文章编号:1005-0523(2009)06-0074-05

基于综合特征的字符模板库的建立与训练

蒋先刚,赵 莹,李 林

(华东交通大学 基础科学学院,江西 南昌 330013)

摘要:叙述邮政编码自动分拣系统的软件处理的原理以及程序设计方法,对于倾斜放置信封的邮编区域的定位问题提出解决方案,利用信息熵理论对字符的综合特征的选取提出融合方案,建立训练了邮政编码识别系统中的手写数字模板库,并设计与之相关的可视数据库系统,该系统大大提高信封的自动分拣率以及可行性。

关 键 词:字符特征;邮编识别;数字模板库

中图分类号:TP391.41 **文献标识码**:A

在邮政编码识别系统中,邮编区域定位是整个识别系统识别率的关键;模板库的合理建立及训练也可大大提高软件的识别率;软件及数据库的合理利用也关系到整个系统能否方便应用以提高工作效率。本文将对这几个问题的具体实现与改进进行探讨,并着重探讨字符模板库的建立训练技术。

1 邮政编码识别系统软件框架设计

邮编识别过程包括邮编矩形框的定位和框内文字的识别两大步骤。软件的具体流程见图1。

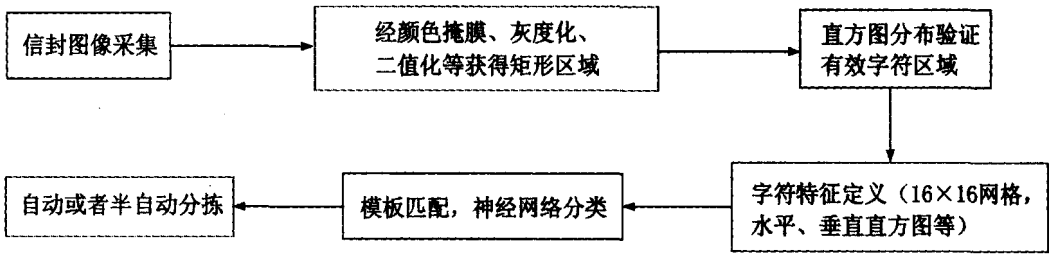

图1 邮政编码识别软件系统流程图

邮政编码识别与分拣系统主要由四大模块组成:

(1) 数据采集模块。对信封图像的获取主要由CCD摄像机、图像采集卡组成,当信封处于摄像头下时,由摄像头获取一幅图像,并通过图像采集卡转换成数字图像信息输入到计算机中。

(2) 邮编区域定位模块。主要通过颜色掩膜的方法来对邮编的位置进行定位,因为我国通用的信封邮编区域标准为6个红色的矩形框,那么邮编区域的特征就是红色区域,且为6个连续矩形框。由于不同印刷厂的油墨差异使得不同信封的红色定义不同,我们通过对50个不同信封的图像采集,用数理统计的方法定义出适合于大部分信封的红色彩色值的定义范围,从而提高其鲁棒性。为了避免同一信封上有多处红色区域,在颜色掩膜处理后,我们用直方图分布特征或hough技术来进一步验证邮编区域定位是否正确。

(3) 邮编框内的数字识别模块。首先对0-9数字进行字符特征定义,主要方法有:网格特征、水平和垂直方向投影直方图特征以及左右扫描线轮廓特征等,然后建立数字模板库,通过将信封上的字符与模板库里的字符进行比较或其它分类方法来进行识别。

(4) 信封分拣的机械执行模块。建立1个机械指令操作数据库,将全国各地邮编的地名与机械手移动的方位和距离建立一一对应关系,且机械手的移动方位和距离可以人为任意调整,实现系统的个性化设

收稿日期:2009-09-07

作者简介:蒋先刚(1958-),男,湖南永州人,教授,研究方向为图像处理与软件技术。

计从而达到邮件自动分拣。

2　邮政编码图像区域定位

2.1　水平放置信封的邮编区域定位

对于水平放置的信封,邮政编码的定位主要核心技术是掩膜处理,也就是只让信封上的红色矩形区域可见。这里对红色的定义就很重要,我们对拍摄的50个信封红色区域像素点进行数理统计,得出红色的HSL空间[1]中心定义为 $h \leqslant 310, l = 0, s = 0$,然后进行颜色掩膜处理,得出6个红色矩形框的位置[2]。

