代　号　10701　　学　号　0320421151

分类号　TP391.4　　密　级　公开

西安電子科技大學

硕士学位论文

题（中、英文）目　文档图像逻辑结构分析方法研究

The Research on Logical Structure Analysis of Document Image

作 者 姓 名　王来敬　指导教师姓名、职务　王泉　副教授

学 科 门 类　工学　学科、专业　计算机系统结构

提交论文日期　二〇〇六年一月

摘要

在普通文档图像中存在着各式各样的表格，对文档图像中的表格进行自动定位、分析和内容识别是DIA领域的研究重点之一。

本文在大量实践工作的基础上，对表格图像的版面结构分析及逻辑结构分析进行了有益的探索。在版面结构分析部分，论文采用了直接抽取构成表格的线条，并计算线条的交点，进而获取各个单元格信息的方法。该方法以线条交点矩阵表示表格版面结构分析结果，不仅降低了对问题描述的难度，体现了表格版面结构的全局行列特征，更易于检索，这为表格图像的后续处理提供了极大方便；在版面结构分析部分，论文提出了完整的表格结构表示方法，并通过嵌套链表描述表格之间的嵌套关系，这为复杂的表格结构描述提供了便捷的描述方式。该方法充分利用了标题域与数据域之间的依赖关系以及基本布局结构的直线交点特征，不仅能够实现对已填充表格的逻辑结构分析，而且可以将表格按照基本的布局结构进行分割。

实践证明，论文所述方法有较好的处理效果，可以满足表格结构自动处理的实际应用需求。

关键字：版面结构分析 逻辑结构分析 倾斜校正 直线交点特征 嵌套结构

Abstract

There are various kinds of tables in the ordinary documents. How to realize the automatic positioning, the analysis and the content recognition of table image has become an important branch of DIA domain.

Based on massive practice, this paper has carried on the beneficial exploration to the table layout structure analysis and the table logical structure analysis. In the table layout structure analysis part, this paper extracts the lines from the table image at first, then gets the intersection points of lines, and at last gains each unit information. The result of table layout structure analysis was represented through the matrix of intersection point of lines. This method not only reduced the difficulty of question description, moreover has manifested the overall characteristic of table layout structure. All this has provided the enormous convenience for the following processing. In the table logical structure analysis part, this paper proposed a new table structure expression method, and through nesting chain can describe the relationships between the nesting structures. And all this has provided the convenient description way for the complex form structure description. This method fully used the dependent relations between the title field and the data field, as well as line intersection characteristic in each basic layout structure. This method not only can analysis filled-in table image's logical layout, but also may segment table image according to the basic layout structure.

The practice proved that, this paper stated the method to obtain the good processing effect, has satisfied the form structure automatic reduction practical application request.

Keywords: layout structure analysis logical structure analysis
skew correction line intersection feature nesting structure

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，议文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若不实之处，本人承担一切相关责任。

本人签名：张荣超　　日期：2006.01.23

关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文中的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密的论文在解密后遵守此规定）

本人签名：张荣超　　日期：2006.01.23

导师签名：[illegible]　　日期：2006.1.23

第一章 绪论

1.1 引言

随着计算机的日益普及，越来越多的人习惯于电子化办公，很多文件、资料被直接保存为数字化格式。由于数字化文档的保存、检索都远远比传统的纸质文档方便，这种电子办公方式正逐渐成为人们被人们所接受并成为日常办公的主流。但是，另一方面，现实生活中，纸质文档并没用因为数字文档的出现而退出历史舞台，反而在文明高度发达、信息快速膨胀的现代社会，纸质文档已呈现出呈指数增长的趋势。更重要的是，纸质文档历来是人们存储和传递信息的最主要媒介，由于价格低廉、便于携带，以及人们长久以来形成的阅读习惯，因而，时至今日，纸质文档仍然是记录与传递信息的最重要载体，依然有着不可替代的重要作用。

然而，纸质文档也有很多的缺点。比如说信息繁杂、不易于检索、容易破损、不适于长期保存等等，而电子文档却有着易于存储、编辑、检索和管理的优点，二者都有着各自的优缺点，无法相互取代。如何能让机器能够自动识别不同类型的文档，并以有效的方法获取、理解文档中所记录的各种信息和数据？这一直是学者们的重要研究工作，这样，也就产生了文档图像分析(Document Image Analysis，简称 DIA)领域。

DIA 领域是人工智能(Artificial Intelligence)领域的重要分支，其理论基础是数字图像处理技术(Digital Image Processing)与模式识别技术(Patten Recognition)。DIA 是一个工程领域，是人工智能(Artificial Intelligence)领域的重要分支，其目的是研究各种理论、探索各种方法，以便将文档图像中的信息转换成计算机能够处理的数字信息（文档、图像、表格等）[1]。长期以来，国内外学者对文档图像分析技术进行了深入而细致的研究。图像倾斜角计算[2]、游程分类[3]、版面切分[4]等基本算法、图像版面结构与逻辑结构的概念[5]、版面结构分析与表示[6,7,8]及其性能评估[9,10]算法等等都得到了广泛研究，一些实际处理系统[11,12]也得以应用。

在 DIA 领域中存在一种比较特殊的文档类型—表格文档图像，其区别于其他普通文档图像的地方就在于表格文档图像中存在表格。正因为表格具有卓越的数据表现力，表格往往具有较普通文档复杂的逻辑结构，这就使得此类文档图像更难于处理。

1.2 研究现状

表格是数据最直观的组织和表达形式，在人们的日常生活和工作中有着广泛的应用，它的形式千差万别、繁简不一，广泛用于信息收集、数据统计以及资料保存。表格所具备的强大数据组织功能和便于查询、统计、计算的能力使其成为计算机数据存储、交换、检索的基本格式，因而对文档图像中的表格进行自动处理便成了实际应用的广泛需要。

表格自动处理是文档图像分析的一个分支，表格的处理方法主要有模板匹配法、全自动处理、半自动处理法等。模板匹配法一般适用于特定类型的文档，而对于普通文档图像中的表格处理，由于没有固定的可以参照的模板，因而难以取得理想的处理效果。半自动处理的方法虽然易于实现，但却需要较多的人工参与，并且效率也不高，这与人们所孜孜以求的自动化处理有相当大的差距。而要对普通文档图像中的表格实现全自动处理便显得更加困难，比如说文档中的表格个数是不确定的，有可能有一个表格，也有可能有多个表格，或是没有表格，甚至一页图像就是由一个表格构成。而且文档图像可能模糊不清，残缺不全，这样，如何合理地对图像进行校验和补全，对于计算机来说也是十分困难的。而且人本身的思维歧义多变，对于同一表格所体现的二维或多维逻辑结构不同人可能会有着不同的理解，于是就没有一个统一的或固定的准则和标准令计算机参考，这样也增加了表格自动处理的难度。

虽然面临着很多的困难，但是国内外学者还是对表格处理进行了广泛而深入的研究，表格版面结构和逻辑结构的概念[14,15,16]以及许多表格结构的描述模型[13,17,18,19,20]表格结构定位与分析[21,22, 23,24,25,26]，及表格结构判定与识别[27,28,29,30,31]，方法被提出，针对表格自动处理的实际系统也得到了广泛的应用。

1.3 表格文档处理概述

1.3.1 功能模块划分

为了实现表格文档的自动化处理，首先需要对表格进行学习，分析表格图像的版面结构及逻辑结构，并将与表格相关的信息存入数据库中，为自动化处理所用。实现表格的自动化处理，需要实现以下几个功能：

- 表格图像的预处理。表格图像的预处理主要是用来消除由于图像质量的退化及扫描过程所带来的噪声，并对倾斜表格图像进行倾斜校正。
- 表格图像的版面结构分析。表格图像的版面结构是指线条、数据区域、字

符、标记、图片等版面元素的位置、大小等物理信息。本文采用直线交点阵为载体对表格的版面结构进行表达。表格的直线交点阵不仅可以体现出表格的整体布局特征，更能体现出表格中每一个单元格的局部特征。同时，直线交点阵以矩阵为基础，便于表格信息的编辑、检索。

- 表格图像的逻辑结构分析。表格图像逻辑结构是指表格版面元素的逻辑属性与各版面元素之间的逻辑关系。通过对表格结构的逻辑分析，可以了解表格基元之间相互关系，为表格结构之间的匹配提供了另一种方法。本文以直线交点阵为基础，提出了一种简便的表格逻辑结构分析方法。

1.3.2 处理模式概述

根据不同研究工作所采用的应用方式，对文档图像中表格结构的自动处理基本可以分为以下三种模式[13]：

1) 模板匹配法：预知表格的版面结构和表格在文档图像中的位置，通过人机界面描绘或形式语言定义表格版面结构的模板。在参考文献[32]中，贝尔实验室的工程师详细描述了基于 GUI 界面制作表格模板的过程。在参考文献[33]中，描述了基于表格模板库对文档图像中表格结构进行匹配定位的方法。图 1.1 描述了基于模板匹配方法的表格版面结构定位与分析流程。

模板匹配法有两个模块，其分别为表格学习模块和表格自动化处理模块。在表格学习的过程中，首先要进行的是表格版面结构的分析，在此基础上进行逻辑结构的分析，并将以逻辑结构表达的表格基元信息以一定的方式存储于数据库中。

在表格自动化处理的过程中，首先对表格进行逻辑结构分析，并将所分析的结果于数据库中的表格结构进行相似性匹配，从中找出最相似的逻辑结构，如果相似度大于一定阈值的话，就将此逻辑结构模型作为待识别表格的逻辑模型，并以此逻辑模型对表格进行自动化的信息提取。

2) 自动定位：根据图像信息，抽取线条与文字区域，根据版面特征与启发式规则，定位并分析表格版面结构。根据线条的类别，又能够进一步分为有线表自动定位和无线表自动定位，在参考文献[26,30]中描述了如何自动定位文档图像中无线表的方法。图 1.2 描述了自动定位方法的处理流程。

3) 人机交互半自动定位：通过 GUI 界面，由操作者划定表格区域，将其标识为表格，而后，进行自动定位和版面结构分析。采用这种方法时，处理系统将使用强制性规则进行表格版面结构分析，保证表格结构定位与分析的完备性。图 1.3 描述了人机交互半自动定位方法的处理流程。

传统的模版匹配法处理流程为：

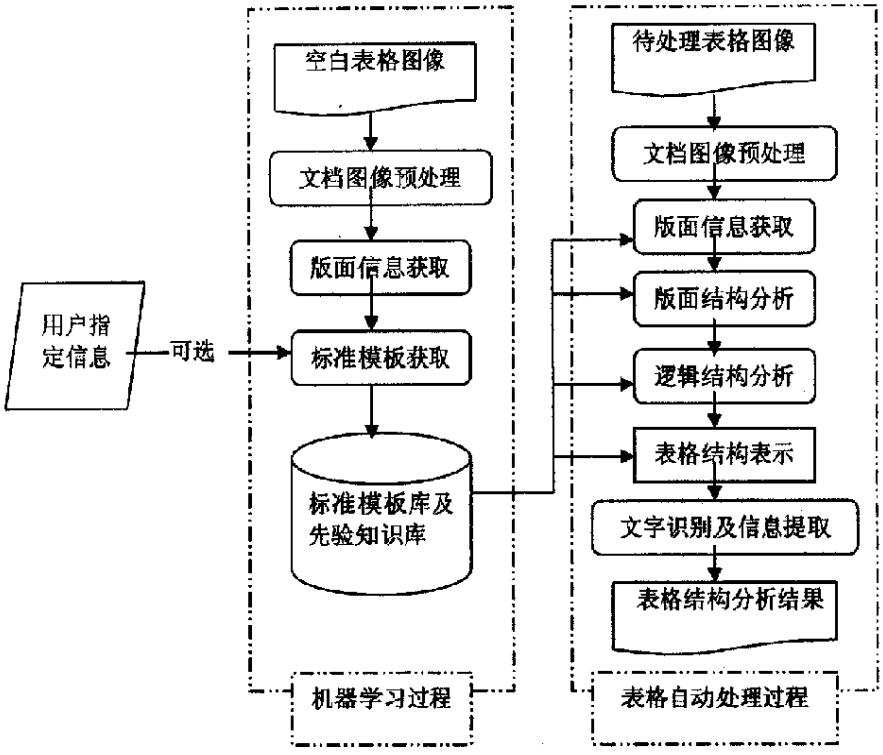


图 1.1 模板匹配法处理流程

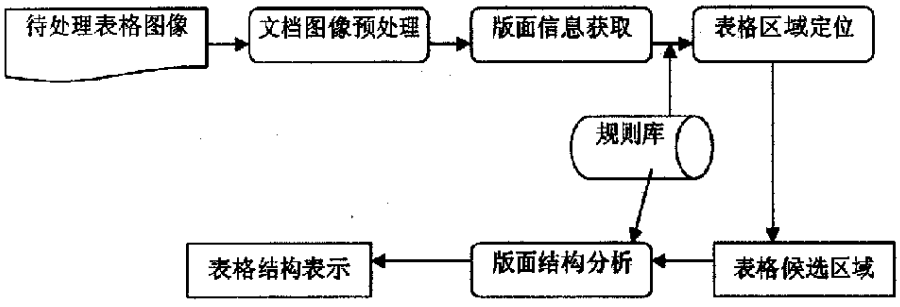


图 1.2 自动分析流程

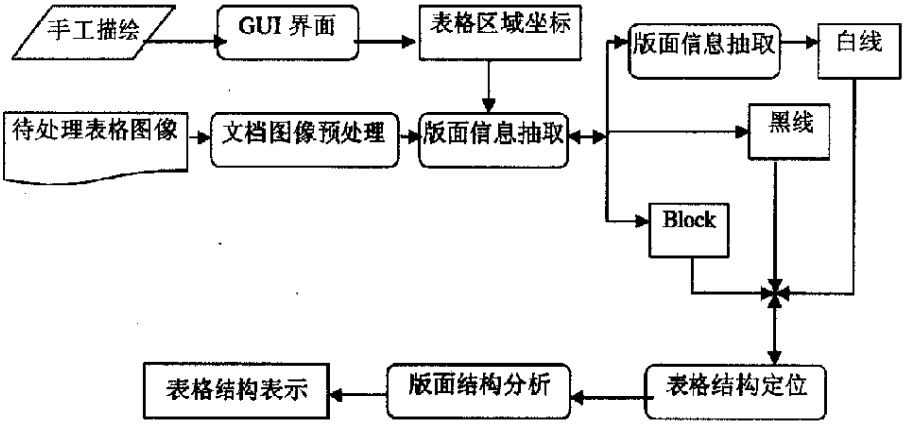


图 1.3 人机交互半自动定位处理流程

1.4 研究范围的界定

表格强大的数据表达能力决定了表格类型的多样性。针对如此众多的表格类型，要想做出一个涵盖所有表格类型的通用方法是非常困难的，因此本文所研究的对象的特点，对研究范围进行了界定。

本文所适用的表格具有以下特征：

a) 表格具有规则的矩形边界。

b) 表格内部的单元格均为规则矩形。

有一些表格不具有上述结构特征，如元素周期表、财务报表等，由于在普通的文档图像中，这些表格并非常见表格，本文未将其作为主要的研究对象。

1.5 研究目标

表格文档处理有着非常广泛的应用前景，表格文档版面理解研究的滞后不仅影响了 OCR 识别的效率，而且已经成为制约文档信息处理系统性能及推广应用的重要因素，表格文档图像的版面理解主要存在以下不足：

1. 由于文档版面理解和 OCR 识别算法对文档倾斜很敏感，文档图像的倾斜校正是一个重要的预处理，许多倾斜校正方法是基于先验知识。还没有一种较为通用的文档图像倾斜校正方法。
2. 表格处理虽然在一些领域中得到实际应用，由于表格处理系统只能处理特定版面的表格、表格版面设计复杂、对纸张质量要求高、要求填写规范。日常生活中的表格是多种多样的，例如报纸上的问卷表纸张质量较差，表

格图像包含畸变、噪声，目前还没有一个比较好的系统来处理这类表格。

3. 传统的表格图像分割方法仅是针对单元格级别，并没有将具有相互依赖关系的单元格划分在同一个区块中。

本文对表格文档图像版面理解中尚未解决的一些问题展开研究。研究的重点是：

1. 研究一种适于表格文档图像的倾斜校正方法，根据不同的文档内容，选用相应的策略进行文档倾斜角度估计，并对倾斜角度的误差进行估计，使得倾斜角度的误差控制在允许的范围内。
2. 研究一种健壮的表格版面信息提取方法，该方法将版面分析结果以矩阵的形式存储，为后续的信息提取、检索提供方便。
3. 研究灵活的表格逻辑结构分析方法，该方法能克服传统方法仅针对空白表格局限性。建立方便灵活的表格文档结构描述方法，便于表格结构之间的匹配。
4. 研究一种基于直线交点特征的表格图像分割方法，该方法区别于普通分割方法的地方在于它能够将由相互依赖关系的单元格划分在同一个区块中，而普通的表格图像分割方法仅是针对单元格进行划分。

