文章编号:1671-4598(2023)01-0282-07 DOI:10.16526/j.cnki.11—4762/tp.2023.01.042 中图分类号:TP391.41 文献标识码:A

基于 OCR 技术的航天器材料及器件试验数据识别系统

陆俊杰[1]，魏亚东[2]，李晓峰[3]，王 成[1]，李洪普[1]，李 锋[3]
(1. 中船重工奥蓝托无锡软件技术有限公司，江苏 无锡 214000；
2. 哈尔滨工业大学 材料科学与工程学院，哈尔滨 150001；
3. 中国船舶科学研究中心，江苏 无锡 214000)

摘要：航天器材料及器件数据库需要海量国内外试验报告数据的支撑，其中表格作为最普遍的数据存储形式含有的数据量最为庞大，然而面对人工识别提取表格数据工作繁琐且易出错的难点，以 PDF 文档的表格为研究对象，提出基于 OCR 技术的航天器材料及器件试验数据识别系统；采用了 B/S 架构，基于 EXT、JAVA、Python 等技术语言进行开发，系统具备 PDF 文档转换、表格识别、数据提取、数据编辑等功能；依据系统设计采用版面分析和 PDFPlumber 表格检测的关键技术和方法以达导准确有效识别 PDF 文档表格的目的，采用 EXT 表格控件形式展现提取的数据经试验测试实现了对 PDF 文档内规整表格的批量识别和数据提取；验证了设计方案的可行性，满足了试验数据试别系统的高识别准确率、快速识别等特点。

关键词：航天器材料与器件；数据识别系统；OCR；PDF 文档；表格识别

Test Data Identification System for Spacecraft Material and Device Based on OCR Technology

LU Junjie[1], WEI Yadong[2], Li Xiaofeng[3], WANG Cheng[1], LI Hongpu[1], LI Feng[3]
(1. Wuxi Orient Software Technology Co., Ltd., Wuxi 214000, China;
2. School of Materials Science and Engineering, Harbin Institute of Technology, Harbing 150001, China;
3. China Ship Scientific Research Center, Wuxi 214000, China)

Abstract: The database of spacecraft materials and devices needs the support of massive test reports at home and abroad. As the most common form of data storage, the table contains the largest amount of data. However, faced with the tedious and error-prone work of manual identification and extraction of table data, the table of PDF document is taken as the research object. A data identification system of spacecraft material and device test based on OCR technology is proposed. Using B/S architecture, based on the developments of EXT, JAVA, Python and other technical languages, the system has the functions such as PDF document conversion, form recognition, data extraction, data editing; According to the system design, the key technologies and methods for the layout analysis and PDFPlumber form inspection are used to identify PDF document forms accurately and effectively. The extracted data are displayed in the EXT form control. The batch identification and data extraction of regular forms in PDF documents are realized through the test. The feasibility of the design scheme is verified to meet the high accuracy and fast recognition of the system.

Keywords: spacecraft materials and devices; data identification system; OCR; PDF; form recognition

0 引言

表格作为一种高度精炼的信息表达手段，以简单明了、规范、便携等鲜明特点广泛存在于各类试验报告文档种。表格信息的识别和提取在计算机处理以前只是依靠人眼识别判断，工作繁琐还容易出错。表格文档识别作为一类独特的数据内容智能识别技术，与常规的文字、符号、图像等形式相比较，其内容含有复杂的层级关系。在对表格内部线条进行识别过程中，最影响识别结果的因素有光线昏暗、分辨率、规范程度等。

文档中对于各种数据呈现形式的表格，除了上述外部因素对识别结果的影响外，表格内部同样也会因为文字与线的重叠等问题干扰整体的识别正确性。在进行表格字符识别过程中，经常出现内部线条不规范从而无法区分的问题，再加上需要提取识别的数据常常是文档内的核心，因此需要更为精准、快速的数据识别技术。

检测表格线对表格单元格中的文本识别非常重要。目前针对不同表格具有不同的单元格提取方法，Zheng 等人采用基于有向单项链检测表格线，通过确定黑色游程长度

收稿日期:2022-06-17； **修回日期**:2022-07-13。

作者简介:陆俊杰(1995-)，男，江苏无锡人，硕士，工程师，主要从事试验数据采集及计算机应用方向的研究。

引用格式:陆俊杰，魏亚东，李晓峰，等.基于 OCR 技术的航天器材料及器件试验数据识别系统[J].计算机测量与控制，2023，31(1)：282-288，293.

