文章编号:1004-5694(2005)04-0451-05

一种有效的多字体印刷字符识别系统实现*

赵全友,潘保昌,姚锦秀,郑胜林,陈箫枫
(广东工业大学 信息工程学院,广东 广州 510640)

摘　要:提出了一种基于神经网络和浮动模板的多字体印刷字符识别方法。在研究大量的多字体印刷字符图像后,给出了一种有效的预处理方法,并在综合抽取宏观特征与微观特征后,送入神经网络的浮动模板法分类器进行识别。实验证明该方法具有相当高的识别率,应用前景十分广泛。

关键词:多字体识别;预处理;特征提取;神经网络;浮动模板法
中图分类号:TP391.43　　**文献标识码**:A

0　引　言

随着办公自动化的发展,印刷体字符识别技术已经越来越受到人们的重视。印刷体字符识别在不同领域有着广泛的应用,比较典型的有邮政编码自动识别、印刷电路板、工业元器件字符标识识别及各种中英文 OCR 表格与文本处理。在本文中我们研究的对象就是一种税务系统通用票据的识别,由于其应用的场合特殊,故要求其有相当高的正确识别率、尽量低的误识率、允许一定的拒识率。但是票据是全国各地不同的商家或厂商自行打印的,所以打印的字体、高度、大小千变万化,打印质量良莠不齐,加上在票据传送和收集过程中会污损,并且其字符集(由数字、部分大写字母及特殊字符"+","-","*","/",">","<"等组成)较大,这给识别造成了一定的困难。我们设计的识别系统如图 1 所示。

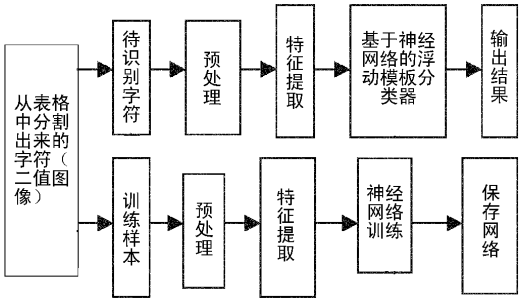

图 1　识别系统

Fig. 1 System of recognition

1　预处理

由于分割后获得的单字符大小不一,可能存在随机噪声、干扰,必须要对其进行一系列的处理以利于后面的识别,如除噪、消除断裂、字符平滑、归一化等。这里我们所采用的预处理过程大致为以下几个顺序连贯的步骤。

1.1　采用不对称结构元素的数学形态学处理

由于打印的字符常常断裂,并且同时存在小噪声,采用不对称结构元素的数学形态学处理恰好可以满足此要求。膨胀和腐蚀是 2 种最基本和最重要的变换或运算,是其它变换或运算的基础。

经过实验,确定结构元素如图 2 所示。

1	1	1
1	1	1
1	1	1

膨胀结构元素D

0	1	0
1	1	1
0	1	0

腐蚀结构元素R

图 2　非对称的膨胀结构元素与腐蚀结构元素

Fig. 2 Dissymmetrical dilated and eroded structure elements

由腐蚀、膨胀这 2 种最基本的数学形态学运算出发,采用不对称结构元素可构造出形态学处理算法:

$$A = (A \Theta R) \oplus D$$

其中,A 为待处理点阵。经过上述处理有效地消除了微断裂,同时去除了小噪声。

* 收稿日期:2004-05-12　　修订日期:2004-08-30

作者简介:赵全友(1980-),男,湖南永州人,博士研究生,研究方向为图像信息技术及系统、计算机模式识别,E-mail:quanyzhao@163.com;潘保昌,男,教授、博士生导师。

1.2 平滑笔画

经过上述处理后,图像点阵可能会产生毛刺或者小的凹陷。为了使点阵更为平滑,采用图3所示的模板进行处理。当图3中的像素 A,B 至少各有一个为1,即合乎该模板条件,则可判 P 为1。

A	0	A
0	P	0
B	0	B

A	A	0
A	P	B
0	B	B

图3 平滑模板

Fig. 3 Smoothing template

1.3 采用连通区域标号法消除大块噪声

由于要识别的印刷体字符集都是单连通的,前面的形态学操作及平滑操作进一步消除了断裂的可能性,但对于较大的噪声却无能为力。为此,可采用区域标号法可以去除较大的噪声块,当标号大于等于2时,即存在2个或者2个以上的区域,删除较小的图像块,保留面积较大的图像块。我们要处理的图像是二值图像,对图像点阵的标号除噪步骤如下。

