基于组合特征的高效数字识别算法*

孔月萍[1,2]，曾 平[1]，李智杰[2]，郑海红[1]，徐培培[1]
(1.西安电子科技大学 计算机外部设备研究所，陕西 西安 710071；2.西安建筑科技大学 信息与控制工程学院，陕西 西安 710055)

摘 要：针对监控屏幕中的数字字符提出了一种高效的识别算法。该算法利用字符图像的欧拉数、凹陷区、水平和垂直穿线等组合特征完成级联分类，无须对待识别字符进行规整、细化和轮廓提取处理，降低了算法复杂度，减少了因细化变形、轮廓断裂引起的误识和拒识。在以此算法为基础实现的监控信息自动采集与记录系统中，对 5 000多个屏幕显示数字字符进行识别测试，平均每秒处理 125个数字，正确识别率达到 98.70%，误识率仅为 1.30%。实验表明该算法在处理速度、识别精度、抗干扰性方面表现良好。

关键词：数字识别；欧拉数；凹陷区；穿线
中图法分类号：TN911.73；TP391.41 **文献标识码**：A **文章编号**：1001-3695(2006)10-0172-02

Effective Recognition Algorithm for Digits Based on Combination Features

KONG Yue-ping[1,2], ZENG Ping[1], LI Zhi-jie[2], ZHENG Hai-hong[1], XU Pei-pei[1]
(1. *Research Institute of Computer Peripherals, Xidian University, Xi'an Shanxi* 710071, *China*; 2. *College of Information & Control, Xi'an University of Architecture & Technology, Xi'an Shanxi* 710055, *China*)

Abstract: An effective recognition algorithm is proposed for digits on monitor screen. By using the features of euler number, concave field, horizontal and vertical crossing Line, digits can be classified and recognized. The algorithm doesn't needs any regularization, thinning and outlining operations on the digits image. So it achieves low memory and computation. With the proposed method a monitor information catching and record system is realized. More than 5000 digits are involved in the experiments and a conclusion is drawn that the average recognition rate is 98.70%, the error rate is 1.30%, and the recognition speed is 125 digits per second. Experiments show that the algorithm is fast with high recognition rate and strong anti-interference ability.

Key words: Digits Recognition; Euler Number; Concave Field; Crossing Line

数字识别是光学字符识别(OCR)的一个重要分支，在脱机自动记录、车牌号码、身份证号码、支票号码、邮政编码以及其他编号识别方面具有重要的实用价值。传统 OCR过程大都包含二值化、去噪、规整、细化、轮廓提取、特征提取、字体字符分类等处理步骤[1,2]，系统运行效率较低。许多数字字符识别系统为了提高识别率，还需对字符笔画进行大量的形状分析或笔画拟合，寻找各笔画所包含的线段、弧、钝角、锐角、圈等，有的还引入了投影运算[3]、神经网络[3,4]，甚至级联分组神经网络[5]，使识别系统变得非常庞大和复杂，无法满足监控信息自动采集与记录、车牌识别等实时处理的要求。基于整体凹陷特征的手写数字识别算法[6]，避免了规整、细化、轮廓提取等图像处理步骤，但其凹陷区生成算法仍有冗余，且未考虑字符图像采样不理想时产生的虚假凹陷特征。本文从快速、准确的屏幕监控数字识别目标出发，设计了基于字符图像欧拉数、凹陷区、水平和垂直穿线等组合特征的数字字符级联分类器，仅需对待识别对象进行多级分类就可达到识别的目的。

1 数字识别算法

字符分类特征的选择决定着识别的效果，因此所选特征应充分反映字符的本质并遵循下述原则[7,8]：①特征的分类能力强，足以区分各个字符；②特征稳定，受字形变化影响越小越好；③特征便于提取，抽取速度快；④特征数量尽可能少。

