

文章编号:1001-9081(2006)10-2450-03

一种基于结构的票据定位新方法

卫志华, 苗夺谦, 夏富春, 张红云

(1. 同济大学 计算机科学与工程系, 上海 200092;
2. 国家高性能计算机工程中心同济分中心, 上海 200092)
(fuchunxia2003@citiz.net)

摘要:为了提高票据识别的准确率,对票据定位进行了研究。提出了一种基于结构的定位方法。它利用票据主边框上交叉点的个数和相对位置作为票据的特征信息,并将这些信息通过相应的分类相似函数集与已定义的模版进行匹配,从而实现对票据的分类,最终提取出当前票据的识别区域。实验结果表明:该算法取得了很好的定位效果。

关键词:定位; 特征提取; 相似函数

中图分类号: TP18 **文献标识码:**A

New structure-based bill location technique

WEI Zhi-hua, MIAO Duo-qian, XIA Fu-chun, ZHANG Hong-yun

(1. Department of computer Science and Engineering, Tongji University, Shanghai 200092, China;
2. Tongji Branch, National Engineering & Technology Center of High Performance Computer, Shanghai 200092, China)

Abstract: In order to enhance the accuracy in bill recognition, the bill location was studied. A new locating technique based on structure was proposed in this paper. It used the number and the relative position of intersecting points on the main margin of checks as the structural feature of the check, classified the bills with the similarity functions we defined, and extracted the recognition region of the bill at last. The experimental results demonstrate that this algorithm works well in bill location.

Key words: bill recognition; feature extraction; similarity function

0 引言

票据定位指的是从一个票据扫描图像(bmp 和 jpg 或 jpeg 格式居多)中,通过某种策略和方法寻找到用户感兴趣的区域,并从图像中提取出来,以图像文件的格式保存起来,作为票据定位后处理的输入。这些后处理包括二值化、边框处理、切割、识别等。在这一系列连续的处理过程中,票据定位非常重要。因为它不仅仅为后面的处理提供输入,更重要的是票据定位的效率和性能决定了整个票据识别的性能。

在参考别人定位算法的基础上^[1~5],通过观察票据自身的特点,发现不同种类的票据在结构上存在很大差别。这些差异主要体现在票据的大小、主边框上交叉点、主边框的宽度和高度等。在这基础上我们提出了一种基于结构的票据定位新方法。该方法主要由两个部分组成,一是票据的特征提取;二是根据特征集对票据分类。在特征提取的过程中,关键是在票据的众多信息中如何有效地选取票据的特征,并从当前票据中提取出来。在票据分类的过程中,关键是如何根据当前的票据特征集,依照某种匹配算法,最终将票据的识别区域提取出来。

该定位算法具体如下:

1) 对当前的票据进行一些常规的预处理。它包括票据摆正、初步二值化值选定等。

- 2) 提取当前票据的结构特征。
- 3) 根据当前票据的特征点信息,使用分类函数对票据进行分类,并提取票据的识别区域。
- 4) 对识别区域进行精确定位。

整个定位过程的处理流程如图 1。

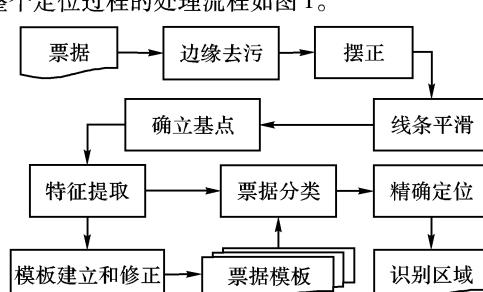


图 1 票据定位流程图

1 预处理

票据的预处理是票据识别前的一个规范化操作。从处理特性上分为常规处理和特殊处理。常规处理主要包括票据的摆正、边界的去污、滤噪音等。特殊处理指的是初步二值化值确定等一系列操作。

1.1 票据摆正

虽然尽可能在扫描时摆正票据,但还是不可避免的出现

收稿日期:2006-04-10; 修订日期:2006-06-20 基金项目:国家自然科学基金资助项目(60175016)

