第32卷 第13期 计 算 机 工 程 2006年7月
Vol.32 № 13 Computer Engineering July 2006
·博士论文· 文章编号：1000—3428(2006)13—0009—03 文献标识码：A 中图分类号：TP391.4

一种新的表格单元格矩形识别算法

陈优广，顾国庆，张 薇，许彦冰
(华东师范大学信息科学技术学院，上海 200062)

摘 要：现有的表格识别算法速度较慢，且仅能容忍表格线的微小断线，文章给出了基于顶点链编码的表格单元格矩形识别算法，利用边界标定自动机，标定表格单元格内环边界并生成顶点链编码，利用顶点链编码特性，有效地去除表格框线上的锯齿，修复断裂的框线，通过搜索单元格矩形4个角的顶点链编码来获得表格单元格的矩形区域。实验证明本算法具有速度快、鲁棒性高、抗表格框线断裂等优点。
关键词：顶点链编码；表格识别；边界标定自动机

A New Form Cell Rectangle Recognition Algorithm

CHEN Youguang, GU Guoqing, ZHANG Wei, XU Yanbing
(School of Information Science and Technology, East China Normal University, Shanghai 200062)

【Abstract】 The form recognition algorithms in existence are inefficient, and only can abide tiny broken lines. This paper presents an algorithm based on vertex chain code for form cell rectangle recognition, the algorithm uses region-labeling robot to label the inner border of a form cell to get its vertex chain code, using the characters of the vertex chain code, the algorithm can remove the sawteeth on the form frame line efficiently and restore the form frame lines and get the region of the form cell by searching the vertex chain code of the four angles of the cell. Experiments prove that the algorithm has the advantages of high speed, high robustness and being able to resist broken form frame lines.
【Key words】 Vertex chain code; Form recognition; Region-labeling robot

1 概述

表格文本分析与识别是计算机文档处理中的一个重要项目。在商业和政府机构等单位中表格扮演着重要的角色，表格是文档中常用的数据资料载体，大量的文档信息是以简明的信息表达方式——表格形式存在(如税务财务报表、数据处理)，且广泛应用于各种场合。因此表格分析和识别有着很大的研究和应用价值。

现在已经提出和发展了很多分析和识别表格矩形块的方法[3～6]。大多表格矩形块识别算法是先检测表格框线，在提取的表格框线的基础上获得表格矩形块，也有采用细化后再取特征点的方法等。通常表格分析方法有投影法、搜索法、细化等。投影法对表格图像纵、横向进行投影，根据得到的投影值中的峰值变化来判断表格线。这种方法很难处理表格线较细且稍有歪斜或复杂表格。搜索法对毛刺、断线和线与栏目中字符粘连则很难处理。表格图像细化的方法通常在交点、折点等处产生畸变，对断线干扰抵抗力差。表格框线的检测算法中，Hough变换是一种较为成熟的检测直线的算法，它已演变出了很多快速算法。Hough变换作为一种全局的检测方法，有利于检测虚线和断裂的直线。矢量化算法是另一类应用较广的直线检测算法，但这两种方法运算量较大，且运算速度较慢。

在日常生活中，人们常常需要填写大量图文表格，且又要保留图文表格上的原样，这时只需要对表格单元格或表格局部进行识别，又考虑到算法速度和计算机实时识别要求，上面提到的表格识别算法就很难适用。本文给出了一种新的表格单元格矩形的识别算法，其中包括表格单元格内环边界的标定、基于顶点链编码的边界光滑和修复处理、寻找单元格角点特征点。实验结果表明本算法具有速度快、抗断裂和对表格进行实时识别的优点。

2 顶点链编码和表格图像的预处理

2.1 顶点链编码

Bribiesca提出了用边界像素的顶点来标记图像的方法[1]。对于正四边形点阵上的图像，可以有3种不同性质的顶点，如图1(a)所示，分别用代码1，2，3来标记。沿着图像边界像素的顶点行走一周，依次记录图像边界像素顶点的代码，所有这些代码构成的序列就是图像边界的顶点链编码。

