第7卷第10期
1996年10月

软 件 学 报
JOURNAL OF SOFTWARE

Vol. 7, No. 10
Oct. 1996

表格图象数据抽取柔性匹配方法*

刘江宁 杨 嵘 张 剑

(长沙工学院计算机系 长沙 410073)

摘要 数据抽取是表格图象处理中重要的一环.本文给出了一种基于锚点柔性匹配进行数据抽取的方法,并讨论了该方法中锚点查找、组织、定位、填写域定位参数确定及数据抽取过程的思想.

关键词 表格图象处理,数据抽取,匹配.

商用表格(Form)信息的智能处理,是伴随着计算机小型化的趋势及办公自动化的应用要求应运而生的一门新的计算机智能应用领域.作为一种信息表达手段,表格以其简明、规范、便于填写和处理等鲜明特点,已经日益深入人们的信息生活之中.如何利用计算机高速自动地获取、存储和管理数量巨大的表格信息已越来越成为人们关注的焦点,并逐渐成为计算机模式识别与图象处理领域的热门研究课题.

表格图象主要是指人们日常生活中使用的表格和档案经扫描仪或传真机输入计算机的黑白二值图象.表格信息处理系统旨在研究表格图象的获取与存储、特征信息的学习与表示以及有效信息的抽取、传输、识别与管理等方面的内容.物理上,一幅表格图象由定域和变域2部分构成.其中定域为印制在表格上的固定信息,如线、图形、条码、说明文字和栏目名等;变域是要求用户填写数据的区域,这些区域可能填写字符、数字或者特殊符号如"√"、"×"等信息.我们将填过信息的表格称为实表,而没有填写信息的表格称为空表.

表格处理的核心机制由7个主要方面构成[1]:(1)实表的预处理:对实表进行去噪、歪斜校正等一系列预处理工作,以方便表格识别等后续过程;(2)表格识别:接受去噪、校正处理后的实表图象文件,通过查询表格模板信息,确定实表属于哪一类表格;(3)表格定位:根据识别的结果,将匹配的一对空表和实表进行对正处理,以校正两者在水平方向和垂直方向的偏移;(4)变域数据抽取:从实表中删除表格中的固有信息,生成用户填入的数据图象.我们称仅由变域数据构成的图象文件为差表;(5)表格重构:按一定的格式重新合成空表与差表,构成一张高质量的实表图象文件.表格重构主要应用于网络环境和表格信息的传输.据统计,一张实表的变域仅占整幅表格空间的2~10%,基于表格重构机制,表格发送端只需传输变域构成的表格,从而大大减少了通信开销,且不会损失图象的有效信息,这种语义上的

---

* 作者**刘江宁**,1966年生,讲师,主要研究领域为人工智能与专家系统,图象处理.**杨嵘**,女,1969年生,讲师,主要研究领域为图象处理.**张剑**,1961年生,副教授,主要研究领域为软件工程,图象处理.

本文通讯联系人:刘江宁,长沙 410073,长沙工学院计算机系

本文1995-09-18收到修改稿

图象压缩技术是目前各种压缩方法所难以匹敌的.(6)OCR:识别差表中的字符,得到相应的文本文件.(7)后处理:对生成的文本文件进行拼写检查、属性验证.

下面,我们讨论表格图象处理中表格定位及变域数据抽取的基本方法.

1 数据抽取基本过程

表格图象的数据抽取划分为学习和处理2部分.表格定位参数的学习部分完成空表锚点的查找、锚点的组织和空表的模化.其中前2项针对表格定位,后1项针对数据抽取.处理部分完成表格的定位和定义域内用户填写数据的抽取.

根据实表质量的差异,表格定位可以划分为刚性匹配和柔性匹配2种.

(1)刚性匹配.如果表格在水平方向和垂直方向均不存在畸变,则用一对参数 $\Delta x,\Delta y$ 即可表示空表与实表在水平方向和垂直方向上的偏移量,视觉上可通过空表固有信息残留量来判断2张表格的匹配程度,显然,此种残留越少,则表格匹配度越高.$\Delta x,\Delta y$ 可以通过空表线与实表线的匹配得到.

(2)柔性匹配.如果表格在水平方向或垂直方向存在畸变,则用一对参数 $\Delta x,\Delta y$ 不足以表示空表与实表的偏移量,例如在实表是传真件、复印件或经受潮烘干或多次折叠已严重变形的情况下,刚性匹配无疑将失败.这时可将图象划分为多个区域,针对每一区域选定偏移量,在每一小区域内认为图象不存在畸变.