(1) 进行从左到右、从上到下扫描。在同一行中不连通的行程(即像素值为1的点)标上不同的号,不同的列标上不同的号。

(2) 从左上到右下扫描,如果2个相邻的行中有相连通的行程则下行的号改为上行的号。

(3) 从右下到左上扫描,如果2个相邻的行中有相连通的行程则上行的号改为下行的号。

(4) 对标过的号进行排列。

(5) 对排列后的标号进行像素点统计,最多像素点的标号记为 MaxSign。

(6) 统计标号为 MaxSign 的点阵块的像素点数目记为 MaxNum。

(7) 删除噪声点,即所有标号不为 MaxSign 且点阵块的像素点数目小于 MaxNum ×25% 的图像块像素赋值为0,否则赋值为1。

1.4 归一化

由于打印字符图像点阵大小不一,为了抽取特征,必须对该字符点阵进行缩放,我们使其归一化为 32×32 大小的点阵。对于数字“1”和特殊符号“—”需要进行特别处理,采用重心归一化;其他字符采用外框归一化。即当原字符点阵的宽高比大于1.6或者小于0.8时,缩放不再是外框填满式缩放,外框中宽高较大者仍然按外框填满式缩放,而较小者缩放比例同较大者,之后使其重心居中。

2 特征提取

特征提取是指找出用于表示输入模式的合适特征,以便在特征空间中增大来自不同类的模式之间的差别的过程。通常,预处理获得的字符输入神经网络时,数据空间的维数会很高,神经网络的结构会很复杂,不适于直接识别。特征提取即是一个降维的过程,那些重要的变量即作为特征,次要的变量可忽略不计。特征提取需要将数据空间(模式空间)变换为特征空间。特征空间要尽可能地代表全体变量,虽然同最初的数据空间相比,特征空间维数减少很多,但它仍然保留了数据内容的大多数本质信息。单一的特征提取方法存在对某一干扰特别敏感的情况,并且不同的干扰对各种特征提取所造成的影响也有很大的区别。我们采取的是一种基于宏观特征与微观特征相结合的方法提取特征。

2.1 宏观特征

(1) 粗网格特征提取。粗网格划分方法如图4所示。把归一化的点阵划分成 8×8 的粗网格,统计每个网格里的黑点数,当点数小于总点数的25%时为0,大于25%小于75%为1,大于75%为2。这里共64个特征值。

(2) 七段框架投影值。七段框架投影形状类似于LED七段显示器的形状(图5),投影方法为将任意点向最近的框边投影。最后统计每个框边的投影点数,这样就形成了7个数,分别计算这7个数与字符总点数之比,然后归一化。归一化的策略如下:如果该比值小于0.08,则记为0;大于0.075同时小于0.15,则记为1;若大于0.15,则记为2。这里共7个特征值。

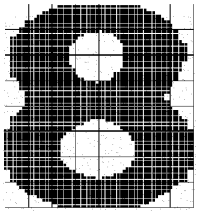

图4 粗网络划分方法示意图

Fig. 4 Sketch map of coarse-gridding partition method

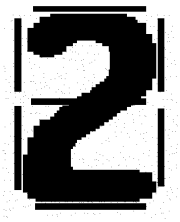

图5 七段框架投影示意图

Fig. 5 Sketch map of sevense-gment frame projection

2.2 微观特征

(1) 网孔数特征。统计字符中的闭合孔洞数,如“0”、“6”、“9”的孔洞数为1,“8”的孔洞数为2。这里为1个特征值。另外孔洞所在位置为一个特征

值,记孔洞在上部为 1,中部为 2,下部为 3,上下皆有为 4。共计特征值为 2 个。

(2) 线特征。分别统计水平、垂直,左下 45 度、右下 45 度 4 个方向的投影,当水平和垂直存在连续的宽度为归一化后点阵宽或者高1/16 且点数大于总宽度的 80% 时为 1,否则为 0 ;当左下 45 度和右下 45 度存在连续的宽度为归一化后点阵宽或者高 1/16 且点数大于总宽度的 90% 时为 1,否则为 0。这里共 4 个特征值。

(3) 凹特征。字符的凹特征是指字符边缘相邻两最外点距所夹最内点水平或垂直方向的距离向中心凹进去的程度。当凹深度≥给定的阈值时视为 1,否则为 0。分别计算左上、左下、右上、右下方向的凹特征,共计 4 个特征值。

(4) 凸特征。字符的凸特征是指在上、下、左、右 4 个方向,从字符边缘曲线检测存在 2 点之间曲线向外凸的程度。当凸深度≥给定的阈值时视为 1,否则为 0。分别计算上、下、左、右 4 个方向的凸特征,共计 4 个特征值。