2.2　倾斜放置信封的邮编区域定位

对于随意放置的信封,通过颜色掩膜处理得到的是1个倾斜的邮编区域,这对于后续字符的正确识别有很大的影响,因此需要对倾斜的邮编区域做1个几何变换,将之旋转到水平位置。要进行这样的几何变换就必须得知邮编区域的倾斜角度,这里用Hough检测邮编区域的倾斜角度。

Hough变换[3]是1种检测、定位直线和解析曲线的有效方法。它是把二值图变换到Hough参数空间,在参数空间用极值点的检测来完成目标的检测。在这里我们用的是直线检测,因为6个邮编框的底边一定是在一条直线上的,我们只要检测出底边倾斜的角度,就可以对图像进行几何变换。直线检测模拟情况如下:

对一直角坐标系中的直线,其方程可以写成

$$\rho = x\cos\theta + y\sin\theta \tag{1}$$

式中:ρ 是原点到直线的距离;θ 是该直线的法线与 X 轴的夹角。参数 ρ 和 θ 可以唯一地确定一条直线。以式(1)作为 $X-Y$ 坐标向 $\rho-\theta$ 坐标的变换方程,进行 $X-Y$ 平面内点集的映射。

用Hough变化检测出倾斜邮编区域后,可以按照检测结果对图像做任意角度旋转,将邮编区域图像置为水平。

3　邮政编码特征定义

在定位出邮政编码的位置后,就要对邮政编码里的字符进行特征提取和识别。为了提高识别效率需要建立1个数字模板库,通过信封待识别数字与模板库内字符的对比,选出相似程度最大的数字为目标识别的值。

为了建立模板库,首先要对数字样本进行特征定义,我们主要使用的方法是:网格特征,水平、垂直投影直方图特征及左右轮廓特征等。

3.1　邮政编码字符网格特征的定义[4]

首先找到每个手写数字样本的连通域的最小矩形区域,在此附近搜索该样品图像的有效宽度和高度,将每个样品的长度和宽度分成 $M \times N$ 等份,构成1个 $M \times N$ 均匀的小区域,对于每1个小区域内的黑色像素个数进行统计,除以该小区域黑像素的面积总数,即得特征值。这样针对同一形状、不同大小的样品得到的特征值相差不大,对同一形状,不同大小的样品视为同一类别的数字。其中 M 和 N 即为网格数,所选择的网格数应该根据技术要求和计算机运算速度综合考虑,M 和 N 值越大,模板也越大,特征数目越多,其区分不同数字的能力越强。因为随着特征数目的增多,运算比较分类的计算时间将增加,系统中需要记录的样品库文件尺寸也需要成倍增加。如果 M 和 N 取值过小,则不利于各数字特征的区别。本系统中,我们取 $M \times N$ 为 16×16,这样既可以满足准确的区分不同字符,又没有过多的计算消耗。

3.2　水平、垂直直方图[5]及左、右轮廓特征定义

直方图是用来表示一幅图像的灰度值分布情况的统计图表。水平、垂直直方图的横坐标是位置,一般用 t 表示,纵坐标为该位置的像素个数。该直方图具有下列性质:

(1) 该直方图反映了图像中各水平(或垂直方向)、各位置的灰度像素个数,并反映了各灰度值像素所在水平(或垂直)的位置,即该直方图反映出了目标的空间信息,可大致定位目标;

(2) 对任意一幅图像,可以唯一地算出一幅与它对应的直方图;

(3) 由于直方图是通过对各水平(或垂直)位置具有灰度值的像素个数进行统计计数得到的。因此一幅图像各子区的直方图之和就等于该图全图的直方图。

因此,我们可以通过对0-9各个数字图像的水平或垂直直方图的统计,来定义字符特征,从而达到识别的效果。

为了建立更好的模板库,我们还引入了水平方向扫描线从左边和右边第1次遇到黑点的序号数特征来进行另1个特征定义,这样可以更好的区分易混淆字体。

3.3 字符综合特征定义

用上述方法对手写邮政编码字符进行定义,特征数就有 $16\times16+16\times4=320$,为了达到更好的识别效果,我们将互信息作为字符的相似性测度,这样可以很好的将字符 16×16 网格特征,水平、垂直直方图特征以及字符左、右轮廓特征融合起来。

