基于有向单连通链的表格框线检测算法*

郑冶枫, 刘长松, 丁晓青, 潘世言

(清华大学 电子工程系,北京 100084)

E-mail: zhengyf@cfar.umd.edu; lcs@ee.tsinghua.edu.cn; dingxq@tsinghua.edu.cn

http://ocrserv.ee.tsinghua.edu.cn

摘要: 表格框线检测是表格识别的基础.现有的表格框线检测算法或者速度慢,或者鲁棒性差,而且没有充分利用表格框线之间的约束信息.提出了一种基于所定义的图像结构基元"有向单连通链"的自底向上表格框线检测算法.在此算法中,有向单连通链是一种黑像素游程序列,作为非常合适的矢量基元,在引入一定表格框线约束信息的条件下合并单连通链,有效地去除伪框线,补全断裂的框线,提高了算法的鲁棒性,可以准确而快速地提取表格框线.通过滤除噪声单连通链,加快单连通链的合并速度,算法速度提高了 3~10 倍,满足了实用要求.实验证明,该算法具有速度较快、鲁棒性高、抗任意角度的倾斜、抗断裂等优点.

关 键 词: 表格识别;图像分析;直线检测;OCR(光学字符识别);智能文档处理

中图法分类号: TP391 文献标识码: A

表格是一种很常见的文档形式.它作为一种高度精炼、集中的信息表达手段,以其简明、规范、便于填写和处理等特点,被广泛地应用在国民经济和日常生活的各个方面.表格的自动输入、存储、管理已经成为文档智能处理领域的一个重要组成部分.

表格由一些有一定约束关系的横线、竖线和少量的斜线组成.为了构成表格单元,直线之间存在相互约束关系.我们称表格中这种相互之间存在约束关系的直线为表格框线,以区别一般的直线.直线检测是图像分析领域中最基本的、不断研究探讨的问题之一.其中较为成熟的算法是 Hough 变换以及繁多的快速算法[1].虽然 Hough 变换作为一种全局的检测方法,对线段的连通性没有要求,有利于检测虚线和断裂的直线.但由于难以确定直线的起点和终点,运算量过大,它在具体的工程实践中的应用却受到了限制.表格中的框线绝大多数集中在水平和垂直两个方向,这提示我们可以将Hough变换中(ρ,θ)空间的θ分量的搜索范围大大地减小,从而大幅度地减少运算量.这种特殊的 Hough 变换等效于实际中经常使用的投影算法[2].但投影法不能提取斜线,而且抗图像倾斜的能力有限,当图像出现较大角度(大于 5°)的倾斜时,算法就会失效.

矢量化算法(vectorization)是另一类应用较广的直线检测算法[3~5].直接对光栅图像的各个像素进行处理,存储量大,而且因为不能利用像素间的位置关系,很不方便.而矢量化过程作为目标识别的预处理过程,将输入的光栅图像转化成矢量基元(比如直线段、圆弧段等等).它一方面使处理对象由像素变成矢量基元,数目下降一个数量级,另一方面选择合适的矢量基元可以使后续的目标识别过程转化成较简单的矢量基元的生长、合并过程,难度大大降低.因为矢量基元的选择决定了目标检测算法的性能,所以它必须容易提取,大小合适,反映待检测目标的最本质的特性.我们构造了一种称为"有向单连通链"的图像结构作为直线检测的矢量化基元,它具有定义简单,物理意义明确,易于检测、存储和处理等优点.在一定约束条件下合并有向单连通链,可以快速、准确

* 收稿日期: 2000-05-11; 修改日期: 2000-10-09

基金项目: 国家自然科学基金资助项目(69972024); 863 高科技发展计划基金资助项目(863-306-ZT03-03-1)

作者简介: 郑冶枫(1975–),男,浙江江山人,硕士,主要研究领域为文本图像处理;刘长松(1969–),男,山东文登人,讲师,主要研究领域为图像处理,模式识别,智能信息处理;丁晓青(1939–),女,江苏睢宁人,教授,博士生导师,主要研究领域为图像处理,模式识别,智能图文信息处理;潘世言(1973–),男,安徽桐城人,博士生,主要研究领域为图像处理,模式识别.