1.6 论文结构及内容组织

本文所解决的问题主要是解决表格图像的版面结构分析、逻辑结构分析、表格逻辑结构的识别以及表格图像分割等工作。

在第一章，分析了表格图像处理的现状，对本文所研究的范围进行了界定，并通过分析传统处理方法的不足，提出了本文的研究目标。

在第二章，阐述了表格图像预处理的流程，并重点介绍了表格图像处理区别于其他文档处理的地方—倾斜检测及校正。本文介绍了一种基于内容的文档倾斜自动校正方法，该方法综合利用了基于参考线[39]方法和基于文字行[41]方法的优点，根据文档的内容采取不同的倾斜校正方案，更具通用性。同时在进行倾斜角度检测的过程中，采用金字塔算法[46]降低图像的分辨率，以提高算法的效率。

在第三章，详细介绍普通文档图像中表格版面结构分析方法。本文所设计的方法采用了直接抽取构成表格的线条，并计算线条的交点，进而获取各个单元格的思路。该方法以线条交点矩阵表示表格版面结构分析结果，不仅降低了对问题描述的难度，更易于检索。在表格框线的定位上，该方法以现存的有向单连通链为基础，通过对其进行规则化而得到表格框线，方便有效。

在第四章，讨论了表格图像逻辑结构分析的方法。本文提出了一种以直线交点特征为基础，充分利用标题域与数据域之间的依赖关系以及基本布局结构的直

线交点特征的方法。该方法不仅能够实现对已填充表格的逻辑结构分析，而且可以将表格按照基本的布局结构进行分割。同时，本文还提出了表格中嵌套结构的描述方法。

最后是对本文工作的总结及对今后工作的展望。

第二章 表格图像预处理

图像预处理是表格文档理解的第一步，处理效果的好坏直接将影响到后续工作的完成。

本文介绍了一种基于内容的文档倾斜自动校正方法，该方法综合利用了基于参考线[39]方法和基于文字行[41]方法的优点，根据文档的内容采取不同的倾斜校正方案，更具通用性。同时在进行倾斜角度检测的过程中，采用金字塔算法[46]降低图像的分辨率，以提高算法的效率。

2.1 引言

在进行表格图像自动化处理工作之前，我们必须首先获得高质量的数字图像，然而这在表格图像处理的过程中，却是比较困难的一个环节。其主要原因在于：

1. 待扫描的纸质图像受到污染，比如纸张受潮霉变、意外的墨迹污染或者人为在表格图像上划线等，都将会影响到表格图像处理的效果。
2. 纸质图像在转化为数字图像的过程中往往会有不同程度的噪声污染或表格倾斜。表格图像扫描时，一方面由于受表格图纸本身的绘制质量、光电扫描时的光照度不均匀以及扫描系统带宽限制等因素的影响，其图像一般都夹杂着噪声和缺陷；另一方面，由于纸张边缘不平，纸张摆放不平整或者扫描仪的纠偏性能不稳定等因素，会使扫描图像存在倾斜的情况。

这些误差都将会为表格图像的自动化处理带来困难，鉴于此，表格图像的预处理工作是在整个处理流程中至关重要的一步。表格图像处理的好坏也将直接决定着整个表格图像处理分析正确率的高低。

由于表格图像区别于其他文本图像的最大特点在于绝大部分表格图像是用直线作为单元格的分界线，而一般的文本图像则通常以空白区域作为不同部分内容之间的分隔。同时，表格框线中水平线的方向可以代表表格图像的方向，这样，通过检测表格图像中水平线的方向就可以检测到表格图像的倾斜角。这就是表格图像与其他类型文本图像在倾斜角检测方法上的差异所在。

表格图像与处理主要包括去除噪声、二值化、倾斜检测与校正等内容。由于表格图像的噪声去除与二值化与其它类型的文本图像相同，故本文在此不再赘述。本章将主要将围绕表格图像的倾斜检测与校正问题进行讨论。

2.2 倾斜检测

纸质文档通过图像获取设备(如数码相机、高速扫描仪等)转化为文档图像，由于人为因素和扫描仪走纸机构的机械误差的影响，文档图像普遍存在一定的倾斜角度[36,37]。倾斜校正是一项重要的文档图像预处理技术，其基本原因在于:

1. 倾斜的文档图像影响了版面分析，使得对文字区、图形区和图像区的分割产生误差;

2. 倾斜的文档图像使字符分割发生困难;

3. 引起字符明显变形，使得 OCR 识别率降低;

4. 影响文档版面正确理解。

文档处理是建立在对版面理解基础上的，版面分析算法对文档的倾斜非常敏感。因此，对文档图像的倾斜校正就显得十分重要。倾斜校正一般分为手动校正和自动校正。手动校正，即系统提供某种人机交互手段，实现文档图像的倾斜校正。自动校正，即由计算机自动分析文档图像的版面特征，估计图像的倾斜斜角度，从而实现文档倾斜校正。由于大量的文档需要计算机来处理，倾斜图像的手动校正需要人工干预，不仅浪费了人力，而且效率很低。所以计算机自动校正成了文档处理领域的研究热点[37,38,39]。目前已经有许多的二值文档图像的倾斜校正算法，这些算法能较好的处理规范的仅包含印刷体文本的文档。由于中文文档的种类繁多、文档版面复杂，文档中包含了文本、图像和图形，以及可能包含了不规范的手写体文字和数学公式等，有的文档大部分版面是图像或者图形，扫描后的文档图像边缘可能会出现大段黑区或噪声，这些因素增加了文档图像倾斜校正难度[38]，文档图像倾斜校正是文档预处理的难点问题，许多研究者对文档倾斜校正做了大量的研究，每年都有一些新的成果发表。迄今，人们已经提出了许多不同的文档倾斜校正算法。其中，绝大多数的文档倾斜校正方法是基于规则和先验知识的，它们利用所处理文档的特点作为先验知识，使算法的精度和效率都得到提高，如基于参考线[39]、基于文字行的方法[41]等。但是，参考线法仅适用于文档存在较长参考线的场合，而文本行方法在每行字符较少时，倾斜校正效果不理想。因此这类方法在应用范围上的具有很大的局限性，不能适应目前大量复杂版面文档处理的需要。有的倾斜校正方法是与文档的先验知识无关，因而具有广泛的适应性(例如 Hough 变换的方法[40,41,42])。 这些算法的运算量很大，而且精确度得不到保证。本文介绍了一种基于内容的文档倾斜自动校正方法，该方法综合利用了基于参考线方法和基于文字行方法的优点，根据文档的内容采取不同的倾斜校正方案，更具通用性。

2.2.1 倾斜检测的基本思想

文档倾斜检测与校正基本思路是基于以下假定的：

1. 文档图像中文字行的倾斜方向与文档的倾斜方向一致；
2. 表格文档图像中的表格框线的倾斜方向与文档的倾斜方向一致。

正因为普通的文档图像均符合上述假定，因此可以通过检测文档中的线条和文字行对文档的倾斜角度进行估计，由此实现对文档倾斜角度的检测。

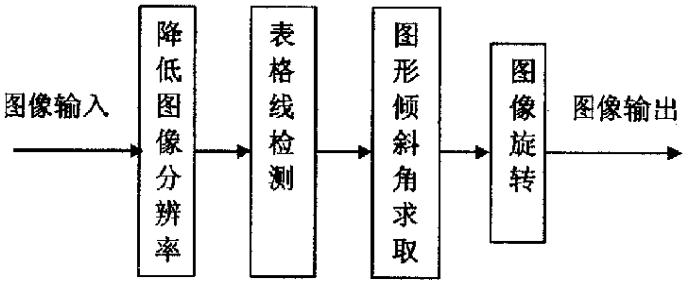


图 2.1 图像的倾斜校正流程图

2.2.2 倾斜检测的常用方法

人们提出了许多的倾斜校正算法，其中常用的方法主要有以下三大类：

1. 基于矩形块的文档图像倾斜校正方法[44,45]

矩形子块结构是构成文档版面的最小基元。这些矩形子块的上、下边为水平线条，左、右边为垂直线条。当文档图像产生倾斜时，矩形子块的边线也会产生相应的倾斜角度。通过检测矩形子块边线，然后计算出边线的方程，就可以得到文档图像是的倾斜角度。

2. 基于水平线和垂直线的文档倾斜校正方法[36,46]

水平线条和垂直线条在文档中也比较普遍(如版面基元间的分隔线等)，特别是在表格文档中(如票据、报表等)，水平线条和垂直线运用更为广泛。通过检测这些线条，计算出它的直线方程，然后算出文档的倾斜角度。

3. 基于文本行的文档倾斜校正方法[43,47]

通常文本行是沿水平方向的排列的，且相邻文本行之间的距离相对固定，因此检测页面图像的倾角不必对整个图像进行扫描计算，只需选择合适的文本子区域，其文本行的方向角对应于整个文档图像的倾斜角。

2.2.3 基于内容的表格文档图像倾斜估计

文档的版面是多种多样的，大多数文档版面中都包含水平方向或垂直方向的线条和文字行。利用线条和文字行来估计文档的倾斜角度，具有通用性好的特点。由于文字行中的字符(例如标点符号和上下标等)不完全共线，所以，采用文字行进行倾斜估计会产生较大的误差，而采用水平线和垂直线条求倾斜角度时，可以倾斜估计的误差较小。

本文设计了一种基于文档内容的表格文档校正方法，首先检测出所有水平和垂直方向上的线条，通过计算这些线条的倾斜角度来估计文档的倾斜角度。当表格文档中没有线条时，将水平和垂直的文字行通过游长平滑和细化处理，将文字行转化为水平或垂直线条。最后利用最小二乘法估计这些直线的参数，然后计算出直线倾斜角。下图为倾斜角检测的流程。

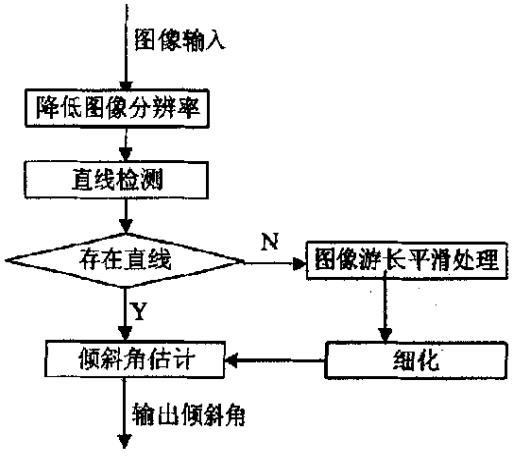


图 2.2 倾斜角检测流程

2.2.3.1 降低图像分辨率

由于直线的检测对图像分辨率的要求不高，为了减小计算量，降低图像分析时的计算复杂度，系统采用金字塔形的方块算法[46] (Pyramidal quadtree structure)降低图像的分辨率。

设变换后的低分辨率图像的尺度为 n，图像中的每一个象素点 $S^{(n)}$ 与尺度为 n-1 图像中对应的四个象素点: $S_i^{(n-1)}$ (i=1,2,3,4)的取值有关。如果 4 个象素点 $S_i^{(n-1)}$ 中有一个点是黑色，则它的子象素点 $S^{(n)}$ 也为黑色:

$$S^{(n)} = \bigcup_{i=1}^{4} S_i^{(n-1)} \tag{2.1}$$

这样 $S^{(n)}$ 中的象素点的数目为 $S^{(n-1)}$ 中象素数目的 1/4。原始图像为 $S^{(0)}$，实际采用 $S^{(1)}$ 对图像降低图像的分辨率。对于更高分辨率的扫描图像，可以采用更大的尺度以降低分辨率。如图 2.3 所示：

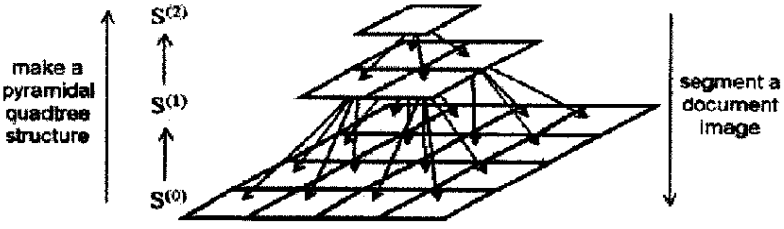


图 2.3 分辨率降低模型图[46]

2.3.3.2 直线检测

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[49]。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(r,q)空间的 q 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[50]。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[48]。

矢量化算法(vectorization)是另一类应用较广的直线检测算法。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。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。

本文采用了一种基于有向单连通链[48]的直线检测算法，此算法以“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,可以快速准确地提取直线。详细的方法如文献[48]所述，本文仅将算法的主要思想描述如下：

有向单连通链的定义

以横向单连通链为例：横向单连通链 C_h 为图像游程序列 $R_1R_2\ldots R_m$。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2.4 所示)，记为：

$$R_i(x_i, ys_i, ye_i) = \left\{(x,y)\middle| p(x,y)=1, x=x_i, y\in\middle|ys_i, ye_i\middle| \; p(x_i, ys-1)=p(x_i, ye+1)=0\right\} \tag{2.2}$$

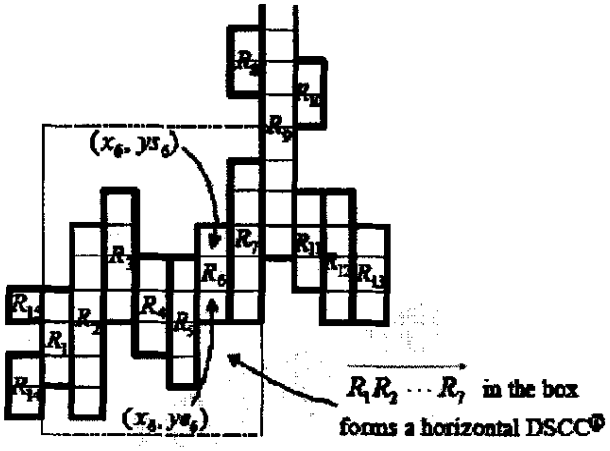


图 2.4 横向单连通链示意图[48]

其中 $p(x,y)$代表坐标(x,y)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：连通游程中每个游程的左侧或者右侧有且仅有一个游程与之连通。

基于有向单连通链的直线检测流程如图 2.5 所示：

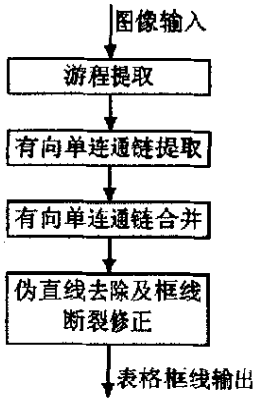


图 2.5 表格框线检测流程

有向单连通链的合并

在实际的表格图像中，根据有向单连通链的定义所提取出来的直线在水平线与垂直线相交的地方都出现了断裂的现象，这是为了使所检测到的直线具有单向连通性所采取的选择。为了使断裂的直线恢复原来的面貌，需要对有向单连通链进行合并处理。

依然以横向为例对有向单连通链进行合并，处理流程如下所示：

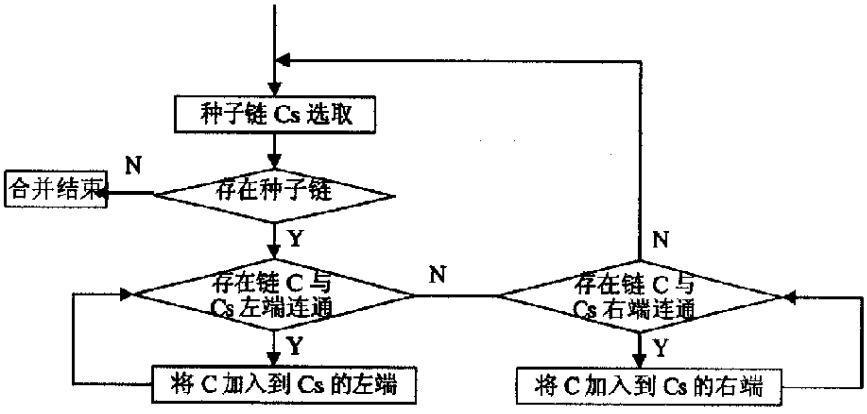


图 2.6 有向单量通链的合并

其中：

- 选择具有最多有效游程且未处理过的有向单连通链作为种子链。
- 如果两条有向单连通链的形状及相对的空间位置比较接近则认为此两有向单连通链为同一条直线。

伪直线的去除及断裂框线的补全

经过有向单连通链的合并，大部分直线都准确地提取出来了，但是还存在两类错误，一类是由于字符笔画误合并所产生的“伪”直线，另一类是直线的断裂。文献[48]引入表格框线之间的约束信息，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。最终所有未用于组成表格单元的直线都认为是伪直线而滤掉，同时断裂很严重的直线也得到一定程度上的补全。

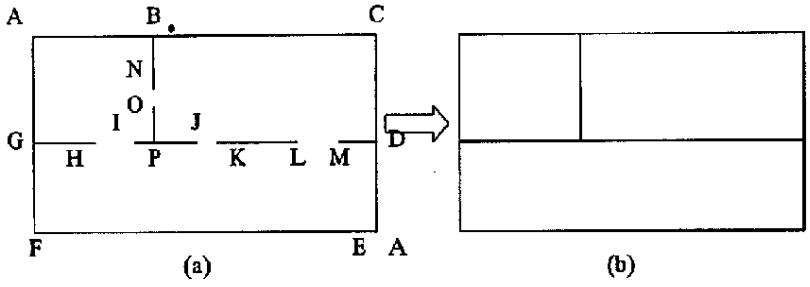


图 2.7 表格断裂补全示意图

2.2.3.3 图像游长平滑处理(Run-Length Smoothing)

通过游长平滑算法[43]对文档图像进行预处理可以将文字行合并成为同一个区域，以方便文字行的检测。游长平滑算法是将图像上长度小于某一阈值的连续白点转换成黑点的算法。图 2.8 为游长平滑示意图，图像经过游长平滑处理后，距离相近的黑点连接成了一个较大的连通成份。在中文文档中，文本以横排为主，以水平游长平滑算法为例对文档图像预处理。

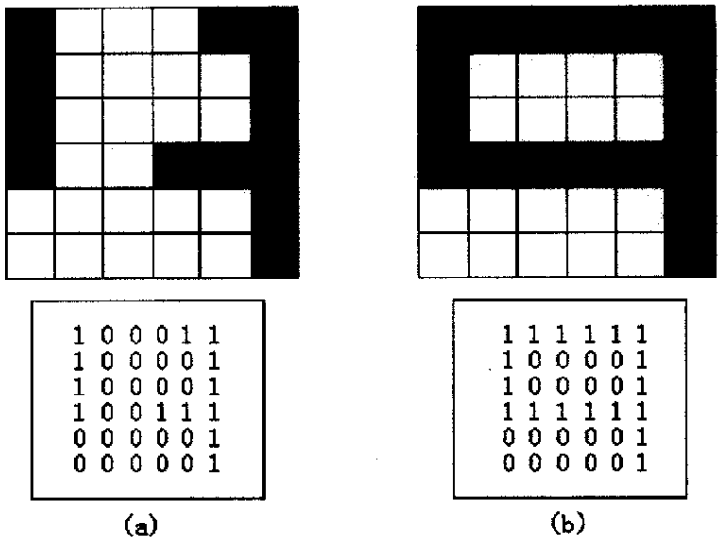


图 2.8 游长平滑示意图(阈值为 3)

文档图像经过游长平滑预处理后，文字行就连成了一个连通的区域，将该图像进行细化，图像中的文字行就变成了一条线，如图 2.9(c).

