关键技术之一。根据汉字笔画分布特点及拓扑结构的相关性,提出了一种新的基于弹性网格及其相关模糊特征的提取方法。该方法使特征向量的信息量增加,特征更加稳定。对银行支票图像大写金额的识别率达到 97.64%,实验结果证明本文方法比其他网格方向特征更有效。

关键词: 人工智能;模式识别;弹性网格;相关模糊特征;手写体汉字识别

中图分类号: TP391.43 **文献标识码**: A

A Feature Extracting Method for Handwritten Chinese Character Recognition Based on Elastic Mesh and Fuzzy Feature of Block

LIU Wei, ZHU Ning-bo, HE Hao-zhi, LI De-xin, SUN Fa-jun

(Computer and Communication College, Hunan University, Changsha, Hunan 410082, China)

Abstract: The directional feature is considered suitable for handwriten chinese character recognition, and it has been widely used as one of the main feature extraction method. Meshing method is one of the key factors of meshing direction feature. According to stroke distributing characterisitic and topologic correlation of chinese characters, we present a new method based on elastic mesh and related fuzzy feature. A more stable feature vector with more information is extracted. The experiment based on the handwritten legal amount on Chinese bank check shows that the method is more effective than other meshing direction features, and recognition rate has been up to 97.64%.

Key words: artificial intelligence; pattern recognition; elastic meshing; correlation fuzzy feature; handwritten Chinese characters recognition

1 引言

文字识别是模式识别的一个重要研究方面,在信息处理,办公自动化,邮政系统等多方面有着重要的使用价值,并且囊括了模式识别领域中所有典型的问题,如特征的选择、分类器的选择以及样本集的选择,因此对于脱机手写体识别的研究具有深刻的理论意义和实用价值。而特征抽取是一个手写体汉字识别系统最为关键的环节之一。良好的特征必须能反映汉字的本质特征,能容忍手写体各种书写风格的变形和随意性。特征提取一直是手写体汉字识别中的一个研究重点[1~3],已经提出许多特征提取方法。近年来大量的研究实验发现,方向特征是一种较好的手写体汉字特征,有许多方向特征已成功应用于手写体汉字识别系统中[4~6]。一般而言,方向特征提取方法可用图 1 所示的流程图来表示。

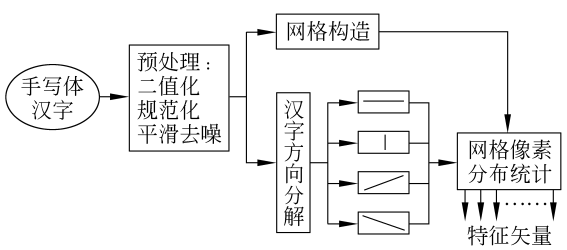

图 1 网格方向分解特征的提取框图

不难看出,提取网格方向特征的两个关键技术

收稿日期: 2006-06-07 **定稿日期**: 2007-01-26

作者简介: 刘伟(1982—),男,硕士生,研究方向为图像处理、汉字识别。

是网格的构造及方向分解方法。

常用的网格构造方法有:模糊网格构造[4,5]和弹性网格构造[6]。

如果网格的划分是确定的。即对于某一网格来说,点(i,j)或者属于该网格,或者不属于该网格,是一个确定性的问题。然而由于书写引起的位移等因素的影响,同一笔画可能会落入不同的网格,其结果使得笔画的微小位移可能会导致特征的突变。而在模糊网格中,由于引入了模糊隶属度的思想,一定程度上克服了笔画位置变化对特征抽取的影响。但是模糊网格没有考虑到笔画之间的关系,没有体现汉字拓扑结构上的相关性。

不同人书写的汉字其差异尽管很大,但是每个字的拓扑结构是相对稳定的。根据笔画分布所构造的弹性网格,对于克服因笔画变形而带来的识别困难,是一个有效的方法。

但是现有的弹性网格特征大多孤立地统计每个子块的特征,没有考虑汉字结构上的相关性[6],这样将导致特征不是很稳定。文献[7]在统计网格特征的时候,考虑了周围子块的影响,识别率得到一定程度的提高,但是它的不足之处是在统计子块横方向特征、竖方向特征、撇方向特征、捺方向特征都加上该子块上下左右子块相应的影响,没有完全体现汉字笔画拓扑结构上的相关性。本文提出的基于弹性网格的子块及其相关模糊特征提取方法,既考虑了手写体汉字笔画变形大的特点,又很好地体现了汉字拓扑结构上的相关性。

