一种表格框线检测和字线分离算法 [1]

刘长松 潘世言 郑冶枫 丁晓青

(清华大学电子工程系智能技术与系统国家重点实验室 北京 100084)

摘 要 该文提出了一种基于有向单连通链的表格框线检测算法，能够合理地利用单连通链边沿的全局统计特性和单连通链之间的局部位置关系，精确地提取表格框线，具有抗倾斜，抗断裂，抗字线交叠等优点。在此基础上，提出了一种能够分离交叠字线的表格框线去除算法。并成功应用于实际的表格识别系统中。

关键词 表格识别，图像分析，直线检测，字符识别

中图号 TP391

1 引 言

表格作为一种高度结构化的特殊文档，被广泛地应用在国民经济和日常生活的各个方面。表格的计算机自动识别是文档智能处理领域的一个重要组成部分 [1]。

表格图像与一般的文本图像最大的区别是以表格框线作为分割表格单元的主要依据。因此，对表格框线的正确提取是划分表格单元的基础，是表格识别最关键的环节之一。

提取框线实际上是一个直线检测的问题。直线检测算法较为成熟的理论是 Hough 变换 [2,3]，以及围绕此理论衍生出的众多的快速算法。但它在具体的工程实践中的应用却受到以下几个不利因素的限制：(1) 运算量大；(2) 只适合于检测直线而得不到端点；(3) 判决门限难以确定。不可能找到一个适用于所有图像的统一门限，而对不同的应用选取各自合理的门限又是一个相当棘手的自适应问题。

如果假设表格线都在水平或垂直方向附近，可以通过缩小角度搜索范围来减少运算量，但以牺牲斜线检测为代价，另外，它也并未解决 Hough 变换的上述 (2)、(3) 的限制。而且在检测某些长度较短但对表格域分割起重要作用的表格竖线时很容易被文字信息淹没而造成漏检。

其它有代表性的表格框线检测算法还有连通域分析法 [4] 和交叉点特征法 [5−7] 等。这些方法在满足各自的约束条件下能够取得好的效果，但对表格线断裂、倾斜等情况难以适应。

我们构造了一种称为有向单连通链 (DSCC, Directional Single-Connected Chain) 的图像结构作为线检测的基元，它具有定义简单，物理意义明确，易于存储和处理等优点。在一定约束条件下合并有向单连通链，我们可以快速准确地提取直线。Hough 变换注重全局信息而没有利用局部信息，这种算法能够合理利用局部和全局的图像信息，具有抗倾斜，抗断裂，抗字线交叠等特点。

为了消除表格框线对表格域字符分割和识别的影响，在识别前必须将检测到的框线从表格图像中去掉。如果待识字符笔画与框线交叠，还必须采用特殊的办法将二者分离，并保证交叠处的字符笔画形状不产生过大的畸变。本文提出基于 DSCC 算法字线分离方法。

2 有向单连通链的定义

对应于横线和竖线，有向单连通链分为横向单连通链和纵向单连通链两种，分别用于检测横线 (包括倾斜角小于 45° 的斜线) 和竖线 (包括倾斜角大于 45° 的斜线)。以横向单连通链 (C_h) 为例：C_h 为图像游程序列 $\{R_1 R_2 \cdots R_m\}$，序列中每一个游程项 R_i 都是横向宽度为一个像

