文章编号：1005-8451（2010）03-0012-05

基于统计结构模式的特种票据字符识别技术

林砺宗，周罗善
（华东理工大学 机械与动力工程学院，上海 200237）

摘 要：本文提出一种基于统计结构模式的多体印刷票据的字符识别方法。论述特种票据图像上字符的分割技术，以满足后续识别要求。重点介绍基于复杂指数和四边码的字符预分类技术和基于汉字特征点的汉字字符识别方法，同时针对提取特征点时存在的干扰点问题，提出一种基于笔画方向性的判断准则。实验证明，此识别技术处理速度快，不受环境噪声的影响，能够识别车票常用汉字。

关键词：统计结构模式；特种票据；汉字字符识别；汉字特征点；笔画方向性
中图分类号：U293.22 文献标识码：A

Character recognition technology of special tickets based on statistical and structure pattern

LIN Li-zong, ZHOU Luo-shan
(School of Mechanical and Power Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: A characters recognition method of multi-font printed tickers based on statistical and structure pattern was present in this paper. The segmentation technology of special tickets was briefly described, so as to satisfy with requires of post-recognition. It was introduced emphatically the character presort technology based on complex index and four-side code and Chinese characters recognizing method based on feature points. Meanwhile aimed to the problem which existed disturbed point in extracting features, the criteria on basis of stroke directionality was present. Experiments verified, this recognition technology had fast velocity of processing, could identify the staple Chinese characters in tickets, could not be effected by environment noises.

Key words: statistical structure pattern; special tickets; Chinese character recognition; feature points; stroke directionality

特种票据字符识别是一项新兴的研究热点[1]，其目的是为了让计算机能够“认票”，从而可对假票进行自动化鉴定。但是，特种票据种类繁多，而且字体种类多样。常见特种票据有火车票、汽车票、银行账单、汇票、明信片等，字符分为手写体和印刷体两种，同时包含汉字、拉丁字母、标点等多种字符，单一的模式识别方法或统计识别方法均难以取得满意的效果。

单一的统计模式虽然容易提取特征，抗干扰性强，但是对于字符种类的变化适应性差，对于相似字符不易识别；单一的结构模式存在提取特征难，抗干扰性差的缺点，不能满足高精度字符识别的要求，通过结合统计模式和结构模式可使识别性能提高，满足识别要求。

本文针对单印刷多字体火车票提出一种基于统计结构模式的特种票据字符识别方法。

收稿日期：2009-07-06
作者简介：林砺宗，教授；周罗善，在读硕士研究生。

1 票据识别系统简介

火车票通过摄像头摄取，以位图文件存入计算机，待后续处理。对于这类票据的识别，需将票据字符分割成单个字符，而后进行字符决策。分割技术包括：滤波处理；二值处理；行、字切割技术；归一化处理等。分割质量直接影响字符的特征提取和识别质量。图1为火车票识别流程图[2]。

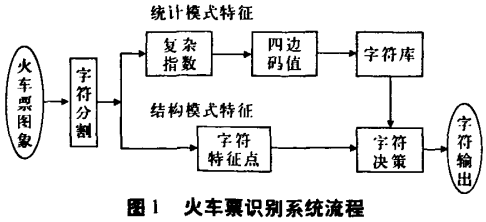

图1 火车票识别系统流程

为提高字符识别速度，需对字符分类，而字符分类需满足如下要求：

（1）正确分类率和分类稳定性高。

（2）分类速度快，且为主要目的。

（3）分类特性平坦，即每类文字数大致相等，充分发挥分类的功效。

（4）分类特征简单，各级分类方法要协调。

根据识别字符类型和方法不同，可选择合适的分类特征和方法达到理想效果。笔者采用基于统计模式的复杂指数和汉字四边码进行字符预分类，采用基于结构模式的汉字特征点作为识别特征，以欧式距离作为判断准则，进行字符最终决策，结合相应的编程软件，给出汉字字符分类和识别的实现过程。