2 模糊网格特征

设函数 $F(i,j)$ 对应于汉字的 $N\times N$ 二值化图像点阵:

$$F(i,j)=\begin{cases}0 & \text{白像素}\\ 1 & \text{黑像素}\end{cases} \tag{1}$$

定义1 将点阵 $F(i,j)$ 均匀地划分为 $m\times m$ 个网格,第 k 个网格记为 $N_k(k=0,1,2\cdots,m\times m-1)$。点 $(i,j)\in N_k$ 当且仅当 $k=fix\left(\frac{i\times m}{N}\right)\times m+fix\left(\frac{j\times m}{N}\right)$,其中函数 $fix(x)$ 表示对 x 取整。

定义2 设网格边长为 $2n$, $a\in[0,n]$,论域 U 为全体点阵像素点,以网格 N_k 左上角为原点,定义第 k 个模糊网格 $FN_k(k=0,1,\cdots,m\times m-1)$ 的隶属函数为:

$$\mu_{FN_k}(i,j)=\begin{cases}1 & a\le i\le 2n-a,a\le j\le 2n-a\\ \dfrac{2n+a-j}{2a} & 2n-a\le i\le 2n+a,2n-i\le j\le i\\ \dfrac{a+j}{2a} & -a\le i\le a,i\le j\le 2n-i\\ \dfrac{2n+a-j}{2a} & 2n-a\le j\le 2n+a,2n-j\le i\le j\\ \dfrac{a+j}{2a} & -a\le j\le a,j\le i\le 2n-j\\ 0 & \text{其他}\end{cases} \tag{2}$$

其形状如图2所示。在三维空间,考虑相邻两个网格对应的模糊网格,当 $a\neq 0$ 时,具有不为零隶属度的图像点集相交。因此,处于网格边界附近的笔画上的点是否属于某一网格不再是一个确定关系,而具有一定的模糊性。并且由于隶属函数是连续的,排除了特征突变的可能性。当 $a=0$ 时,模糊网格退化为一般意义的网格。

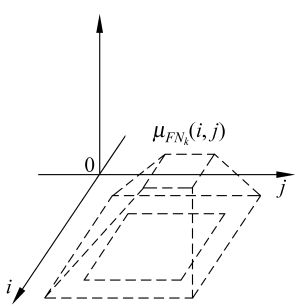

图2 隶属函数 $\mu_{FN_k}(i,j)$

3 已有的弹性网格特征提取方法

构造弹性网格就是用一组假想的网线对汉字图像的区域划分。如图3是采用8×8均匀网格将一个汉字切分为64个区域。所谓的均匀网格是指每一个网格的面积是均等的。对一个汉字二值图像在水平、垂直两个方向上的直方图均匀等分实际上就是对汉字图像非均匀等分,这种划分方法所形成的网格就是非均匀的,称为弹性网格,如图4所示。即若纵横方向的网线分别为 N_1, N_2,则当满足式(3)与式(4)时,所形成的网格就是弹性网格。

$$\int_{j}^{N_1}\int_{i}^{i+1}F(x,y)\,dx\,dy=\int_{j}^{N_1}\int_{i}^{i+1}F(x,y)\,dx\,dy$$
$$\forall i,k=1,2,\cdots,N_1-1 \tag{3}$$

$$\int_{i}^{i+1}\int_{j}^{N_2}F(x,y)\,dx\,dy=\int_{k}^{k+1}\int_{j}^{N_2}F(x,y)\,dx\,dy$$
$$\forall i,k=1,2,\cdots,N_2-1 \tag{4}$$