[1] 2000-10-08 收到，2001-06-14 定稿
国家 863 计划及国家自然科学基金资助

素，纵向由连续的黑像素段形成的游程 (如图 1)，记为：$R_i(x_i, ys_i, ye_i) = \{(x,y)|\forall p(x,y)=1, x=x_i, y\in[ys_i, ye_i]$ 且 $p(x_i, ys-1)=p(x_i, ye+1)=0\}$。其中 $p(x,y)$ 代表坐标 (x,y) 处的像素值，1 代表黑 (前景) 像素点，0 代表白 (背景) 像素点；x_i, ys_i 和 ye_i 分别表示游程 R_i 的 x 坐标，起始 y 坐标和终止 y 坐标；C_h 中的各个 R_i 在 x 方向 (横向) 上排列成一个序列，且序列中任意相邻的两个游程 R_i 和 R_{i+1} 横向单连通，即除了 C_h 两端的游程 R_l 和 R_m 外，任何 R_i 的两侧都有且仅有一个游程与其连通。对于 R_l 的左侧和 R_m 的右侧，要么不存在任何连通游程 (如 R_{13} 的右侧)，要么存在一个以上的连通游程 (如 R_1 的左侧)，要么虽然只有一个连通游程，但这个连通游程同时还与处于 R_l 或 R_m 同一列的其它游程连通 (如 R_9)。

纵向单连通链 (C_v) 的定义与 C_h 非常相似，不再赘述。

通过对输入图像进行有向单连通链的提取，可以得到大量的单连通链。每一小段有向单连通链都可能是某条框线的一部分，也可能是文字等其它对象的一部分。我们进一步的目标是找到能组合成直线的单连通链的集合，从而得到线段的精确位置。

3 表格线检测

实际的表格图像中，每根表格线都是由排列成一直线且相互之间没有交叠的若干有向单连通链组成。通过合并这些有向单连通链，就可以最终得到表格框线。为了选择参与合并的单连通链，我们定义了有向单连通链之间的“同线距离”。两个单连通链的形状及相对空间位置越接近一条直线，二者的同线距离就越小。具体的定义如下 (以横向单连通链为例，如图 2)：

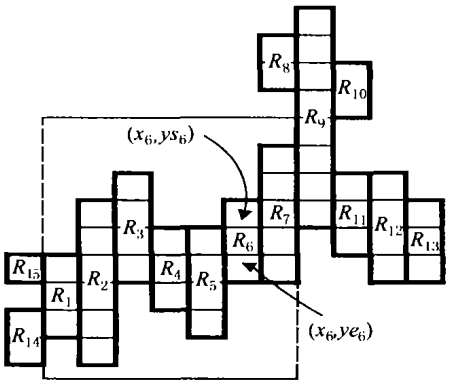


图 1　横向单连通链示意图

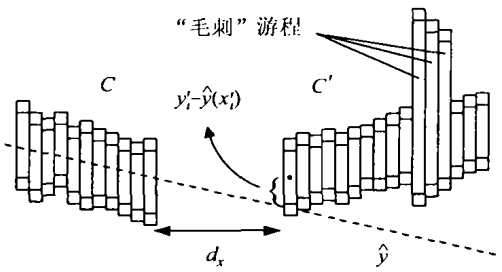


图 2　横向单连通链的同线距离

假定已获得横向单连通链 $C=\{R_1R_2\cdots R_n\}$ 的中心点拟合曲线 $\hat{y}(x)$。另一横向单连通链 $C'=\{R_1'R_2'\cdots R_m'\}$ 到 C 的“同线距离”定义为 (假设 C' 中参与计算的有效游程数目为 m 个)：

$$d_{C'C}=\begin{cases}\infty, & \text{当} d_x\le 0\text{时}\\ d_x+\dfrac{\sum_{i=1}^{m}(y_i'-\hat{y}(x_i'))^2}{m}, & \text{当} d_x>0\text{时}\end{cases}$$

其中 y_i' 是第 i 个游程的中点纵坐标，$d_x=\max(x_1,x_1')-\min(x_n,x_m')$。若 $d_x\le 0$，表示 C 和 C' 在纵向存在交叠部分，此时 C 和 C' 不可能属于同一条直线，所以设定其距离为无穷大。若 $d_x>0$，d_x 的数值代表 C 和 C' 内侧两个端点游程的横向距离，$d_{C'C}$ 和式中的第 2 项代表

C' 各中心点到 C 延长线的均方误差。这一项越小， C 和 C' 越有可能处在同一条直线上。我们采用最小二乘拟合法延伸 C 。只有长度小于两倍游程平均长度的游程才作为“有效游程”，参与拟合，这样可以排除“毛刺游程”的干扰。