Shannon 互信息[6]的定义如下

假定 A 的概率分布密度函数 $p_i=P(a=i)$,其中 $i=1,2,\cdots,m$,那么 A 的 Shannon 熵为:$H(A)=-\sum_i p_i\log_2 p_i$,式中 H 表示熵;A 表示图像;a 表示图像灰度值。

对于两幅图像 A 和 B,它们之间的图像互信息为

$$I(A,B)=\sum_{i,j}p_{ij}\log_2\frac{p_{ij}}{p_i p_j}=H(A)+H(B)-H(A,B) \tag{2}$$

式中:p_i 和 p_j 是 A 和 B 的灰度概率分布;p_{ij} 是联合灰度概率分布;$H(A,B)$ 是 A 和 B 的 Shannon 联合熵。

这样我们将样本字符与待识别字符的网格特征的互信息定义为 $I_g(A,B)$,其中 A 为样本字符图像,B 为待识别字符图像,同理我们将样本与待识别字符图像的水平、垂直直方图特征以及左右轮廓特征的互信息分别定义为 $I_s(A,B)$,$I_p(A,B)$,$I_l(A,B)$,$I_r(A,B)$。

两幅图像的总体互信息为

$$I^*(A,B)=\alpha I_g(A,B)+\beta I_s(A,B)+\gamma I_p(A,B)+\mu I_l(A,B)+\nu I_r(A,B) \tag{3}$$

(3)式中:$0\leqslant\alpha,\beta,\gamma,\mu,\nu\leqslant1$,$\alpha+\beta+\gamma+\mu+\nu=1$。

将这些互信息分别计算出来再把它们加权平均,这样既可以反映字符图像的全局特征,又考虑了局部细节特征,可以使识别率得到有效提高。本文通过实验对比取各参数值分别为 0.35,0.10,0.10,0.35,0.10。

4 手写模板库的建立与训练模块的设计

为了建立手写数字模板库,我们用黑色签字笔在 A4 纸上均匀写上 $6\times4=24$ 个数字,包含 0-9 内数字。共取 40 张样本训练模板库。模板库建立流程如图 2。

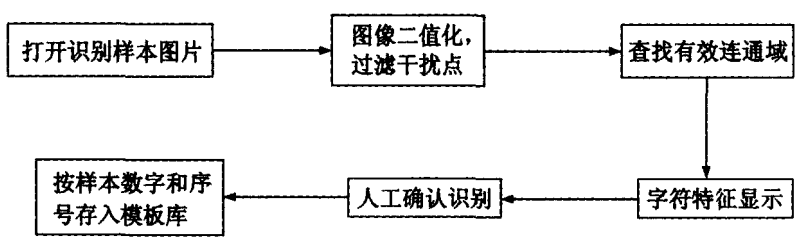

图2 手写字符样本库的建立与训练流程

打开任意1张样本图,对样本图进行二值化等预处理,然后在图中查找有效连通域,将查找到的连通域返回到1个标签中,此时,这个字符图像将通过归一化[7],划分为 16×16 的网格,然后统计每个网格的黑像素的个数,即网格内含有字符笔画像素的个数,当含有笔画的个数达到网格总像素的20%时,将这些像素当成干扰信号,将这个网格背景设为信息0,否则将其设为字符信息1,得出的0,1组成的矩阵即为字

符的特征向量。同时，统计出该字符的水平、垂直直方图特征以及左、右轮廓特征。由人工识别出数字并输入此数字，这时程序就将字符以及它的各个特征保存到模板库中，将查找到的连通域置为背景色，继续寻找下一个连通域。模板库的建立过程中的一些特征和处理过程如图 3 所示。