由于阿拉伯数字均由一些曲线构成，因此环(或称孔洞)是最基本的拓扑特征；此外，最高效的识别器——人眼在识别数字时，根本不关心字符的直线数、凹弧凸弧数、钝角锐角数，往往只观察字符的整体凹凸形状和上下、左右结构关系就可以高效地判断其所属[6,9]。本文基于以上特征分析和选择原则，以数字字符的欧拉数、凹陷区、全数字列、水平穿线作为拓扑不变属性，提出了基于上述组合特征的快速数字识别算法，系统整体框架如图 1所示。由于该算法依据字符的组合特征完成识别，与字符的大小、位置无关，因此，预处理阶段仅需二值化、去噪、字符分割即可，核心由后续的多级特征分类器组成，即利用数字字符的欧拉数特征进行一级分类；其次，利用字符图像的凹陷区特征、全数字列、水平穿线特征达到最终分类识别的目的。

图 1 数字识别算法框架

1.1 数字字符的欧拉数特征

欧拉数的定义可表示为式(1)，其中 C表示图像中数字字

收稿日期：2005-08-26；**修返日期**：2006-08-01
基金项目：陕西省自然科学基金资助项目(2004F32)

符前景区域的连通分支数,H表示字符前景区域内的孔洞数。

$$E = C - H \quad (1)$$

对数字字符而言,其中的 0,4,6,8,9均存在一个或两个孔洞,而 1,2,3,5,7则没有孔洞。因此,利用欧拉数特征值 E可将数字字符集合划分成三个子集,以缩小判别空间,利于下一步识别处理。

1.2 数字字符的凹陷区特征

在待识别字符图像中的任意两点间画直线,直线中不属于字符前景部分所在的区域即为图像的凹陷区。左凹陷区指字符图像的任意背景点右边均存在字符笔画的凹陷区;右凹陷区指字符图像的任意背景点左边均存在字符笔画的凹陷区[1,6]。

本文对文献[6]中提取图像凹陷区的算法进行了改进,即沿水平、垂直、45 和 135 对角线四个方向发出射线,扫描、判断射线是否与字符前景相交,若有 x条射线与字符相交,则该背景点标记为 x,从而得到字符图像的凹陷区赋值背景场。算法描述如下:

{初始化所有背景点标记为 0;沿水平、垂直、45°、135 射线方向依次作如下处理;
　{取出一个图像行 (射线)形成行向量 L_i;
　　if L_i中包含数字字符部分且仅有一个连通区域
　　then所有背景点标记增 1;
　　if L_i中包含数字字符部分且有一个以上连通区域
　　then L_i中连通区域之间的所有背景点标记增 2,其余背景点标记增 1;}
}

在凹陷区赋值背景场标记图像基础上,扫描字符图像每个凹陷区的标记背景场,若每个背景点右边均存在数字字符部分,则该数字字符具有左凹陷区。右凹陷区的判断方法与此类似,不再赘述。图 2即是数字字符 9,6,1的赋值背景场标记图像。图 2(a)中标记为垂直线段的部分即为字符 9的右凹陷区;图 2(b)中标记为垂直线段的部分即为字符 6的左凹陷区。数字字符的凹陷区数目 CN及左、右位置 CL, CR反映了字符的凹凸特征,依此特征可实现在欧拉数分类基础上更精细的二级分类。

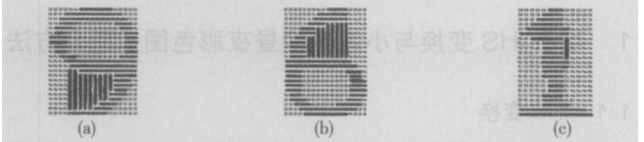

图 2 数字字符的赋值背景场、凹陷区、左凹陷区、右凹陷区标记图像

1.3 数字字符的水平垂直穿线特征

由于图像的采样条件不一定理想,会出现虚假凹陷区 (如图 2(c)中垂线部分所示的凹陷区),因此需要设置凹陷区尺度阈值,判断并消除虚假凹陷区,阈值可根据图像大小来确定;但设定不合理时,仍会消去某些真正的凹陷区,如数字字符 4的右凹陷区,为弥补此缺陷将在全数字列、水平穿线特征分类上加以完善。