若 C' 可以合并入 C ，它必须同时满足以下两个合并准则：

(1) 线性延伸条件： $\sqrt{d_{C'C}-d_x}<W$ ， W 为 C 的平均宽度；

(2) 间隙条件：考查位于 C 和 C' 内侧两个端点之间，长度为 d_x ，宽度为 W 的图像区域，可能出现以下 3 类情况：

(a) 空白。设定门限 T_1(实验中取经验值 $T_1=15$)，若 $d_x\le T_1$ ，我们认为空白是表格线的正常断裂， C 和 C' 仍属于同一直线，应合并；若 $d_x>T_1$ 则说明 C 和 C' 相距过远，所以不合并为一条直线。

(b) 存在其他单连通链，其宽度小于两倍 C 的宽度。处理方法同情况 (a) 。

(c) 存在其他单连通链，其宽度大于两倍 C 的宽度。此时 C 和 C' 之间存在直线或字符笔划。设定一个较小的门限 T_2(实验中取经验值 $T_2=8$)，若 $d_x\le T_2$ ，合并 C 和 C' ，否则不合并。

合并算法的第一步是选定一条合适的单连通链 C_s 作为“种子链”。首先在 C_s 的某个单侧寻找距离 C_s 最近的一系列 C'_i ，然后按同线距离从小到大的顺序依次判定是否满足上述合并条件。若找到可以合并的 C'_k ，则将 C_s 和 C'_k 中的所有有效游程 $R_i(i=1,2,\cdots,n)$ 和 $R'_j(j=1,2,\cdots,m)$ 放在一起，做最小二乘拟合，继续进行搜索和合并。处理完一侧，处理另一侧，直到 C_s 的两侧都找不到可以合并的 C' 为止。从剩余的所有未经合并的单连通链中选取新的初始“种子链”，用同样的方法可以检测出其它直线。重复上述过程，直到再也无法找到合适的初始“种子”链为止。

4 表格框线的去除和交迭文字的自动保留

表格识别过程中的另一个重要任务是将已检测出的框线从原始表格图像中去除。去除框线的难点在于当字符笔画和表格框线相交叠时如何在去除框线的同时完整地保留字符笔画。

文献报道的很少几种字线分离算法 [4,8] 都有局限性，而且它们都没有和各自的表格框线检测算法结合起来。

由于有向单连通链对表格线的描述提供了相当丰富的信息，充分利用这些信息，我们提出与框线检测算法紧密结合的新的字线分离方法。

我们将字线相交方式分成 3 大类：

第 1 类：笔画与直线的相交角度较大。如图 3 。

这种情况的图像特点是笔画部分构成的单连通链夹在其两侧的框线单连通链之间并且与它们都连通，但形状和边沿轨迹则与框线单连通链有明显的差异。在单连通链的合并过程中，这样的笔画单连通链一般不能满足合并的线性延伸条件，因而合并算法会自动“跳过”它并直接与它另一侧的框线单连通链合并 (例如图 3 中“种子”链 C_s 向右延伸时跳过 C_1 和 C_2 而直接与 C_3 合并，原因是 C_1 和 C_2 的纵向宽度过大)，这样，无需再作任何后续的分析处理，笔画就已经与框线自动分离开来。这种情况在实际的字线相交中占绝大多数。