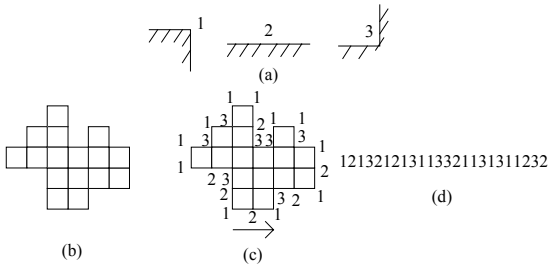

图1 图像的顶点链编码

图1中，(a)表示3种不同性质的顶点；(b)表示正方形点阵上的图像；(c)表示顶点链编码元素示例；(d)表示图像边界的顶点链编码。

2.2 表格图像的预处理

表格图像的预处理主要包括图像的二值化、倾斜矫正和表格最外层边框的修复。倾斜矫正可用文献[8]中提供的算法来实现。修复最外层边框可通过投影的方法来实现，通过投

作者简介：陈优广(1971—)，男，博士生、讲师，主研方向：图像处理与模式识别；顾国庆，研究员、博导；张 薇，博士生；许彦冰，硕士生
收稿日期：2005-08-02 **E-mail**：ygchen@cc.ecnu.edu.cn

影获得表格的最小外接矩形，把最小外接矩形的 4 条边上的像素点设定为黑色，使得表格是一封闭区域。本文后面的算法是在经过预处理后的表格上进行的。

3 标定单元格内环边界

利用文献[7]提供的边界标定自动机可以标定单元格内环边界并获得边界的顶点链编码。自动机从单元格内的一白色像素出发，按一定方向行走，当自动机走到黑色像素边缘，则标定黑色区域边界，判定标定的边界是否是单元格内环边界，若不是，自动机继续按给定的方向行走，直到找到并标定了单元格内环边界为止。

以自动机起始点为原点建立坐标系，当自动机沿图像区域边界行走时，用 $\Delta\delta$ 记录象限变化时的角度增量，设其初期值为 0。当自动机沿边界曲线从第 1 象限移动到第 2 象限、从第 2 象限移动到第 3 象限、从第 3 象限移动到第 4 象限或从第 4 象限移动到第 1 象限时，$\Delta\delta$ 增加 90°，即 $\Delta\delta=\Delta\delta+90^\circ$；若是相反方向，那么 $\Delta\delta=\Delta\delta-90^\circ$。因图像区域边界是一连通曲线，自动机沿图像边界行走一周时，有 $\Delta\delta=0^\circ$ 或 $\Delta\delta=\pm360^\circ$。当 $\Delta\delta=\pm360^\circ$ 时，则认为该边界包含坐标原点，否则边界不包含坐标原点，如图 2 所示。

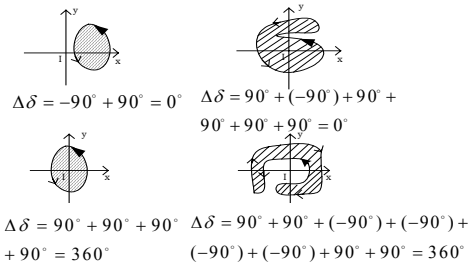

图 2 $\Delta\delta$ 变化示意图

规则 当 $\Delta\delta=\pm360^\circ$ 且顶点链编码的长度大于给定的阈值时，则认为标定的边界为单元格内环边界，否则标定的边界不是目标单元格内环边界。

自动机沿单元格内环行走一圈的同时，通过自动机的输出，可以获得单元格内环边界的顶点链编码。对理想状态下的表格，即表格框线是光滑且没有断裂和与字符粘连的情况，假设从矩形一角点出发记录其顶点链编码，则其顶点链编码为 $32^a32^b32^a32^b$，如图 3 所示。