地提取直线.单连通链的合并结果还有少量的错误.一类是字符笔划的误合并,即存在“伪”直线;一类是直线断裂.表格框线约束信息的引入可以帮助去除伪直线,补全断裂的直线.我们称这种引入表格框线约束信息的直线检测算法为表格框线检测算法.

本文第 1 节给出有向单连通链的定义.第 2 节讨论基于有向单连通链的框线检测算法.第 3 节讨论算法的加速问题,加速后我们的算法的速度与投影法的速度相当.最后,我们将通过实验,验证本算法的有效性.实验表明,我们提出的基于有向单连通链的表格框线检测算法具有速度较快、抗任意角度的倾斜、抗断裂等特点.

1 有向单连通链的定义

分别对应于横线和竖线,有向单连通链也分为横向单连通链和纵向单连通链两种.横向单连通链用于检测横线和倾斜角度小于 45°的斜线;纵向单连通链用于检测竖线和倾斜角度大于 45°的斜线.以横向单连通链为例:横向单连通链 C_h 为图像游程序列 $\overrightarrow{R_1R_2...R_m}$.序列中每一个游程项 R_i 都是横向宽度为一个像素、纵向由连续的黑像素段形成的游程(如图 1 所示),记为

$$R_i(x_i, ys_i, ye_i) = \{(x, y) \mid p(x, y) = 1, x = x_i, y \in [ys_i, ye_i], p(x_i, ys - 1) = p(x_i, ye + 1) = 0\}.$$

其中 $p(x,y)$代表坐标(x,y)处的像素值,1 代表黑像素点,0 代表白像素点;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_1 和 R_m 以外,任何 R_i 的两侧都有且仅有一个游程与其连通.对 R_1 的右侧和 R_m 的左侧也是如此.但对于 R_1 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 1 个以上的连通游程(如 R_1 的左侧有 R_{15} 和 R_{14} 同时与之连通),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_1 或 R_m 同一列的其他游程连通(如 R_9).

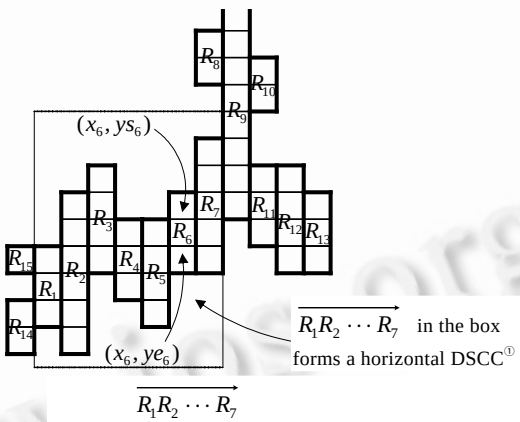


Fig.1 Horizontal DSCC

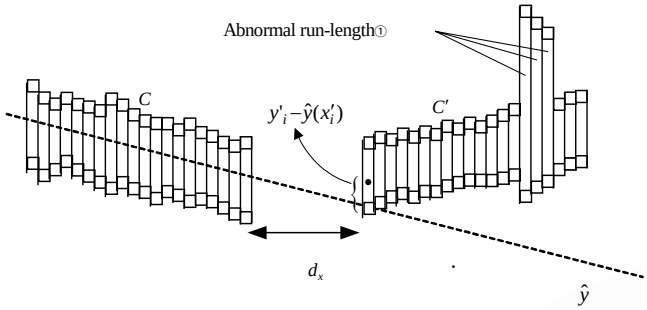
图 1 横向单连通链示意图

纵向单连通链 C_v 的定义与横向单连通链非常相似,对其详细的定义此处不再赘述.

2 表格框线检测

2.1 有向单连通链的合并

在实际的表格图像中,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 所示):



①“毛刺”游程.

Fig.2 Co-Linear distance of two horizontal DSCCs

图 2 两横向单连通链的同线距离

假定已获得横向单连通链 $C=\overrightarrow{R_1R_2...R_n}$ 的中心点拟合 $\hat{y}\ (x)$,则另一横向单连通链 $C'=\overrightarrow{R'_1R'_2...R'_m}$ 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 N 个):

$$d_{C'C}=\begin{cases}\infty, & \text{当 } d_X\le 0 \text{ 时 (即 } C \text{ 和 } C' \text{ 有交叠时),}\\ d_x+\dfrac{\sum_{i=1,R'_i\in B}^{m}\left(y'_i-\hat{y}(x'_i)\right)^2}{N}, & \text{当 } d_X>0 \text{ 时.}\end{cases}$$

其中 $d_x=\max(x_1,x'_1)-\min(x_n,x'_m)$.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 $d_x>0$,则表示 C 和 C'在纵向没有交叠,d_x 的数值代表 C 和 C'内侧两个端点游程的横向距离.此时 $d_{CC'}$ 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,表明 C'越贴近 C 的延伸部分,即 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.式中 B 表示有效游程集合.

