一种简单有效的手写体数字特征提取算法①

高宏宾，赵智彬

(五邑大学 计算机学院，江门 529020)

摘 要：综合统计特征提取法与结构特征提取法的优点，设计了一种新的基于圈、左右轮廓特征与行分段特征来获取数字字符特征向量的方法。该办法用较少的特征向量就能够保留数字字符拓扑结构中的关键信息，适应性很强。仿真实验中，首先根据字符容易获取的结构特征（圈）对字符进行大体分类，然后利用基于级联结构的 AD AdaBoost 神经网络根据余下的特征值进行逐层淘汰识别。结果表明，该办法在识别速度与识别正确率方面都有所改进。

关键词：数字识别；特征提取；左右轮廓特征；行分段特征；AD AdaBoost；级联结构

A Simple Effective Extraction Algorithm for the Features of Handwritten Digits

GAO Hong-Bin, ZHAO Zhi-Bin

(College of Computer, Wuyi University, Jiangmen 529020, China)

Abstract: A new method to capture the feature vectors of numeric alphabetic based on circles, lr- sides contour and row subsection is proposed by intergating the opinions of Statistic Feature Extraction and Structural Feature Extraction. It uses less feature vectors to keep the substantial information in the topologic structure of a numeric alphabetic, which has a very strong adaptability. In the imitating experiment, first the feature vectors extracted are to be assorted roughly based on the Structural Features (circle) of the numeric alphabetic, and then the other feature vectors is to be recognized and eliminated gradually making use of the AD AdaBoost neural network. The result shows that, this method not only has a faster speed of recognition, but also has a higher ratio of correct recognition.

Key words: digits recognition; feature extraction; lr- sides contour; row subsection; AD AdaBoost; cascade structure

1 引言

阿拉伯数字按联机书写与脱机书写的区别，可分为联机手写体数字识别与脱机手写体数字识别两种方式。联机手写体数字识别是一种通过联机光电板把数字实时输入计算机的方法,它被处理的是一维点(坐标)串，这些点(坐标)串中所包含的笔划走向、起终点和笔划顺序等信息都很容易被正确提取，作为识别的关键信息,故目前联机手写体字符的识别技术已经很成熟，识别率高达 99%以上；而脱机手写体数字识别处理的仅仅是一些经过光电仪器扫描所得到的二维点阵图像，不包含任何实时信息。因此，脱机手写体数字识别要比联机手写体数字识别难得多，目前存在的各种算法，就其效率及可靠性而言，依然无法达到理想的程度，仍需进一步探讨研究。

2 相关研究

一般而言，脱机手写体数字识别包括两个基本步骤：一、是提取字符的特征；二、是根据特征进行分类。其中第一步是基础。

目前特征提取的方法多种多样，按使用特征的不同，这些方法可以分为三类：①直接对图像点阵根据某种算法进行降维的方法[1]；②基于统计特征（通常包括点密度的测量、矩、特征区域等等）的方法[2]；③基于结构特征（包括圈、端点、交叉点、笔画、轮廓等等）的方法[3-5]。一般来说，每类特征都有其优劣。相对而言，直接对图像点阵进行降维的方法是获取特

① 收稿时间:2010-12-23;收到修改稿时间:2011-02-25

征的最简单办法，但这种办法提取出来的特征向量维数特高，对神经网络依赖性过大，如果降维过多，失真又太大，往往不能采用；统计特征提取法相对简单快速，时间效率高，但要提取出充分反映字符拓扑结构的低维特征集比较困难，因此，识别正确率一般不高。文献[2]给出了一种较好的统计特征提取办法——13 个点特征提取法,该方法从每个字符中提取出关键的 13 个统计数据作为特征集,其主要特点是特征提取简单有效,节省了特征提取时间,提高了识别系统的运行速度。尽管该办法在正确率方面也有所改进，但依然不高，只有 91.5%；结构特征提取法是模拟人眼识别数字字符的功能，直接分析字符的拓扑结构特征，容易适应手写体数字书写随意性的特点，一般来说，所提取的特征向量维数较低，识别可靠性相对较高。但是由于计算机是一维线性处理器，面对扫描后的二维字符图像，字符的某些结构特征（如叉点、端点、凹凸区等）是很难被正确识别的。尽管很多结构特征提取算法都是先采用某种细化算法提取骨架，再做分析，在一定程度上有利于计算机对结构特征的正确判断。但是，细化算法极易受到字符图像毛刺和伪分支的影响，而这些对于扫描所得到的字符图像来说，却是无法避免的。