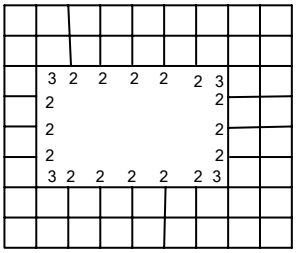

图 3 理想矩形内环的顶点链编码

对于理想表格，其内环边界的顶点链编码中编码为 3 的个数有且仅有 4 个，并且对应于矩形的 4 个顶点，在顶点链编码中寻找链码为 3 的顶点就确定了矩形区域。但是大多表格线在预处理后，会产生表格线断裂，且表格框线是不光滑的，线的边缘有许多毛刺和凹陷。当表格线不光滑且有断裂时，自动机标定单元格内环时就会走出目标单元格，标定的区域除目标单元格外也标定了其他相邻的单元格，且获得的链编码中除链码 2 和 4 个角的顶点链码 3 外，还有很多的链码 3 和 1。为确定矩形 4 个角的位置，必须对标定区域的顶点链编码进行分析，包括平滑框线和找出断裂位置并修复。本文后面称表格线断裂处为断笔。

4 表格框线的平滑

分析框线上的毛刺和凹陷处的顶点链编码，构成毛刺和凹陷的子链是 32^a1，12^b3，1133，3311 这 4 种基本子链构成，a 和 b 决定了毛刺或凹陷的长度，一般情况 $a,b\le3$。利用顶点链编码平滑框线就是去除链编码中的对应毛刺和凹陷的子链，通过遍历顶点链编码，容易删除链中形如 32^a1、12^b3、1133、3311 的子链。根据顶点连编码的特性，上面的子链删除，不会删除矩形角点的链码 3 和断裂处的链码 1。这里称经过平滑处理后的顶点链编码为修正顶点链编码。

5 断笔的修复

5.1 定义特征点

为了便于描述，这里称表格框线断裂处为断笔。分析边框矩形发生的断裂情况，可以归结为如下 3 种类型：

(1) 边框线在中间断开，形成两个对等的孤立端点，如图 4(a)所示。

(2) 边框线在边框矩形的顶点处断笔，形成一个孤立端点，如图 4(b)所示。

(3) 断裂处将原来 90°度直角点变成了一个孤立端点和一个-90°直角点，如图 4(c)所示。

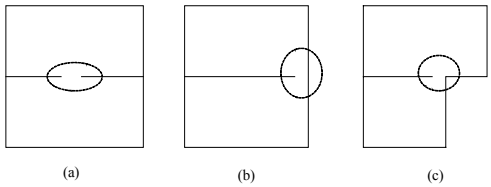

图 4 边框线的断笔情况

根据上面的断笔类型，定义 3 种特征点类型分别为 n_{90}、n_{180} 和 n_{-90}，如图 5 所示。

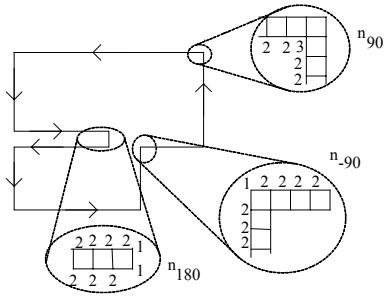

图 5 3 种特征点位置及其对应修正顶点链编码的子链形式

显然边框矩形的 4 个顶点一定是特征点。用 $Count(S)$ 表示特征点 S 的个数，那么可验证 3 种特征点存在下面的关系：

$$Cout\ (n_{90})=2Count\ (n_{180})+Count\ (n_{-90})+4$$

由顶点链编码的特点，在特征点 n_{90} 附近的子链为 2^m32^n，在特征点 n_{180} 附近的子链为 $2^m12^a12^n$；在特征点 n_{-90} 附近的子链为 2^m12^n，其中 m，n，a 为正整数，并且 a 的值很小(一般为框线的宽度大小)。