图 3　8 ×8 均匀网格

(a) 8×8弹性网格

(b) 4×4弹性网格

图 4　弹性网格

其中, $F(x,y)$ 表示汉字二值图像,

$$F(x,y)=\begin{cases}1 & \text{表示黑像素}\\ 0 & \text{表示白像素}\end{cases}$$

在汉字为未解之前,首先对汉字图像按上述的方法进行网格构造。假设得到的网格为 B_1, B_2, …B_n,然后将汉字进行四方向分解后,我们用 $F_H(x,y)$、$F_S(x,y)$、$F_P(x,y)$、$F_W(x,y)$ 分别表示分解后"横"、"竖"、"撇"、"捺"四个方向的汉字子图像,则第 i 个子块内的弹性网格特征由下列各式给出:

$$D_H^i=\frac{\iint_{B^i}F_H(x,y)\,dx\,dy}{\iint F(x,y)\,dx\,dy} \tag{5}$$

$$D_S^i=\frac{\iint_{B^i}F_S(x,y)\,dx\,dy}{\iint F(x,y)\,dx\,dy} \tag{6}$$

$$D_P^i=\frac{\iint_{B^i}F_P(x,y)\,dx\,dy}{\iint F(x,y)\,dx\,dy} \tag{7}$$

$$D_W^i=\frac{\iint_{B^i}F_W(x,y)\,dx\,dy}{\iint F(x,y)\,dx\,dy} \tag{8}$$

采用弹性网格对汉字图像进行划分,目的是使不同书写风格的同一类汉字,在相对应网格内的笔画分布概率接近,以使网格内的特征向量保持相对稳定,这可在一定程度上容忍因书写的风格不同所产生的笔画变形。但是,这样没有考虑汉字笔画拓扑结构上的相关性。另外不同人书写的同一类汉字的同一笔画,其空间位置的差异也是很大的,在相对应网格内的笔画分布概率不可能相同,也会与周围网格相关联。

针对这点不足,孙立民在文献[7]中提出:在统计第 i 个子块特征时(包括横、竖、撇、捺),都加上第 i 块的上下左右子块相应的影响,这在一定程度上使得特征的稳定性得到提高,但是它的不足之处是在统计子块横方向特征、竖方向特征、撇方向特征、捺四方向特征都加上该子块上下左右子块相应的影响,没有完全体现汉字笔画拓扑结构上的相关性。

4　本文基于弹性网格的特征提取方法

设子块 B^i 的 8 邻域子块分别为 B_1^i、B_2^i、B_3^i、B_4^i、B_5^i、B_6^i、B_7^i、B_8^i,如图 5 所示。本文的方法是对横方向特征、竖方向特征、撇方向特征、捺方向特征区别对待。在统计第 i 个子块的横方向特征矢量时把它的左边子块 B_1^i 和右边子块 B_3^i 的子块特征作一合理考虑,在统计第 i 个子块的竖方向特征矢量时把它的上边子块 B_2^i 和下边子块 B_4^i 的子块特征作一合理考虑,在统计第 i 个子块的撇方向特征矢量时把它的右上方子块 B_5^i 和左下方子块 B_7^i 的子块特征作一合理考虑,在统计第 i 个子块的捺方向特征矢量时把它的左上方子块 B_6^i 和右下方子块 B_8^i 的子块特征作一合理考虑,无疑会使子块特征的信息量增加,同时降低因笔画变异大造成的子块特征向量的不稳定性。

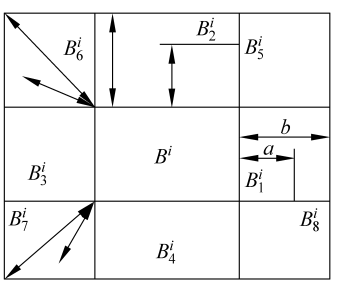

图 5　笔画关联示意图

如何合理地统计周围子块特征向量的贡献,是需要慎重考虑的。因书写风格迥异,同一笔画落在 B^i 内的长短会有变化,而这种变化是和周围子块相关的。显然,8 邻域内的黑像素点离子块 B^i 的距离越近,则相关的程度越大,反之,则小。而且,笔画拓扑结构上的相关性也有相同规律。因此,我们采用拟正态分布的模糊隶属度函数来刻画这种相关性[7],如图 6 所示。

这里假设 B^i 8 邻域特征向量隶属于 B^i,但其隶属度随黑像素与 B^i 的距离变化,这种变化用正态分布函数的右半部分表示,离 B^i 越近,其隶属度越大;反之,则越小。