鉴于以上分析，本文综合统计特征提取法与结构特征提取法的优点，设计了一种新的基于圈（circle）、左右轮廓特征(lr- sides contour)与行分段特征(row subsection)来获取数字字符特征向量的方法。该办法不需要在预处理阶段对数字图像进行复杂的细化等运算，从而在提高系统运行速度的同时，也避免了因细化形变而造成的误识。仿真实验中，首先根据字符容易获取的结构特征（圈）对字符进行大体分类，然后利用基于级联结构的 AD AdaBoost 神经网络根据统计特征（左右轮廓特征和行分段特征）进行逐层淘汰识别。多次仿真结果表明，该办法在识别速度与识别正确率方面都有所改进。

3 特征提取

3.1 预处理

市场上现有的数字字符识别系统在预处理阶段通常包括：图像的灰度化、二值化、平滑去噪、断笔补偿、近距离连通域连接、字符分割、规范化(归一化)、倾斜矫正、细化等步骤[6]。本文采取基于圈、左右轮廓特征与行分段特征来获取字符特征向量的办法，不需要进行复杂的细化和倾斜矫正等操作，从而简化了预处理的过程，加快了算法的速度。

为了方便起见，在下文中均假定已对字符图像进行必要的预处理，尺寸规范化后其外框大小为 20*20（以像素为单位），顶行与底行均有黑色像素。

3.2 确定字符的圈（circle）

数字字符的圈是人眼识别字符的重要信息，本文提出一种不需要经过细化就可以确定字符圈的办法。算法的基本思想是：按照自顶向下逐行寻找字符图像中位于黑色像素之间的白色像素段，然后判断该白色像素段上方是否封闭。如果封闭，则逐行下降沿着白色像素寻找到达边界的路径，如果不存在这样的路径则认为该字符有圈。

为了便于描述,在下面的算法流程中，引进了函数 R(,j, k)，其定义为：当函数的扫描位置超过字符图像的边界时，函数返回值为“-1”；否则，如果在字符图像的第 i 行像素中，从第 j 列到第 k 列均是黑色像素，函数返回值为“0”；否则，函数的返回值是第 i 行像素中，在第 j 列与第 k 列之间任一个白色像素的列号。具体算法描述如下：

(1) 初始化: $r \leftarrow 1$。

(2) 如果 $r >$ H-2 (H 为字符图像的高度)，转(11)；否则，转(3)。

(3) 扫描字符图像的第 r 行，用数 $C[N][2]$ 组按顺序记录该行中所有在相邻两段黑色像素之间的白色像素段的位置，例如：用 $C[j][0]$ 记录该行中第段白色像素的起始列号，$C[j][1]$ 记录该行中第段白色像素的终止列号；另外，用变量 T 记录该行中一共被记录的白色像素段的数目。转(4)。

(4) 如果 T =0，令 $r \leftarrow r+1$，转(2)；否则转(5)。

(5) 令 $k \leftarrow r$， $j \leftarrow 1$，转(6)。

(6) 如果 $j <= T$，令 $r \leftarrow k$，转(7)；否则，令 $r \leftarrow k+1$，转(2)。

(7) 判断第 j 段白色像素的上方是否封闭，即调用函数 R($r-1, C[j][0], C[j][1]$)，如果返回值为“0”，转(8)；否则，令 $j \leftarrow j+1$，转(6)。

(8) 调用函数 R($r+1, C[j][0], C[j][1]$)，$r \leftarrow r+1$ 如果 r 返回值为“-1”， 令，转(6);否则，如果返回值为“0”，转(10)；否则，转(9)。

(9) 令 R($r+1, C[j][0], C[j][1]$)，$r \leftarrow r+1$。在第 r 行

中，以像素 $O(r,Mcol)$ 为始点，分别向左、向右扫描，如果包含像素 $O(r,Mcol)$ 的这段白色像素有一端到达边界，令 $j \leftarrow j+1$，转(6)；否则，用 $C[j][0]$ 记录该段白色像素的起始列号，$C[j][1]$ 记录该段白色像素的终止列号，转(8)。