文字行经过游长平滑、细化处理后转换成了线条，线条的倾斜方向就是文档图像的倾斜方向。利用该线条计算文档图像的倾斜角度可以大大的降低计算量。一般情况下，线条的长度越长，计算出图像倾斜角度的误差就越小，所以首先采

用最长的线条计算文档图像的倾斜角度。

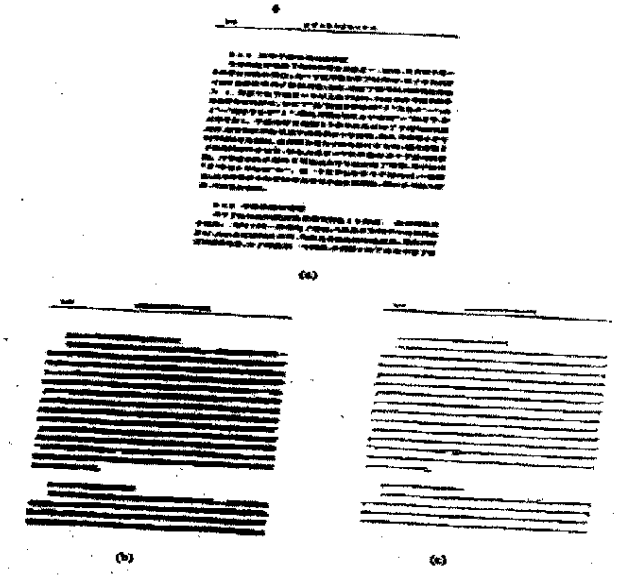


图 2.9 游长平滑及细化示例

2.2.3.4 倾角估计

文档中的水平文字行经过游长平滑等预处理也变成了水平和垂直的线条。通过投影可以找到一条最长的直线。在数字图像中，这条直线实际上是一组位于直线上的点。这些点中包含了噪声。从这组包含噪声的点中求出直线方程，就能通过直线方程求出整个文档的倾斜角度。采用线性参数的最小二乘法来估计直线方程的参数，可以得到最可信赖的直线方程，从而求得倾斜角度值。

最小二乘法原理

为了能从这组包含噪声的点拟合出最可信赖的直线方程。本文采取最小二乘法对所得到的点进行一元线性回归分析，具体方法如下：

文档中各有向单连通链中，每一游程中点坐标的 x 和 y 坐标之间形成了一组线性关系，但是由于一些随机因素的影响，使得他们之间偏离了原来的线性关系。这种偏离可以用下式来描述：

$$y_t = \beta_0 + \beta x_t + \varepsilon_t \qquad t=1,2,\cdots, N \tag{2.3}$$

其中，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_N$ 表示随机因素对这种线性关系的影响。

一般，假设 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_N$ 是一组独立的，并且服从正态分布的随机变量；而对于变量 x，假设它是可以准确测量的。于是，变量 y 就是服从 $N(\beta_0+\beta x_t,\sigma)$ 的随机

变量。设b_0,b分别是β_0,β的最小二乘估计，于是得到一元线性回归的回归方程：

$$\hat{y}=b_0+bx \tag{2.4}$$

根据最小二乘一元线性回归分析原理b_0,b，可以通过下面两个式子计算得到：

$$b=\frac{N\sum_{i=1}^{N}x_iy_i-\left(\sum_{i=1}^{N}y_i\right)\left(\sum_{i=1}^{N}y_i\right)}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2}$$
$$b=\frac{\left(\sum_{i=1}^{N}x_i^2\right)\left(\sum_{i=1}^{N}y_i\right)-\left(\sum_{i=1}^{N}x_i\right)\sum_{i=1}^{N}x_iy_i}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2} \tag{2.5}$$

而直线的倾斜角为：

$$\theta_0=\tan^{-1}(b_0) \tag{2.6}$$

这样，文档的倾斜角度为：

$$\theta=\sum_{n=1}^{n=M}\theta_n \Big/ M \tag{2.7}$$

2.2.3.5 倾角检测误差保证

设原图像的倾斜角为θ，而经倾斜检测得到的倾斜角为θ'，则倾斜校正的误差为$\Delta\theta=|\theta-\theta'|$。

为了保证文档版面理解的精度，倾斜校正后的误差应该小于一个阈值θ_T。(如图 2.10 所示)。即应满足：

$$\Delta\theta<\theta_T=\tan^{-1}\Delta d/L \tag{2.7}$$

其中△d 文本行之间的距离，L为文本行的长度。从(2.7)式可以看出：文字行的长度越长，对倾斜校正要求的精度越高。

由于文字行的长度 L 难以确定，我们直接采用文档图像的长度 W 来代替。△d 通常也取一个经验值，采用需要处理文档中文字的最小行间距△d'。

实际的倾斜校正的误差应满足：

$$\Delta\theta<\theta_T'=\tan^{-1}\Delta d'/W \tag{2.8}$$

就能正确的分割出文字图像。

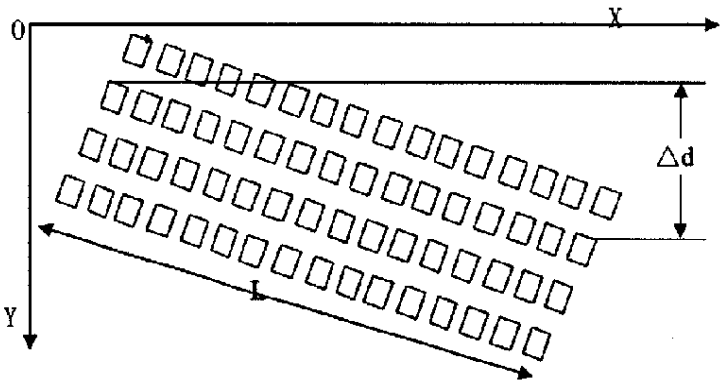


图 2.10 文档倾斜的最大允许误差

多数倾斜估计方法选用一行文字或一个局部区域来求倾斜角度，由于文挡图像容易受到噪声等干扰，倾斜估计的精度得不到保证。也有的倾斜估计方法采用整篇文档信息来估计倾斜角度，对于图像占多数的文档估计的误差也比较大。本文通过预处理，将文字行变换为水平或垂直的线条，用最小二乘法估计线条的倾斜角度和估计角度误差，如果精度满足文档处理的要求，则停止计算倾斜角度。当文档中噪声较多、文字行、线条较短、图像污损严重时，计算出的倾斜角度的误差不满足式(2.8)，则再加入其它的线条进行倾斜角度的计算。这样有效的减小倾斜估计的误差，使精度满足文档的处理要求。倾角估计流程如下如所示：

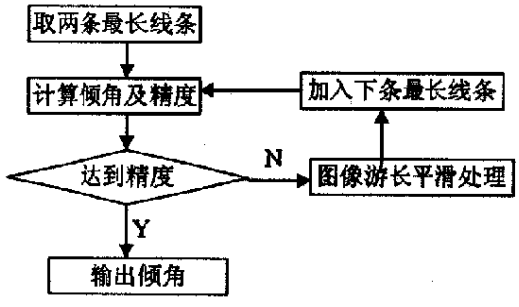


图 2.11 倾角估计流程

2.3 实验

下面以图(2.9)所示的文档为例，计算文档图像的倾斜角度。首先将文档图像进行游长平滑预处理，文字行就变成了线条，选取线条长度大于 1/2 文档图像宽度的

线条 $L_1,L_2,\cdots,L_M$ ，由式(2.6)计算这些线条的倾斜角度 $\theta_1,\theta_2,\cdots,\theta_M$ 通过式(2.7)计算文档的倾斜角度 θ。下表为实验测试结果：

表 2.1 测试结果

序号	实际角度	检测角度	误差
1	-15	-14.999 231	0.000769
2	-12	-12.001 608	0.001 608
3	-9	-9.000 441	0.000441
4	-6	-5.999 725	0.000 275
5	-3	-2.999 138	0.000 862
6	0	0.000 000	0.000 000
7	+3	+2.994 211	0.005 789
8	+6	+5.998 582	0.001 418
9	+9	+9.001 144	0.001 144
10	+12	+11.998 529	0.001 471
11	+15	+15.000073	0.000 073

注：上表中所用单位为度（°）

2.4 小结

本章阐述了表格图像预处理的方法。首先介绍了表格图像预处理的必要性；接着对表格图像的倾斜校正进行了详细的论述。由于表格图像区别于其他文本图像的最大特点在于，绝大部分表格图像是用直线作为单元格的框架线的，同时，表格框线中水平线的方向可以代表表格图像的方向，这样，通过检测表格图像中水平线的方向就可以检测到表格图像的倾斜角。在总结前人研究成果的基础上，论文设计了一种基于文档内容的文档图像倾斜校正方法，通过水平和垂直方向上的线条和文本行来对文档的倾斜角度进行估计。由于线条估计文档的倾斜角度的误差较小，所以优先考虑用线条来估计文档的倾斜角度。为了减小提取直线的计算量，本文采用金字塔形算法降低图像的分辨率。对于只有文字行的文档，采用游长平滑和细化预处理，将文字行处理成水平和垂直的直线。

第三章 表格版面结构分析

表格种类繁多、版面结构复杂，表格版面中包含许多文字识别技术不能处理的对象如线条、图形、图像等。因此，如何将表格中填写的信息从表格的背景信息中分离出来，即表格的版面理解问题，是表格自动化处理的关键。

本文所设计的方法采取直接抽取构成表格的线条，并计算线条的交点，进而获取各个单元格信息的表格版面分析思路。本文以线条交点矩阵表示表格版面结构分析结果，不仅降低了对问题描述的难度，更易于检索，这为表格图像的后续处理提供了方便。同时表格框线的提取直接以先期图像倾斜校正时得到的有向单连通链为基础，通过对其规则化而得到表格框线，该方法方便有效。

3.1 引言

版面分析是印刷体汉字识别系统的重要组成部分，与字符识别具有同等重要的地位。它是利用计算机自动地对印刷体文档图像进行分析，提取出文本、图像、图形、表格等区域，并确定其逻辑关系。这就使系统用户避免了手工画框标识文本块的繁琐操作，减少人机交互的时间，从而提高识别系统的自动化程度和输入效率。作为汉字识别的预处理过程，正确合理的版面分析结果是后续版面识别工作的必要条件。因此，研究能够适应各种版面特点的通用版面分析方法，具有十分重要的意义。

目前大多数的表格处理系统中，通过在表格版面上添加定位标记[55,56]，来实现表格图像的倾斜校正和表格定位。该方法只能处理特定的表格，对纸张质量的要求高，不利于处理系统的推广应用。丁晓青和吴右寿[50]等人采用基于投影的方法检测表格线的位置，用寻找角点的办法来定位填写区域的位置。Fan[57]等人采用细化后提取特征点的方法，但细化算法会产生畸变，而且没有好方法将字符中和表格线上的特征点区分开。还有一些表格处理采用提取所有水平和垂直的表格线方法定位实现表格定位[58]。Yu 和 Jain[59]用块相邻图抽取表格框架，这些方法都没有考虑表格线断线的情况。日常生活中，表格的印刷误差、图像扫描设备机械误差、光学误差的影响下，表格图像不可避免的会产生水平方向和垂直方向上的畸变，而且噪声较多，对于这些图像质量较差的文档，还没有一种较为通用的表格的处理方法。

3.2 表格版面结构分析的目标

表格版面结构分析是为了提取表格中各种基本基元的信息，并以合适的数据结构进行存储为后续的版面结构检索工作提供方便。对表格版面结构进行自动分析，需达到以下处理目标:

1. 获得完整、详细的表格版面结构描述.通过版面结构分析，能够得到表格结构内部所有数据的排列、组织关系。版面结构的分析结果将直接作为表格逻辑结构分析的依据信息，结合各种预定义的先验知识和启发式规则，对表格内部的数据进行抽取、分类、存储、检索。

2. 修正、补偿图像中各种噪音造成的数据分隔符断裂、缺失错误。保证表格版面结构分析结果的完备性和有效性。

3. 通过版面结构自动分析，能够获取完整的全局版面特征和局部版面信息。

3.3 表格版面结构分析的思路

对表格的分析已经有较多可行的方法，其主要思路大概分为以下两类:

1. 认为表格是其内部数据逻辑关系的一种可视化图形表示，通过抽取组成表格的线条、分析各个单元格的位置、大小及单元格中字符串的版面信息，能够得到完整的表格版面结构并进一步进行逻辑结构分析。
2. 直接抽取构成表格的线条，并计算线条的交点，进而获取各个单元格，并将单元格作为逻辑结构进行分析处理。本文所采取的就是这种思路。

3.4 表格版面结构分析的难点

表格版面结构的定位与分析系统是文档图像分析系统的重要部分。由于表格版面结构本身的复杂性，图像中表格结构的定位与分析是一项很困难的工作，有大量的文献说明了这方面的研究工作从二十世纪九十年代之前就开始了，并且是一个热点[38]。造成表格版面结构的定位与分析困难的原因很多，主要有:

第一，表格版面结构的分析是基于图像的分析。

通过前面的分析可知，表格的定位与分析是在表格图像的预处理之后进行的。同普通的图像一样，表格文档图像也有各种噪声。这是由于文档在打印、扫描等过程中，不可避免地会使得文档图像的质量受到影响。有些表格就变得不是很清楚，有些线条不清楚，甚至出现缺失、变形等现象。这些都会使得表格的结构发生变化。同时，表格的定位与分析的对象是一个没有任何先验知识与文本信息的数字图像。换句话说，获得表格结构中的文本信息是在定位出数据区域，并进行

OCR 识别之后的事，所以，在这之前的表格结构的定位与分析并不能从文本信息中获得帮助。

第二，表格结构的复杂性。

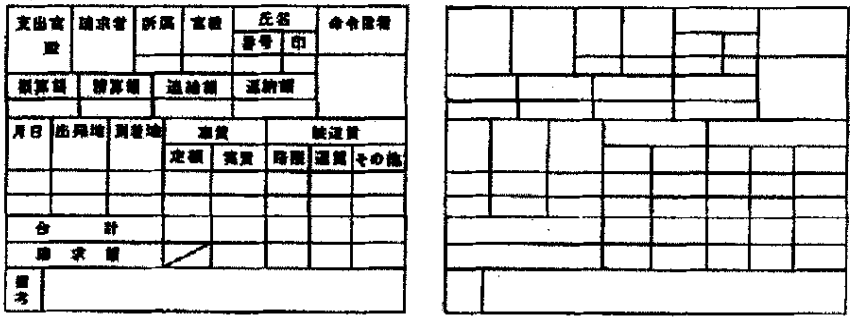
由于表格结构的规范化和表格的排版规则，处理系统会对预处理结果中某些线条的位置进行调整。而且，考虑到数据区域中的文本和单元格边框之间的间距，所以正确结果中的线条存在一个合理浮动范围，表格定位与分析系统的分析结果只要在正确结果的浮动范围之内，就认为是正确的。这显然给表格结构的分析增加了一定的难度。

3.5 表格版面结构分析

3.5.1 表格框线的定义及特征

3.5.1.1 数据分割符

数据分隔符：分割表格内部数据的版面元素，表现为表格图像中的水平线条、竖直线条和分割行、列的空白区域。在表格结构中，线条有着不同的作用，只有分割数据区域和围绕表格边界的线条才能被称为数据分隔符，作为表格版面结构分析的依据。