(5) 穿线数特征。分别统计垂直中线,水平上中线,水平下中线的穿线数,把过线数作为特征值。这里共 3 个特征值。

由上可知,该抽取特征的方法不用细化,速度快,而且避免了细化产生的伪特征,更为可靠。把以上的各种宏观特征和微观特征共计 88 维向量作为神经网络的输入。

3　神经网络的浮动模板法分类器

由于字体的种类繁多,加上随机污染的引入,虽然经过一系列的预处理后有效的去除了许多干扰和形变,同时消除了部分字体间的差异,但是仍然难以使用传统的模板匹配及结构模式(如分类树)识别方法。实验中我们采用了改进的模板匹配法去识别,但是由于字体繁多,需要不断地增加新模板,这使得识别速度下降并且识别率维持在 93% 左右,难以提高。采用树分类器,其各级分类的标准难以选取,而且分类标准的鲁棒性值得商榷。而神经网络具有很强的学习和记忆能力,被广泛地应用到模式识别领域。基于以上原因,我们采用神经网络识别方法,而基于宏观特征与微观特征相结合的方法提取的特征送入神经网络训练,它能有效地减少网络规模。同时由于神经网络具有并行性的特点,字符的全局信息(即宏观特征)恰能迎合此种并行性的要求。为了缩短神经网络的训练时间,我们结合实际工程经验,给出了有效的训练方法及流程图。并且在分类器的设计上结合了浮动模板法,能较大幅度地提高识别率。

3.1　神经网络设计

BP 神经网络原理和算法,很多相关文献[4,5,9,14]作了详细的叙述,我们这里不再赘述。

3.1.1　网络参数确定

我们研究的待分类模式共 16 个,输入向量为 88。可确定神经网络的输入层 N 为 88,输出层 K 为 16,根据经验公式隐节点数可按 $L=\sqrt{N+K}+a$ 或 $L=\mathrm{lb}\ N$ 确定大致范围(式中 L 为隐含层节点数,N 为输入层节点数,K 为输出层节点数,a 为 0 ~ 10 之间的整数)。这里我们选用 $L=\sqrt{N+K}+a$ 来计算 $L=\sqrt{N+K}+a=\sqrt{88+16}+a=10.2+a$,由于选取是整数,经试验后选取 $a=6$,则 $L=16$。

3.1.2　改进的神经网络的训练

由于 BP 网络的训练时间长,有时候网络会陷入局部极小值或者不能收敛。为了加快训练速度,同时避免陷入局部极小值和改善其他能力。训练时采用如下策略改善训练。

(1) 采用带动量因子算法。动量因子的使用,有助于使网络从误差曲面的局部极小值中跳出。

(2) 采用自适应学习速率。在误差曲面比较平坦的区域,加快搜索速度,加大学习速率 ;当遇到谷地区域,放慢搜索速度,减少学习速率。

(3) 采用子集分步学习。如果某子集收敛,保存该子集训练权值。再逐步扩大这个子集进行训练,读入保存的权值,继续训练,从而避免不收敛。

(4) 跳读。如果训练过程中误差太少,则跳过不予反向更新权值。这样有助于加快网络训练。

其改进的神经网络训练流程图如图 6 所示,采用改进的神经网络训练算法能明显加快网络训练速度,保证网络能够收敛。

3.2　神经网络浮动模板法识别器

在实际的票据图片中,由于存在表格线,随机噪声污染等因素影响,在分割时不可避免地会带入一些,这样就会造成字符归一化偏移中心,并使字符变形以致识别错误。我们采用的浮动模板法能有效地消除这些影响。