(10) 该字符有圈，算法结束。

(11) 该字符无圈，算法结束。

容易证明，上述算法的时间复杂度为 $O(n)$，其中是每幅字符图像的像素数目。在实际应用中，当圈内白色像素总数小于一定阈值时，该圈为无效圈。

3.3 确定左右轮廓(lr- sides contour)特征

定义 1. 对数字二值化点阵图像中某一行，以数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，所经过的像素数, 称为该行的左轮廓特征值。n 个左轮廓特征值从上到下按顺序排列所构成的特征向量, 称为左轮廓特征。与左轮廓特征相类似，可定义数字的右轮廓特征。

图 1（b）是将从图 1（a）中提取出来的左右轮廓特征按照相应的顺序重新拟合字符图像所得的结果。可以看出，尽管拟合的结果相对原字符有所变形，但它保留了原字符图像拓扑结构中的关键信息，基本上不影响对字符的正确识别分类。

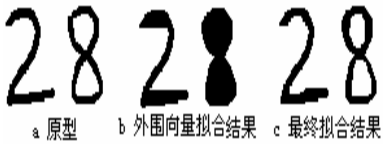

图 1

3.4 确定行分段(row subsection)特征

为了让识别网络进一步适应字符形状及笔画粗细的差异，提高识别的精度，本文引进了行分段特征。

定义 2. 对数字二值化点阵图像中的某一行，以数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) ，称为该行的行分段特征值。n 个行分段特征值从上到下按顺序排列所构成的特征向量，称为行分段特征。

对于一幅宽不变，高为 20 个像素的字符图像，一个字符共可提取 20 个行分段特征值。对图 1（b）利用从图 1（a）中提取出来的行分段特征按照一定的原则进一步拟合，得出结果如图 1（c）所示。

4 系统设计与仿真分析

4.1 实验系统设计

自 1997 年,Freund 和 Schapire 提出了 AdaBoost 算法[7] 以来，,该算法得到了广泛的应用，其中最著名的是 Viola 和 Jones 提出的基于级联结构的 AdaBoost 算法，其结构如图 2 所示，它在人脸检测的研究中取得了很好的分类效果。

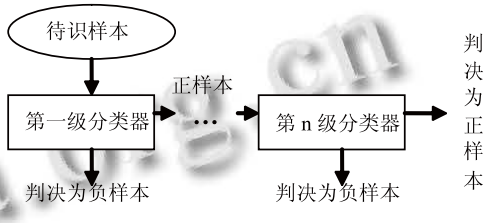

图 2　级联结构分类器模型

在级联结构中 ,强分类器的复杂程度随着级数的增加而增加。大多数的负样本都在前面比较简单的级次中被淘汰,而使几乎全部的正样本顺利通过.尽管级数越高，识别样本所需要的特征向量越多，但是越往后需要识别的候选样本越少 ,因此使系统具有很好的实时性。

李闯等人在级联结构的基础上进一步对 AdaBoost 算法作了改进，提出 AD AdBoost 算法[8]。该算法主要的改进是：在级联结构的各个层次中，在保证正样本高通过率的同时，有效地降低了负样本的错误率。

本文利用 AD AdaBoost 算法来根据手写体数字提取出来的特征进行分类识别。AD AdaBoost 算法是二值分类算法，而手写体数字识别问题是十值分类问题，因此必须对 AD AdaBoost 分类器进行组合变形，使其满足多分类要求。在本系统中，AD AdaBoost 分类器采用了两级级联结构，第一级相对简单，仅以 40 个左右轮廓特征值，归一化后作为输入变量；第二级比较复杂，输入变量有 60 个，除了左右轮廓特征值之外，还包括 20 个行分段特征值。分类系统框架如图 3 所示。其中 Ci 是第一级分类器;Di 是第二级分类器。系统中任一分类器都是只有二分类功能，则只输出“是这类字符”和“不是这类字符”。

在图 3 所示的系统中，首先根据字符的圈将字符分为两大类，目的在于减少每大类中类别的个数，从而提高系统的速度和精度。如果字符被判断有圈，则让 C1 识别，C1 的输出结果有两种可能：“是 0”或“非

