第7卷第7期 软件学报 Vol. 7, No. 7
1996年7月 JOURNAL OF SOFTWARE Jul. 1996

通用表格处理系统中定位方法的研究*

刘 真 吴泉源

(国防科技大学计算机系 长沙 410073)

摘要 本文讨论了通用表格处理系统的基本结构和流程,提出了一种全新的表格定位方法——四角定位法,并且阐明了该方法具体实现时应遵循的基本原则.实验表明,四角定位法是一种通用、快速、准确的定位方法.

关键词 表格定位,四角定位法,表格自动化处理,定位点.

表格是信息高度精炼、集中的一种表达形式,它广泛地应用于人们日常工作和生活之中.表格信息的计算机处理,以前只能依靠人工录入来完成,这不仅非常繁琐,而且容易出错.随着文字识别技术的发展和成熟,表格信息的计算机自动化处理已逐渐成为可能.但是由于表格种类繁多、结构复杂,且包含许多文字识别技术不能处理的对象如长线条、图形等.因此,如何将表格中填入的信息从表格的背景信息中分离出来,是表格自动化处理的关键.最原始的分离方法是将表格印刷成某种不同于填入信息的彩色(比如红色),然后在扫描输入时滤除该种颜色,这样做的最大缺憾是无法知道填入的信息属于哪一个信息域.因此,要准确地处理表格中的填入信息,就必须先进行表格定位,然后再从表格中抽取填入的信息.

表格定位在表格的自动化处理中扮演着一个非常重要的角色,它决定了信息抽取的准确度.但由于表格的形式千变万化,扫描输入的表格图象也存在着较大差异、浓淡不一,加之已填入信息的表格(称为实表)与未填入信息的表格(称为空表)具有较大的差异性,所以要将实表图象与空表图象按象素点对齐(称为表格定位),以抽取填入的信息具有极大的难度.

早期的表格定位方法通常采用约束条件下的特殊处理,例如在表格的特定位置印刷用于定位的十字交叉线,因此这样的处理方法通用性不强,使处理的对象受到限制.近年来,国外投入了大量人力、物力进行通用表格处理系统的研究,取得了很大的进展,推出了一些近于实用的系统,其中较典型的系统有 Sequoia Data Corporation 研制的 ScanFix 系统;由 IBM 研制的基于线条定位的 IFP 系统[1]以及由 T.I.S. 公司最新推出的 FORMOUT.特别是 FORMOUT 的功能非常强大,甚至能够处理变形的表格(如传真).

国内对表格自动化处理的研究起步较晚,处理的对象多是一些由横、竖线组成矩形框架结构、形式简单的表格(准确地只能称为 table),而对于不含有线条或具有许多文字提示行

---

* 作者刘真,1964年生,讲师,主要研究领域为图象处理,模式识别.吴泉源,1942年生,教授,博士导师,主要研究领域为软件工程,知识工程与智能应用.

本文通讯联系人:刘真,长沙 410073,国防科技大学计算机系

本文 1995-04-26 收到修改稿

/段等形式复杂的表格(准确地应称为 form)却缺乏研究.为了高起点地跟踪与赶超世界上通用表格图象处理的最新技术,我们研究开发出一个通用的表格处理系统.

1 通用表格处理系统的结构和流程

通用表格处理系统分为表格学习过程和表格处理过程 2 大部分.其结构示于图 1.

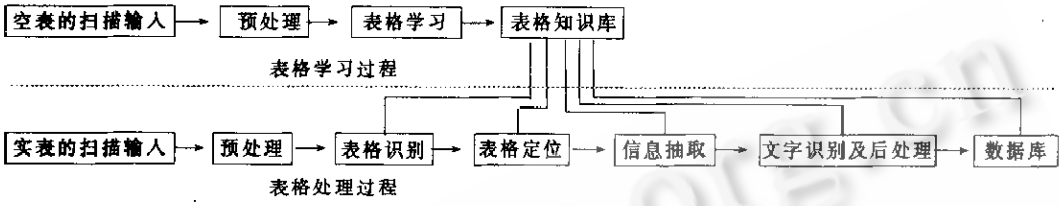


图1 通用表格处理系统

1.1 表格学习过程

处理某种表格前,必须先进入学习过程,学习有关的表格知识,以备处理该类表格时使用.

(1)空表的扫描输入

当扫描输入空表图象时,应该尽可能地使空表图象水平,没有噪声,且没有文字/线条的粘连和断缺,以利于准确地获取表格知识.表格图象一般采用二值图象.

(2)预处理

为后继处理作准备.依需要决定进行何种操作,它通常包括去噪、歪斜图象校正等功能.