设 $\mathrm{origin}_{M\times N}$为待识别数字字符的二值图像,定义字符前景点位置集合为式 (2),全数字列标志为式 (3),水平穿线与字符前景的交点集合为式 (4),该交点集与参考点的位置标志为式 (5)。

$$F = \{ (i, j) \mid (i, j) \text{为字符前景像素点的位置} \} \quad (2)$$

$$T_c = \{ t \mid t = \text{True, if } j = a \text{(常数)} \wedge \forall (i, j) \in F \} \quad (3)$$

$$H_r = \{ (i, j) \mid r = (b, c) \wedge i = b \wedge (i, j) \in F \} \quad (4)$$

$$LH_r = \{ v \mid v = \text{True, if } \forall (x, y) \in H_r, \mid (x, y) - (b, c) \mid \leq 0 \} \quad (5)$$

其中,$r = (b, c)$为水平穿线参考点,$i = 1, 2, \cdots, M$; $j = 1, 2, \cdots, N$。显然对于数字字符 1,4满足全数字列穿线特征,而 0,2,3,5,6,7,9则不满足此条件。此外,对于数字字符 2,3,5分别以其顶部、底部凹陷区的第一行最右像素点 $r1 = (c_1, d_1)$, $r2 = (c_2, d_2)$为基准作两条水平穿线,直线和字符前景的交点集 H_{r1}, H_{r2}与参考点的位置关系 LH_{r1}, LH_{r2}表现各异,据此即可完成欧拉数、凹陷区分类基础之上的最终识别。

1.4 数字识别算法描述

综合上述欧拉数、凹陷区、全数字列、水平穿线特征的定义和分析,可得出基于字符图像组合特征的数字判别逻辑关系如式 (6)。

$$X = \begin{cases} 0 & E = 0, CN = 0 \\ 1 & E = 1, CN \leq 1, T_c = \text{True} \\ 2 & E = 1, CN > 1, LH_{r1} = \text{False}, LH_{r2} = \text{True} \\ 3 & E = 1, CN > 1, LH_{r1} = \text{False}, LH_{r2} = \text{False} \\ 4 & E = 0, CN > 0, T_c = \text{True} \\ 5 & E = 1, CN > 1, LH_{r1} = \text{True} \\ 6 & E = 0, CN > 0, T_c = \text{False}, CL = \text{False} \\ 7 & E = 1, CN \leq 1, T_c = \text{False} \\ 8 & E = -1 \\ 9 & E = 0, CN > 0, T_c = \text{False}, CL = \text{True} \end{cases} \quad (6)$$

其中,CN表示凹陷区数目, CL, CR表示左、右凹陷区存在标志。因此,基于组合特征数字识别算法的形式描述为

{原始采集图像二值化、去噪、字符分割,得到待识别数字字符图像 $\mathrm{origin}_{M\times N}$;
　由 $\mathrm{origin}_{M\times N}$计算字符前景像素点集合 F;
　由 $\mathrm{origin}_{M\times N}$计算字符赋值背景场图像 $\mathrm{back\_sign}_{M\times N}$;
　由 $\mathrm{back_sign}_{M\times N}$计算字符图像的凹陷区数目 CN,左、右凹陷区存在标志 CL, CR;
　由 $\mathrm{origin}_{M\times N}$, F计算字符图像的 E, T_c, LH_{r1}, LH_{r2};
　根据式 (6)判别数字值 X;
}

2 实验结果及讨论

以此算法为基础,在 Intel Pentium (R) 1.6GHz处理器、256MB内存的计算机上以 Visual C++6.0为平台设计,实现了监控信息自动采集与记录系统,对采集到的 5 222个屏幕显示数字字符进行测试,识别结果及系统执行时间统计如表 1、表 2所示。实验数据表明,该算法充分利用了数字字符结构简单、判别集有限的特点,避免了细化、轮廓提取等图像处理步骤,节省了存储空间和运行时间,减少了误差,达到了快速准确识别数字字符的效果,其性能可靠、抗干扰性较强,是解决数字识别问题的有效方法之一。