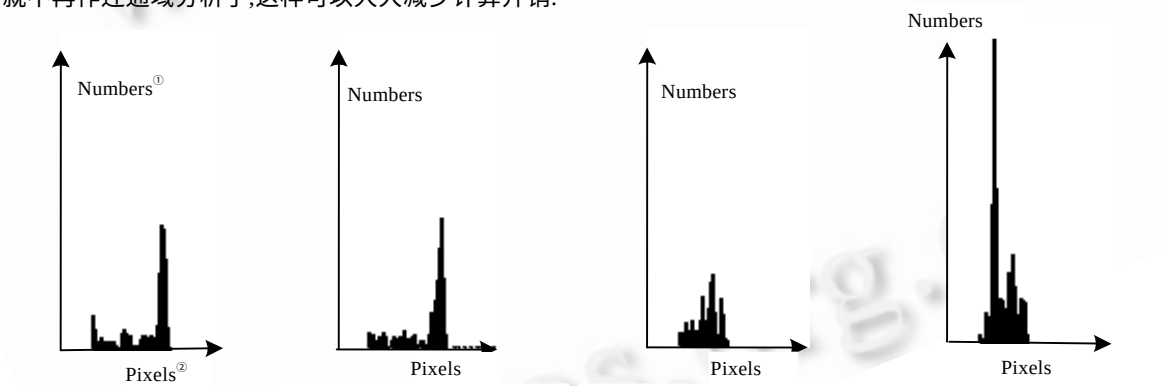
若 C'可以合并入 C,它必须同时满足以下两个合并准则:

1. 线性延伸条件:$\sqrt{d_{C'C}-d_x}<W$,W 为单连通链 C 的平均宽度;
2. 间隙条件:考察位于 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 3 类情况:
   (1) 空白.设定门限 T_1(实验中 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,不应再视为一条直线,所以不合并.
   (2) 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 1.
   (3) 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 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 1 步是选定一条合适的单连通链 C_s 作为“种子链”(我们选择有效游程最多的作为“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 $R_i(i=1,2,...,n)$和 $R_j'(j=1,2,...,m)$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,再处理另一侧,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他直线.重复上述过程,直到再也无法找到合适的初始“种子”链为止.由于合并前单连通链比较短,其统计特性不够稳定,为了防止发生不可弥补的误合并,我们在第 1 次合并时,门限设得比较小.经过初始合并后,连续性比较好的直线就可以完整地提取出来,而断裂比较严重的直线被识别成若干条线段.接着,我们加入字符尺寸的信息,将门限放宽到等于字符尺寸,进行第 2 次合并.经过第 2 次合并后,我们滤除小于字符宽度的横线和小于字符高度的竖线,排除单个字符笔划产生的直线.