(a) (b)

图 3.1 表格分隔符 其中(a)为原始图像，(b)为(a)的表格分割符

3.5.1.2 框架线

框架线：对表格内部数据区域 R，贯通 R 边界的数据分隔符被称为 R 的框架线。所谓贯通，是指数据分隔符的封闭端点位于 R 的边界。

在对表格版面结构进行图像信息抽取与分析的过程中，针对数据分隔符采取

如下处理原则：

- 所有的数据分隔符都是框架线。对任何一个数据分隔符，都能够找到一个区域，使其在这个区域内部成为框架线。依据这一原则，能够将表格版面结构自顶向下分为多层子区域，直到包含每一个单元格为止。
- 所有的数据分隔符都必须拥有封闭端点。依据这一原则，能够对由于图像质量低下和噪音干扰引起的线条断裂、扭曲、残缺进行修正。

3.5.1.3 表格框线特征

经过对常见表格类型的分析，发现表格框架线具有以下特征：

- 表格中的框线必须具有封闭的端点，并能够作为某一区域的框架线。所有的表格框线都必须作为某一数据区域的框架线。通过抽取框架线，能够对框线赋以逻辑属性，进而形成更高的逻辑元素—单元格。这使得自顶向下的表格版面结构分析成为一个可计算的处理过程。
- 任何结构复杂的表格都可以由单元格通过一定的逻辑结构组织而成。根据表格结构描述模型，表格能够表示为单元格集合，而基于表格框线全局特征对单元格的组织反映了内部数据的全局分割关系。根据这种处理思想，能够对表格中的版面信息进行自顶向下的组织、分析。
- 表格框线具有全局版面特征和局部版面特征，不仅相交、围绕形成各个单元格的局部边界；同时还承担着子表分割、行列分割等全局数据分割作用。
- 在表格中，所有表格框线均具有封闭端点。基于这一原则，能够正确的剔除表格中的非框线线条，保证表格结构分析的准确。

3.5.2 表格框线提取

根据表格框线的特征，求出表格的水平线和垂直线，并对直线中的断裂、缺失情况进行修复，从而得到完整、有效的表格区域。

抽取图像中的黑像素线条已经是非常成熟的图象处理技术，常见的方法包括 Hough 变换、RunLength 编码、区域投影等[34,53,54] 。本文采用直接以先期图像倾斜校正时得到的有向单连通链为基础，进行表格框线的提取。

3.5.2.1 表格框线的检测

经过表格图像的预处理，表格图像中的线条已被矢量化，并存储在有向单连

通链中。但是，经过图像的旋转后，有向单连通链与表格框线之间就失去了匹配关系。据此本文设计了一种基于有向单连通链的表格框线规则化算法，通过将倾斜文档中的单向连通链作进一步的旋转，使之与校正过的文档相匹配，从而得到图像旋转后的表格框线。

3.5.2.1.1 有向单连通链的旋转

设存在一有向单连通链 $DSCC=\{R_i(x_i,ys_i,ye_i),i=1,2,\cdots,N\}$，其与倾斜校正前的表格框线相匹配。但经过倾斜校正，DSCC 与表格框线失去了相匹配的关系，为此，需要对 DSCC 进行规则化，使之与旋转后的表格框线相匹配。记旋转后的有向单连通链为 DSCC'

设有向单连通链中 DSCC 中游程 $R_i(x_i,ys_i,ye_i)$ 的起始点为 (x_i,ys_i)，终止点为 (x_i,ye_i)。经倾斜检测得到的倾斜角为 θ，则旋转后的起始点为 (x_i',ys_i')，终止点为 (x_i',ye_i')，可以通过下式计算得到：

$$\begin{bmatrix} x_i' \\ ys_i' \\ 1 \end{bmatrix}=\begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x_i \\ ys_i \\ 1 \end{bmatrix} \tag{3.1}$$

$$\begin{bmatrix} x_i' \\ ye_i' \\ 1 \end{bmatrix}=\begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x_i \\ ye_i \\ 1 \end{bmatrix} \tag{3.2}$$

有向单连通链的长度是表格框线在水平线上的投影长度，当表格框线经过倾斜校正后，与之相匹配的有向单连通链的长度将会变长（如图所示）。

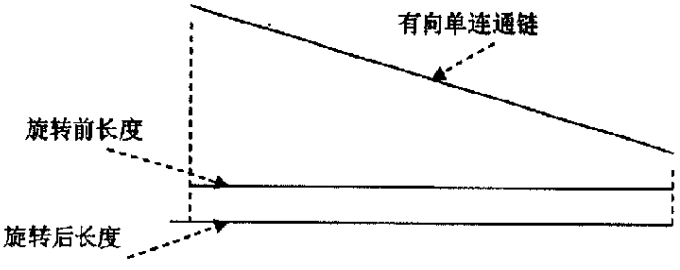


图 3.2 旋转前后有向单连通链长度的变化

DSCC 中 $R_i(x_i,ys_i,ye_i)$ 的起始坐标为 (x_i,ys_i)，通过公式(3.1)转换得 (x_i',ys_i')。同时，DSCC 中 $R_k(x_k,ys_k,ye_k)$ 的终止坐标为 (x_k,ye_k)，通过公式(3.2)转换得 (x_k',ye_k')。

如果$x_k'=x_i'$，则说明点(x_i',ys_i')和点(x_k',ye_k')在同一垂直线上，此时称起始坐标(x_i,ys_i)与终止坐标(x_k,ye_k)相匹配。

如果起始坐标(x_i,ys_i)与终止坐标(x_k,ye_k)相匹配，则根据点(x_i',ys_i')和点(x_k',ye_k')生成一个新的游程。

如果DSCC中所有游程的终止点坐标均不与起始点坐标(x_i,ys_i)相匹配，则直接根据点(x_i',ys_i')和点$(x_i',ys_i'+W_{DSCC})$生成一个新的游程，其中W_{DSCC}为DSCC的平均宽度。

如果DSCC中游程的终止点坐标(x_k,ye_k)不存在于之相匹配的起始点坐标，则直接根据点$(x_k',ye_k'-W_{DSCC})$和(x_k',ye_k')点生成一个新的游程。

对所得到的新游程，按照 x 坐标升序进行排列所得到的游程序列就组成了一有向单连通链。

然而，经过旋转后的单向连通链内部仍包含一定的随机噪声，如图 3.3(a)所示。为此，需要对此连通链进行规则化变换，使之成为一外边框为矩形的规则连通链。下面以横向单向连通链为例进行规则化处理。

3.5.2.1.2 有向单连通链的规则化

设单向连通链为 DSCC，其所包含的游程组成一垂直游程集合 RLSet，RLSet=$\{R_i(x_i,ys_i,ye_i)\}$ i=1,2,…,N。提取集合中各游程的起始坐标ys_i和终止坐标ye_i，分别组成起始坐标序列$\{ys_1,ys_2,\cdots,ys_N\}$和终止坐标序列$\{ye_1,ye_2,\cdots,ye_N\}$。设由 DSCC 转换后的规则单向连通链的外边界为：上边界 Top，下边界 Down，左边界 Left，右边界 Right。

规则单向连通链的上下边界坐标的决定取决于起始坐标序列和终止坐标序列，这里分别用序列$\{ys_1,ys_2,\cdots,ys_N\}$的均值和序列$\{ye_1,ye_2,\cdots,ye_N\}$的均值作为规则单向连通链的上下边界坐标。

规则单向连通链的左右边界坐标的决定取决于$R_1(x_1,ys_1,ye_1)$和$R_N(x_N,ys_N,ye_N)$的水平坐标x_1、x_N。

则：

$$\text{Top}=\frac{1}{N}\sum_{i=1}^{N}ys_i \quad \text{Down}=\frac{1}{N}\sum_{i=1}^{N}ye_i$$
$$\text{Left}=x_1 \qquad \text{Right}=x_N \tag{3.3}$$

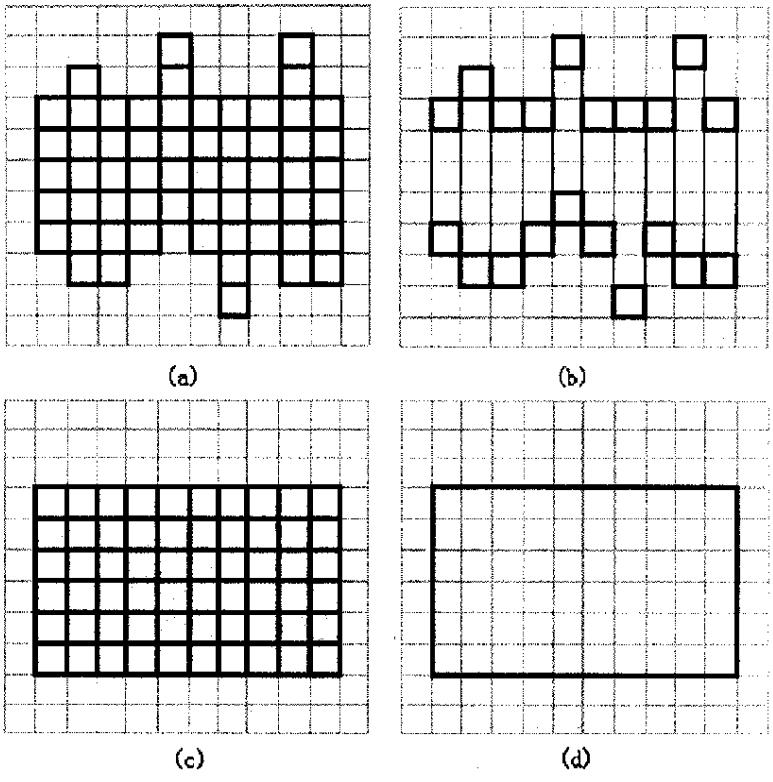


图 3.3 单向连通链规则化（以横向为例）

其中图 3.3(a)为带有噪声的有向单连通链，图 3.3(b)为从有向单连通链中提取出的起始坐标序列和终止坐标序列，图 3.3(c)为规则化后的有向单连通链，图 3.3(d)为规则化后有向单连通链的外边框。

3.5.2.1.3 表格框线(Frame Line)的表示

通过有向单连通链规则化后，每个有向单连通链的外边框内部所包含的黑像素就构成了一条表格框线。本文采用表格框线的左上角点和右下角点的坐标来对表格框线进行表示。

```
FrameLine=
{
TopLeftPoint; //表格框线的左上角点;
DownRightPoint; //表格框线的右下角点;
}
```

图 3.4 表格框线的表示

如上图所示，表格框线 AB，设左上角顶点坐标为(xl,yu)，右下角坐标为(xr,yd)。那么表格框线 AB 就表示为((xl,yu), (xr,yd))。

3.5.3 线条交点阵的生成

为体现表格版面结构与表格逻辑结构之间的独立性，同时便于表格版面结构的存储、重现、格式转换。我们引入了线条交点阵及交点类型阵，其中保存了表格版面结构的全局与局部信息，同时为后继的逻辑结构分析提供了规范化的基本信息。

线条交点(Intersection Point)：表格版面结构中水平、竖直数据分隔符相互交会、贯通的位置，称为线条交点。线条交点反映了数据分隔符的相交情况，并记录了数据分隔符的全局和局部版面结构特征。

表格版面结构分析结束后，表格被拆解成多个基本简单区域。由于直接依据框架线信息对结构进行分解，而线条信息的残缺、断裂会使得一些线条在处理过程中只能被修正，而不能作为结构分解的依据。这样的分析结果完全依据图像信息，无法作为表格版面结构表示和表格逻辑结构分析的有序输入信息(图像质量影响分解顺序，因此，简单区域集合是无序的)。

所以，必须对表格版面结构分析结果进行重新组织与整理。基于人的直观认识，使用数据分隔符、Block 信息即能够表示表格版面结构信息。但是，数据分隔符具有的全局、局部版面特征使得对表格版面结构的表示过于笼统。而且，不利于检索(复杂表格结构中，很难确定数据分隔符的排序顺序)。

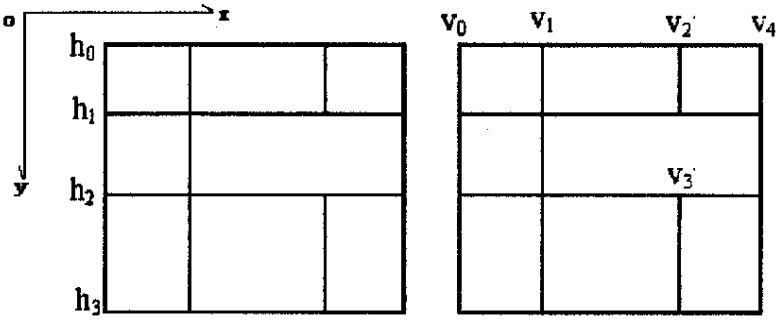
基于以上分析，我们使用线条交点来记录数据分隔符的全局、局部版面特征，使用线条交点矩阵表示表格版面结构分析结果。以点代替线，降低了对问题描述的难度。更为重要的是，线条交点矩阵体现了表格版面结构规范性的全局行列特征，而且易于检索，为更高层次的处理、应用提供了格式良好的数据。

线条交点矩阵生成的过程就是对表格版面结构分析结果整理的过程，将所有被切分的线段合并成线条，根据规则计算线条交点的位置和属性。

3.5.3.1 线条排列及行列数估计

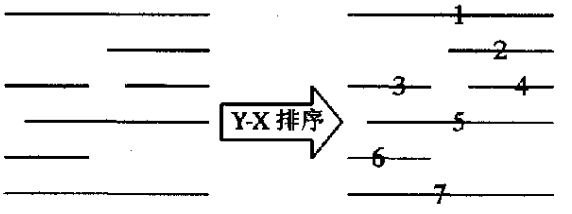
我们分别用 hi 和 vi 表示表格中的水平线和垂直线，为了定义水平线序列 H 和垂直线序列 V，需要将水平线按 Y-X 排列，同时将垂直线按 X-Y 排列（Y-X 排序及 X-Y 排序的示意图如图 3.6 所示）。H、V 的定义如下：

$$H=\{h_0,h_1,\cdots,h_i,\cdots,h_{n_{H-1}}\},y_i\le y_j \ \ if \ \ i<j;$$
$$V=\{v_0,v_1,\cdots,v_i,\cdots,v_{n_{V-1}}\},x_i\le x_j \ \ if \ \ i<j; \tag{3.4}$$

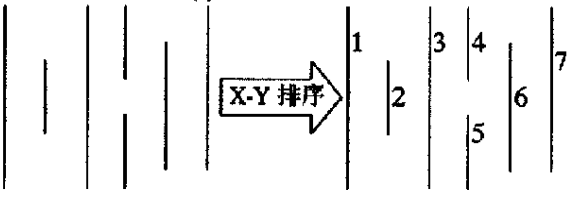


水平框线　　垂直框线

图 3.5 表格框线排序结果



(a)水平线 Y-X 排序



(b)垂直线 X-Y 排序

图 3.6 线条排序示意

为了确定线条交点阵的行列数，需要对经过排序的线条序列 H、V 进行行列数判断。这里依据动态阈值对行列数进行判断，取阈值为最小文本行宽度，记为

MinTextSize，线条之间的距离记为 Dis，当行列之间的距离大于阈值时，行列数加 1。判断示意图如图 3.7 所示。

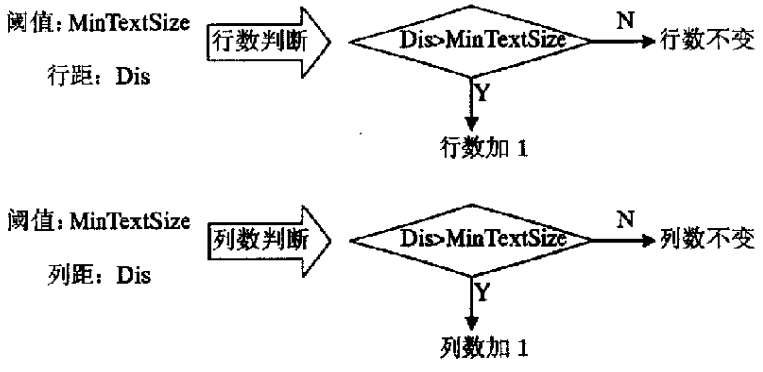


图 3.7 行列数估算示意

经过行列数判断， 就可以确定待生成的线条交点阵的行列数。

3.5.3.2 交点阵的生成

本文所采用的表格框线是具有一定宽度的，而表格框线的交点是单像素的，为此本文在线条交点的选择上以两条相交表格框线的中线交点为准。表格框线交点的计算如下图所示：

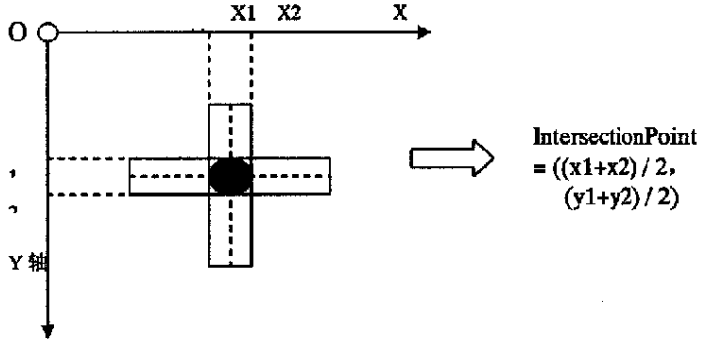


图 3.8 交点坐标的计算

如下图所示，表格的版面结构存在三类九种线条交点。

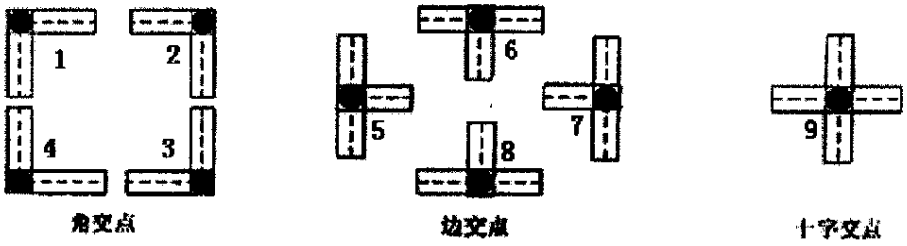


图 3.9 交点的类型及类型编码

交点类型编码的具有以下属性：

表 3.1 交点类型编码属性

类型编码	属性
0	此处不存在交点
1	交点处向右、向下连通
2	交点处向左、向下连通
3	交点处向左、向上连通
4	交点处向右、向上连通
5	交点处向上、向右、向下连通
6	交点处向左、向右、向下连通
7	交点处向上、向左、向下连通
8	交点处向左、向右、向上连通
9	交点处各方向均连通