作者简介:卫志华(1979-),女,山西人,讲师,博士研究生,主要研究方向:智能信息处理; 苗夺谦(1964-),男,山西人,教授,博士生导师,主要研究方向:智能信息处理、主曲线、Rough Set、数据挖掘等; 夏富春(1982-),男,江苏人,硕士研究生,主要研究方向:Rough Set、图像处理; 张红云(1972-),女,江苏人,讲师,博士,主要研究方向:主曲线、Rough Set、数据挖掘、图像处理等。

票据扫描图像倾斜的情况。进行票据定位前要计算出票据大致的倾斜角度，并进行旋转。这样处理的目的是：降低由于票据倾斜所导致的特征点提取失败或识别区域提取不精确的概率，进而降低票据识别失败的可能性。

在票据最左边的边框上，依次采取多个有一定间距的点。一般任意选取两个点，求其连线的斜率得到倾斜角度^[6]。为了提高点采样间的独立性和计算的精确度，使用最小二乘法对采样点进行一阶拟合，求出其斜率，进而求出倾斜角度。

如果计算得到角度大于 θ_{\max} 时，拒绝定位和识别该票据，因为当前的票据在经过票据旋转后，失真较大，对后期的特征点提取和匹配影响较大。而对于角度小于 θ_{\min} 时，可以不进行票据旋转，因为此时票据倾斜对票据的定位和识别影响较小。

1.2 初步二值化值确定

确定初步二值化值的主要目的是将票据的背景和票据图像中的主边框分离。由于背景灰度值和边框的灰度值相差一般很大，因此此处二值化时所使用的分界值（阈值）并不是一个非常严格的值。可以简单的对当前票据进行灰度统计，选取与背景和边框相应的灰度峰值，使用调和方法得到一个合适的值作为背景和边框的分界值。

假定背景的灰度峰值为 G_{back} ，边框的灰度峰值为 G_{line} ，设背景的调和值为 δ_{back} ，边框的调和值为 δ_{line} ，那么二值化的分界值 G_{sep} 为：

$$G_{sep} = G_{back} * \delta_{back} + G_{line} * \delta_{line}$$

在票据图像中背景的峰宽比边框的峰宽来得宽，因此在选取 δ_{back} 和 δ_{line} 时，一般有 $\delta_{back} > \delta_{line}$ 。

2 特征提取

首先必须指出所谓票据的特征点主要指的是当前票据图像主边框上的交叉点，这其中不包含由于手写数字、汉字、英文等产生的与边框的交叉点；也不包含其他诸如由光学扫描仪噪声引起的交叉点。把那些有噪音或手写数字和文字等干扰形成的交叉点称为伪特征点。在这些伪特征点当中，存在一部分是可以通过算法判别并去除，这部分点称为可去伪特征点，其余的称为不可去伪特征点。

从这个角度上讲，交叉点分成三类：特征点、可去伪特征点和不可去伪特征点。在本文中，特征点理解为狭义上的特征点和不可去伪特征点。

2.1 交叉点提取算法

在基于结构的票据定位技术中，特征点提取的主要思想是：沿着当前票据主边框，按一定的方向，依次寻找所有的交叉点，判定并去除可去的伪特征点。具体的算法步骤如下：

1) 在当前的票据图形的左下角的某个足够大的区域($W \times H$)，使用预处理中确定的二值化值，分别向 X 和 Y 方向上投影，对黑色像素点个数进行统计。并把结果存放在两个数组 $X[W], Y[H]$ 中。

2) 对投影结果的两个数组 $X[W]$ 和 $Y[H]$ 进行中值平滑，平滑宽度为 l_w (Line Width 线宽)。

$$\begin{cases} X[i] = \sum_{j=-lw/2}^{lw/2} X[i+j]/lw \\ Y[k] = \sum_{j=-lw/2}^{lw/2} Y[k+j]/lw \end{cases}$$

3) 对平滑后的两个数组分别取峰值(最大值)，这样确定的坐标 (x_{\max}, y_{\max}) 为该区域的基点，也是整个票据图像的基点(参考点)。当然该坐标还需要修正到票据坐标中。