(3)表格学习

表格学习完成空表特征的抽取、定位信息的获取和信息域的描述.表格要学习的知识应该是不受填入内容影响的稳定信息.它一般包括自动学习和交互式学习 2 部分.

(4)表格知识库

表格学习到的信息以及空表图象都被存入表格知识库.对表格知识库的操作应该方便可靠.

1.2 表格处理过程

(1)实表的扫描输入

实表的扫描输入没有太多的限制,一般只要求在一定的范围之内扫入图象的倾斜度.

(2)预处理

类似于表格学习过程中的预处理.

(3)表格识别

当一次要处理的表格有多种时,需要分辨当前处理的是哪一种表格.

(4)表格定位

由于扫入实表的位置不可能与空表完全一致,因此相对空表来说,实表图象具有一定角度的倾斜和位移,要从实表图象中抽出填入的信息,就必须将实表图象与空表图象对正,以确定何处是填入的信息.表格定位是通用表格处理系统中最关键的技术之一.

(5)信息抽取

当实表图象与空表图象对正以后,用实表图象"减去"空表图象,剩下的就是填入的信

息.当填入信息与表格固有信息发生相连或相交时,删去表格固有信息,就会造成填入信息的丢失,因此需要采用修补技术.

(6)文字识别及后处理

首先,对从实表图象中抽取出来的信息,采用目前较成熟的文字识别技术,将图象信息变成文字信息.然后,利用表格学习过程中得到的域的描述(字符、数字、地址、姓名等特性),对文字识别结果进行有效性验证,即后处理.

(7)数据库

利用空表学习过程中得到的域的描述,将文字识别结果存入数据库.

2　表格定位方法研究

实表图象相对空表图象来说,一般存在一定角度的倾斜和位移,要将两者对齐就需要进行角度和位移的校正.特别地,有些实表甚至可能是复印件或传真件,机械和光学的作用导致这样的实表具有一定的形变,这时仅仅依靠角度和位移的校正不可能将实表图象与空表图象对齐,对于这种具有较大畸变的实表图象,本文不作讨论,仅仅假定处理的对象没有畸变或只具有很小的畸变,因此整幅实表图象可以按照同一的角度和位移进行校正处理,从而实现表格定位.

表格定位看似简单,但由于实表图象的质量可能较差,实现起来却很不容易.比较直观的想法是利用表格中的线条或文字行来实现定位.基于线(簇)的表格定位方法参见文献[2],实验结果表明,这种方法存在容易受到线条断缺的影响且对于没有线条的表格图象无能为力的缺点.而基于文字行的表格定位方法,则由于实表中填入的文字信息与背景文字信息混杂在一起,因此难于利用实表中固有的背景文字信息进行定位.本文正是在这样的背景下,探讨了一种不依赖于线条、文字行等特定因素,较为通用的表格定位方法——四角定位法.

四角定位法的思想源自早期的在表格图象左上、右上、左下、右下的4个角绘制十字交叉线作为定位点的方法.由于表格图象的4个角可能已不存在十字交叉线,因此,用什么替代十字交叉线作为定位点是问题的关键.一个必然的选择是使用最靠近图象角落顶点的表格的有效信息点.这里有2个待明确的概念:一个是"最靠近"所依赖的是何种距离;另一个是什么样的点才是"表格的有效信息点".这可分2种情况来讨论.

2.1　实表图象水平时四角定位点的求取

因为空表图象在扫描输入时要求尽量水平放置,因此可以认为空表图象是水平的(即使有歪斜,也可用程序校正或以空表图象的角度作为参照系零度).实表图象一般需要进行歪斜度校正,以确保实表图象水平.

定义1. 表格的有效信息点

在理想状态下,表格图象的黑象素点都是表格的有效信息点.但是由于表格图象容易受到噪声的干扰,所以表格图象中的某些黑象素并不一定是有效信息点.为了避免干扰,实际上选取图象的 $m\times m$ 窗口内黑点数大于 c 时(m,c 为自然数)的窗口左上角坐标点作为表格的有效信息点.

从上述定义可知,表格的有效信息点也可能是一个白象素点.

定义 2. 距离

以图象左上角顶点作为坐标系的原点,图象右侧为 X 坐标的增加方向,图象下侧为 Y 坐标的增加方向.

对于直线簇 $x+y=n$ $(n=0,1,2,3,\cdots\cdots)$

定义同一直线上的所有点到原点具有相同的距离.n 值越大,距离越远.