2　字符分割

火车票图像输入过程中，各种内外因素导致火车票图像不整洁、有折痕或者出现断开等，给识别增加难度，车票是彩色图像，字符又是排列在一起，所以对车票字符的识别前，需对车票进行预处理，提高后续特征提取的质量。

直接识别彩色的火车票计算量大、速度慢、效率低，因此将彩色空间中的火车票转换到灰度空间中，如图2。

图2　火车票灰度图

为了解决灰度图中出现的断点和漏点等问题，需要对图像进行第1次平滑处理，可采用中值滤波法。完成第1次滤波处理后，采用如图3的灰度直方图可以对灰度图像进行二值处理，得到一张二值图像。对图像进行第2次平滑处理，消去二值图像上面的噪声和断点，本文采用数学形态学中先膨胀后腐蚀的闭运算对图像进行平滑，效果好，不会删去有用信息。

得到二值图像后就要对其进行行、字分割，目

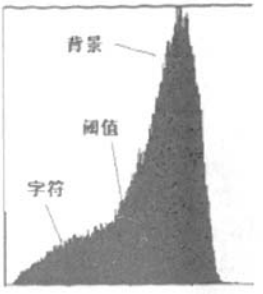

图3　灰度直方图

的就是能够将里面的字符分割[3]成为单独的一个字符图片。目前识别的是火车票中常用字，因此分割出车票中的“限乘当日当次票”这行字，并且将这行字进行字分割。行分割利用了横向空白间隙，采用横向投影的方法，将图片中的每一行表示出来，从图中我们观测到要分割的字符位于第6行，只需将第6行的字分割出来即可；然后应用平均字宽和竖直投影结合的方法进行字分割，如图4。

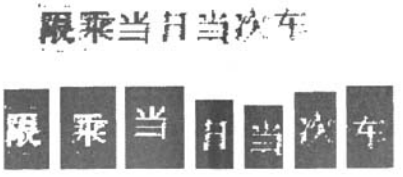

图4　通过行、字分割后的字符

分割出来的字符图片大小不一，会对后续的处理带来困难，必须进行归一化处理，包括位置归一化和大小归一化。为了使计算加快，将字符位置尽量朝图像的左上角靠，使字符大小成统一的64×64规格。操作是先求出分割后字符的重心坐标，通过计算同目标重心坐标的差值，将字符移动到目标位置，再通过基于外框的尺寸规范化方法，使任意大小字符的像素为64×64规格、位置统一的字符图片，这样可使识别时间开销减少，提高字符特征提取质量。

3　预分类

完成字符分割后的字符决策包括预分类和识别。在预分类处理中，本文重点叙述基于统计模式的复杂指数和四边码作为分类特征值。此判断方

法简单易行，抗干扰性能强，适应能力好。

3.1 复杂指数

复杂指数[4]是根据文字复杂指数的大小，将文字分为若干类的分类方法。目前的复杂指数是文字在单位质心二次矩中的笔划长度，反映了文字纵向或横向笔划的复杂程度。

文字的横向(x)和纵向(y)方向上的复杂指数定义为：

$$c_x = \frac{L_x}{\sigma_x} \tag{1}$$

$$c_y = \frac{L_y}{\sigma_y} \tag{2}$$

式中，c_x、c_y分别为横向和纵向复杂指数。

L_x、L_y分别为横向和纵向文字线段的总长度。

σ_x、σ_y分别为横向和纵向质心二次矩的平方根。

L_x、L_y可由文字2 bit × 2 bit网格中16种分布状态近似求出，而σ_x、σ_y定义为：

$$\sigma_x = \left[\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(i-G_I)^2 c(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}c(i,j)}\right]^{\frac{1}{2}} \tag{3}$$

$$\sigma_y = \left[\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(j-G_J)^2 c(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}c(i,j)}\right]^{\frac{1}{2}} \tag{4}$$