4) 从由3)确定的基点开始，按顺时针方向，沿着主边框寻找下一个交叉点，判断是否是可去伪特征点。如果不是伪特征点，则将其添加到特征点集合中。

5) 重复4)依次对主边框的左边、上边、右边、底边，直到所有的边都经过处理。

2.2 交叉点判定算法

根据统计，现实中票据的边框是有宽度，并且在交叉点处的边框可能有断点(如图2中的(g)和(h))。此外还有噪音的影响(如图2中的(d),(e),(f))。

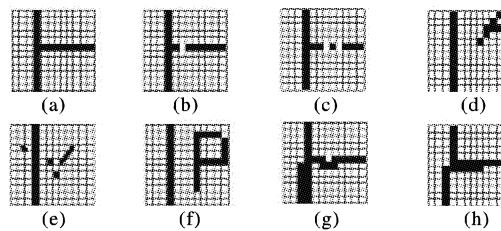


图2 票据图像中交叉点的各种情形

图2中(a)为理想交叉点，(b)和(c)为出现断点的交叉点，(d)、(e)、(f)为有噪音或手写英文等的交叉点，(g)和(h)是实际情形的交叉点。据此交叉点的判定算法的基本思想：在可能为交叉点 P 处，在票据的内侧(包括边框线)，沿法线方向选取连续的 $2N+1$ 点(N 一般比边框线宽稍大)。根据预处理确定的初步二值化值，将选取的点转化成0或1，并存储到数组 $Pix[2N+1]$ 中。定义一个数组 $Weight[2N+1]$ ，其中 $Weight[i]$ 表示判定 P 为交叉点时， $Pix[i]$ 的权值。

定义：

$V = Pix * Weight$, 其中 * 为数组相乘

交叉点的判定如下：

$$\begin{cases} V = 0 \rightarrow P \text{ 点为噪音点, 位于空白区域} \\ V < V_o \rightarrow P \text{ 点为交叉点} \\ \text{其他} \rightarrow P \text{ 点不是交叉点} \end{cases}$$

在判定规则中 V_o 为交叉点判定阈值。在此过程中交叉点判定权值数组非常重要，它往往决定了交叉点判定是否成功。它的设计规则如下：

- 1) 离边框线越近(包括边框上的点)，权值越小；
- 2) 离边框稍远的点，权值很大；
- 3) 离边框较远的点，权值非常小。

2.3 伪特征点去除算法

当寻找某个交叉点时，在边框法线方向上，边框内侧，以交叉点为中心，选取一个宽度为 r ，高度为 t 的区域。通过预处理确定的初步二值化值，把该区域转化成 $r \times t$ 的由0或1组成的数组 $A_{r \times t}$ 。

根据统计和分析，当交叉点为特征点时，数组 A 中至少含有一行和一列，行中有连续的长度为 t' 的“1”(黑色像素)，列中有连续的长度为 r' 的“1”(黑色像素)，并且相交。在此 t' 越接近 t ， r' 越接近 r ，判别的效果越好。

3 票据分类

票据分类是票据定位的核心。只有对票据进行正确分类，才能从票据识别模板中读取识别区域信息，并将识别区域从票据图像中提取出来。票据分类是票据自动识别中后续处理的前提。

票据分类的主要工作是根据提取出来特征点信息，匹配系统的票据模板。即在这些模板当中，寻找一个最佳匹配作为票据分类的结果，最后读取所匹配模板中的识别区域信息，

并根据此信息来提取当前票据的识别区域。

3.1 模板库构建

在对当前票据进行分类的时候,使用了系统的票据模板库。它是由许多票据模板组成的。每个模板中,除了包含匹配时所需要的特征点信息,还应该包含识别区域的信息(相对于基点的位置,识别区域宽度和高度),票据的名称等信息。这些信息都是在构建模板库的时候,从模板票据中获得的。

作为模板的票据要求:1) 空白票据,交叉点必须清晰且连贯。2) 背景和边框的比较清晰,区分度高。3) 待识别区域涂成深黑色方块。4) 扫描时尽量不要倾斜或折叠。

假设识别区域仅有一个。如下的方法可以获得识别区域信息:

- 1) 使用预处理确定的初步二值化值的 $1/N$ (N 为灰度参考比,此处设为 5) 对模板票据进行 Y 和 X 方向上的投影,统计结果保存到 $PX[Width]$, $PY[height]$ 两个数组中($width$ 和 $Height$ 为模板票据的宽度和高度)。