第 2 类：笔画与直线的相交角度较小。如图 4 。

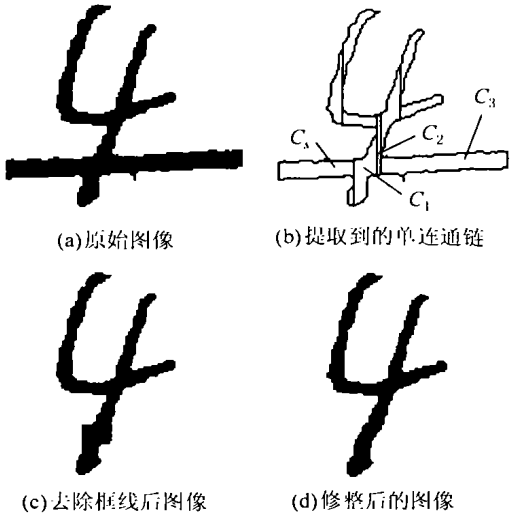


图 3　字符笔画与直线相交角度较大时的字线分离

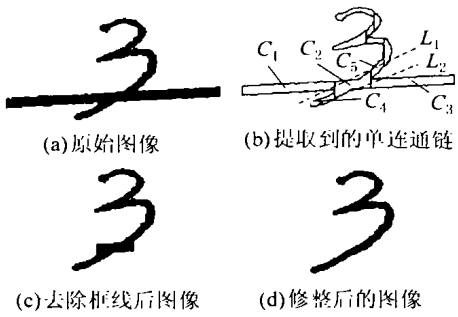


图 4　字符笔画和直线相交角度较小时的字线分离

这种情况下，笔画单连通链能够满足合并的线性延伸条件从而被并入直线中，但它们与周围的框线单连通链还是存在不同之处．以图 4(b) 为例，图中的 $\{C_1C_2C_3\}$ 是某直线的一段，$\{C_4C_2C_5\}$ 是笔画轨迹．容易看出 C_2 是笔画与直线相交形成的单连通链．相交单连通链有以下几个特点：

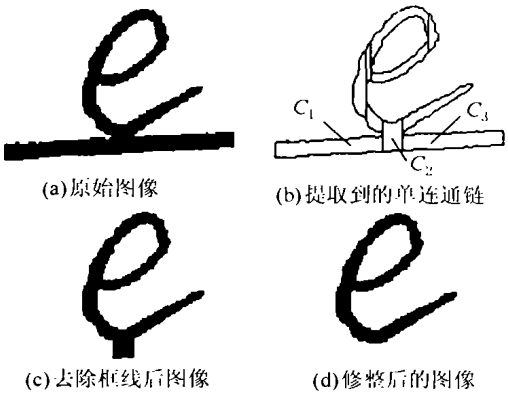
(1) 相对其两侧的单连通链，相交单连通链本身的长度一般都较短．

(2) 由于相交角度小，在其两侧除了被合并的框线单连通链 (C_1 和 C_3) 以外，还可以找到明显的笔画“切入”及“切出”单连通链对 (C_4 和 C_5)．

(3) 当笔画的“切入”及“切出”单连通链分别位于框线的上下两侧时 (即笔画穿越框线的情形)，与相交单连通链“距离”最近的单连通链有可能是“切入”或“切出”单连通链，而不是被合并的框线单连通链．例如在图 4 中，虽然合并运算时与 C_2 “距离”最近的的确是 C_1 和 C_3，但它们并不是 C_2 本身轨迹的延伸．它们之所以能够与 C_2 合并是受当时已经合并的所有其它框线单连通链共同作用的结果．如果我们用单独的 C_2 作“种子”链重新计算，就会发现此时与其“距离”最近的单连通链是 C_4 或 C_5．这样，框线合并与单独合并所选取的最近“距离”单连通链产生了不一致，以此就可判断 C_2 是相交单连通链而不是框线单连通链．如果被检框线的单连通链串中存在某一个同时具有以上 3 个特征，我们可以比较有把握地认定它是相交单连通链，并在框线去除过程中加以保留．

第 3 类是笔画与直线几乎相切．

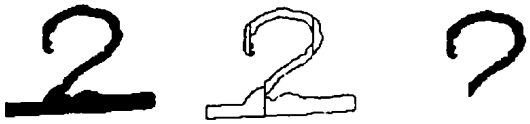
这里又分成两种情况：第 1 种情况是字线相切部分的长度较短且笔画的大部分仍没有融入框线图像内，如图 5．这时相切部分的单连通链 (C_2) 的宽度大约是笔画粗细和框线粗细之和，在框线合并过程中由于不满足线性延伸条件而被“跳过”．第 2 种情况是字线相切部分的长度较长且笔画的大部分已经融入框线图像内，如图 6．这时，上文讨论的第 2 类字线相交方式中的 3 个条件只能部分满足或全都不能满足，此时，依靠单纯的图像几何特征已不足以完成可靠的判决，仍需要寻找更有效的解决方法．