这样，就可以使用线条交点在图像上的版面位置坐标及线条交点类型来对表格框线交点的属性进行描述：

Intersection Point =

{

Point=(x,y);//线条交点的版面位置坐标。

Type;//线条交点类型。

}

其中

- Type 为线条交点类型，记录了表格框线在水平方向和垂直方向上的版面特征。
- Point 为线条交点的版面坐标，当交点的类型编码为 0 时，Point 的值设为 (-1,-1)。

3.5.3.3 交点类型的判断

设存在表格框架线 h 和 v，分别表示为$((x_{hs},y_{hs}),(x_{he},y_{he}))$和$((x_{vs},y_{vs}),(x_{ve},y_{ve}))$。

当 $x_{hs}=x_{vs}$ 且 $y_{hs}=y_{vs}$，交点类型为 1；

当 $x_{he}=x_{ve}$ 且 $y_{hs}=y_{vs}$，交点类型为 2；

当 $x_{he}=x_{ve}$ 且 $y_{he}=y_{ve}$，交点类型为 3；

当 $x_{hs}=x_{vs}$ 且 $y_{he}=y_{ve}$，交点类型为 4；

当 $x_{hs}=x_{vs}$，$y_{hs}>y_{vs}$，且 $y_{he}<y_{ve}$ 交点类型为 5；

当 $y_{hs}=y_{vs}$，$x_{hs}<x_{vs}$，且 $x_{he}>x_{ve}$ 交点类型为 6；

当 $x_{he}=x_{ve}$，$y_{hs}>y_{vs}$，且 $y_{he}<y_{ve}$ 交点类型为 7；

当 $y_{he}=y_{ve}$，$x_{hs}<x_{vs}$，且 $x_{he}>x_{ve}$ 交点类型为 8；

当 $y_{hs}>y_{vs}$，$y_{he}<y_{ve}$，$y_{he}=y_{ve}$，$x_{hs}<x_{vs}$，且 $x_{he}>x_{ve}$ 交点类型为 9；

3.5.3.4 交点类型的错误检测及纠正

当文档图像的质量比较差时，往往会造成表格框线的断裂。表格框线的断裂分为两种情况：交点处的断裂和非交点处的断裂。

对于非交点处的表格框线断裂，并不会影响到线条交点阵，也就是说并不会影响到表格版面结构的分析。这是因为此种表格框线的断裂可以根据交点类型进行恢复。

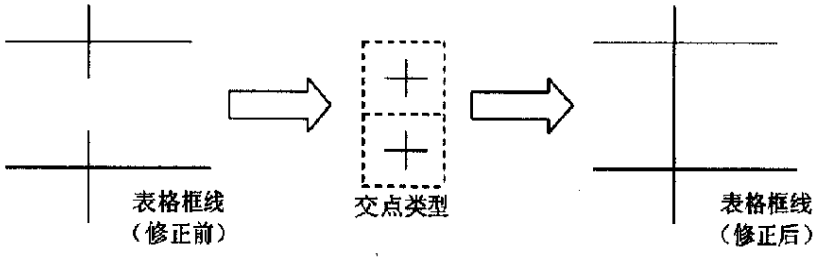


图 3.10 基于交点类型的表格框线恢复示例（非交点处断裂）

对于交点处断裂，需要根据实际情况分析，交点类型识别错误的位置并进行相应的修正（如图 3.13 所示）。几种常见的交点焦点错误类型如图 3.1 所示，关于如何对交点类型进行错误检测及纠正，在文献[62]中有详细论述，在此不作赘述。

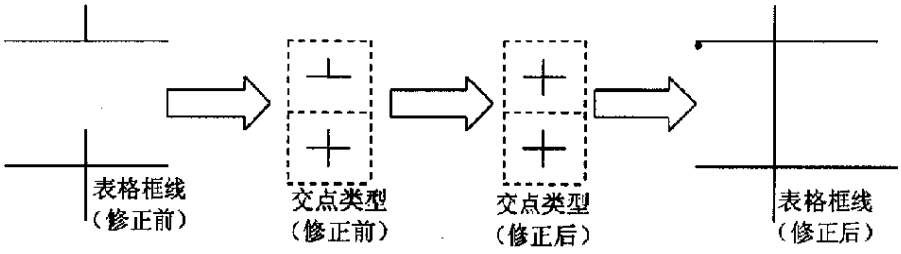


图 3.11 基于交点类型的表格框线恢复示例（交点处断裂）

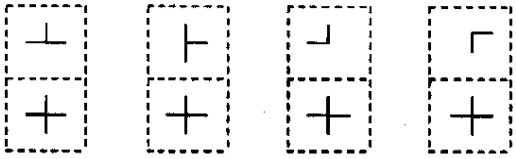


图 3.12 几种常见的交点类型错误（以垂直方向为例）

3.5.3.5 线条交点阵的生成

线条交点阵记为 PtMatrix，线条交点阵的行列数分别记为 RowNum 和 ColNum，表格框线的行列数分别记为 HNum 和 VNum。正如上文所述，线条交点阵的行列数与表格框线的行列数相对应，则：

$$\begin{aligned} RowNum &= HNum \\ ColNum &= VNum \end{aligned} \tag{3.5}$$

线条交点阵的矩阵表达形式为：

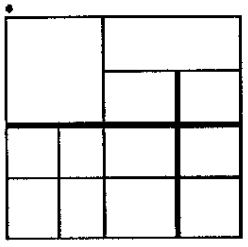
$$PtMatrix=\begin{bmatrix} pt_{1,1} & pt_{1,2} & \cdots & pt_{1,ColNum} \\ pt_{2,1} & pt_{2,1} & \cdots & pt_{2,ColNum} \\ \vdots & \vdots & \ddots & \vdots \\ pt_{RowNum,1} & pt_{RowNum,1} & \cdots & pt_{RowNum,1} \end{bmatrix} \tag{3.6}$$

其中，$pt_{i,j}$ 为第 i 行水平线与第 j 列垂直线的交点。

3.5.3.6 交点类型阵的生成

为了方便后续工作的数学计算，本文将交点的类型属性单独提取出来，并以矩阵的形式描述。可见交点类型阵是专门用来存储交点类型的矩阵，交点类型阵是线条交点阵的一个子集。

通过交点类型阵，可以唯一的确定表格的结构，同时同一个表格唯一确定一个交点阵。交点阵与表格之间是一一对应关系。下图为一表格的交点类型阵。



1	0	6	0	2
0	0	5	6	7
5	6	9	9	7
5	9	9	9	7
4	8	8	8	3

(a) (b)

图 3.13 交点类型阵

记交点类型阵为 PtTypeMatrix。交点类型阵的矩阵表达形式为：

$$PtTypeMatrix=\begin{bmatrix} type_{1,1} & type_{1,2} & \cdots & type_{1,ColNum} \\ type_{2,1} & type_{2,2} & \cdots & type_{2,ColNum} \\ \vdots & \vdots & \ddots & \vdots \\ type_{RowNum,1} & type_{RowNum,2} & \cdots & type_{RowNum,ColNum} \end{bmatrix}_{m\times n} \tag{3.7}$$

其中 $type_{i,j}$ 为第 i 水平线与第 j 垂直线的交点类型。当 $type_{i,j}$ 为 0 时，表示此处没有交点。而当 M(i,j)为非 0 时，表示第 i 水平线与第 j 垂直线存在一类型编码为 $type_{i,j}$ 的交点。

3.5.4 单元格信息获取

单元格的定位是表格图像自动化处理流程中非常重要的一步，其定位的好坏是决定信息是否可以正确提取的关键。本文设计了基于线条交点阵的单元格信息获取方法。

具体方法描述如下：

Step1 初始化所有点的状态为未处理；

Step2 从 PtMatrix 中选择第一个状态为未处理的点 p；

Step3 沿 p 点向右寻找与 p 相邻接且未处理过的点 p1；如果找不到，则记点 p 为已处理，转向 Step2)；

Step4 沿 p 点向下寻找与 p 相邻接且未处理过的点 p2；如果找不到，则记点 p 为已处理，转向 Step2)；

Step5 如果沿 p1 向下与沿 p2 向右均可以遍历到点 p3，并且点 p3 的状态为未处理，则出直线 pp1、pp2、p1p3、p2p3 所围成的矩形区域就构成了一个单元格，记录之；并标记点 p 为已处理，转向 Step2)；

直至 PtMatrix 中点的状态均为已处理时，就完成了单元格的抽取工作。实践

证明，此方法是有效的。

3.6 小结

本章介绍了一种表格版面结构分析的方法，此方法采取直接抽取构成表格的线条，并计算线条的交点，进而获取各个单元格信息的思路。具有以下特点：

- 表格框线的提取直接以先期图像倾斜校正时得到的有向单连通链为基础，方便快捷。
- 以线条交点矩阵表示表格版面结构分析结果。以点代替线，不仅降低了对问题描述的难度，体现了表格版面结构规范性的全局行列特征，更易于检索，为更高层次的处理、应用提供了格式良好的数据。

第四章 表格逻辑结构分析

经过表格图像预处理后，文档图像中所有表格中的交点被获取，并且得到反映交点类型情况的交点阵。在此基础上，能够通过交点阵对表格进行进一步的结构分析，判断出表格的逻辑结构。这种方法，既对版面结构信息进行了良好的组织(易于检索和重现)。同时，又将其抽象为更高层次的表示形式，便于更高处理级别(可编辑、可检索、可理解)的分析与处理。

本文提出了一种基于直线交点特征的自底而上表格图像分割方法。该方法充分利用了标题域与数据域之间的依赖关系以及基本布局结构的直线交点特征，不仅能够实现对已填充表格的逻辑结构分析，而且可以将表格按照基本的布局结构进行分割。同时本文还提出了表格中嵌套结构的描述方法。实验证明，本文所述的表格图像分割方法是有效的。

4.1 引言

近几年来，国内外已提出了许多关于表格文件图像分析的方法，但其中关于表格逻辑结构分析的方法却很少。其中表格图像逻辑结构分析的方法大致可以分为两类。第一类以空白表格为处理对象，以表格的规范性、矩形性为前提，通过分析表格中标题域与数据域之间的依赖关系来对表格进行逻辑分析[15][63]。此类处理方法，将数据域与标题域之间的依赖关系分为 4 种，依此将表格的基本布局结构划分为 11 种简单的布局结构[15]。通过根据标题域与数据域之间的依赖关系实现对表格中基本布局机构的识别。此类方法的优点在于能对表格按照具体内容进行分割，将同一个基本布局结构中的单元格划分在一起，缺点在于，此类算法仅适用于对空白表格的处理，而对已填充数据的表格之逻辑结构无法进行分析。另一类方法主要通过单元格之间的逻辑位置关系来对表格的逻辑结构进行表达[13]。此方法由于没有考虑到单元格之间的依赖关系，使得表格的逻辑划分不能直接体现出表格的基本布局特征和表格之间的嵌套关系。但是此类方法却具有不受表格内容限制的特征。

是否存在一种既能将表格划分为基本布局结构的组合，又能适用于已填充表格的方法呢？本文从直线的交点阵出发，以标题域与数据域之间的依赖关系和基本布局结构的直线交点特征为基础，尝试对此问题进行解决。

4.2 表格逻辑结构分析目标

表格逻辑结构反映了表格数据的类别与内在逻辑联系，是位于表格版面结构之上的深层结构。表格版面结构分析的目的是为了获取表格深层的逻辑结构，同时，根据先验的逻辑结构知识，能够对表格版面结构进行更精确的分析与处理。对表格逻辑结构进行自动分析，需要达到以下处理目标：

- 获得完整、详细的逻辑结构描述。通过逻辑结构分析，能够得到表格结构内部所有数据的排列、组织关系。版面结构的分析结果将直接作为表格逻辑结构分析的依据信息，结合各种与定义的先验知识，对表格内部的数据进行抽取、分类、存储、检索。
- 根据逻辑结构分析过程中得到的信息，对表格中子表类型的判别。通过逻辑结构自动分析能够获得完整的全局版面特征和局部版面信息，基于这些信息设计判别机制，能够剔除类似于表格的非表格结构，不仅保持了非表格结构的完整，同时，提高了逻辑结构分析的可信度。
- 描述表格版面的数据结构应合理。表格版面结构的描述是表格逻辑结构识别的基础，好的表格逻辑结构描述应该能够便于表格结构的检索、匹配。

4.3 表格逻辑结构与表格版面结构之间的关系

作为数据组织与表达形式，表格图像中可显示的版面结构与表达数据关系的深层次逻辑结构之间存在着对应关系。但是，这种对应关系是不确定的，同一种排版方式能够使用与不同类型数据的组织。同一数据组织关系能够以不同的表格版面形式表达出来。如图 4.1(a)、图 4.1(b)表示的是同一种数据组织关系，却有完全不同的版面结构；图 4.1(c)、图 4.1(d)对应同样的逻辑结构，而版面结构却存在差异。

表格版面结构具有固定的二维行列特征，而表格逻辑结构却存在一维、二维甚至多维特征。表格逻辑结构的维度是由内部标题域与数据域之间的映射关系确定的。

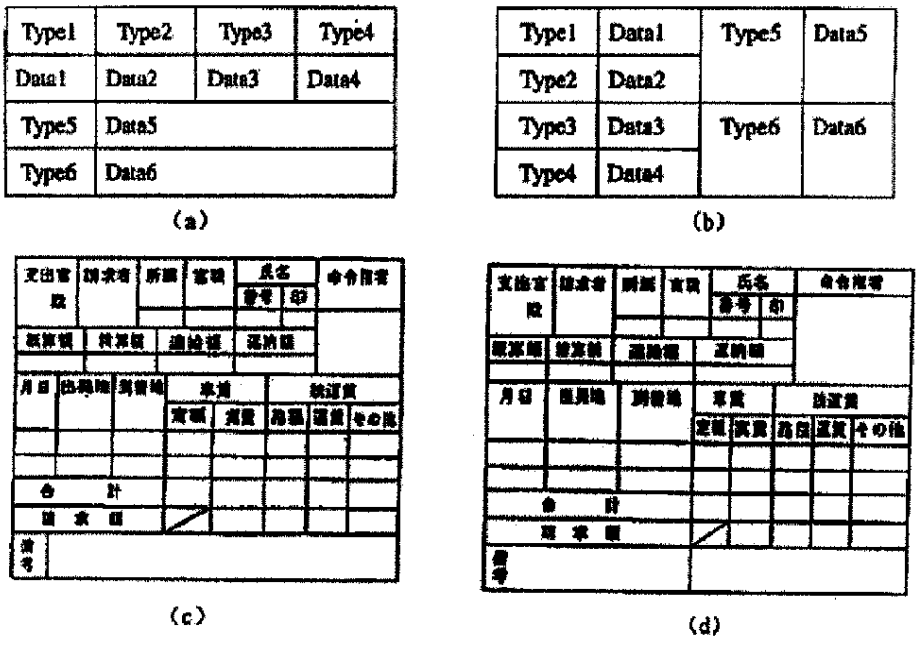


图 4.1 表格逻辑结构与版面结构之间的不稳定关系

4.4 基于直线交点特征的表格结构分析方法

4.4.1 基础理论及假定

表格在布局结构上有以下两个基本特征[63]：

1. 标题域的主导性。一个标题域的所属子表（或所属数据域）总是构成一矩形区域，此标题域与此矩形区域之间的相对位置关系是直接相邻的左右或上下关系，或者说一标题域的所属子表（或数据域）直接位于此标题域的下方或右方，并与此标题域共同构成一个矩形区域（如图 4.2 所示）。称表格的这一布局结构特点为“标题域的主导性”。
2. 标题的矩形性。一个规范表由数据域和标题域组成，并共同构成一个矩形区域，在基本表中，其数据域一般也聚集在一起，且相互之间直接邻接，也构成一个矩形区域（如图 4.3 所示）。 本文将表格的这两个矩形性特征统称为“规范表的矩形性”。 因而由直接相邻接的数据域组成，并构成局部最大的矩形区域称为矩形数据区。 这也就是数据域的聚集性。