2.2 字符尺寸的估计

在直线检测算法中需要设定最短直线长度门限,大于该门限的直线保留,小于该门限的直线被认为是字符笔划而滤除.最短直线长度门限实际代表了字符的尺寸.由于表格的不同,扫描分辨率的不同,字符尺寸变化很大,从十几个像素到上百个像素.很多文献中该门限或者设定为一个固定值[2],或者作为一个参数需要用户输入[3,5].若能自动估计字符的尺寸,则可以提高直线检测算法的自适应能力.我们提出一种基于连通域分析的自动估计方法.利用在生成单连通链时提取的黑像素游程,对这些游程作连通域分析,就可以统计得到连通域宽度和高度的直方图,如图 3 所示.对于单一字号汉字占多数的表格,直方图只出现一个明显的峰,该峰即对应字符的尺寸.而在实际表格中,汉字、数字、英文常常同时出现,不同的字号也同时出现.直方图中出现多个峰,此时我们取高度大于一定门限的最大峰作为字符尺寸的估计.图 3 是两个表格样张的连通域宽度和高度直方图.表格样张 1 只有单一字号的汉字,在直方图中形成一个非常明显的峰,如图 3(a)和图 3(b)所示,我们取最高的峰作为字符尺寸的估计.表格样张 2 中同时存在汉字和数字,数字的宽度是汉字的一半左右,其连通域宽度的直方图比较分散,形成多个明显的峰,如图 3(c)所示.我们取最右边的峰作为字符尺寸的估计.对于左右和上下结构的汉字,理论上必须将各个连通域合并后才能估计出字符的尺寸.这种相邻连通域的合并一方面计算量大,另一方面比较困难.它相当于作字符切分,而中英文混排的字符切分是比较困难的,在图像粘连严重时就更难了.幸好,实验结果表明,这种连通域合并是不需要的.以左右结构的汉字为例,其各个连通域在高度直方图中有助于形成正确的峰,而在宽度直方图中,会造成一些干扰,但在统计足够多的字符后,这种干扰不会引起字符尺寸估计错误.我们的连通域分析利用了已经提取的黑像素游程,比从像素级上开始作连通域分析快很多(实验结果表明可以快 5~10 倍).通过实验,我们发现,统计 100 个连通域就可以得到相当精确的结果,所以我们统计完 100 个连通域后就不再作连通域分析了,这样可以大大减少计算开销.



(a) Connected component width histogram of form sample 1 (a) 表格样张 1 的连通域宽度直方图

(b) Connected component height histogram of form sample 1 (b) 表格样张 1 的连通域高度直方图

(c) Connected component width histogram of form sample 2 (c) 表格样张 2 的连通域宽度直方图

(d) Connected component height histogram of form sample 2 (d) 表格样张 2 的连通域高度直方图

Fig.3 Connected component width and height histograms

图 3 连通域的宽度和高度直方图

2.3 伪直线的去除和断裂框线的补全

经过两次单连通链合并,大部分直线都准确地提取出来了,但还存在两类错误,一类是由字符笔划误合并产生的“伪”直线,另一类是直线的断裂.若不引入表格框线之间的约束信息,我们很难进一步得到准确的直线信息.为了利用表格框线之间的约束信息,我们先利用所提取的直线搜索表格单元[3].由于存在直线的断裂,几个表格单元可能合并成一个单元,如图 4 所示,合并成 ACEF 单元.若单元内的线段或位于同一直线上的线段组合的长度大于单元尺寸的 4/5,就将该单元分解为两个表格单元.执行这样的分解,直到再也不能分解下去为止.比如图 4 中线段 GH,IJ,KL,MD 长度大于 0.8 倍 AC,可以将表格单元 ACEF 分割成 ACDG 和 GDEF 两个单元.线段 BN,OP 又可以进一步将单元 ACDG 分割成 ABPG 和 BCDP 两个表格单元.

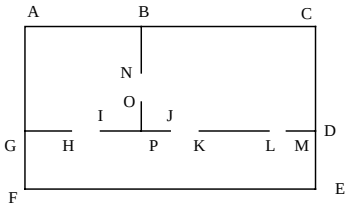


Fig.4 Cell with broken lines

图 4 含有断裂框线的表格单元

引入表格框线的约束信息后,所有未用于组成表格单元的直线都被认为是伪直线而滤除掉,同时断裂很严重的直线也可以得到一定程度的补全.

3 算法的加速

上述基于单连通链的直线检测算法比 Hough 变换快一个数量级,但比起倾斜投影法[2]要慢 5~10 倍.单连通链的数目巨大是造成算法速度较慢的原因.一张典型的 1000×1000 像素的表格,单连通链的数目在 5~6 万.我们采取了一些措施非常有效地提高了算法的速度,同时保持了算法的鲁棒性.

3.1 游程平滑

在提取黑像素游程时,若两个游程距离小于等于两个像素或两个长游程(大于 15 个像素)的距离小于等于 5 个像素,则合并这两个游程.如图 5 所示,由于印刷和扫描的原因,直线内部出现很多孔.在作游程平滑后,直线中的缺陷得到弥补.实验结果表明,游程平滑一方面可以减少单连通链的数目(减少 20%),另一方面可以提高算法抗直线断裂的能力.