5.2 断笔处特征点的分布类型

根据断笔类型，断笔处必有 n_{180} 特征点，用 P 表示断笔处 n_{180} 特征点，如图 6 所示，(其中 P 所在的边用阴影表示)，在断笔处有一条边或一个特征点与 P 相对应，这里称之为对等边或对等点，有以下 4 种类型：

类型 1 P 与另一个 n_{180} 特征点形成对等点，二者所在的曲线相互平行且方向相反，是因边框线在中间断开造成的。如图 6(a)所示。

类型 2 P 与另一个 n_{180} 特征点形成对等点，二者所在的曲线相互垂直且方向相对，是因边框矩形在顶点处断开造成的，断笔连接的方法与二者所在的曲线在标定区域中的位置有关系。如果标定区域在两者夹角的内部，则这两个端点消失，重新构建出一个边框矩形的顶点，如图 6(b)所示；如果标定区域在两者夹角的外部，则这两个特征点消失，重新构建出一个 n_{-90} 的特征点，并且这个 n_{-90} 的特征点的断笔类型需要进一步判断。如图 6(c)所示。

类型 3 P 所在的边与另一个 n_{180} 特征点所在的边垂直，形成对等边，这种情况是因边框矩形的两条垂直边都出现断裂造成的。断笔连接后 P 消失，重新构建出一个边框矩形的顶点。这种情况只能去掉一个 n_{180} 的特征点，另一个 n_{180} 的特征点的断笔类型需要进一步进行判断。如图 6(d)所示。

类型 4 P 所在的边与边框矩形的一条边垂直，形成对等边，这种情况是因边框线在端点处断开造成的。断笔连接后，P 消失，重新构建出一个边框矩形的顶点。如图 6(e) 所示。

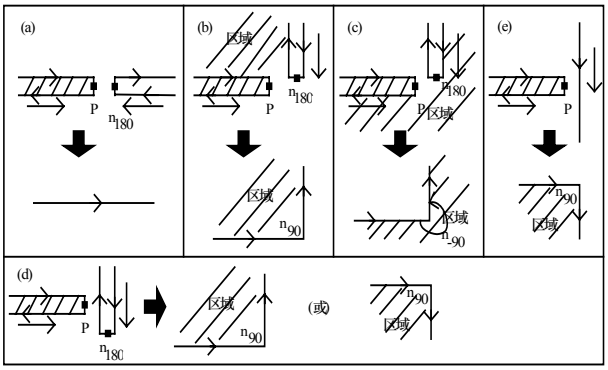

图 6 断笔处特征点的分布类型

根据 P 与相应的对等点近，还是与对等边近，可以判定 P 点所属的类型。遍历整个顶点链编码，计算所有与 P 能构成断笔类型的对等点或对等边的欧几里德距离，并求出这些距离中的最小值 $D_{\min}$，则与 P 的距离为 $D_{\min}$ 的对等点或对等边构成的类型就是断笔 P 处的断笔类型。

5.3 断笔的连接方法

断笔连接并不是对断开的边框线真正连起来，而是为找到目标单元格的 4 个角特征点对修正顶点链编码作的处理。下面根据断笔处特征点的分布类型给出相应的连接方法。

类型 1 在修正的顶点链编码中，把 P 所对应的子链开始到与它对等的 n_{180} 特征点所对应的子链为止的链编码段删除。如图 7(a)所示。

类型 2 求出二者所在边的交点，在修正的顶点链编码中，把 P 所对应的子链开始到与它对等的 n_{180} 特征点所对应的子链止链编码段用“3”或“1”代替，“3”或“1”对应于求出的交点，如图 7(b)、图 7 (c)所示。

类型 3 求出 P 点所在的边与另一个 n_{180} 的特征点所在的边交点，在修正的顶点链编码中，把 P 所对应的子链起到另一个 n_{180} 特征点所对应的子链的链编码段之前用“3”来代替，“3”对应于求出的交点。如图 7(d)所示。