(a)原始图像 (b)提取到的单连通链

(c)去除框线后图像 (d)修整后的图像

图 5 字符笔画和框线相切但相切部分较短时的字线分离



(a)原始图像 (b)提取到的单连通链 (c)去除框线后图像

图 6 字符笔画完全融入框线时难以完成字线分离

5 字线分离后对字符笔画的修整

观察图 3(c), 图 4(c) 和图 5(c)，我们发现保留下来的字线相交单连通链仍然保留有框线的痕迹，这些畸变会给字符识别带来不利影响，必需去除.

框线单连通链被去除之后，相交单连通链两侧只剩下笔画单连通链，且与之相连通. 例如图 4(b) 中当框线单连通链 C_1 和 C_3 被去除之后，相交单连通链 C_2 和其两侧原来的笔画单连通链 C_4 和 C_5 合并成为一个大的单连通链 (如图 4(c))．我们把它放大显示在图 7 中.

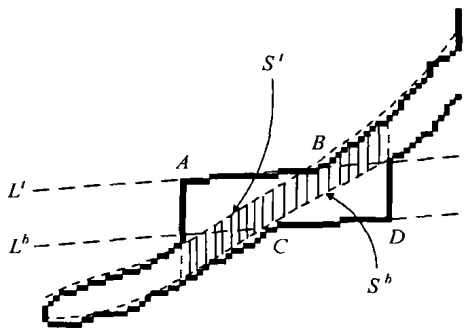


图 7 对字线分离后的新笔画单连通链边沿的修整

为了去除直角突起，我们首先要找到这个大单连通链的上下边沿中不属于笔画边沿的点，即图 7 中 AB 和 CD 之间的点. 一般情况下，这些点就是那些原来框线连通链延伸出来的点，它们的特点是与框线边沿的拟合直线的距离很小. 据此，我们就可以将它们与笔画边沿点区分开来. 具体方法是：计算大单连通链上边沿的每一个点到被去除的框线上边沿拟合直线的垂直距离，如果该距离大于框线上边沿拟合方差的两倍且该点位于框线上边沿的上方，则保留，否则就去除该点. 对所有被保留下来的点进行 3 次多项式拟合，得到笔画上边沿的拟合曲线 S^t．用类似的方法可以得到笔画下边沿拟合曲线 S^b.

获得笔画上下边沿的拟合曲线 S^t 和 S^b 之后，我们就可以去除相交单连通链 (C_2) 中位于 S^t 和 S^b 之外的黑像素点，而保留了如图 7 中影线部分所示的笔画部分。而对相交单连通链两侧的原笔画单连通链 (C_4 及 C_5) 则不需要作任何改动。

6 实验结果

我们把本文提出的算法用于增值税发票表格识别。

图 8 是一张经过表格线检测并去除表格线的实际对比图像。

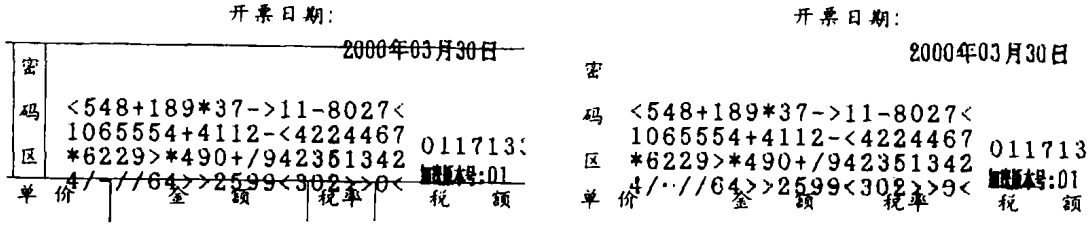


图 8　增值税发票的去线结果比较

在增值税发票识别过程中，表格线检测为我们提供了定位信息。表格线去除后，对字符分割有很大帮助，并对于打印到表格线上的文字识别有了极大的提高。我们对 1016 幅实际增值税发票样本密码区的识别率进行统计，如果不去除表格线，共有 58 幅图像密码区识别有错误，去除表格线以后，只有 31 幅图像密码区有识别错误。