由于采用的是弹性网格,所以 B^i 8 邻域的面积是不等的,如果采用相同的的模糊隶属度函数曲线,

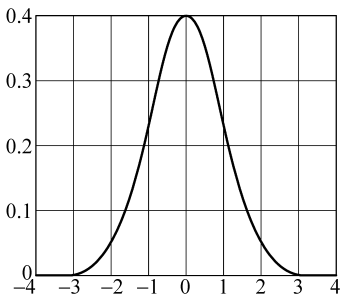

图 6 拟正态分布图

则不能正确地反映这种相关性。我们知道,正态分布函数 $\phi(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$ 当 $\sigma=1$, $x=3$ 时, $\phi(x)$ 值已经下降为峰值的 1 %。所以,我们把隶属度函数设定为:

$$\mu(a) = \exp\left(-\frac{(3a/b)^2}{2}\right) \quad a \geq 0 \tag{9}$$

综合考虑 8 邻域的相关性后,则第 i 个子块内的弹性网格特征由下列各式给出:

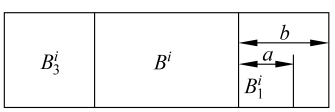

图 7 横笔画关联示意图

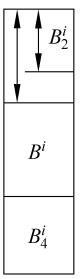

图 8 竖笔画关联示意图

对于第 i 个子块内的横方向特征,

$$D_H^i = \frac{\iint_{B_i} F_H(x,y)\,dx\,dy}{\iint F(x,y)\,dx\,dy} + \frac{\iint_{B_1^i} F_H(x,y)\,\mu(a)\,dx\,dy}{\iint F(x,y)\,dx\,dy} + \frac{\iint_{B_3^i} F_H(x,y)\,\mu(a)\,dx\,dy}{\iint F(x,y)\,dx\,dy} \tag{10}$$

其中 a 表示 B_1^i 或 B_3^i 内黑像素到 B^i 的距离, b 表示 B_1^i 或 B_3^i 的远端边缘到 B^i 的距离。(见图 7)。

对于第 i 个子块内的竖方向特征,

$$D_S^i = \frac{\iint_{B_i} F_S(x,y)\,dx\,dy}{\iint F(x,y)\,dx\,dy} + \frac{\iint_{B_2^i} F_S(x,y)\,\mu(a)\,dx\,dy}{\iint F(x,y)\,dx\,dy} + \frac{\iint_{B_4^i} F_S(x,y)\,\mu(a)\,dx\,dy}{\iint F(x,y)\,dx\,dy} \tag{11}$$

其中 a 表示 B_2^i 或 B_4^i 内黑像素到 B^i 的距离, b 表示 B_2^i 或 B_4^i 的远端边缘到 B^i 的距离。(见图 8)。

对于第 i 个子块内的撇方向特征,

$$D_P^i = \frac{\iint_{B_i} F_P(x,y)\,dx\,dy}{\iint F(x,y)\,dx\,dy} + \frac{\iint_{B_5^i} F_P(x,y)\,\mu(a)\,dx\,dy}{\iint F(x,y)\,dx\,dy} + \frac{\iint_{B_7^i} F_P(x,y)\,\mu(a)\,dx\,dy}{\iint F(x,y)\,dx\,dy} \tag{12}$$

其中 a 表示 B_5^i 或 B_7^i 内黑像素到 B^i 的距离, b 表示 B_5^i 或 B_7^i 的对角线距离(见图 9)。

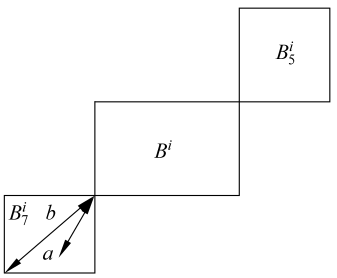

图 9 撇笔画关联示意图

对于第 i 个子块内的捺方向特征,

$$D_W^i = \frac{\iint_{B_i} F_W(x,y)\,dx\,dy}{\iint F(x,y)\,dx\,dy} + \frac{\iint_{B_6^i} F_W(x,y)\,\mu(a)\,dx\,dy}{\iint F(x,y)\,dx\,dy} + \frac{\iint_{B_8^i} F_W(x,y)\,\mu(a)\,dx\,dy}{\iint F(x,y)\,dx\,dy} \tag{13}$$