- 2) 对 PX 和 PY 数组使用宽度 $lw (= 3)$ 进行平滑(同特征提取中的 2)步),得到数组 PX' 和 PY' 。

- 3) 在 PX' 寻找一个宽度至少为 W_{min} 的波峰,并标记起始点 X_{start} 和宽度 W 。

- 4) 在 PY' 寻找一个宽度至少为 H_{min} 的波峰,并标记起始点 Y_{start} 和宽度 H 。

- 5) 由此得到在模板票据中存在一个矩形识别区域,它位于 (X_{start}, Y_{start}) ,宽度为 W ,高度为 H 。

- 6) 根据特征提取中找到的基点位置,求出点 $Q(X_{start}, Y_{start})$ 相对于基点的坐标 $M(X_{start}', Y_{start}')$ 。

宽度 W ,高度 H 和点 M 就是所求的识别区域信息。

说明投影时灰度值阈值越小越好,但不可以低于涂黑方块的灰度值,即 N 必需适中。

3.2 票据分类算法

根据特征提取中提取的特征点信息,票据匹配的算法如下:

- 1) 对当前票据某个边上的特征点集,分别与目标模板中相应边上特征点集进行匹配,并记录匹配点的个数。

- 2) 将边上匹配点数除以较大集合数,再与边匹配度 ew 相乘得到边相似度。

- 3) 依次对左边,上边,右边,和底边应用 1) 和 2) 后,将所有边的匹配度相加,这样得到该模板的匹配度。

- 4) 对系统中所有的模板应用 3),同时记录匹配度最大值和相应的模板。

最后得到最大匹配度对应的那个模板即为所求。

3.3 票据分类函数定义

如上的算法,从本质上就是票据匹配依次调用边匹配,而边匹配依次调用点匹配。下面从数学的角度定义匹配函数和结果函数。

定义 1 点相似函数。设点 $P1$ (当前票据), $P2$ (目标模板),这两个点的相似函数 $p(P1, P2)$ 定义为:

$$p(P1, P2) =$$

$$\begin{cases} 1 \leftarrow P1 \text{ 在以 } P2 \text{ 为中心的 } m \times n \text{ 的矩阵中} \\ 0 \leftarrow \text{其他} \end{cases}$$

其中 m 和 n 为点偏离参数。

定义 2 边相似函数。假设边 $E1$ (当前票据), $E2$ (目标模板),这两个边的相似函数 $e(E1, E2)$ 定义为:

$$e(E1, E2) =$$

$$\begin{cases} ew * \sum \max(p(P, Q)) \\ \max(\text{size}(E1), \text{size}(E2)) \mid p \in \text{set}(E1), Q \in \text{set}(E2) \end{cases}$$

其中 ew 表示目标模板中当前边的权值。

定义 3 模板相似函数。设 $M1$ (当前票据), $M2$ (目标模板),当前票据相似函数 $m(M1, M2)$ 定义为:

$$m(M1, M2) =$$

$$\left\{ \sum e(E, D) \mid E \in \text{edge}(M1), D \in \text{edge}(M2) \right\} / m(M1, M2)$$

$$m(M1, M2) = \text{onSameEdge}(E, D) \&\& \text{LikeSize}(M1, M2)$$

定义 4 (结果函数):

$$\text{result}(cur, Tmps) = \{\max(m(cur, tmp)) \mid tmp \in \text{set}(Tmps)\}$$

其中 cur 表示当前票据的特征点集合, $Tmps$ 表示系统模板集合。

对以上的四个函数作如下的说明:

- 1) 为了区分边上的点对模板匹配的影响不同,定义了数组 $ew[4]$, 分别表示各个边的影响权值。

- 2) 各个函数中使用的 $\text{set}(x)$ 表示 x 集合, $\text{size}(x)$ 表示集合中元素的个数, $\text{edge}(x)$ 表示 x 中的某个边, $\text{onSameEdge}(x, y)$ 表示 x 和 y 是否属于同一类型边(类型指的是左边,上边,右边和底边), $\text{LikeSize}(A, B)$ 表示票据 A 和 B 在主边框上相似, 即主边框在高度和宽度上分别几乎相等。