(a) Original straight line (b) Straight line after run-length smoothing

(a) 原始直线 (b) 游程平滑后的直线

Fig.5 Run-Length smoothing

图 5 游程平滑

3.2 滤除尺寸小的单连通链

长度小于等于 3 个像素的单连通链的数目占一半以上.它们大部分是由字符或噪声产生的,少量是由虚线或断裂的直线产生的.为了去掉噪声和字符产生的单连通链而不影响算法抗断裂的能力,我们采取的策略是滤除两类小尺寸的单连通链:

(1) 长度小于等于两个像素的单连通链(对应噪声产生的单连通链).

(2) 长度小于等于 4 个像素且其一侧或两侧存在其他相连通的单连通链(对应字符产生的单连通链).

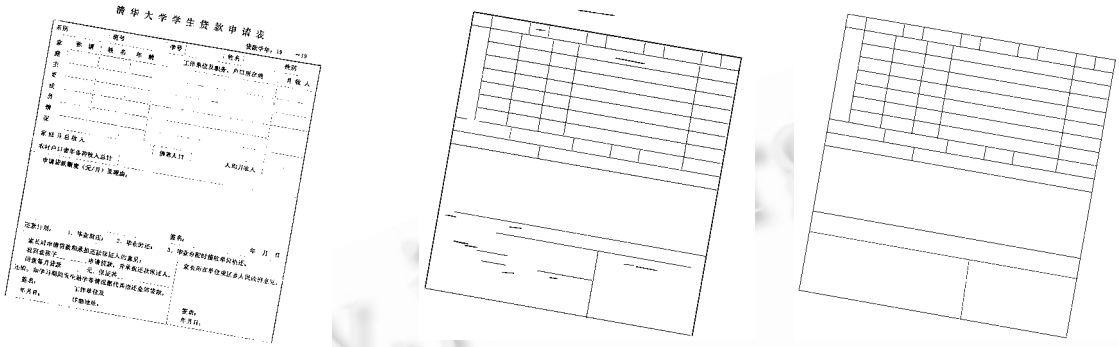
虚线或断裂直线产生的较短的单连通链,绝大多数长度大于两个像素且是孤立的,不符合上述两个条件,因而得到保留.实验表明,由于滤除了噪声,同时保留了虚线和断裂直线的信息,系统的鲁棒性基本没有降低,而速度一般可以提高将近一倍.

3.3 减少单连通链合并时的搜索范围

在单连通链合并时,只需在局部的范围搜索,没必要将全部单连通链搜索一遍.为了减少搜索范围,我们将表格图像分成若干个等宽的条带.在每个条带内提取框线.提取横线时,将图像分割成等宽的横条;提取竖线时,将图像分割成等宽的竖条.假设单连通链的个数与表格的面积成正比,则计算复杂度为$(k_1\times H\times W)\times(k_2\times H\times W)$,其中 W 为表格宽度,H 为表格的高度,k_1,k_2 是常数.前一项为单连通链的个数,后一项为每次合并时需搜索的单连通链个数.改进后,以提取横线为例,若将表格分成 w 个像素宽的条带,则总的计算复杂度为$(k_1\times k_2\times H^2\times W^2)\times(W\div w)$.取 W 为 2000,w 为 400,则单连通链合并环节的速度提高了 5 倍.

4 实 验

实验 1 检验了本算法抗倾斜、抗直线断裂的能力.原始表格图像经过 10°旋转.从图 6(b)中我们可以看到,单连通链合并后还存在少量伪直线,有几根横线发生断裂.引入表格框线的约束信息后,我们得到最终正确的框线提取结果,如图 6(c)所示.为了较大规模地校验该算法的有效性,我们收集了 200 张表格图像,其中 150 张表格图像质量较好,50 张表格图像质量较差,噪声大,存在严重的框线断裂现象.实验结果表明,对于质量较好的表格,框线检测的正确率在 98%以上;等于质量较差的表格,框线检测的正确率在 93%左右.