对于直线簇 $x-y=n$ $(n=0,\pm1,\pm2,\pm3,\cdots\cdots)$

定义同一直线上的所有点到图象右上角具有相同的距离.n 值越大,距离越小.

从上述定义可知:

(1)最靠近图象左上角的表格的有效信息点是使 $x+y$ 最小的表格的有效信息点.

(2)最靠近图象右下角的表格的有效信息点是使 $x+y$ 最大的表格的有效信息点.

(3)最靠近图象右上角的表格的有效信息点是使 $x-y$ 最大的表格的有效信息点.

(4)最靠近图象左下角的表格的有效信息点是使 $x-y$ 最小的表格的有效信息点.

图 2(a),(b)分别展示了按上述定义求取的空表图象和实表图象左上角的四角定位点.

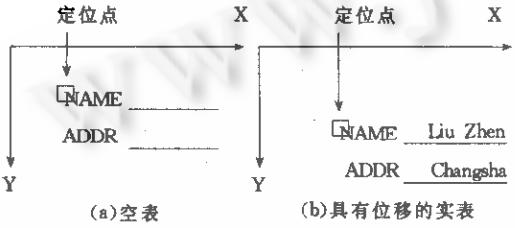


图2 表格图象左上角定位点示意图

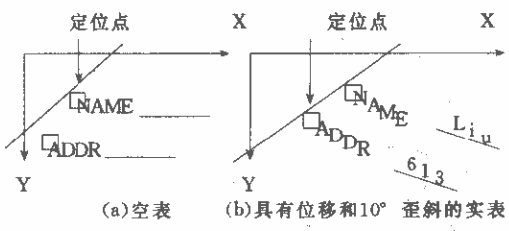


图3 表格图象左上角定位点示意图

2.2 实表图象歪斜时四角定位点的求取

如果对歪斜的实表图象进行校正将回到第 1 种情况.这里要讨论的是不对实表图象进行校正而直接求取四角定位点.这样做的好处有 2 个:①使表格定位方法不需附带要先求取实表图象的角度这个条件;②可以将实表图象的歪斜度校正和位移校正合为 1 个步骤一次性完成,以避免对实表图象的 2 次校正操作,从而大大地节省了处理时间.

在空表图象水平而实表图象歪斜的情况下,如果仍然使用定义 2 求取四角定位点,就会使得空表图象和实表图象求取的四角定位点出现不一致.这是因为,实表图象在水平状态下原本不最靠近图象角落顶点的有效信息点,在歪斜状态下就会变得最靠近图象的角落顶点.图 3 展示了这种情况.

为了避免这种情况的发生,有必要对不同角度的歪斜图象采用相应斜率的等距直线簇.因此,将定义 2 修改如下:

定义 3. 距离

对于直线簇 $ax+by=n$ $(n=0,1,2,3,\cdots;\ a,b$ 为自然数)

定义同一直线上的所有点到图象原点具有相同的距离.在相同的 a,b 参数情况下,n 值越大,距离越远.

对于直线簇 $ax-by=n$ $(n=0,\pm1,\pm2,\pm3,\cdots;\ a,b$ 为自然数)

定义同一直线上的所有点到图象右上角具有相同的距离.在相同的 a,b 参数情况下,n 值越大,距离越小.

从上述定义可知,选用不同的 a,b 参数,就可得到不同斜率情况下的四角定位点.但由于在求取四角定位点以前并不知道实表图象的歪斜角度,因此如何确定 a,b 参数是问题的关键.下面讨论2种确定 a,b 参数的方法.

(1)穷举法

穷举法是对实表图象允许的最大歪斜角度范围内的 a,b 参数,在一定的精度下进行穷举.该方法的缺点是 a,b 参数的取值范围太大.

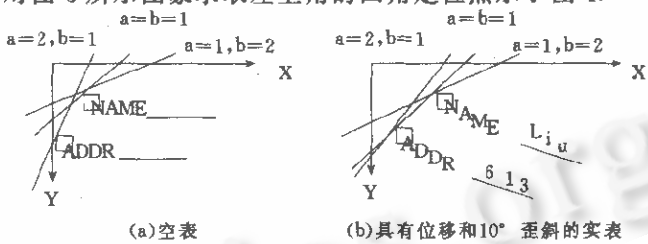
例如实表图象最大歪斜不超过15°时,相对 $a=b=1$ 即斜率为45°的直线簇的最大范围即为30°～60°,亦即 $tg30°\leqslant\frac{a}{b}\leqslant 60°$,这近似于 $\frac{25}{43}\leqslant\frac{a}{b}\leqslant\frac{43}{25}$,则 a,b 的取值范围为25～43的所有整数.如果希望有更高的精度,a,b 取值范围可扩大为250～430的所有整数.