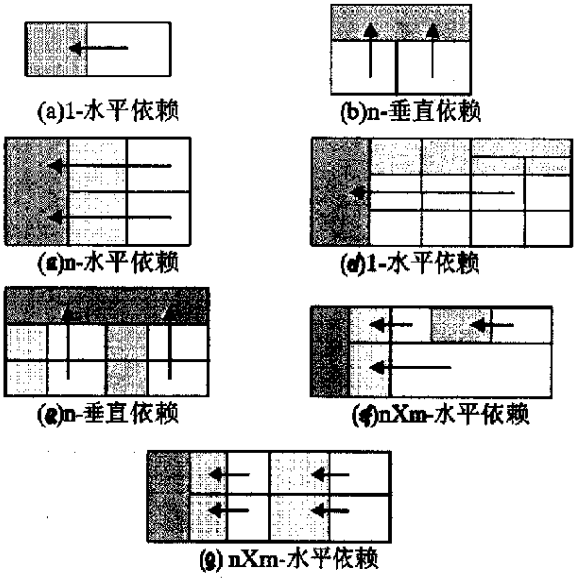


图4.2 标题域的主导性

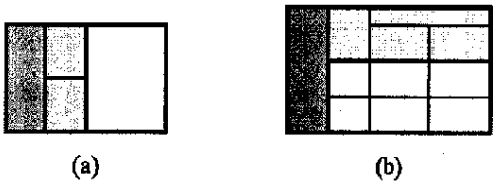


图4.3 数据域的聚集性

4.4.1.1 标题域与数据域之间的依赖关系

表格中数据域与标题域之间存在以下依赖关系：1-水平依赖关系、1-垂直依赖关系、N-水平依赖关系、N-垂直依赖关系，并根据此 4 类依赖关系将表格的布局结构划分为 11 种基本的布局结构[1]（如图 4.4 所示）。如果不考虑单元格的属性(标题域或是数据域)，我们可以将这 11 种基本布局结构归纳为图 2 中的 6 种布局（如图 4.5 所示）。

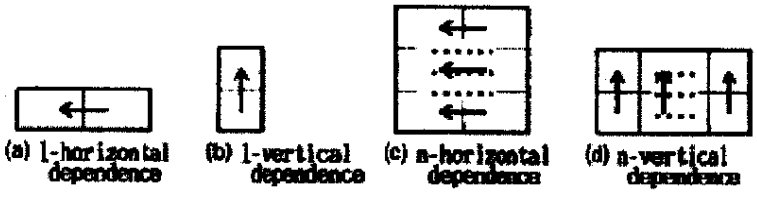


图 4.4 单元格之间的依赖关系[15]

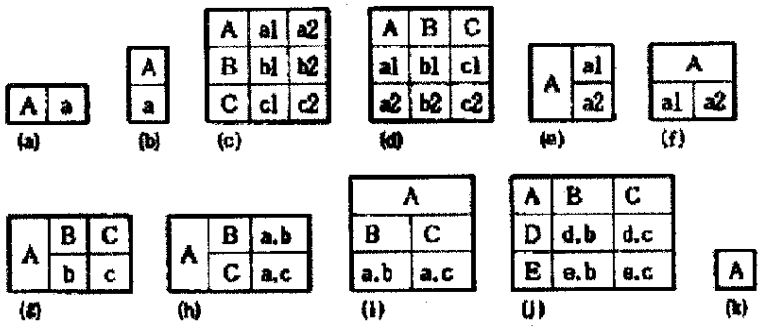


图 4.5 基本结构布局[15]

主导单元格：基本布局结构中，统领整个结构体的单元格称为主导单元格。对于 a、e 布局结构，主导单元格为结构体上方的单元格，对于 b、d 布局结构，主导单元格为左方的单元格，对于 c 布局结构的主导单元格为上方第一行或右方第一列单元格。

根据单元格之间的依赖关系，可以将这 6 种布局结构可以分为三类：一个主导单元格主导 N 行(列)单元格、一个主导单元格主导一行(列)单元格、独立单元格。其中第一类包含 a、b，第二类包含 c、d、e，第三类包含 f。

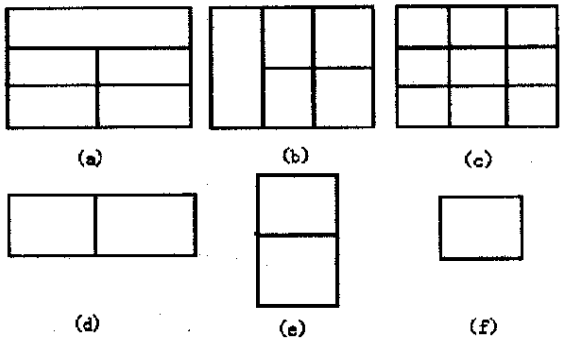


图 4.6 基本布局结构

4.4.1.2 结构体与嵌套结构体

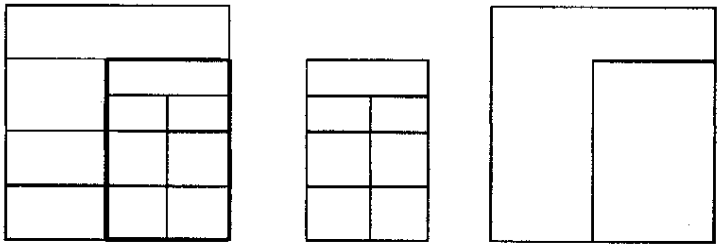
通过对表格基本布局结构的分析以及对标题域与数据域之间的依赖关系的充分考虑，本文提出了结构体与嵌套结构体的概念，并以此为基础实现了一种全新的表格逻辑结构描述方式。

结构体：出具有一定意义和布局结构且其外边界为矩形的单元格集合体，称为结构体。

基本结构体为具有基本布局结构类新的结构体，而复合结构体为基本结构体之间通过组合、嵌套而得到的结构体。

嵌套结构体：如果两个结构体 StructA、StructB 的布局结构为基本布局结构类型，且满足 StructA 包含 StructB，则称 StructB 为 StructA 的嵌套结构体。而称 StructA 为 StructB 的父结构体。

如下图所示，结构体 b 被包含在结构体 a 的内部，则结构体 a 为结构体 b 的父结构体，而结构体 b 为嵌套结构体。



(a)结构体 a　(b)结构体 b　(c)结构体 a,b 之间的包含关系

图 4.7 嵌套结构体

结构体的属性通过布局结构类型、行数/列数以及内部嵌套结构体的情况进行描述。

根据结构体内部是否包含有结构体，可以将结构体分为两类：复合结构体和基本结构体。如图 4.8 所示，结构体 a 内部包含有结构体 b，a 为复合结构体，而 b 为基本结构体。

嵌套关系是构成复杂结构体的一种常见的方式。本文所采取的方式是首先对基本的结构体进行识别，然后消除基本结构体对全局布局的影响，在进行基本结构体的识别，如此反复，直至复杂的结构体被分解成一组基本的结构体。这是一中自底向上的方法。

任何复杂的结构体都是由基本的结构体组合成，同样，任何复杂的结构体都可以通过一定的方式分解为一系列基本的结构体。

4.4.1.3 结构体表示

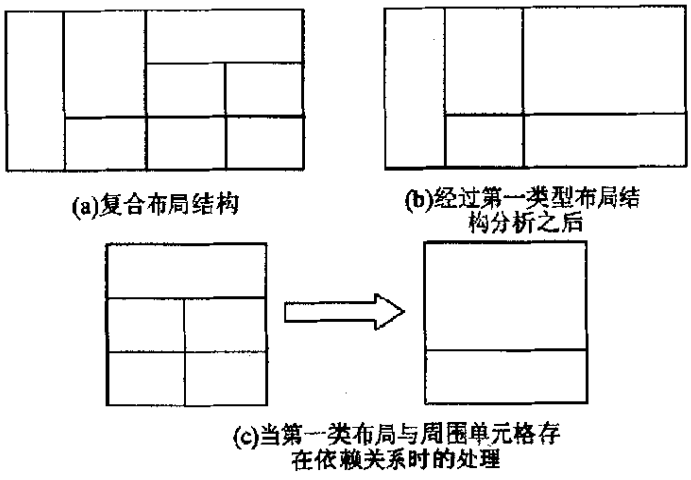


图 4.8 结构体的表示

记结构体为 Struct。

Struct=

{

Type；//结构体全局布局类型;

RowNum；//结构体内部包含单元格的行数；

ColNum；//结构体内部包含单元格的行数；

pStructLink；//指向嵌套结构体的指针；

}

其中：

- pStructLink 反映结构体内部嵌套结构体的情况，当 pStructLink 为空时，表示此区域内部不存在嵌套的结构体，而当指针为非空时，指针指向嵌套体所在的嵌套链表。
- RowNum 为结构体内部包含单元格的行数，当结构体为第一类布局结构时 RowNum=HNum-2，对于其他布局结构，RowNum=HNum-1。其中 HNum 为结构体内水平线行数目。
- ColNum 为结构体内部包含单元格的列数，RowNum=VNum-1。其中 HNum 为结构体内水平线行数目。

在图 4.8 中，图 4.8 (b)为表格的全局布局结构，其中存在嵌套关系，故其全局布局结构表示为(b,2,2,p)，其中，p 为此结构所指向的嵌套结构体链表的指针。说明此区域的全局布局结构为行数为 2，列数为 2 的 b 类型布局结构，并且内部存在

嵌套结构体。

嵌套链表 Nesting Link

嵌套链表是专门用来存储嵌套结构体的链表。

NestingLink=

{

StructLink：//存储嵌套结构体的链表；

pParentStruct：//指向其父结构体的指针；

}

其中：

- StructLink 为存储嵌套结构体的链表，内部依次存储着 pParentStruct 所指向的父结构体内部所包含的嵌套结构体。
- pParentStruct 为指向其父结构体的指针。当 pParentStruct 的值为 NULL 时，表示嵌套链表为非法链表；当 pParentStruct 的值为非空时，通过 pParentStruct 可以获得其父结构体的信息。

嵌套结构体表示

记嵌套结构体为 NestedStruct。

NestedStruct=

{

Type；//结构体全局布局类型；

Position；//嵌套结构体在其父结构体中的位置；

RowColNum：//结构体内部包含单元格的行数或列数；

pStructLink：//指向嵌套结构体的指针；

}

其中：

- Position：当结构体的全局布局结构类型为 a 时，Position 表示嵌套结构体在其父结构体中的列位置，而当当结构体的全局布局结构类型为 a 时，Position 表示嵌套结构体在其父结构体中的列位置（本文所讨论的嵌套结构体均为第一类型布局结构）。
- RowColNum：当结构体的全局布局结构类型为 a 时，RowColNum 表示结构体内部所包含单元格的列数，当结构体的全局布局结构类型为 b 时，RowColNum 表示结构体内部所包含单元格的行数。
- pStructLink：同样是反映结构体内部嵌套结构体的情况，当 pStructLink 为空时，表示此区域内部不存在嵌套的结构体，而当指针为非空时，指针指

向嵌套体所在的嵌套链表。

4.4.2 表格逻辑结构分析概述

本文所提出的表格逻辑结构分析方法采用自底而上的顺序对表格结构进行分析。首先分析表格中的基本布局结构，并消减底层基本布局结构对上层表格布局的影响，然后对表格的上层布局进行分析，如此反复直至布局结构分析结束。

通过对大量不同布局结构表格的分析发现，当表格中不含第一种布局结构的结构体时具有以下特征：

a) Type5、Type6类型的点为表格中基本布局结构的开始标志。

b) Type7、Type8类型的点为表格中基本布局结构的结束标志。

鉴于上述直线交点的特征，本文首先对第一类布局结构进行识别，然后，对第二类布局结构进行识别。对表格进行逻辑结构分析的总体步骤如下：

a) 对第一类布局结构进行分析

1) 对a、b类布局结构进行识别；

2) 分析结构体之间的嵌套关系；

3) 对结构体中的交点类型进行处理；

4) 依次执行1)、2)、3)，直至表格中不存在第一类布局结构。

b) 对第二、三类布局结构进行分析

1) 对c、d、e、f类布局结构进行识别；

2) 对结构体中的交点类型进行处理；

3) 分析结构体之间的嵌套关系。

c) 表格全局逻辑结构表示

经过第一类布局结构分析后的表格中仅存在第二、三类布局结构。同时在对第二、三类结构布局进行分析时，不考虑此两类布局结构之间的相互嵌套关系。首先利用已存在结构体的连通性，其次根据表格中不存在第一类布局结构时的直线交点特征进行布局结构识别。经过第二、三类布局结构分析后的所得到的区块即为表格的全局结构划分，此时的交点阵恰好反映出表格的全局布局结构框架。

4.4.3 基本布局结构的识别

任何复杂的表格结构都是由简单的基本布局结构组成，通过分析基本的布局结构，并消除以检测出布局对全局表格布局的影响，即可检测出更复杂的布局结构。

4.4.3.1 第一类布局结构的识别

下面以基本布局结构 a 为例，介绍第一类布局结构的识别方法。

根据基本布局结构 a 的交点特征进行结构的识别，算法描述如下：

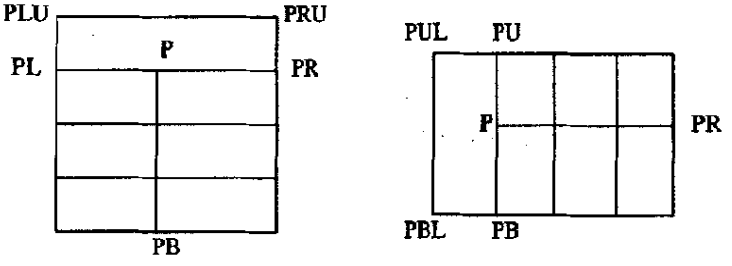
step1) 取出表格交点阵中 Type6 类型的交点 P；

step2) 查找结构体的左边界。沿 P 点所在的水平线上查找左边的第一个非 Type6 类型的交点 PL，如果 PL 为 Type5 类型的交点，则 PL 所在的垂直线为左边界；

step3) 查找结构体的右边界。沿 P 点所在的水平线上查找右边的第一个非 Type6 类型的交点 PR，如果 PR 为 Type7 或 Type9 类型的交点，则 PR 所在的垂直线为右边界；

step4) 查找结构体的上边界。PL 点上方第一个向右连通的点 PLU，PR 点上方第一个向左连通的点 PRU。如果 PLU、PRU 在同一条水平线上，则 PLU/PRU 所在的水平线为上边界；

step5) 查找结构体的下边界。沿 PL、PR 之间交点所处的垂直线，向下找到的第一个为非 Type9 类型且向下非连通的交点 PB，则 PB 所在的水平线为下边界；



(a)基本布局结构 a　　(b)基本布局结构 b

图 4.9 基本布局结构示意图

4.4.3.2 第二、三类布局结构的识别

以基本布局结构 c 为例，介绍第二、三类布局的结构分析过程。

根据基本布局结构 c 的交点特征进行结构的识别，算法描述如下：

Step1) 查找交点阵中 Type5、Type7 类型的交点，标记其所在垂直线为分割线；查找交点阵中 Type6、Type8 类型的交点，标记其所在水平线为分割线；

Step2) 以分割线为界将表格分割为多个区块。

Step3) 如果区块内部含有 Type9 类型的交点，则此区块为 c 布局结构类型；否则判断是否属于其它类型；

- 如果区块左边框上含有 Type5 或右边框上含有 Type7 类型的交点，则此区块为 e 布局结构类型；
- 如果区块上边框上含有 Type6 或下边框上含有 Type8 类型的交点，则此区块为 d 布局结构类型；
- 如果区块边框上不存在 Type5、Type6、Type7、Type8 类型的交点，则此区块为 f 布局结构类型。

4.4.4 结构体之间嵌套关系的分析

4.4.4.1 结构体与周围单元格之间的连通性

如果两个结构体之间存在相互连通的关系，则此两结构体之间有相互依赖关系的可能性就很大。为了研究结构体之间是否存在相互的依赖关系，需要对结构体之间的连通性作进一步的分析。

如果结构体 S 左边界上的点均向左连通，则称此结构体与左侧单元格相联通。

同样，如果结构体 S 右边界上的点均向右连通，则称此结构体与右侧单元格相联通。

如果结构体 S 上边界上的点均向上连通，则称此结构体与上侧单元格相联通。

如果结构体 S 下边界上的点均向下连通，则称此结构体与下侧单元格相联通。

对于布局结构 a 而言，仅考虑结构体与周围单元格之间的左连通和右连通。而对于布局结构 b，则仅需要考虑结构体与周围单元格之间的上连通和下连通。对于二、三类布局结构不需考虑结构体与周围单元格之间的连通性。

如果结构体与上下左右侧的单元格均不联通，则称此结构体为独立结构体。

4.4.4.2 交点类型转换

本文采取自底向上的方法对表格结构进行分析。为了使在分析过底层结构体后，能够使上层的结构体体现出其全局的布局特征，需要对底层内部的交点类型进行转换，从而消减底层结构体对上层结构体的影响。