(a) Form image (b) Lines extracted after DSCC merging (c) Lines extracted ultimately

(a) 表格图像 (b) 单连通链合并后得到的框线 (c) 最终提取的框线

Fig.6 Form frame line detection with DSCC

图 6 表格框线检测的例子

实验 2 比较了倾斜投影法[2]、条带投影法(strip projection)[6]、加速前和加速后的有向单连通链法的速度.测试条件为 Pentium Ⅱ 233,64M 内存,结果见表 1.对于尺寸和黑像素比例比较小的表格,如第 1 个表格,将表格分成条带的措施没有发挥作用,仅仅依靠减少单连通链的方法,速度提高不大.第 4 个表格尺寸比较大,黑像素比例较高,而且扫描时由于亮度设置不合理,产生大量的校验噪声,使单连通链的数目急剧膨胀,改进后速度提高了 20 倍.对于一般表格,改进后速度可以提高 3~10 倍.图像尺寸越大,速度提高越明显.

Table 1 Comparison of speed of some straight line detection algorithms

表 **1** 几种直线检测算法速度的比较

Image size① (Pixels×Pixels)	Horizontal line number②	Vertical line number③	Skew projection④	Strip projection⑤(s)	DSCC before speeding up⑥(s)	DSCC after speeding up⑦(s)
684×650	8	7	0.25s	0.52	0.86	0.58
1816×1112	20	15	0.47s	0.84	2.92	1.03
2400×3438	30	8	0.83s	1.62	10.50	1.93
4198×3165	65	10	1.53s	3.83	109	5.0

①图像大小,②横线数,③竖线数,④倾斜投影法,⑤条带投影法,⑥单连通链法(加速前),⑦单连通链法(加速后).

5 总 结

基于我们自定义的一种称为“有向单连通链”的图像结构基元,本文提出了一种全新的自底向上的表格框线检测算法.表格框线约束信息的引入,帮助我们去除伪直线,补全断裂的直线,提高了算法鲁棒性.实验结果表明,基于有向单连通链的框线检测算法具有抗任意角度的倾斜、抗一定程度断裂的优点.对于断裂非常严重的表格,如何进一步充分利用表格框线之间的约束信息,提高抗断裂的能力,是我们今后改进的方向.

References:

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

[2] Liu, J.H., Ding, X.Q., Wu, Y.S., *et al*. Description and recognition of form and automated form data entry. In: Proceedings of the 3th International Conference on Document Analysis and Recognition. Montreal, Canada, 1995. 579~582.

[3] Liu, W.Y., Dov, D. From raster to vectors: extracting visual information from line drawings. Pattern Analysis and Application, 1999,2(1):10~21.

[4] Yu, B., Jain, A.K. A generic system for form dropout. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996,18(11):1127~1131.

[5] Pan, S.Y. Research and realization of a generic form recognition system [MS. Thesis]. Beijing: Tsinghua University, 1999 (in Chinese).

[6] Chen, J.-L., Lee, H.-J. An efficient algorithm for form structure extraction using strip projection. Pattern Recognition, 1998,31(9):1353~1368.

附中文参考文献:

[5] 潘世言.通用表格识别系统的研究与实现[硕士学位论文].北京:清华大学,1999.

A Form Frame-Line Detection Algorithm Based on Directional Single-Connected Chain*

ZHENG Ye-feng, LIU Chang-song, DING Xiao-qing, PAN Shi-yan

(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

E-mail: zhengyf@cfar.umd.edu; lcs@ee.tsinghua.edu.cn; dingxq@tsinghua.edu.cn

http://ocrserv.ee.tsinghua.edu.cn

Abstract: The existing form frame line detection algorithms are either time consuming or with low robustness. Furthermore, all these approaches do not use the constraint information between form frame lines. In this paper, a novel bottom-up form frame line detection algorithm is proposed based on the directional single-connected chain (DSCC). Defined as an array of black pixel run-lengths, DSCC works very well as an image structure element or a vector in this vectorization algorithm. By merging multiple DSCCs under some constraints, people are able to extract the form frame lines automatically yet fast. With the help of the constraints between form frame lines, the robustness of the approach is increased drastically by getting rid of pseudo lines and completing broken lines. By filtering DSCCs created by noise and speeding up the merging of DSCCs, the speed of this algorithm is comparable with the well-known projection method. Experimental results show that this algorithm is fast, resistant to moderate serious line break and skew of any angle.

Key words: form recognition; image analysis; line detection; optical character recognition (OCR); intelligent document processing

* Received May 11, 2000; accepted October 9, 2000

Supported by the National Natural Science Foundation of China under Grant No.69972024; the National High Technology Development 863 Program of China under Grant No.863-306-ZT03-03-1