0”。如果 C1 认为“是 0”，则让 D1 进一步识别，这时如果 D1 也认为“是 0”，则输出最终结果：‘0’，识别过程结束；如果其中 C1 或 D1 认为“非 0”，则让 C2 及 D2 用相同的办法判断是否为‘2’。依此类推，一直到字符被识别或被拒识。对于被判断无圈的字符，识别过程类似。

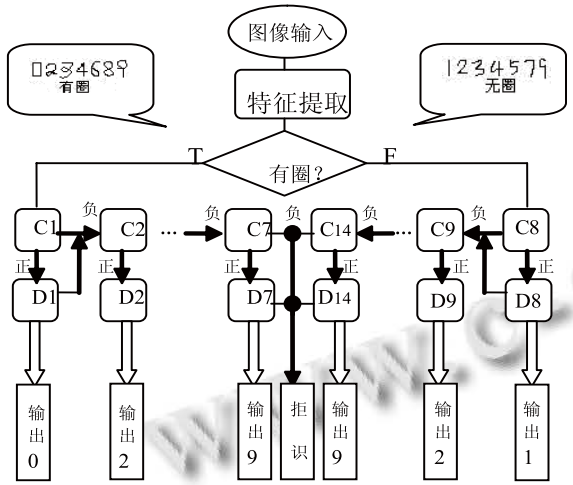

图 3 分类系统框架图

4.2 仿真结果分析

实验数字从实际财务报表上提取，共提取 3000 个手写体数字图像，0~9 每类 300 个。预处理后，每幅图像用 20*20 的向量表示。从中随机抽取 2000 个作为神经网络的训练样本，待学习训练结束后，再用余下的 1000 数字作为测试样本。测试的最终结果如表 1 所示。

表 1 测试结果

字符类别	正确识别	拒识	误识	可靠性(%)
0	97	1	0	99%
1	106	0	0	100%
2	84	1	1	97.6%
3	91	1	0	98.9%
4	110	2	0	98.2%
5	98	1	2	97%
6	89	0	0	100%
7	119	0	1	99%
8	87	0	2	98.1%
9	105	0	2	98.1%
总计	986	6	8	99.2%
比率（%）	98.6%	0.6%	0.8%	

将实验数据分别应用于文献[1~5]所提出的办法与本文办法作比较分析，结果如表 2 所示。

表 2 比较分析结果

办法	拒识率	误识率	正确率	样本数目
[1]	0	10.2%	89.8%	1000
[2]	0	8.3%	91.7%	1000
[3]	0	3.3%	96.7%	1000
[4]	0	7.2%	92.8%	1000
[5]	1.8%	3.5%	94.7%	1000
本文办法	0.6%	0.8%	98.6%	1000

5 结语

文中所设计的手写体数字识别系统取得了比较理想的结果。另外，适当地增大数字图像的大小（以像素为标准）以及 AD AdaBoost 分类器的级联级数，相信能使系统获得更高的识别率。根据文中获取特征向量的方法，可知，特征维度是随着数字图像高度的变化呈线性变化的，这为此改进提供了可能性。

本文创新点：获取特征向量的新办法，不管对于数字签名系统，还是汉字识别系统都有一定的借鉴价值。

参考文献

1 张国华,万钧力.基于主分量分析法的脱机手写数字识别.计算机工程,2007,33(18):219–221.
2 宋曰聪,胡伟.手写体数字识别系统中一种新的特征提取方案计算机科学,2007,34(9):236–239.
3 刘雄华,冯刚.链码在高精度手写体数字识别系统中的应用.计算机应用,2004,24(7):168–169.
4 张红云,苗夺谦,张东星.基于主曲线的脱机手写数字结构特征分析及选取.研究与发展,2005,42(8):1344–1349.
5 叶飞,黎峰.基于整体特征的快速手写体数字字符识别.计算机工程与设计,2006,27(22):4347–4352.
6 张猛,余仲秋,姚绍文.手写体数字识别中图像预处理的研究.算机信息,2006,22(61):256–258.
7 Freund Y, Schapire RE. A Decision-theoretic Generalization of Online Learning and an Application to Boosting. Journal of Computer and System Sciences,1997,55(1):119.
8 李闯,丁晓青,吴佑寿.一种改进的 AdaBoost 算法--AD AdaBoost 算法.计算机学报,2007,30(1):103–109.