考虑到结构体之间的相互依赖关系，对于独立结构体和非独立结构体的交点类型采取不同的转换方法。

就独立结构体而言，交点类型转换方法为：

1) 内部交点全部置为 Type0；

2) 边界上的交点按图 4.10 所示的交点类型转换表进行转换。

而对于非独立结构体，交点类型转换方法为：

1) 内部交点全部置为 Type0；

2) 对于 a 布局结构体，仅需将 PL 和 PR 两点按图 4.10 所示交点类型转换表进行转换。b 布局结构体与 a 布局结构体类似，仅需将 PU 和 PB 两点按图 4.10 所示交点类型转换查找表进行转换。

	交点类型				
	Type5	Type6	Type7	Type8	Type9
左边界	Type0				Type7
右边界			Type0		Type5
上边界		Type0			Type8
下边界				Type0	Type6

图 4.10 边界交点类型转换查找表

交点类型转换

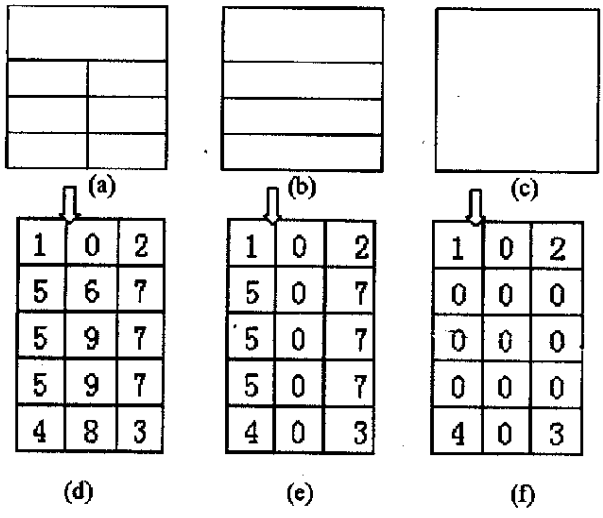


图 4.11 基本结构体交点类型转换

如图 4.11 所示，图 4.116(b)为当图 4.116(a)是非独立结构体时的转换结果，而图 4.11(c)为独立结构体时的转换结果。

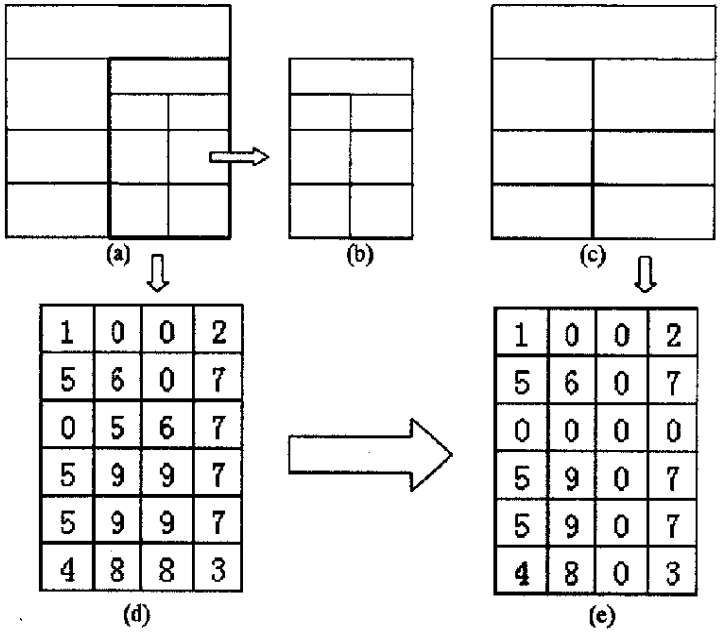


图 4.12 复合结构体交点类型转换

其中，图 4.12 (a)为复合结构体，图 4.12 (b)为图 4.12 (a)的嵌套结构体，图 4.12 (c)为嵌套结构体进行交点类型转换后的结果。

由此可见，复合结构体经过交点类型转换后，可以更加清晰地体现出全局的布局结构类型。

4.4.4.3 嵌套关系分析

经过对各种常见的嵌套结构进行分析发现，第二、三类布局结构与其周围的单元格之间的关系必定不是完全连通的，结构及意义都相对比较独立。基于此本文仅考虑嵌套结构体为第一类布局结构类型的嵌套关系。

通过嵌套结构体的布局类型以及在其父结构体中的位置来对嵌套关系进行描述。如图 4.13 所示，复合结构体 Sa 由 b 布局类型的结构体 Sc 内部嵌套 a 类型的嵌套结构体 Sb 而成。结构体 Sb 在结构体 Sc 中的位置如 4.13(c)所示。

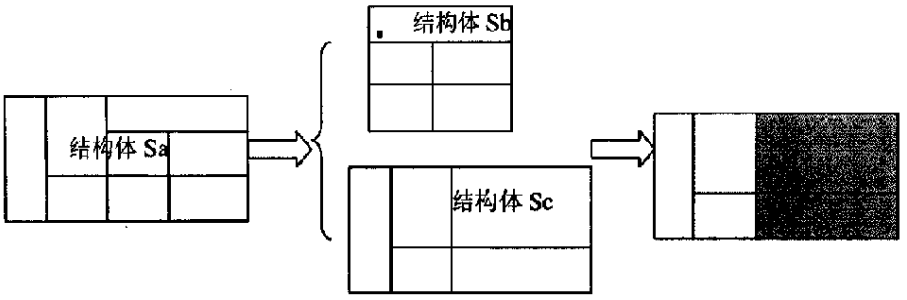


图 4.13 非同种嵌套关系示意图

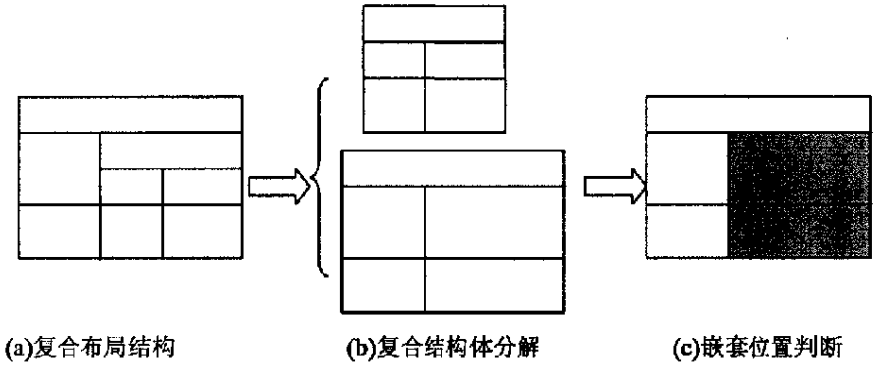


图 4.14 同类型嵌套关系示意图

对结构体之间嵌套关系的判断关键在于嵌套结构体在其父结构体中位置的判断。本文采取基于交点类型阵的判断方法，具体判断方法因嵌套结构体的布局类型不同以及嵌套结构体与其父结构体是否为同种类型而存在一些差别。

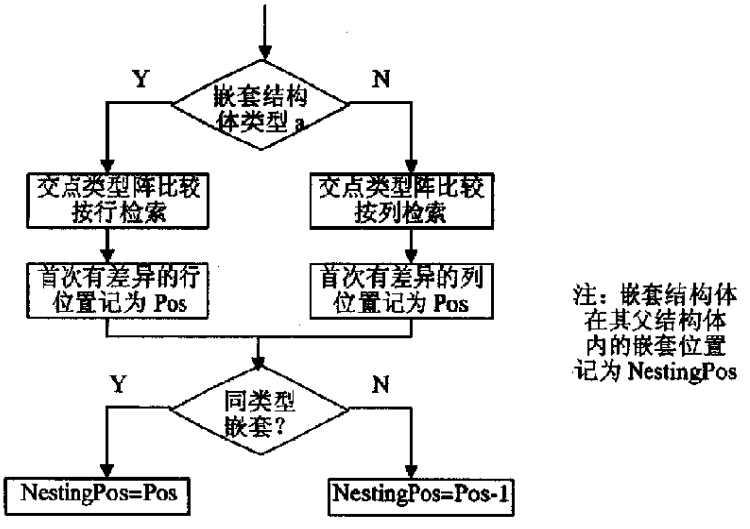


图 4.15 嵌套关系之位置判断流程

图 4.13(a)内的嵌套结构体 Sb 为 a 布局结构类型，按行在两个交点类型阵（如图 4.16 所示）中同步检索，发现第二行第三列的交点类型值存在差异，则记 Pos 为 3。又因为嵌套结构体 Sb 与其父结构体 Sc 之间非同类型嵌套，则嵌套的位置 NestingPos 记为 Pos-1，NestingPos=2，即嵌套结构体在其父结构体内的第 2 列的数据域内存在嵌套关系。

1	6	6	0	2
0	0	5	6	7
0	5	9	9	7
4	8	8	8	3

1	6	6	0	2
0	0	0	0	0
0	5	9	0	7
4	8	8	0	3

图 4.16 位置判断 a)为原交点类型阵

(b)为分解后的全局交点类型阵

结构体之间的嵌套关系的描述如下：

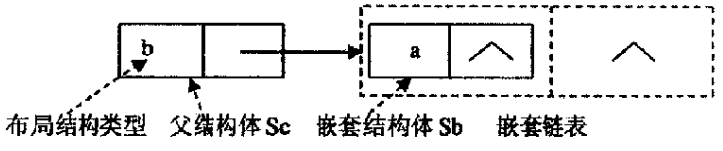


图 4.17 嵌套关系描述

4.4.5 表格全局逻辑结构表示

表格的逻辑结构表示通常采用单元链表[13]、二叉树[15]来描述单元格之间的逻辑关系。采用单元链表描述逻辑结构不能唯一确定出表格的实际逻辑结构，在文献[13]中通过表格的网格矩阵来消除逻辑结构表达的歧义。二叉树方法以左节点表示其下方存在的单元格，而右节点表示右方存在的单元格。全局结构描述树和局部结构描述树均采用二叉树结构来表述。局部结构描述树不能表达出表格中存在嵌套关系，同样此方法也存在歧义，不能实现逻辑表达结构与表格之间的一一对应关系。

X-Y 树[61,62]是一种自上而下的页面布局分析方法。根节点表示整个页面，叶节点为页面中所分割出来的区域，而每层依次表示水平或垂直的分割结果。X-Y 树具有以下特点：

- 根节点表示整个表格；

- 树中的每个节点均代表表格中的一个区块；
- 一个节点的子节点可以通过对其进行水平或者垂直方向上的划分得到。
- 划分方向的选择，采取水平方向和垂直方向依次交替的方法。如果父节点进行了水平方向上的划分，则子节点需进行垂直方向上的划分。并且对根节点进行划分的方向可以任意选取。

本文利用 X-Y 树进行页面分析的算法思想将 X-Y 树结构应用到表格全局逻辑结构的表达，同时采用嵌套关系链表来表示表格中所存在的嵌套关系。具体算法以表格的直线交点阵为基础，描述如下：

Step1 创建一个根节点；

Step2 在感兴趣的区域内查找贯通整个区域的水平线/垂直线；

Step3 在贯通线处进行分割，每进行一次分割，产生一个新的子节点。在每个递归层次，水平和垂直方向上的分割依次进行；

Step4 递归执行 step2 和 step3，直至感兴趣区域内部不存在贯通整个区域的直线。

如图 6(e)为 6(c)所示布局结构的全局逻辑结构描述。

下图是运用 X-Y 树进行表格结构表达的一个示例：

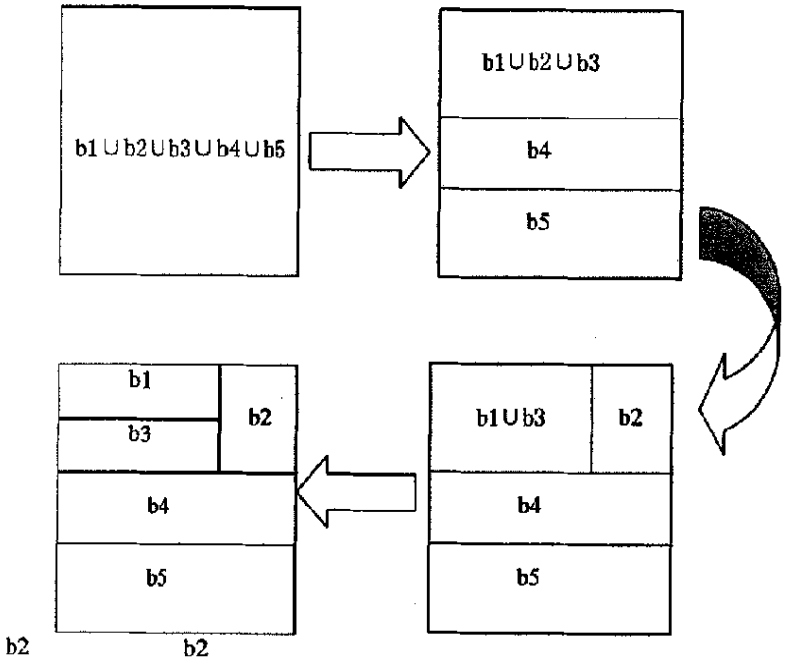


图 4.18 全局布局结构分析过程

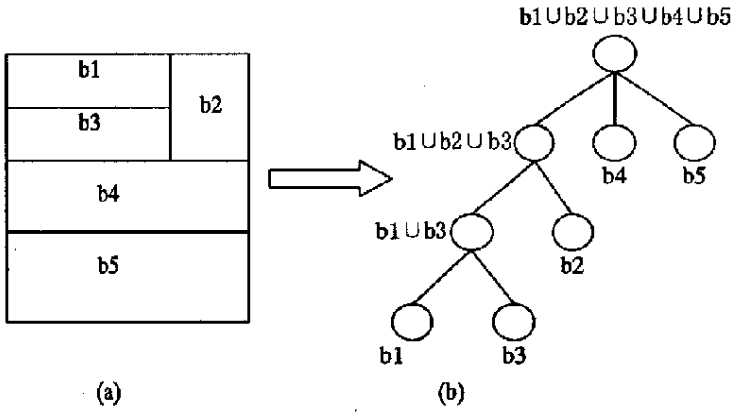


图 4.19 表格全局结构的 X-Y 树表示

4.5 实验结果

为了验证算法的效果，本文实现了一个表格图像分析系统，系统对收集的 50 幅图像进行测试，实验图像分为两类：一类图像的表格较规范，表格中几乎不存在偶然的单元格之间邻接关系，共 30 幅，另一类存在较多的偶然单元格直接邻接关系，共 20 幅。试验分别对全局结构体总数、正确识别数目、结构体之间存在嵌套关系的数目以及正确识别嵌套关系的数目进行统计。实验结果表明，本算法很好地处理表格中偶然的单元格直接邻接关系的影响，将具有相互依赖关系的单元格分隔在同一个结构体中。

表 4.1 试验结果：

类别	全局结构体总数	正确识别数目[1]	正确率[1]	存在嵌套关系数目	正确识别数目[2]	正确率[2]
第一类	171	171	100%	126	126	100%
第二类	143	141	98.6%	87	86	98.8%

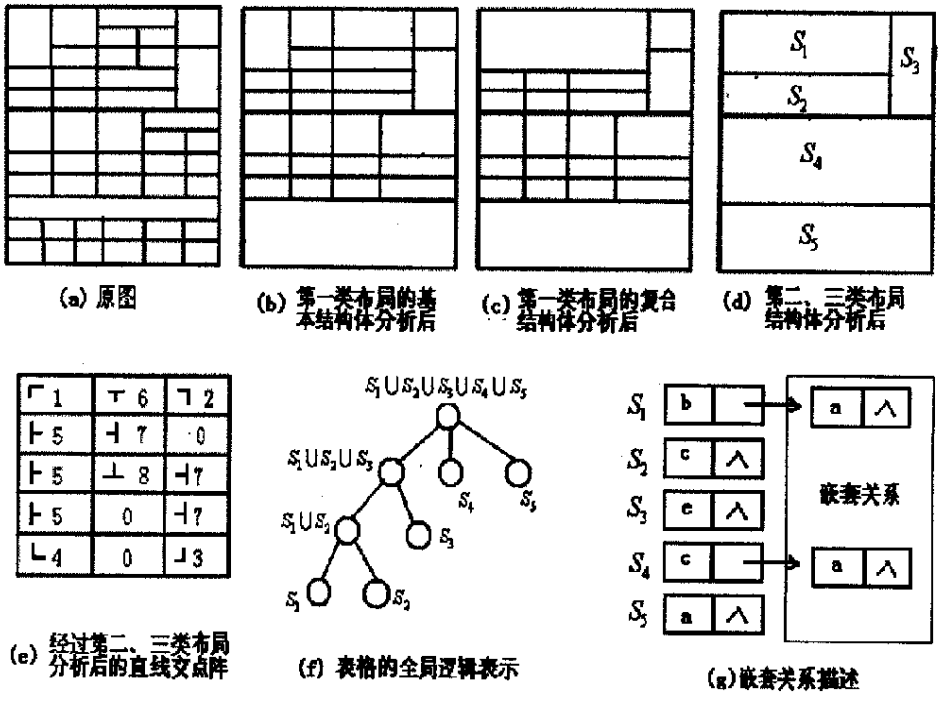


图 4.20 一个实例表格的处理结果

4.6 小结

通过对表格中直线交点特征的分析，本文提出了一种全新的自底向上的表格图像逻辑结构分析方法。此方法主要适用于具有矩形性的规范表[63]。对表格中标题域与数据域之间依赖关系的充分利用，使算法能够根据表格的基本布局结构来对表格图像进行逻辑划分，将具有相互依赖关系的单元格分隔在同一个结构体中，这样的分割方法更具有实际意义。同时由于算法自底向上地对表格结构进行分析特点，使得识别结果不容易受到表格中一些偶然的单元格直接邻接关系的影响。另外，本文还介绍了表格中基本布局之间嵌套关系的表示方法，它使得对复杂结构表格的描述更为简练有效。对于存在错误交点类型的交点阵，如何进一步根据表格基本布局和标题域与数据域之间的依赖关系来提高算法的容错能力，是今后需要改进的方向。