除增值税发票以外，基于本文算法的表格识别系统已经成功应用于清华文通表格识别系统中，广泛应用于工商、交通、税务、统计等行业和部门中，基本能够适应实际应用中的各种表格。具有准确、可靠、抗干扰能力强等特点。

7 结　论

本文提出了一种全新的表格框线检测算法。我们首先定义了一种称为“有向单连通链”的图像结构，在此基础上给出了一种基于自定义“同线距离”的有向单连通链合并算法。

本文还在线检测算法的基础上提出了去除表格框线的方法。详细分析研究了字线交叠情况下的字线分离和字符笔画修整算法。实验和应用表明新的框线检测算法具有抗倾斜，抗断裂，抗字线交叠等优点，基于它的框线去除方法能够在很大程度上解决字线交叠的分离问题。

参　考　文　献

[1] Yuan Y. Tang *et al.*, Automatic document processing: A survey, Pattern Recognition, 1996, 29(12), 1931–1952.

[2] J. Illingworth, J. Kittler, A survey of the Hough transform, Computer Vision, Graphics and Image Processing, 1988, 44(1), 87–116.

[3] Mark C. K. Yang, *et al.*, Hough transform modified by line connectivity and line thickness, IEEE Trans. on PAMI, 1997, 19(8), 905–910.

[4] Bin Yu, Anil K. Jain, A generic system for form dropout, IEEE Trans. on PAMI, 1996, 18(11), 1127–1134.

[5] 刘今晖，印刷表格自动输入数据库的研究与实现，硕士学位论文，清华大学，1992.

[6] Liu Wenyin, Dov Dori, From raster to vectors: Extracting visual information from line drawings, Pattern Analysis and Applications, 1999, 2(2), 10–21.

[7] Chun-Ta Ho, Ling-Hwei Chen, A high-speed algorithm for line detection, Pattern Recognition Letters, 1996, 17(5), 467–473.

[8] Jin-Yong Yoo, *et al.*, Line removal and restoration of handwritten characters on the form documents, Proc. 4th International Conference on Document Analysis and Recognition, Ulm, Germany, 1997, 128–131.

A FRAME LINE DETECTION AND REMOVAL ALGORITHM FOR FORM DOCUMENT RECOGNITION

Liu Changsong Pan Shiyan Zheng Yefeng Ding Xiaoqing

(*State Key Lab. of Intell. Tech. & Sys., Dept. of EE, Tsinghua Univ., Beijing* 100084, *China*)

Abstract A new frame line detection algorithm based on the structural image element—Directional Single-Connected Chain (DSCC) is proposed. Taking advantages of the global statistical property of the edges of the DSCCs, and their local mutual relations, the algorithm is able to accurately extract frame lines from scanned form images. It demonstrates the desired performance of insensitive to line slant, breaks as well as touches from character strokes inside the form cells. Based on this algorithm, a frame line removal approach is presented, by which the frame line can be removed without affecting the touched character strokes.

Key words Form recognition, Image analysis, Line detection, Character recognition

刘长松： 男，1969 年生，讲师，研究方向：文本图像处理，模式识别，计算语言学.
潘世言： 男，1973 年生，硕士生，研究方向：表格处理与识别.
郑冶枫： 男，1975 年生，硕士生，研究方向：表格处理与识别.
丁晓青： 女，1939 年生，教授，研究方向：数字图像处理，模式识别，智能信息处理.