式中，G_I、G_J分别为文字质心位置的I、J坐标值，定义为：

$$G_I = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} i\cdot c(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}c(i,j)} \tag{5}$$

$$G_J = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} j\cdot c(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}c(i,j)} \tag{6}$$

前面所有式子中的N、M为文字点阵的长、宽，$c(i,j)$为像素值。

编制汉字字符横向和纵向复杂指数的计算程序，对于样本字体进行复杂指数计算，得到相应字体的复杂指数值，表1列出了部分样本汉字的复杂指数值。

表1 部分样本文字的复杂指数和四边码值

样本字	横向复杂指数	纵向复杂指数	四边码值	样本字	横向复杂指数	纵向复杂指数	四边码值
车	4	3	1111	有	3	3	1210
乘	3	3	2221	元	1	4	2120
上	5	3	1010	月	6	1	2013
当	3	3	1321	在	4	3	1121
到	3	4	1321	州	3	4	1211
等	4	3	1111	座	3	3	2211
日	5	1	1012	次	3	3	1121
海	3	3	2222	号	2	5	3210

3.2 字符框四边码

从文字点阵的四周边框起，向内取适当宽度，以此宽度分割出文字四周的4部分，根据每个部分内含有黑像素的多少分4级（0，1，2，3）编码，因而在进行四边码分类时，首先要确定用于分级的3个阈值。由于汉字字体类型多样，又分简体和繁体，所以在选择宽度和阈值时要根据不同的字体和尺寸来选择，本次编程基于先验知识确定向内宽度为字宽或字高的1/6，又根据字符图像64 × 64规格，从而选用的阈值为100、225、400，在编程软件中编制计算样本汉字的四边码值的程序，并计算样本的四边码值，如图5，为前面分割出的一个“当”字，其四边码为“1321”。

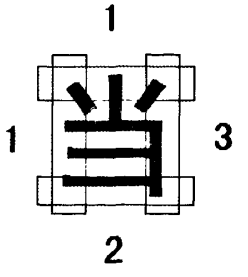

图5 文字四边码举例

通过对火车票常用字的复杂指数计算和四边码计算，发现复杂指数对于汉字位置、汉字尺寸的变换有较强的抗干扰力，稳定性能好，但是对于笔划粗细变化比较敏感。四边码对文字断线有较强的适应性，但对文字大小改变敏感，故在字符识别的预分类中，采用复杂指数和四边码值互补组合使用，分类效果好，优势突出。表1就是部分字符的复杂指数和四边码值，并以此建立字符分类库。

4　汉字字符决策[5~7]

汉字字符基本由直线笔画构成，是一种直线型字符。一个汉字的信息绝大部分集中在汉字骨架上，而汉字骨架信息又大多体现在若干特征点上，一旦确定笔画特征点，也就确定具体的汉字。汉字笔画特征点可分为端点、折点、岐点、交点，如图6。在进行汉字特征点提取前，先要对汉字点阵进行细化处理，得到汉字的骨架，再通过相应的算法寻找骨架中的特征点，提取特征点时考虑用8领域搜寻，假设值为“1”的点是汉字字符内部阵点。

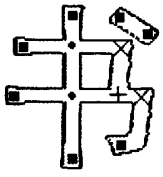

图6　特征点示意图

端点：笔画的起点或者终点且不与别的笔划相接，在端点的8领域中只存在一个值为“1”的点，通过这个点逐一查找端点。

折点：笔画方向变化显著的点，在8领域中可能存在两个点值为“1”，由于折点的种类多，笔者采用反求法，先寻找所有8领域内有两个点值为“1”的疑似折点，再根据情况判断是否为折点，如果满足情况就不是折点；否则认为该点是折点。