2 锚点查找算法

数据抽取基于定位锚点进行.常用的定位锚点包括水平/垂直线、图标、单词等,锚点查找算法以连通块查找算法为基础.通过连通块的分析可以确定块的类型,并得到表格的全部结构信息,从而为表格定位提供依据.

一幅表格图象的物理层描述分为象素、象素段、笔段、块和域5级.直觉上来看,图象可划分为一系列域,如文字行、阴影域、图标或反白域等,它们之间的关系结构构成了表格版面描述的基础;域由连通块构成;连通块由一组相互关联的笔段构成;笔段为一组连续、充分发展的象素段集合,除两端外,中间不出现分叉与合笔.

定义. 象素段:图象一行中一组连续的黑象素构成一个象素段.一个象素段 s 可用其左边界 $l(s)$ 和右边界 $r(s)$ 加以刻画.

定义. 象素段的邻接相关:若相邻行的2个象素段 s_1,s_2 满足

$$\min(r(s_1),r(s_2))-\max(l(s_1),l(s_2))\geqslant -1$$

则称 s_1,s_2 邻接相关.

定义. 象素段 S_1,S_2 称为相关的,若 S_1,S_2 邻接相关,或者存在象素段 S_3,使得 S_1 与 S_3 邻接相关,且 S_2 与 S_3 相关.

定义. 相邻行的2个象素段 s_1,s_2 称为唯一相关的,iff s_1,s_2 邻接相关,且 s_1 所在的行中不存在其它象素段与 s_2 相关,且 s_2 所在的行中不存在其它象素段与 s_1 相关.

定义. 笔段:从第 k 行第 $k+n$ 行的 n 个象素段构成的集合 $\{s_1,\cdots,s_n\}$ 称为一个笔段,若

(1)s_i,s_{i+1} 唯一相关($1\leqslant i\leqslant n-1$);

(2)在第 $k-1$ 行不存在与 s_1 唯一相关的象素段;

(3)在第 $k+n+1$ 行不存在与 s_n 唯一相关的象素段.

称满足条件(1)、(2)的象素段集为开笔段或活跃笔段(Active Stroke),并称其为 s_n 所在行上的开笔段.类似地可定义笔段之间的邻接相关和相关关系.

定义. 第 i 行上的开笔段集 $\{S_1,\cdots,S_m\}$ 与 $i+1$ 行的象素段集 $\{s_1,\cdots,s_n\}$ 之间的笔段——象素段相关图(简记为 $SSRG$)$G=\langle V,E\rangle$ 为

$V=\{S_1,\cdots,S_m\}\cup\{s_1,\cdots,s_n\}$

$E=\{\langle S_i,s_j\rangle|S_i$ 的最后一个象素段与 s_j 邻接相关$\}$

将图象第 i 行的全部开笔段及第 $i+1$ 行的全部象素段按从左至右的顺序进行排列,通过线性遍历 2 张表,可迅速生成 SSRG. SSRG 与经典图象分割技术中采用的 LAG 结构类似.[2,3]

由上面的定义可以看出:象素段的相关关系为表格图象全体象素段构成的集合 I 上的一个等价关系,由此可以确定 I 的一个划分 $P(I)$. 我们将 $P(I)$ 中每一元素称为一个块. 图 1 给出了块、笔段、象素段之间的关系. 图 1 由 2 个连通块组成,每块包含 4 个笔段,笔段之间的相关关系发生在首尾象素段上,如 *Stroke*1 与 *Stroke*2,*Stroke*3,*Stroke*4 均相关,因为其最后一个象素段与后面 3 个笔段的第 1 个象素段相关,*Stroke*2,*Stroke*3 及 *Stroke*4 虽然没有直接相关关系,但通过 *Stroke*1,它们相互关联在一起.