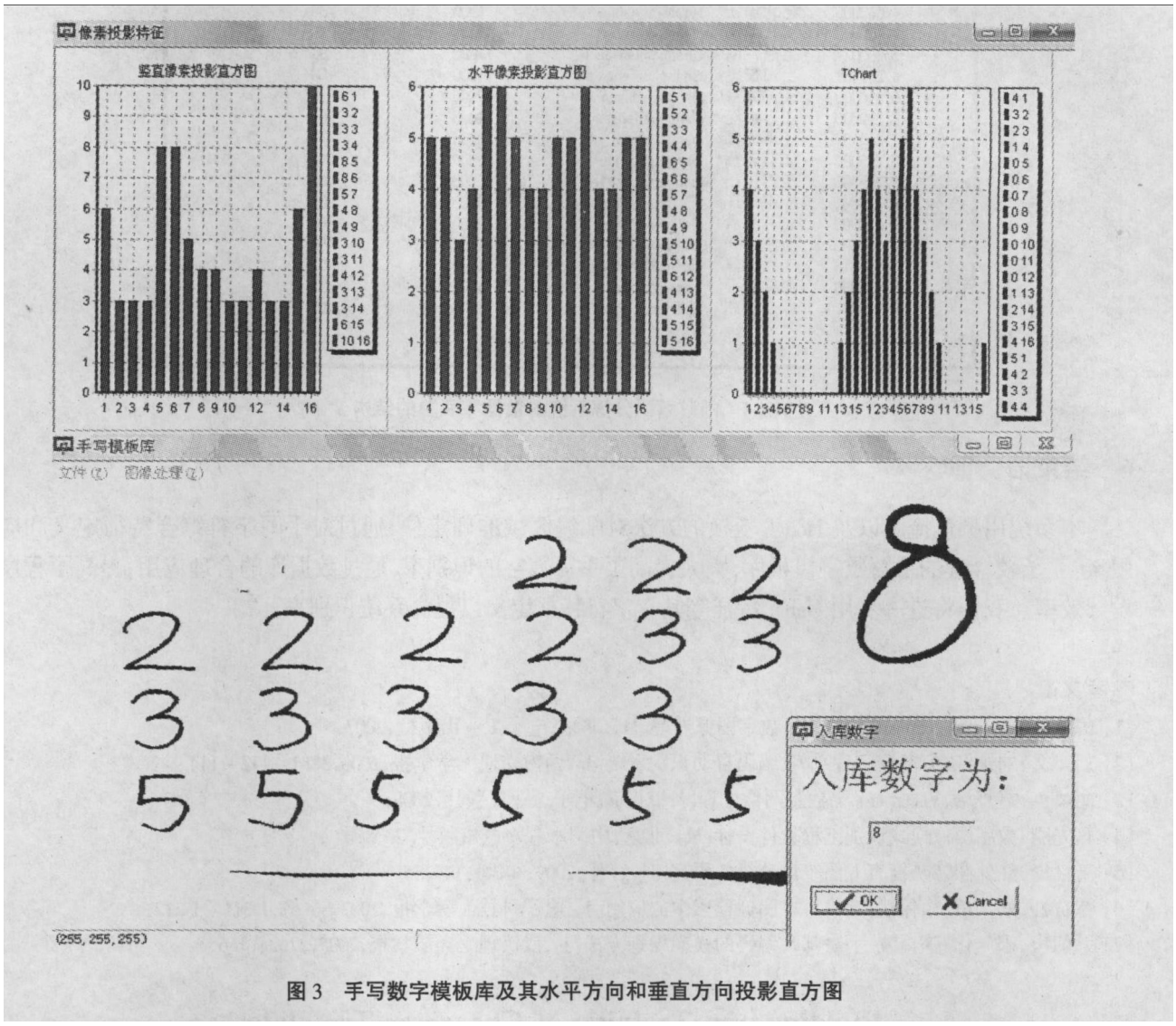

图 3　手写数字模板库及其水平方向和垂直方向投影直方图

5　分拣系统的数据库设计

考虑到我国邮政系统所涉及区域的广泛性、复杂性与程序编制的简约性和模块化，将对应不同邮政编码区域择选所需的语音播放对象及机械执行方位进行有效关联以实现自动分拣或者半自动分拣，我们建立了 1 个邮政编码处理可视数据库，字段包含有邮政编码、对应区域地址、不同地址语音文件名、机械手的移动距离和机械手转动的转角度数，在数据库管理状态下，用户可以输入、修改和显示数据库中的数据。