类型 4 求出 P 点所在的边与边框矩形的一条边的交点。称 P 对应的子链为子链 1，称与它对等的边所对应的形如 32^m3 的子链为子链 2。在修正的顶点链编码中，如果子链 1 在子链 2 的后面，那么将从子链 2 的后一个“3”一直到子链 1 为止的链编码段用“3”代替。“3”对应于求出的交点。如图 7(e)所示。如果子链 1 在子链 2 的前面，那么从子链 1 起直到子链 2 中的前一个“3”为止的链编码段用“3”来代替。如图 7(f)所示。图 7 中的表格框线用直线表示，且只列出了特征点处的编码。

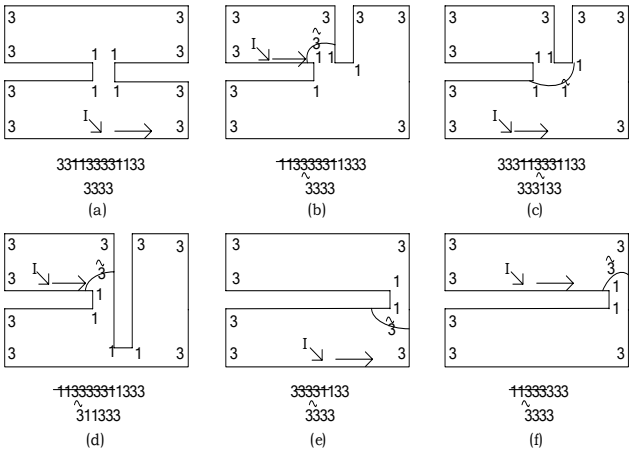

图 7 断笔连接方法

6 确定单元格矩形的角点特征点

循环遍历修正顶点链编码，当遇到 n_{180} 的特征点时，由 5.2 节提供的方法确定其分布类型，若没有与之相符的类型，该特征点为虚假的特征点，直接将它相应子链从修正顶点链编码中删去。若存在与之相符的类型，利用 5.3 节中介绍的方法进行相应的断笔连接，直到顶点链编码中只剩下 4 个“3”为止结束。

另外，由于目标单元格中有缺口，自动机可能对两个或两个以上单元格进行区域标定，对标定边界的顶点链编码进行修正处理后，识别出单元格矩形可能是错误的，对这种情况，需要对修正顶点链编码进行旋转处理，处理方法是使位于初始点右下角并与初始点距离最近的特征点作为顶点链编码的起始点，对修正顶点链编码经过旋转处理后，再利用上面介绍的方法可正确识别出目标单元格的边框矩形。

7 实验结果与结束语

用本文的算法对共 32 张实线表格的单元格进行识别，其中总单元格数为 3 623，识错的单元格数为 194，正确率为 94.6%，其中一些表格人为加上一定的噪音和断线。除去文字与表格线粘连的单元格，正确率在 99%以上。该算法不用对表格框线进行检测，利用边界标定自动机跟踪表格单元格的内环，生成顶点链编码，利用顶点链编码的特性，平滑框线并修复断裂的框线，通过寻找角点特征点的方法来识别表格单元格矩形。实验证明本算法具有算法速度快、识别率高的优点，在表格图像有一定的噪声和表格线断裂的情况下，也能很好地识别。分析识别错误的单元格，发现造成识别错误的原因有两个：一是当表格线断裂很严重或表格线是虚线时，自动机只能跟踪部分单元格内环边界，而很大部分单元格内环边界没有标定到；二是文字与表格线有粘连。对这两种情况，本文的算法还不能很好地识别，现在的扫描仪或数码相机获得的图像经过分割，一般不会出现上面的这两种情况，本算法基本上满足了应用，利用本文提供算法已成功开发了基于图像的填表软件。 （下转第 14 页）