3.2.1　浮动模板设计

浮动模板法,相当于一个可以浮动的略小于归

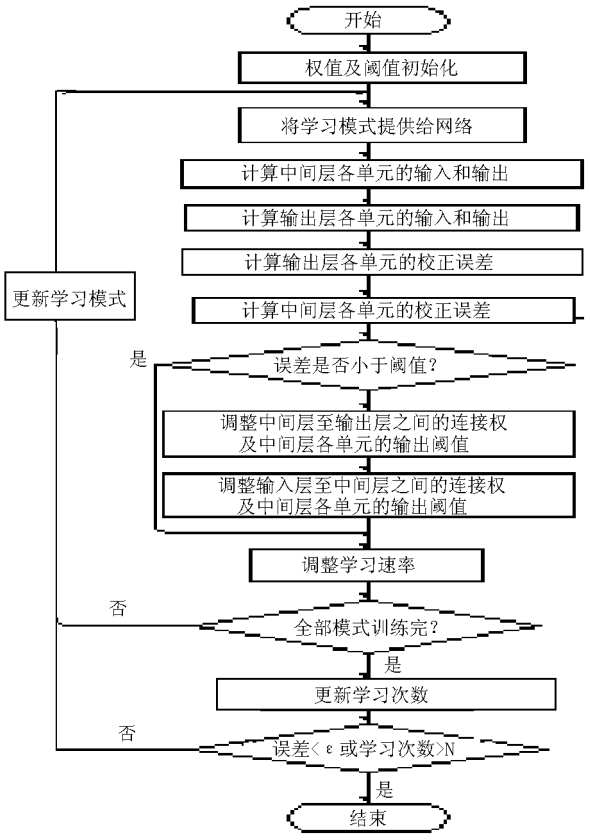

图6 改进神经网络的训练流程图

Fig. 6 Flow chart of ameliorative neural networks training

一化图像点阵大小的模板在预处理后的图像点阵上浮动,取可以被看见的部分,去掉模板外的部分,之后再采用外框归一化方法使其归一化。由于表格线往往存在于字符的上边、下边、左边、右边、左上边、右下边、右上边和左下边,其中下边、右边、右下边这3种情况粘连为最多,而右上边和左下边粘连这2种情况极少。所以我们设计了3种浮动模板来消除粘连表格线,其浮动模板的运动方向如图7所示。

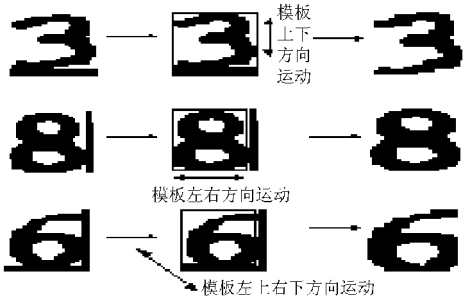

图7 浮动模板运动示意图

Fig. 7 Sketch map of floating mask moving

3.2.2 神经网络浮动模板识别

由于分割后单字符可能存在粘连的情况,或者不存在粘连的情况。首先将预处理过后的单字符经特征提取后再送入神经网络识别,如果误差小于阈值,则直接给出结果,如果误差大于或者等于阈值,则需要依次送入左右型浮动模板神经网络识别、上下型浮动模板神经网络识别、对角型浮动模板神经网络识别,每个识别都有一个误差,比较误差,输出三者误差较小者的识别结果。其全部识别流程如图8所示。

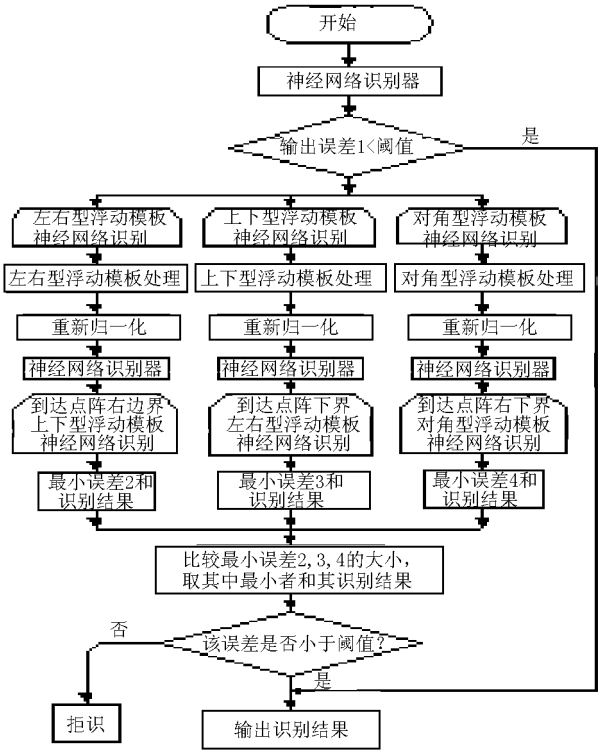

图8 浮动模板法神经网络识别流程图

Fig. 8 Flow chart of floating mask and neural networks recognition method

4 实验结果

以16个输出模式为例,我们从实际采集的税务票据图片中分割出了170 796个单字符图片。每个模式从中随机选取了1 240个单字符作为训练样本,共计19 840个单字符,剩余的150 956个字符作为测试样本。神经网络训练迭代次数在895次时,已经能识别所有的样本,这时的网络泛化能力最好。