其中 a 表示 B_6^i 或 B_8^i 内黑像素到 B^i 的距离, b 表示 B_6^i 或 B_8^i 的对角线距离(见图 10)。

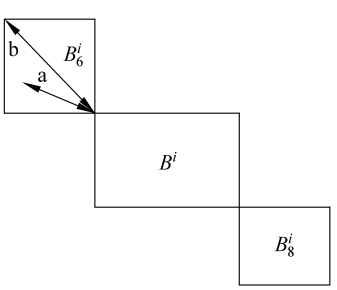

图 10 捺笔画关联示意图

5　实验

实验采用南京理工大学用于手写体汉字金额自动识别研究的 NUST603HW 手写汉字样本,我们选取银行支票上大写金额常用的"零,壹 ··拾,佰,仟,万"14 类汉字样本,每类取 2 000 个(共 28 000 个)样本分成两部分用于实验。1 000 个用于训练样本,1 000 个用于测试样本。模式分类器采用最小距离 Bayes 分类准则[9],其基本原理是:对训练样本进行特征提取作为该字的模板,识别时,比较待识别样本的特征矢量与各模板之间的距离,距离最小的模板作为识别结果。距离测度有欧氏距离、城市距离、相关距离、加权距离等,考虑到速度和容量等因素,在实验中采用了最简单的欧氏距离作为距离测度。实验结果如表 1 所示。

方法 1 采用文献[7]的算法,网格构造用的是弹性网格(8 ×8 个),方法 2 为模糊方向特征方法,网格构造用的是均匀网格(8 ×8 个),本文方法的网格构造采用的是弹性网格(8 ×8 个)。这三种方法的子笔画分解算法统一采用文献[8]提出的方法。

从表 1 可以看出,在 14 类字中,本文的方法每类字的识别率都比方法 1 要高,总的识别率比方法 1 识别率提高了 2.87 %。本文的方法虽然有个别字识别率低于方法 2。但是总的识别率比方法 2 识别率提高了 1.07 %。可见本文方法的有效性。

表 1　"零"到"玖"及"拾""佰""仟""万"共 14 类汉字的识别结果

字符 / 识别率 %	方法 1	方法 2	本文的方法
零	89.80	91.10	95.30
壹	93.90	96.80	97.90
贰	96.90	99.10	99.40
叁	95.50	96.70	98.10
肆	95.40	98.70	98.60
伍	94.30	96.60	95.90
陆	93.00	94.30	96.00
柒	93.90	95.60	96.20
捌	97.80	98.40	99.60
玖	96.10	99.50	98.20

续表

字符 / 识别率 %	方法 1	方法 2	本文的方法
拾	94.80	95.20	96.80
佰	95.30	91.70	97.80
仟	95.50	98.70	98.10
万	94.60	99.80	99.00
合计	94.77	96.57	97.64

6　结论

本文在弹性网格的基础上,提出了一种子块相关模糊特征提取方法。这种方法既考虑了汉字笔画的分布特点,又很好地考虑了汉字拓扑结构上的相关性,是对人认知汉字机理的一种模仿,这对识别书写风格差异大、随意性强、结构变形大的手写体汉字,是一种很好的方法,实验结果令人满意。

参考文献:

[1] Qivind Due Trier, et al. Feature Extraction Methods for Character Recognition: A Survey[J]. Pattern Recogniton, 1996, 29(4): 641-662.

[2] R Plamondon, S N Srihari. On-line and Off-line Handwritting Recogition: A Comprehensive Survey [J]. IEEE Trans on PAMI, 2000, 22(1): 63-81.

[3] 陈津颖,金奕江,马少平. 手写体汉字在特征空间的可视化分析[J]. 中文信息报, 2000, 14(5): 42-48.

[4] 王正群,叶晖,孙兴华,杨静宇. 基于模糊方向特征的手写体汉字识别[J]. 模式识别与人工智能, 2001, 9(3): 318-320.

[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(3): 42-45.

[6] 吴天雷,马少平. 基于重叠动态网格和模糊隶属度的手写体汉字特征抽取[J]. 电子学报, 2004, 2: 187-190.

[7] 孙立民,狄红卫,余英林. 基于子块特征及其相关模糊隶属度特征的手写体汉字识别方法[J]. 通信学报, 1999, 20(12): 82-85.

[8] 王正群. 手写体汉字识别研究[D]. 南京:南京理工大学计算机系, 2001.

[9] 边肇祺. 模式识别[M]. 北京:清华大学出版社, 1998.