表 1 实验结果统计

样本	测试总数	正确识别数	误识数	识别率	误判率
字符 0	525	519	6	98.86%	1.14%
字符 1	514	511	3	99.42%	0.58%
字符 2	489	485	4	99.18%	0.82%
字符 3	577	564	13	97.75%	2.25%
字符 4	539	526	13	97.59%	2.41%
字符 5	495	487	8	98.38%	1.62%
字符 6	513	507	6	98.83%	1.17%
字符 7	488	485	3	99.39%	0.61%
字符 8	534	529	5	99.06%	0.94%
字符 9	548	541	7	98.72%	1.28%
合计 平均	5222	5154	68	98.70%	1.30%

(下转第 182页)

输出层神经元为一个,最大训练次数 600,期望误差为 0.001,初始学习率为 0.01。仿真结果如表 3所示。

表 3 标准 BP人工神经网络与 SVR预测结果对比表

年份	实际客运量(万人)	标准 BP神经网络预测值(万人)	差值(万人)	相对误差%	SVR的预测值(万人)	差值(万人)	相对误差%
1983	1 049 460	1 028 857	- 20 603	- 1.9	1 048 244	- 1 216	- 0.1
1984	1 122 650	1 077 132	- 45 518	- 4.1	1 118 308	- 4 342	- 0.4
1985	1 109 100	1 153 984	44 884	4.1	1 147 819	38 719	3.5
1986	1 073 600	1 075 552	1 952	0.2	1 074 816	1 216	0.1
1987	1 114 139	1 171 980	57 841	5.2	1 104 031	- 10 108	- 0.9
1988	1 216 000	1 150 579	- 65 421	- 5.4	1 175 981	- 40 019	- 3.3
1989	1 128 000	1 118 233	- 9 767	- 0.9	1 129 216	1 216	0.1
1990	948 900	942 764	- 6 136	- 0.6	950 116	1 216	0.1
1991	942 080	933 280	- 8 800	- 0.9	943 296	1 216	0.1
1992	987 900	999 187	11 287	1.1	989 116	1 216	0.1
1993	1 045 800	1 084 064	38 264	3.7	1 055 636	9 836	0.9
1994	1 080 090	1 068 742	- 11 348	- 1.0	1 077 913	- 2 177	- 0.2
1995	1 020 810	1 080 416	59 606	5.8	1 063 033	42 223	4.1
1996	936 000	994 201	58 201	6.2	983 042	47 042	5.0
1997	919 000	979 366	60 366	6.6	867 536	51 464	5.6
1998	930 000	1 004 051	74 051	7.9	989 697	59 697	6.4

从表 3可以看出,在上述两种方法对铁路客运量进行的预测中,两种方法所获得的预测相对误差在一定程度上都能反映铁路客运量时间序列趋势,但在训练误差方面和测试误差方面,基于 ε-SVR的方法明显小于 BPANN,这主要反映了对于训练样本以外的检验样本,基于 ε-SVR方法有更强的泛化预测能力。

4 结论

采用 ε-SVR对铁路客运量时间序列进行预测研究。通过与标准的 BPANN对比表明,ε-SVR对铁路客流数据有更好的预测效果,有很强的自学习性、自适应性,且收敛快、准确性高,说明该预测模型是可信的。本预测方案是完全数据驱动的,是定量的,具有一定局限性。因此在该方法的基础上附加一定的定性分析,以弥补完全数据驱动的不足。另外,到目前为止,成熟的铁路客流量预测方法仍然停留在对运输总量的预测上,但是总运量预测只是运量预测一个方面的内容,还必须考虑客流在具体发到地点和具体线路上的分布问题。对具体运输产品的运量在空间位置分布上的研究和预测,对路网建设、投资决策和经营管理有更实际的参考价值。

参考文献:

[1] 杨浩. 铁路运输组织学 [M]. 北京:中国铁道出版社, 2001. 21-24.

[2] 侯福均,吴祈宗. BP神经网络在铁路客运市场时间序列预测中的应用 [J]. 运筹与管理, 2003, 12(4): 73-75.

[3] Vapnik V. An Overview of Statistical Learning Theory [J]. IEEE Trans on NN., 1999, 10(3): 988-999.