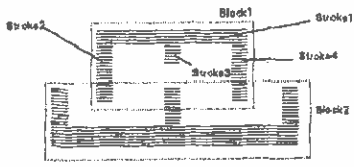


图 1 块、笔段、象素段之间的关系

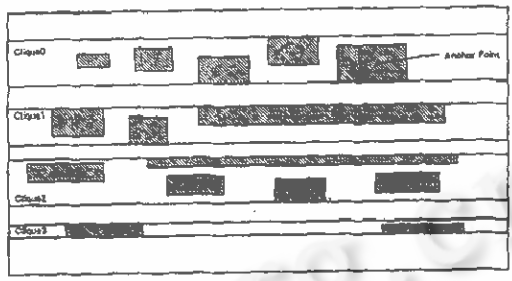


图 2 表格图象锚点组织示意图

对于给定的图象区域,通过自上而下逐行扫描,采用相关分析法可构造出全部连通块.[4~7]由于大部分图象压缩技术通常依据游程的哈夫曼编码来进行,上述算法可以直接在压缩图象的基础上完成. 上述算法的时间复杂性与图象中的游程数成线性关系,这是因为我们对参与相关分析的每张表进行序化的结果. 在实际处理中,若连通块宽度和高度均在一定范围内,则将其作为锚点.

3　锚点的组织与定位方法

由于实表存在角度,从实表中得到的锚点与空表的锚点在大小及相对位置上均存在很大的差异,因此,要从锚点的匹配中准确捕获定位信息,首先必须对实表的锚点进行逻辑旋转[8],由于填写信息的影响、扫描时产生的噪音及畸变,加上逻辑旋转产生的信息损失,空、实表在锚点的个数和结构关系上仍存在很大的差异. 为了达到准确匹配的目的,我们给出下面锚点组织结构:(1)将水平相关的锚点(一条扫描线能同时穿过的 2 个锚点)组织成一个大的结构,称为团. (2)团按从上到下的次序排列. (3)每个团内的锚点按从左往右的次序排列.

团及锚点均用矩形界定其范围.图2给出了一张表格图象上锚点组织示意图.

为了达到准确匹配的目的,一般采用锚点的分层组织法.在锚点的分层组织结构中,高层信息可以指导低层匹配过程,低层的精细匹配可以提供准确的定位参数.通常采用分层组织法的匹配过程更加稳定,效率更高.根据分层的思想,锚点的匹配又可以进一步划分为粗匹配和精细匹配2个部分.粗匹配寻找匹配的入口点;精细匹配查找所有可匹配的锚点对,并得到每一对匹配锚点的定位参数,即水平方向和垂直方向的偏移量.

(1)粗匹配.粗匹配基于团进行,实际上是按团中第1个锚点(最左边的锚点)来进行.团匹配是一个试探的过程,先假定空表的第 i 个团与实表的第 j 个团相匹配,在此基础上得到一个定位入口参数 $\Delta x,\Delta y$.线性遍历空表与实表的所有团,得到可匹配的团的个数.用匹配的团数多少为标准,确定上面的最佳 i,j 值以及初始定位参数 $\Delta x,\Delta y$.在上面的遍历过程中,为了处理表格变形的情形,采用了自适应的思想,即2个团是否匹配,以它们前面2个已匹配的团所得到的定位参数为依据,当空表的一个团与实表的一个团匹配时,即它们的第1个锚点匹配成功(宽度与高度相当,且考虑 $\Delta x,\Delta y$ 后位置接近)时,修改 $\Delta x,\Delta y$ 为当前2个锚点的 $\Delta x,\Delta y$,下面的匹配在新的 $\Delta x,\Delta y$ 的基础上进行.

(2)精细匹配.精细匹配是在锚点基础上的匹配.匹配算法可以描述为:线性遍历2张团表,若空表的某一个团与实表的某一个团相匹配,则通过线性遍历这2个团的2张锚点表,得到团内所有可匹配的锚点对及其定位参数.团的匹配及锚点的匹配均是自适应的.由于粗匹配过程提供了入口定位参数,精细匹配不需要试探.