在经过准确的邮编区域定位、字符识别后，得到了识别出的 6 位邮政编码，在邮政编码数据库中查找该邮编所属的城市和区域，考虑到成本和系统组合的个性化，如果该邮政系统无执行机械设备，可以调用数据库里的地址对应语音文件并播放，工作人员可以通过语音提示来进行分拣工作，这样大大提高了工作效率，也可以减少错误分拣的概率，从而实现半自动分拣。如果使用包含径向角度和半径值控制的机械手进行全自动分拣，可以通过读取数据库里对应的移动距离和角度，由机械实现分拣。如：330013，所属城市为江西南昌，机械手将会把信封吸合并移动到 0°，1 500 mm 处。

组合个性化的系统或该设备安置在不同邮局时，当各区域信封分拣放置的位置需要改变时，可以通过修改数据库表单中的 MachineDistance 和 MachineAngle 的值来调整机械手的移动距离和角度。图 4 显示的

是邮政编码分拣系统的数据库表单的结构。

Table : L:\...\PostalCode.DB

PostalCode	PostalCode	DistrictName	Sound	MachineHandDistance	MachineHandAngle
1	063000	河北唐山	HebeiTangshan	1000	30
2	102800	河北廊坊	HebeiLangfang	1000	45
3	200021	上海卢湾	ShanghaiLuwan	1000	90
4	201401	上海奉贤	ShanghaiFengxian	1000	60
5	210000	江苏南京	JiangsuNanjing	2000	30
6	212000	江苏镇江	JiangsuZhenjiang	2000	0
7	214100	江苏无锡	JiangsuWuxi	2000	45
8	215000	江苏苏州	JiangsuSuzhou	2000	90
9	215300	江苏昆山	JiangsuKunshan	2000	60
10	223800	江苏宿迁	JiangsuSuqian	1500	30
11	223900	江苏泗洪	JiangsuSihong	1500	60
12	310000	浙江杭州	ZhejiangHangzhou	1500	90
13	330013	江西南昌	JiangxiNanchang	1500	0
14	474250	河南镇平	HenanZhenping	1000	75
15	516057	广东惠州	GuangdongHuizhou	1000	0

图4 邮政编码分拣系统的数据库表单的结构

6 结束语

本系统用颜色掩膜以及Hough变换的方法对邮编区域准确定位，通过对手写字符综合特征定义和应用，建立了较合理的手写数字模板库，大大提高了手写数字的识别率，通过数据库的合理应用，提高了程序设计效率。我们将进一步用Gabor特征等其它字符特征定义以提高系统识别率。

参考文献：

[1] Rafael C Gonzalez, Richard E Woods. 数字图像处理[M]. 北京：电子工业出版社，2007.
[2] 刘 文，刘永红，何友全. 一种邮政编码自动识别系统[J]. 西南交通大学学报，2003，38(1)：112－114.
[3] 谢凤英，赵丹培. Visual C++数字图像处理[M]. 北京：电子工业出版社，2008.
[4] 蒋先刚. 数字图像模式识别工程软件设计[M]. 北京：中国水利水电出版社，2008.
[5] 王忆锋. 红外图像灰度直方图统计分析的研究[J]. 红外，2009，30(4)：14－15.
[6] 范自柱，刘二根，徐保根. 互信息在图像检索中的应用[J]. 电子科技大学学报，2007，36(6)：1 311－1 312.
[7] 尹朝庆，宋 化，陈 波. 手写邮政编码的模糊识别方法[J]. 武汉理工大学学报，2005，27(2)：156.

Building and Training of Character Template Litrary Based on Synthetical Features

JIANG Xian-gang, ZHAO Ying, LI Lin

(School of Basic Sciences, East China Jiaotong University, Nanchang 330013, China)

Abstract: This paper describes the principles and software design technologies of automatic sorting postal code system. It proposes a solution to locating the region of a declined envelop and makes an integration program for selecting synthetically feature of character by using the information entropy theory. It establishes and trains a handwriting digital feature library of automatic sorting postal code system. It also designs a reasonable and related database system which greatly increases the recognition rate of the envelope and the feasibility of automatic sorting postal code system.

Key words: character features; postal code recognition; digital template library

(责任编辑：刘棉玲)