[4] Cortes C, Vapnik V. Support Vector Networks[J]. Machine Learning, 1995, 20(4): 273-297.

[5] Osuna E, Freund R, Girosi F. Training Support Vector Machines: An Application to Face Detection [C]. Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, LosAlamitos, CA, USA: IEEE Computer Society, 1997. 130-136.

[6] Guyon I, Weston J, Barnhill S, *et al* Gene Selection for Cancer Classification Using Support Vector Machines [J]. Machine Learning, 2002, 46(6): 389-422.

[7] 邓乃扬,田英杰. 数据挖掘中的新方法——支持向量机 [M]. 北京:科学出版社, 2004. 96-166.

[8] 张学工. 关于统计学习理论与支持向量机 [J]. 自动化学报, 2000, 26(1): 32-42.

[9] The World Bank Transportation, Water and Urban Development Department[EB/OL]. http://www.railinfo.com/gb/db/wbtjsj/default1.asp, 2000-01.

[10] 岳毅宏,韩文秀,程国平. 多变量时间序列相空间重构中参数的确定 [J]. 控制与决策, 2005, 20(3): 290-293.

[11] Ito K, Nakano R. Optimizing Support Vector Regression Hyperparameters Based on Cross-validation [J]. Neural Networks, 2003, (3): 2077-2082.

作者简介:

夏国恩 (1977-),男,四川内江人,博士研究生,主要研究方向为管理信息系统、决策支持系统、商务智能系统;曾绍华 (1969-),男,重庆壁山人,讲师,博士研究生,主要研究方向为计算机网络、数据挖掘;金炜东 (1959-),男,安徽淮南人,教授,博导,博士,主要研究方向为优化理论与优化控制、智能信息处理、系统仿真等。

(上接第 173页)

表 2 运行时间统计

样本图像大小	测试总数	执行时间 (s)	平均执行时间 (s)
40 ×25	112	1.570	1.402 ×10-3
30 ×20	123	1.014	0.824 ×10-3

参考文献:

[1] Milan Sonka, Vaclav Hlavac, Roger Boyle. 图像处理、分析与机器视觉 [M]. 北京: 人民邮电出版社, 2001. 204-211.

[2] 章毓晋. 图像工程——图像处理和分析 [M]. 北京: 清华大学出版社, 2003. 232-233, 244.

[3] Minchul Jung Font Classification and Character Segmentation for Postal Address Reading[D]. USA: Bell & Howell Information and Learning Company, 2001. 30-36, 65-67.

[4] I-Chang Jou, Shih-Shien Yu, Tsay S C. A New Feature Extraction Method by Neural Networks[C]. Circuits and Systems IEEE International Symposium, 1990. 3249-3252.

[5] 王伟, 盛立东. 基于级联分组 BP神经网络的高精度手写体数字识别系统 [J]. 中文信息学报, 2000, 14(2): 60-62.

[6] 龚才春, 刘荣兴. 基于整体特征的快速手写体数字字符识别 [J]. 计算机工程与应用, 2004, (19): 82-83.

[7] Torfinn Taxt Recognition of Handwritten Symbols[J]. Pattern Recognition, 1990, 23(11): 1156-1166.

[8] 杜建强, 陈月林, 刘少媚, 等. 工程图纸上的字符提取和识别系统 [J]. 计算机技术与自动化, 1995, 14(4): 40-42.

[9] 许捍卫, 王成. 一种简单的数字识别方法研究 [J]. 地矿测绘, 2003, (4): 31-32.

作者简介:

孔月萍 (1965-),女,四川人,副教授,博士研究生,主要研究方向为图形图像处理、数据库技术;曾平 (1956-),男,四川人,博导,硕士,主要研究方向为图形图像处理、计算机外部设备;李智杰 (1981-),男,河南人,硕士研究生,主要研究方向为图像处理;郑海红 (1979-),女,河北人,博士研究生,主要研究方向为图形图像处理;徐培培 (1955-),女,上海人,高工,主要研究方向为计算机外部设备。