假设票据匹配度在达到峰值前,向当前票据特征点集合添加点为特征点,达到峰值后添加的点为不可去为特征点。事实上这个假设是成立的。根据边匹配函数,前者在未达到峰值前向特征点集合中添加伪特征点时,边相似函数值是不变的;而后者在特征点集合中添加特征点,要么是重复的,要么就是先前的特征点集合中有误判成特征点的伪特征点。根据如上的假设票据匹配度随着特征点集合中元素的个数增加的变化关系如图 3。

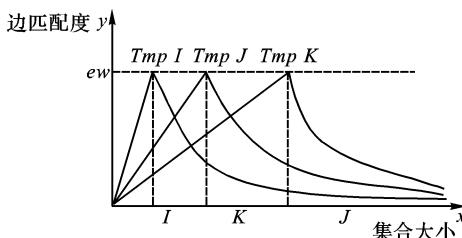
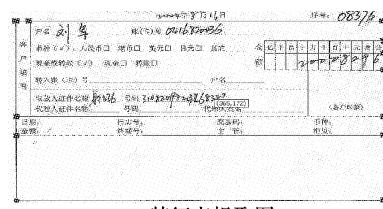
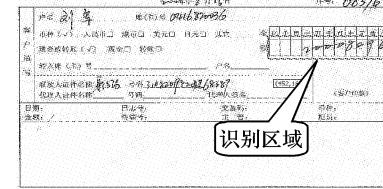


图 3 匹配函数和特征点集合大小变化关系

4 实验结果



(a) 特征点提取图



(b) 票据分类图

图 4 实验结果

本实验是在一个由 Java 开发实现的票据自动识别实验系统中进行的。该系统用 3860 张票据来验证本文提出的基
(下转第 2456 页)

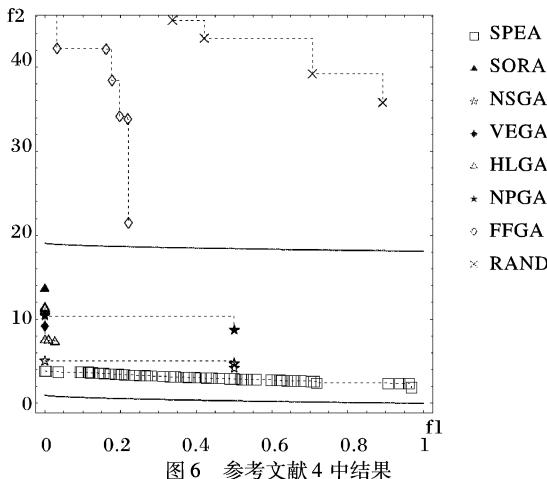


图 6 参考文献 4 中结果

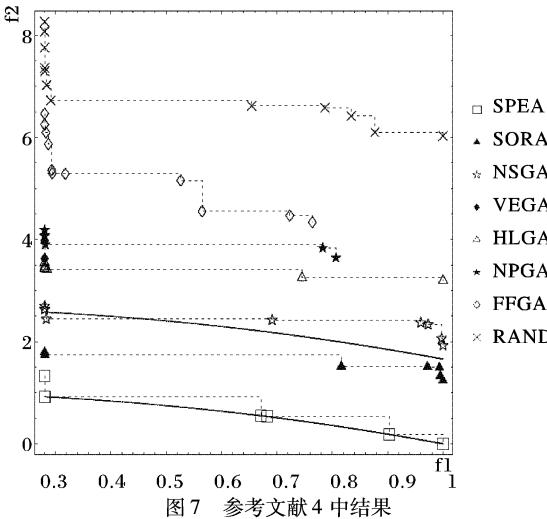


图 7 参考文献 4 中结果

4 结语

论文在分析多目标问题求解方法的基础上,针对 NSGA-II 在求解复杂函数时出现的参数恶化的现象,通过完善其排挤机制,同时,通过改进其交叉算子系数以及第一代子群体的产生方式,使得改进后的算法在逼近性和分布均匀性上都取

(上接第 2452 页)