我们采用Visual C++ 6.0进行编程,实现了上述算法。对测试样本集进行识别,测试识别结果见表1。

5 结 论

实验结果表明,该字符预处理方法十分有效,提取的宏观特征和微观特征都具有较好的鲁棒性。基

于神经网络和浮动模板的识别方法取得的识别精度令人相当满意,达到预期相当高的正确识别率、尽量低的误识率、允许一定的拒识率的目标,可见其应用在税务等高要求场合是可行的。

表 1　神经网络浮动模板法识别测试样本结果统计表

Tab. 1 Testing sample classified result statistics of floating mask and neural networks recognition method

模式	测试样本数/个	误识数/个	误识率/%	拒识数/个	拒识率/%	正确识别率/%
0	10 214	1	0.01	9	0.09	99.90
1	10 357	2	0.02	10	0.10	99.88
2	9 687	0	0.00	9	0.09	99.91
3	10 780	0	0.00	6	0.06	99.94
4	10 628	0	0.00	7	0.07	99.93
5	7 519	0	0.00	4	0.05	99.95
6	7 281	0	0.00	6	0.08	99.92
7	9 280	1	0.01	21	0.23	99.76
8	7 896	0	0.00	7	0.09	99.91
9	7 951	0	0.00	7	0.09	99.91
+	7 819	0	0.00	2	0.03	99.97
−	7 686	4	0.05	6	0.08	99.87
*	10 295	0	0.00	13	0.13	99.87
/	11 741	0	0.00	3	0.03	99.97
>	14 185	0	0.00	1	0.01	99.99
<	7 637	0	0.00	0	0.00	100.00
合计	150 956	8	0.005	111	0.074	99.921

参考文献:

[1] CHANG Pan bao. Floating Mask Method for Extracting Handprinted Character Feature [A]. IEEE proc,8th,ICPR[C]. Paris,1986.

[2] CASTLMAN K R. Digital Image Processing [M]. Prentice Hall,Inc. 1996.

[3] TOM M. Mitchell. Machine Learing [M]. McGraw-Hill Companies,Inc. 1997.

[4] MARTIN T. Hagan, HOWARD B. Demuth, MARK H. Beale. Neural Network Design[M]. PWS Publishing Company. 1996.

[5] Il-Seok Oh, CHING Y. Suen. Distance features for neural network-based recognition of handwritten characters [J]. International Journal on Document Analysis and Recognition. 1998,7:73-88.

[6] FANG Chi, LIU Chang-song. Automatic performance evaluation of printed chinese character recognition systems [J]. International Journal on Document Analysis and Recognition. 2002,3:177-182.

[7] 潘保昌. 浮动模板法—一种抽取字符特征的方法[J]. 计算机学报,1983,6(6):469-477.

[8] 张平,潘保昌. 一种自由手写体数字识别方法研究[J]. 光电工程,1995,22(6):43-46.

[9] 郑南宁,王龙. 胡超,等. BP 神经网络的改进及其用于手写数字识别的研究[J]. 西安交通大学学报,1992,26(2):1-12.

[10] 杜敏,辛大欣. 基于混合特征的提取的手写体数字识别方法的研究[J]. 西安交通大学学报,1996,30(9):94-99.

[11] 严国莉,黄山. 印刷体数字快速识别算法在身份证编号数字识别中的应用[J]. 计算机工程, 2003,29(1):178-179.

[12] 张宏林. Visual C + + 数字图像模式识别技术及工程实践[M]. 北京:人民邮电出版社, 2003.

[13] 谢忠红. 基于组合分类器的自由手写体数字识别方法[D]. 南京:南京气象学院,2003.

[14] 郑胜林,彭明明,潘保昌. 一种基于 Hough 变换的神经网络字符识别方法[J]. 广东工业大学学报,2003,4:73-77.

(责任编辑 刘勇)

Implementation of effective multifont printed character recognition system

ZHAO Quan-you,PAN Bao-chang,YAO Jin-xiu,ZHENG Sheng-lin,CHEN xiao-feng

(*College of Information Engineering,Guangdong University of Technology,Guangzhou* 510640,*P. R. China*)

Abstract:A multifont printed character recognition based on floating mask method and neural network is presented in this paper. After researching on many multifont printed character images, authors give an effective preprocessing method. Then the macro features and the micro features are extracted, which are to be sent to the BP neural network. The classifier of combination of the neural network and the floating mask yields the right results of classification. The experiment proves this method can get very high recognition rate. The prospect of its application in many fields is obvious.

Key words:multifont character recognition; preprocessing; feature extracting; neural network; floating mask method