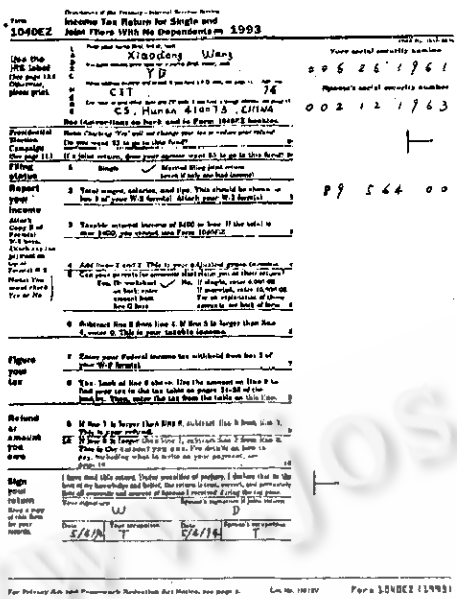
4 填写域定位参数确定与数据抽取

填写域的定位参数通过查找最近邻的匹配锚点对而得到.邻近的锚点可利用Voronoi图的检索来实现.空表锚点Voronoi图的构造在空表学习时完成.由于难以保证每一空表锚点均能在实表中找到对应的匹配锚点(如FAX过程中的信息损失或实表在扫描平板上放置不当而丢失一部分图象),因此,必须针对锚点的匹配情况对空表的Voronoi图进行动态维护.

在理想情况下,根据空表与实表之间的定位参数,由实表减去空表即为差表的内容.但是,即使在刚性匹配的情况下,由于扫描过程中存在的随机干扰及预处理引起的图象信息损失,实表与空表之间的固定信息不可能一致,得到的差表图象往往质量很差.因此,需要为空表引入一个膨胀常数 C,即对空表图象进行膨胀处理,使得空表中的一个黑象素生长到周围 $C\times C$ 大小的区域中.记 E 为空表,F 为实表,E 膨胀后的结果为 $fat(E,C)$,则差表 $D=F-fat(E,C)$,膨胀常数 C 与图象扫描精度有关.

当数据与表格背景信息重迭或非常接近时,空表膨胀法会引起变域信息损失.例如当用户填入的数据紧压表格线或穿越表格线时,移去表格线的同时,将丢失填入的数据信息.因此,在采用膨胀空表法抽取数据时,必须利用连通性及字符的结构特征,采用插值法来完成断缺字符图象的部分修补工作.也可以在后处理程序中通过拼写检查进行纠正.图3给出了数据抽取的一个实例.

实表图象　　差表图象



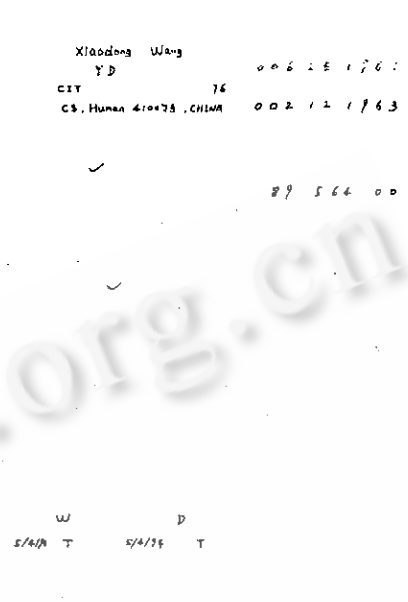


图 3 表格图象处理实例

参考文献

1 吴泉源，张剑，刘江宁. 表格图象处理系统 FPS 的设计与实现. 智能接口与智能应用论文集，1993.

2 帕夫利迪斯著，张寿荃等译. 结构模式识别. 上海：上海科学技术文献出版社，1981.

3 徐建华编著. 图象处理与分析. 北京：科学出版社，1992.

4 Doster W. Designing a document analysis system. Tutorial Presented at 8th Int' 1th Conf. on Pattern Recognition, Paris, France, Oct. 1986.

5 Scherl W *et al*. Automatic separation of text, graphic and picture segments in printed material. In: Pattern Recognition in Practice, Amsterdam: North— Holland, 1980.

6 Srihari S N, Zack G M. Document image analysis. Proceedings of the 8th International Conf. on Pattern Recognition, Paris, 1986.

7 Wahl F M, Wong K Y, Casey R G. Block segmention and text extraction in mixed text/image documents. Computer Graphics and Image Processing, 1982, **20**.

8 刘江宁，时春，张华滨等. 歪斜图象校正技术. 智能接口与智能应用论文集，1993.

AN ELASTIC MATCHING METHOD FOR DATA EXTRACTION OF FORM IMAGE PROCESS

Liu Jiangning　Yang Rong　Zhang Jian

(*Department of Computer Science　Changsha Institute of Technology　Changsha*　410073)

Abstract　The extraction of the filled—in data is an important stage in the form image process system. Based on the alignment anchor points, an elastic matching method for data extraction is approached in this paper. The searching procedure and organizing strategy of the anchor points, the alignment method of the blank form and the filled—in form are discussed also.

Key words　Form image process, data extraction, match.