(2)近似法

由于穷举法在一定的精度下 a,b 参数的取值范围太大,考虑到表格4个角落的实际情况是文字行、线条和图形等元素之间的错落并不是很多,亦即大的轮廓变化并不是很多,而四角定位法的实质就是选取轮廓的凸出点作为定位点,因此降低穷举法的精度,减少 a,b 参数的取值,并不会明显减少求取的四角定位点个数.同理,为了使得在空表图象中求取与实表图象基本相同的定位点,可将相同的 a,b 参数应用到空表图象.这样,对空表图象和实表图象的定位点求取方法变成一致.

对穷举法中的例子,选取 a,b 参数为如下3组:$a=b=1$;$a=1,b=2$;$a=2,b=1$.将这3组 a,b 参数应用到空表和实表图象,各得到12个四角定位点.

用上述方法对图3所示图象求取左上角的四角定位点示于图4.



(a)空表　　(b)具有位移和10°歪斜的实表

图4 表格图象左上角定位点示意图

不论实表图象是水平还是歪斜的,不恰当的四角定位点求取方法将导致计算效率的低下和定位点出现偏差.从理论上来说,对于 $m\times m$ 窗口及窗口内黑点数阈值 c,m 和 c 越大,则噪声或污渍被当作有效信息点的可能性越小,但表格的真实黑象素部分不被认作为有效信息点的可能性也越大;对于窗口移动的步长 R,R 越小,求取的定位点越准确,但时间也越长.因此,在实际实现中,为了既得到较好的准确度,又不耗费过多的时间,最好采用多级处理,即首先采用较大的窗口、黑点数阈值和步长,然后再逐渐缩小这些参数,直至达到要求的定位精度.

2.3 四角定位点的选取

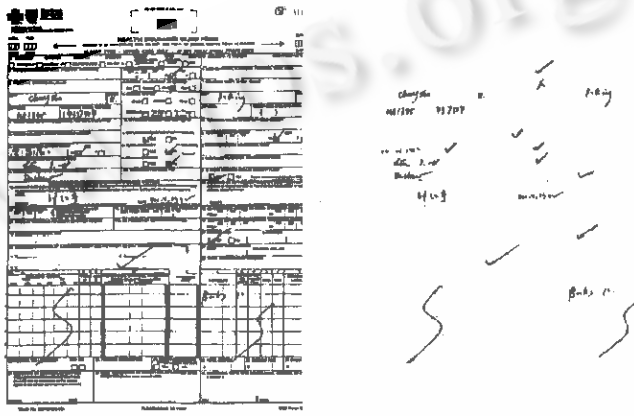
由于实表图象可能受到噪声、污渍等情况的影响,或者由于简化 a,b 参数的选取而造成求取的某些定位点不正确,对于这些不正确的定位点应该去掉.其方法是根据空表的定位点分布情况,检查实表中各对应定位点之间的结构关系,保留结构关系最稳定的2个定位点.

利用实表中的这2个定位点及空表中对应的2个定位点，即可计算出实表图象相对空表图象的歪斜和位移，从而实现实表图象与空表图象的对正.

3 实验结果及结论

在PC486和WINDOWS环境下，a，b参数采用近似法，用C语言实现四角定位法.对70种A4大小的表格共247张实表进行处理，当最大误差为4个象素点时，四角定位点的平均求取时间仅为0.2s.图5展示了一个实表图象和按四角定位法抽取信息后的图象.

理论分析及实验结果表明，四角定位法是一种通用、快速、准确的定位方法，特别是它能对歪斜图象进行处理，使歪斜校正和位移校正可以一次完成，节省了处理时间.



(a)缩小的实表 (b)抽取的信息

图5 四角定位法实例

参考文献

1 Cassy R G, Ferguson D R. Intelligent forms processing. IBM System Journal, 1990, **29**(3): 435～450

2 沈清，吴泉源.表格对正处理.In：高文编，计算机智能接口与智能应用'93论文集，1993.

RESEARCH ON FORMS REGISTRATION IN GENERAL FORMS PROCESSING SYSTEM

Liu Zhen Wu Quanyuan

(*Department of Computer Science National University of Defence Technology Changsha* 410073)

Abstract This paper discusses the basic structure of general forms processing system, suggests a new method——forms registration by anchors in the four corners, and gives the rules to implement this method. It is shown by experiment that forms registration by anchors in the four corners is general, rapid and exact.

Key words Forms registration, forms registration by anchors in the four corners, automatic forms processing, anchors.