岐点：三叉点且其中两个笔画方向相同，因为是一个三叉点，所以这类点的8领域中存在3个点值为“1”，而且有两个笔画是同方向的，因此我们只要先找到那些在8领域内有3个点值为“1”的疑似岐点，再判断其中两点是否位于对称位置即可。

交点：四叉点且有两对相等的对顶角，交点的判断比较容易，只要一个点在其8领域内存在4个点值为“1”，并且这4点两两位于对称位置即可。

实际编程过程中，由于细化会产生一些噪声，将会影响点的判断，仅仅靠上面的判断方法不能够准确地判断各点的种类，特别是在判断交点、岐点和折点时，笔者在编制特征点提取程序时，把笔画方向性考虑到逻辑决策中，即当发现存在的四叉点、三叉点或二叉点时，还要判断沿笔画方向上是否存在值为“1”的点。如果存在，那么就确定这个点是特征点；否者就是干扰点，予以剔除。

例如对于交点8领域中的一种情况：由图7(a)得到1点应该是一个四叉点，即有可能是交点，再看1点的25领域（图7 (b)），如果不加任何改进的话，点2和点4就可能被计算机判断认为是岐点，但事实上这两个点不是特征点，因而在结果中就加上了干扰点。具体改进措施如下：

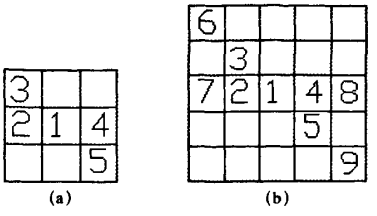

图7　判断举例

先寻找汉字骨架点阵中的疑似二叉点、三叉点或者四叉点，如果找到就做如下判断：在沿点阵方向上判断是否还存在点的值为“1”，例如图7的b图方向2-1-4上还存在点7和点8，方向3-1-5上还存在点6和点9，那么我们就可以判断点1确实是交点，而对于点2来讲，首先是一个三叉点，有一个是疑似岐点，但是在方向32上不存在其他点值为“1”的，所以点2不是一个岐点，而是干扰点，同理点4也不是岐点，这样就可区分特征点和干扰点，事实证明，这个改进措施确实有效，能够提高特征点提取的准确率。判断笔划方向上的点个数可以根据实际情况而定，取决于细化的质量，对于Hilditch细化算法而言，一个方向上判断存在1个~3个点就足够了。笔者以“当”字为例，编程时采用Hilditch细化算法和沿笔画方向上取两个点的方法，其细化图及特征点识别结果如图8。

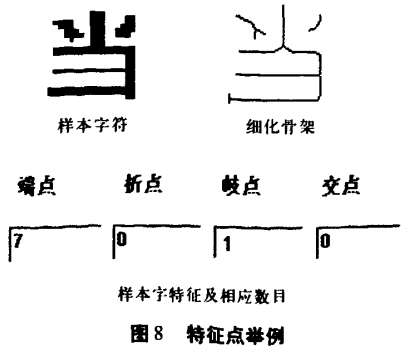

图8　特征点举例

汉字笔画特征点集中了主要的汉字结构信息，

端点和折点决定了一个汉字的笔画位置和形状；岐点和交点决定不同笔划间的相互连接关系，因而它可用在多体印刷汉字字符识别中作为字符决策特征，可以大大减少存储量，提高识别速度。

完成汉字字符特征提取后，进行汉字字符最终决策。对于字符的决策采用常用的明考夫斯基距离，即以距离作为判断准则：

$$D(X,G)=\left[\sum_{i=1}^{m}|x_i-g_i|^q\right]^{1/q} \quad (7)$$

在式（7）中，X表示输入未知汉字的特征向量，$X=(x_1,x_2,x_3,\cdots x_m)$；G为字库中某一个标准文字特征向量，$G=(g_1,g_2,g_3,\cdots g_m)$。