结束语

表格文档图像的结构分析是 DIA 领域中的研究难点和热点，本文围绕表格图像理解技术方面作了一些研究工作。下面对本文的主要内容进行简要的总结：

- 表格文档倾斜校正：本文介绍了一种基于内容的文档倾斜自动校正方法，该方法综合利用了基于参考线[39]方法和基于文字行[41]方法的优点，根据文档的内容采取不同的倾斜校正方案，更具通用性。同时在进行倾斜角度检测的过程中，采用金字塔算法[46]降低图像的分辨率，以提高算法的效率。
- 表格版面结构分析：本文所设计的方法采取直接抽取构成表格的线条，并计算线条的交点，进而获取各个单元格信息的表格版面分析思路。具有以下特点：
  - ✧ 本文以线条交点矩阵表示表格版面结构分析结果。以点代替线，不仅降低了对问题描述的难度，体现了表格版面结构的全局行列特征，更易于检索，这为表格图像的后续处理提供了方便。
  - ✧ 表格框线的提取直接以先期图像倾斜校正时得到的有向单连通链为基础，通过对其进行旋转、规则化，从而得到表格框线，该方法方便有效。
- 表格逻辑结构分析：通过对表格中直线交点特征的分析，本文提出了一种全新的自底向上的表格图像逻辑结构分析方法。此方法主要适用于具有矩形性[63]的规范表。具有以下特点：
  - ✧ 对表格中标题域与数据域之间依赖关系的充分利用，使算法能够根据表格的基本布局结构来对表格图像进行逻辑划分，将具有相互依赖关系的单元格分隔在同一个结构体中，这样的分割方法更具有实际意义。
  - ✧ 由于算法自底向上地对表格结构进行分析的特点，使得识别结果不容易受到表格中一些偶然的单元格直接邻接关系的影响。
  - ✧ 本文采用嵌套链表描述表格中基本布局之间嵌套关系，它使得对复杂结构表格的描述更为简练有效。

本文主要对在表格文档图象识别研究领域中的两个核心技术—表格版面结构识别和表格逻辑结构识别技术作了一些有益的探索，基本能够满足表格文档结构分析识别的需要，然而，其中算法同样存在需要商榷、优化的地方：

- 本文所述方法存在一定的局限性，仅针对于具有矩形性[63]的表格图像，而对不具有矩形性的表格无法进行分析。
  对不具有矩形性的表格无法进行分析。

- 本文的目标是不仅对空白表格进行逻辑结构分析，同样也可以对已填充表格进行分析。但由于表格结构的复杂性以及目前方法对标题域与数据域之间依赖关系认识的局限性，使得本文所述方法的通用性有所降低。如何进一步根据表格基本布局和标题域与数据域之间的依赖关系来提高算法的通用性是今后需要改进的方向。

致谢

衷心感谢我的导师王泉副教授。本论文从选题到完成的整个过程中，始终得到王老师的悉心教导，在此谨致以最诚挚的谢意。三年的研究生学习生活转瞬即逝，王老师那孜孜以求，诲人不倦的治学态度和高尚的人格魅力给我留下了极其深刻的印象，会时刻影响、激励我走好自己的人生之路。

感谢计算机外部设备研究所的所有老师及同学，是他们为我的学习创造了良好的环境。特别感谢田玉敏老师及万波老师等在学习及工作上对我的帮助。

感谢刘有利、张文刚、丁凰、阳山、王昊等同学给予我的帮助和支持。

感谢王霞女士及其家人给予我的关心和帮助，在此谨致以衷心的感谢。

深深的感谢我的家人，正是他们长期不懈的鼓励和支持，我才能够顺利完成我的学业，他们永远是我最敬爱的人。

最后，我想借此机会向所有关心、教育、指导过我的院各级领导、教师、同学们致以诚挚的谢意。

参考文献

[1] George Nagy. "Twenty Years of Document Image Analysis in PAMI..". January, 2000. IEEE transactions on Pattern Analysis and Machine Intelligence, vol.22, No.1. pp.38-62.
[2] B.B.Chaudhuri and U.Pal, "Skew Angle Detection of Digitized Script Documents", Feb1997 IEEE Trans. Pattern Analysis and Machine Intelligence, vo.19 ,no.2, pp.182-186.
[3] S. DZenzo, L.Cinque, and S. Levialdi, "Run-Based Algorithms for Binary Image Analysis and Processing ",J an.1 996, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.18 ,no .1, pp .83 -88.
[4] K. Etemad, D. Doerman, and R. Chellappa, "Multiscale Segmentation of Unstructured Document Pages Using Soft Decision Integration", Jan. 1997, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.19 ,no .1, pp .92 -96.
[5] George Nagy. "What does a Machine need to know to read a document" March, 1992. Proceedings of Symposium on document analysis and information retrieval. Las Vegas, Nevada, USA. pp.1-10.
[6] C.A.Cabrelli and U .M.Molter, "Automatic Representation of Binary Images", Dec.1990 IEEE Trans. Pattern Analysis and Machine Intelligence, vol.12, no.12, pp.1190-1195.
[7] R.Cattoni,T.Coianiz, S.Messelodi,C.M.Modena. "Geometric Layout Analysis Techniques for document image understanding: a review." 1998. Technical Report 9703-09, ITC-IRST.
[8] AndreasD engel,Rainer Bleisinger,Rainer Hoch,Frank Fein,FrankHones. "From Paper to Office Document Standard Representation" July,1992.IEEE Computer. pp.63-67.
[9] G.Nagy."Document image analysis: Automated performance evaluation."1995.In A.L. Spitz and A.Dengel, editors, Document Analysis Systems, World Scientific, Singapore.pp.137-156.
[10] VF.Margner,PKarcher,A.K.Pawlowski. "On Bench Marking of Document Analysis System." August, 1997. Proceedings of 4th International Conference Document Analysis and Recognition, Ulm, Germany(ICDAR'97).pp .331-336.
[11] YH. LIU-GONG B. DUBUISSON,H.N. PHAM. "A General Analysis System for Document's Layout Structure Recognition." August, 1995. Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Canada(ICDAR'95).pp.597-600.
[12] J. Hochberg, P. Kelly, T. Thomas, and L. Kems, "Automatic Script Identification from Document Images Using Cluster-Based Templates", Feb. 1997, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.19 ,no .2, pp .17 6-181.
[13]史广顺. "文档图像中表格结构的自动定位与分析", 2003. 4, 博士学位论文, 南开大学。
[14] Xinxin Wang. "Tabular Abstraction, Editing, and Formatting." 1996. Phd thesis, University of Waterloo, Ontario, Canada.
[15] T.Watanabe, Q.L.Quo, and N.Sugie." Layout recognition of multi-kinds of table-form documents." 1995, I EEE Trans. on Pat. Anal. And Machine Intel.,vol.17(4),pp.432-445.
[16] John H. Shamilian, Henry S. Baird,Thomas L. Wood. "A Retargetable Table

Reader." August, 1997. Proceedings of 4th International Conference Document Analysis and Recognition, Ulm, Gertnany(ICDAR'97).pp .158-163.
[17] E. Green and M. Krishnamoorthy. "Model-based analysis of printed tables", August 1995,In Proceedings of 3rd International Conference on Document Analysis and Recognition, pp.214-217,Montreal,Canada.
[18] TB.Haas. "The development of a prototype knowledge-based table-processing system." December, 1997.Master's thesis, Brigham Young University.
[19] AlessandroL .K oerich,LuanL ingL ee. "AutomaticS torage,R etrievalan dV isualization Of Bank Check Images."September,1999.Proceedings of 5th International Conference on Document Analysis and Recognition, Bangalore, India(ICDAR'99). pp .111-114.
[20] Osamu Hori,David S. Doermann. "Robust Table-form Structure Analysis Based on Box-Driven Reasoning."August,1995. Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Canada(ICDAR'95). pp.218-220.
[21] MatthewHurst and ShonaDouglas. "Layout and Language: Preliminary investigations in recognizing the structure of tables",1997,40' International Conference Document Analysis and Recognition, pp.1043-1047(Ulm,Germany).
[221 AsadA. Abu-Tarif," TABLEP ROCESSINGA ND UNDERSTANDING",1998. Master's thesis ,Rensselaer Polytechnic Institute Troy, NewYork.
[23] JuanF .A rias,Atul Chhabra and VishaiMisra, "Interpretinga ndR epresentingT abular Documents ." 1996 Conference on Computer Vision and Pattern Recognition(C VPR'96), pp.600-605.
[241 Watanabe,T. ,H .Naruse,Q.Luo,N.Sugie. "Structure Analysis of Table-form Documents on the Basis of the Recognition of Vertical and Horizontal Line Segments." September,1991. Proc. of the 1st Int. Conf. on Document analysis and recognition, Saint-Malo, France. pp.283-293.
[25] A.Laurentini, PViada." Identifying and understanding tabularm ateriali n compound documents." 1992.In Proceedings of the Eleventh International Conference on Pattern Recognition(ICPR'92),The Hague, The Netherlands. pp .405-409
[26] S.Chandran, R .Kasturi. "Structural recognition of tabulated data." October, 1993.In Proc. of 2nd Int. Conf. on Document Analysis and recognition, Tsukuba Science City, Japan. pp.516 -519.
[27] T.Watanabe,Q.Luo. "A multiplayer recognition method for understanding table-form documents." 1996.Int.Journal of imaging Systems and Technology, Vol.7, pp .279-288.
[28] K.Itonori. "Table Structure Recognition based on Text block Arrangement and Ruled Line Positions" October,1993. Proc., IAPR 2nd Int. Conf. on Document Analysis and Recognition, Tsukuba Science City, Japan. pp .765-768.
[29] Yalin Wang, Ihsin T. P hillips,Robert Haralick. " Table Detection Via Probability Optimization "A ugust,2002. D.Lopresti,J. Hu, and R .Kashi(Eds). Document Image Analysis System ,5th International Workshop DAS2002,Princeton,NJ, USA. pp .272-282
[30] E.Green,M.Krishnamoorthy. "Recognition of tables using tables grammars." April,1995.In Proceedings of the Fourth Annual Symposiumon Document Analysis and Information Retrieval, University of Nevada, Las Vegas, pp .261-277.
[31] Watanabe. T, Fukumura.T. "A framework for validating recognized results in under standing table-form document images." August, 1995. Proceedings of 3rd International Conference on Document Analysis and Re cognition, Montreal, Canada(ICDAR'95),pp.536-539.
[32] John H. Shamilian, Henry S. Baird,Thomas L. Wood. "A Retargetable Table

Reader.", August, 1997.Proceedings of 4th International Conference Document Analysis and Recognition, Ulm, Germany(ICDAR'97), pp.158-163.
[33] Kuo-ChinFan,Yuan-KaiWang, Mei-LinChang. "Form Document Identification Using Line Structure Based Features." September, 2001. Proceedings of 6th International Conference on Document Analysis an d Recognition, Seattle, Washington, USA(ICDAR '01), pp.704-708.
[34] YY Tang, H. Ma, J. Liu, B.F. Li, D. Xi. "Multi resolution Analysis in Extraction of Reference Lines from Documents with Gray Level Background. ", 1997, IEEE PAMI, vol.19(8),pp.921-925.
[35] 梁虹、周继勤、柏正尧、杨汉春."基于数学形态学的表格自动处理系统." 2000.7. 计算机应用,vol.20(7).pp.17-20
[36] 刘建胜, 汪同庆, 王贵新, 居炎, 袁样辉. 基于边框线的版面分析方法应用于选票处理, 计算机工程与应用, pp.248-256, vol.38, 2002.
[37] 王朱华, 李佐, 蔡士杰. 基于连续性的页面倾斜检测与校正. 计算机辅助设计与图形学学报,2001,vol.13(8), pp.735-739.
[38] G. Nagy. Twenty Years of Document Image Analysis in PAMI. IEEE Trans. On Pattern Analysis and Machine Intelligence. 2000,vol.22 (1),pp.38-82.
[39] Yi-Kai Chen, Jhing-Fa.Wang "Skew detection and reconstruction based on maximization of variance of transition-counts. ", Pattern recognition, 2000,vol.33, pp.195-208.
[40] 杨波. 基于内容的文档图像压缩方法研究. 博士学位论文, 重庆:重庆大学, 2002.
[41] R. Smith. "A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation." In Proc. of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, August 1995,pp.1145-1148.
[42] B Yu.A K Jain. "A robust and fast skew detection algorithm for generic documents. " Pattern Recognition,1996, vol.29(10), pp.1599-1629.
[43] B. Gatos. N. Papamaikos. and C. Chamzas. "Skew Detection and Text Line Position Determination in Digitized Documents." Pattern Recognition, 1997.vol.30(9), pp.1505-1519.
[44] HK Kwag.S. H.K im.S. HJeong.G.S.Lee. "Efficient Skew Estimation and Correction Algorithm for Document Images." Image and Vision Computing,2002, vol.20, pp.25-35.
[45] 明底烈, 柳健. 小角度倾斜图像的倾斜快速检测和校正方法. 华中理工大学学报, 2000,vol.28(5), pp.66-68.
[46] Seong-WhanLee. Parameter-Free Geometric Document Layout Analysis. Pattern Analysis and Machine Intelligence, 2001,vol.23(11),pp.1240-1255.
[47] S.Chen,R.M.Haralick. An Automatic Algorithm for Text Skew Estimation in Document Images Using Recursive Morphological Transforms. In Proc. of the 1st IEEE International Conference on Image Processing, Austin,Texas, 1994,pp.139-143.
[48] 郑冶枫, 刘长松, 丁晓青, 潘世言. 基于有向单连通链的表格框线检测算法. 软件学报, 2002, vol. 13, pp. 790-796.
[49] Illingworth, J., Kittler, J. A survey of the hough transform. Computer Vision, Graphics, and Image Processing, 1988,vol.44(1), pp.87-116.
[50] 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 3rd International Conference on Document Analysis and Recognition. Montreal, Canada, 1995,pp.579-582.
[51] Liu, W.Y., Dov, D. From raster to vectors: extracting visual information from line drawings. Pattern Analysis and Application, 1999,vol.2(1),pp.10-21.

[52] Yu, B., Jain, A.K. A generic system for form dropout. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996,vol.18(11),pp.1127-1131.
[53] T.Pavlidis." A vectorizer and feature extractor for document recognition". Computer Vision, Graphics, and Image Processing, 1986,vol.35, pp.111-127.
[54] B.Kong,S.Chen,R.M.Haralick. "Automatic Line Detection in Document Images Using Recursive Morphological Transforms." 1995,SPIE,vol.2442,pp .163-17
[55] 陈自利. 基于小波与神经网络的文字识别系统研究. 博士学位论文, 重庆:重庆大学, 1999.
[56] 蔡雷. 手写字符自适应识别的研究. 博士学位论文. 重庆:重庆大学, 1999.
[57] FanK C .Lu J M .WangJ Y. A Feature Pint Approach to the Segmentation of Form Documents. Proceedings of the 3'a International Conference on Document Analysis and Recognition. Washington D.C.:IEEE Computer Press,1995,pp.623-626.
[58] Jiun-Lin Chen, His-Jian Lee. Field Data Extraction for Form Document Procession Using a Gravitation-based algorithm. Pattern Recognition.2001,vol.34, pp.1741-1750.
[59] Yu B. Jain A K. Generic A. System for Form Dropout. IEEE Transaction on Pattern Analysis and Machine Interlligence.1996,18(11),pp.1127-1134.
[60] 李星源. 高文一种鲁棒性的结构未知表格分析方法. 软件学报. 1999, vol.10(11), pp.1216-1224.
[61] G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. In: 7th Proceedings of ICPR, vol.1, Montreal, Canada, 1984, pp. 347-349.
[62] Luiz Antonio Pereira Neves, Jacques Facon. Methodology of automatic extraction of table-form cells. In: Proceedings XIII Brazilian Symposium on Computer Graphics and Image Processing, 2000, pp.15-21.
[63] 刘冰等. 表格文件图像逻辑结构提取方法. 中国图象图形学报, 2000, vol.5(A), pp. 678-682.

在读期间的研究成果

1. 王来敬, 王泉. 表格图像分割. 西安电子科技大学学术年会, 2005.