于结构的票据定位技术。特征点提取和票据分类的实验结果如图 4,实验结果见表 1。

表 1 票据定位效果表

测试总数(T)	正确分类(C)	拒绝识别(R)	定位出错原因	
			扫描(S)	程序(P)
3 860	3 650	95	53	39

定义票据的定位准确率和可靠性:

$$\text{准确率} = \frac{C}{T} * 100\% ; \text{可靠性} = \frac{C+R}{T} * 100\%$$

计算得到:

$$\text{准确率} = \frac{3 650}{3 860} * 100\% = 95.15\% ;$$

$$\text{可靠性} = \frac{3 650 + 95}{3 860} * 100\% = 97.6\%$$

5 结语

本文中提出的基于结构的票据定位技术,利用票据主边框上交叉点的个数和相对位置作为票据分类的依据。这种票

得了很好的结果。这表明该算法对多目标问题求解做出了有益的探索,还可以在求解目标更多的问题以及工程实际中作进一步的研究。

参考文献:

- [1] DEB K, PRATAP A, AGRAWAL S, et al. A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II[R]. KanGAL Report No. 200001. India, 2000.
- [2] ZITZLER E. Evolutionary Algorithms for Multi-objective Optimization: Methods and Applications[D]. A dissertation submitted to the Swiss Federal Institute of Technology Zurich for the degree of Doctor of Technical Sciences. 1999.
- [3] SCHAFFER JD. Multiple objective optimization with vector evaluated genetic algorithms[A]. In Proceedings of an International Conference on Genetic Algorithms and Their Applications[C]. Pittsburgh, PA., 1995.
- [4] 黄邵. Web 服务组合的形式化建模与相容性分析[D]. 长沙: 湖南师范大学, 2006.
- [5] ZITZLER E, DEB K, THIELE L. Comparison of Multi-objective Evolutionary Algorithms: Empirical Results[J]. Evolutionary Computation, 2000, 8(2): 173 - 195.
- [6] HAJELA P, LIN C-Y. Genetic Search Strategies in Multi-criterion Optimal Design[J]. Optimization, 1992, (4): 99 - 107.
- [7] FONSECA CM. Multi-objective Genetic Algorithms with Application to Control Engineering Problems[D]. Department of Automatic Control and Systems Engineering, The University of Sheffield, UK, 1995.
- [8] HORN J, NAFPLIOTIS N, GOLDBERG DE. A Niched Pareto Genetic Algorithm for Multi-objective Optimization[A]. In Proceedings of the First IEEE Conference on Evolutionary Computation[C]. Piscataway, NJ. IEEE, 1994. 82 - 87.
- [9] DAVID A, VAN VELDHUIZEN, LAMONT GB. Multi-objective Evolutionary Algorithms: Analyzing the State-of-the-Art[J]. Evolutionary Computation, 2000, 8(2): 125 - 147.
- [10] SRINIVAS N, DEB K. Multi-Objective Function Optimization Using Non-dominated Sorting Genetic Algorithm[J]. Evolutionary Computation, 1995, 2(3): 221 - 248.

据定位方法与票据是否为手写或印刷无关。从这种意义上讲,该技术是非常通用。特别是在中国,票据种类非常多,并且使用频繁,这种基于结构的票据定位在票据自动识别上有较大的应用价值。

参考文献:

- [1] SMITHIES S, NOVINS K, ARVO J. A Handwriting-based Equation Editor[A]. Process of Graphic Interface(GL) 99[C]. June 1999. 84 - 90.
- [2] CHOU PA. Recognition of equations using a two-dimensional stochastic context-free grammar[J]. Proceedings SPIE Visual Communications and Image Processing IV, November 1989, 11: 852 - 863.
- [3] LAM L, LEE S-W, SUEN CY. Structural Classification and Relaxation Matching of Today Unconstrained Handwriting Zip-code Number [J]. Pattern Recognition, 1988, 1(21): 19 - 23.
- [4] 叶晨洲. 车辆牌照字符识别[J]. 上海交通大学学报, 2000, 5(34): 88 - 92.
- [5] 韩永强, 李世祥. 基于线处理的子图像定位算法[J]. 计算机与现代化, 1998, 2: 10 - 12.
- [6] 马洪波. 手写票据识别预处理算法研究[D]. 吉林大学, 2000.