当q=2时，为欧式距离，记作：

$$D(X,G)=\sqrt{\sum_{i=1}^{m}|x_i-g_i|^2} \quad (8)$$

利用距离准则来判断时，当输入文字的特征向量X和字库中的某一标准文字的特征向量G_m相同时，即$D(X,G)=0$，但是由于输入的图像存在干扰或噪声，所计算的距离不可能是0，所以需要事先设定一个阈值δ，$\delta\geqslant 0$。分别计算输入文字特征向量X和某一分类集中的所有标准字符特征向量G_1，G_2，…，G_m之间的距离$D(X,G_1)$，$D(X,G_2)$，…$D(X,G_m)$，求出最小的值$D(X,G_i)$，若$D(X,G_j)\leqslant\delta$，即可判断出输入的未知文字。本文中预分类将识别字符划分到近似分类集中，汉字字符的4个特征点作为4个特征向量值，利用欧式距离判断准则，实现文字识别的目的，算法简单，编程易行，达到快速、精确识别，基本符合特征票据字符识别要求。

5 结束语

通过理论分析，结合编程软件，我们设计出了满足要求的字符识别系统。为了验证系统的功能性，我们采用火车票上经常使用的字符进行验证，如前文中提到的“限乘当日当次车”这句话中的6个不同字，并且使用不同字号和对图片加噪，得到如图9所示的结果。经过试验验证，得出如下结论：

（1）对于小号字体，字符识别率较低，这是由于小号字体在进行归一化运算时，尺寸归一化算法不合理造成，但是对于火车票的字体大小，一般

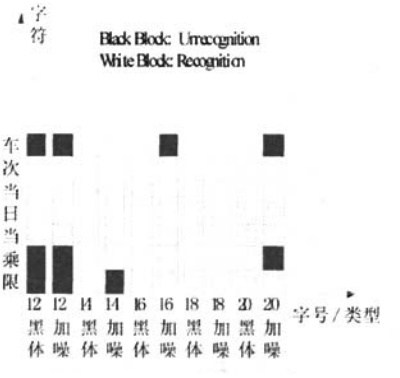

图9 验证结果数据图

在16号左右，图中可以看到识别较高，满足要求。

（2）加噪处理后的字符，通过中值滤波基本可以消除噪声，识别率仍然较高，从而说明这个系统的抗干扰能力强。

（3）鉴于火车票上有些字符组合是地名，可以考虑将地名自动联系识别算法应用到火车票字符识别中[8]，可以提高识别率。

（4）对于火车票上的非汉字字符的识别，可以考虑不采用二级识别方法，直接使用本身的结构快速识别。

参考文献：

[1] 陆发春，李晓辉. 残损文献的文字图像处理及识别技术[J]. 国家图书馆学刊，2003（3）：69-73.

[2] 张博洋，吴晓娟，张 青，等. 一种适用于特定票据字符识别的分割技术[J]. 山东大学学报，2004，34（6）：51-55.

[3] 张炘中. 汉字识别技术[M]. 北京：清华大学出版社，1992.

[4] 秦岐华，向旭宇. 汉字复杂指数特征提取技术的实现及其改进[J]. 计算机工程与设计，2006，27（2）：265-267.

[5] 王 恺，靳简明，史广顺，等. 基于特征点的汉字字体识别研究[J]. 电子与信息学报，2008，30（2）：272-276.

[6] 熊 伟，谢剑薇，曾 峦. 检测骨架图形特征点的新方法[J]. 红外和激光工程，2002，31（4）：301-304.

[7] Xinhua You, Bin Fang, Xinge You, etc. Skeleton Representation of Character Based on Multiscale Approach[C]. In CIS 2005 Part II LNAT 3802, 2005: 1060-1067.

[8] 谭红叶，郑家恒，刘开瑛，等. 中国地名自动识别系统的设计与实现[J]. 计算机工程，2002，28（8）：128-129，270.

[9] 陈兵旗，孙 明. Visual C++实用图像处理专业教程[M]. 北京：清华大学出版社，2004.