来检测线条，并且如果满足某些条件，就可以与其他线条合并[2]；Zuo 等人提出了一种鲁棒的表格登记方法，用于更好地理解来自扫描表格图像的结构化信息，其中使用了基于卷积的表格线检测方法，但前提是提供了表结构，其中包括表头、行和列的数量，甚至是近似的单元格大小[3]；Tian 等人根据表格特征定位表区域，并基于 Hough 变换方法检测表格线；刘昱首先采用投影法确定表格框线坐标，然后利用搜索法确定表格框线的长度[4]；夏禾采用基于区域生长的表格定位方法；何柳提出了一种分步提取表格特征的表格识别方法，首先利用基于连通域的表格轮廓提取方法提取表格轮廓，确认表格所在区域，然后通过基于数学形态的表格线提取方法提取表格中的表格线，最后根据所提取表格线交点特征获取表格单元信息，完成表格的识别[5]。

现如今国内外航天器材料及器件数据库系统的数据录入主要通过自定义文件模板导入或键盘输入，需要依赖人工去处理原始数据，工作繁琐且浪费时间；然而，主流的原始数据都以表格的文件存储形式，其中最普遍的是 PDF 文档表格数据。因而为了提高工作效率、优化数据录入方式、提升系统智能化、自动化操作，提出基于 OCR 技术的航天器材料及器件试验数据识别系统，自动识别提取 PDF 文档中的表格数据，为仿真模型的建立、仿真结果比对以及退化模型的建立提供参考依据。

1　航天器材料及器件试验数据试别系统结构及原理

光学字符识别，简称为 OCR（optical character recognition），是指电子设备（例如扫描仪或数码相机）检查纸上打印的字符，经过检测暗、亮的模式肯定其形状，而后用字符识别方法将形状翻译成计算机文字的过程。它是模式识别领域中最富有挑战性的研究课题，从 20 世纪 50 年代开始，许多的研究者剧针对该领域开展可广泛的探索，时至今日，人们在正规化理论、特征抽取理论、匹配理论等方面取得了众多成果，正逐渐成为计算机视觉领域的重要任务，在名片、车牌、身份证、表格表单、发票识别系统等广泛应用[6]。

典型的 OCR 识别系统通常由以下几部分组成：

1）预处理。

预处理通常包含对原始图像的去噪、倾斜校正和各种滤波处理。

2）版面分析。

版面分析是根据内容的排版布局进行分割的技术。其原理是为了得到满足要求的字符块，遵循预先设计的划分规则去解析，判别划分出来的字符块是否满足需求。

3）字符切分。

根据上述版面分析的原理设计相应的内容分割算法，发展至今常用的算法例如基于结构分析的切分算法、基于识别结果的切分算法、整体切分算法、综合切分算法等[7]。

4）特征提取。

特征提取作为识别系统中相对核心的步骤，其原理是分析所切分出来的字符块，总结统计其通用特性的过程。

5）分类识别。

分类识别其原理是在特征提取得到的数据库中，判断是否含有与现有字符高度匹配的类型。

6）后处理。

后处理的原理是通过常规的话语、词组等整理生成的先验知识，修改识别得到的字符，提高准确性。

航天器材料及器件数据库需要集中汇入海量的国内外数据作为支撑，其主要的数据文件是 PDF 文档。在 PDF 文档中，数据的展现形式包含表格数据、文字、曲线数据、图像数据等，其中数量最多的是表格数据，因此最大的需求是针对 PDF 文档内的表格需要高效的方法进行识别提取。

因此，针对需要批量处理的固定格式的表格，本文提出设计基于 OCR 的航天器材料及器件试验数据识别系统，采用如今较为流行的 Python 语言进行系统开发。

在试验数据识别系统中，采用 B/S 架构实现前后端操作交互，在客户端进行 PDF 文件的上传和审阅操作；在服务端，对 PDF 文件进行表格检测、数据识别、数据提取等。

本系统采用奥蓝托公司自主研发产品 TDM 作为基础平台，该平台吸取国外相近产品的框架设计及优点，主体框架使用跨平台语言 JAVA 进行编码开发，数据处理使用 QT（C＋＋）进行编码开发，可以在 Windows、Unix、Linux 等不同操作系统运行。系统多次经过第三方测评，成熟可靠。在 TDM 的基础上进行功能模块的定制化和优化工作，有助于整个系统开发工作的开展，其优点有提高系统扩充模块的能力；减少系统功能之间的耦合度；优化系统对外的封闭程度。

试验数据识别系统所搭建的网络结构是服务器/客户端模式，其优势在于具备普适性、高性能和严格的规则。此模式根据客户端种类具体划分为客户机服务器模式（Clietn/Server，简称 C/S）和浏览器/服务器模式（Brower/Server，简称 B/S）。具体功能描述如下文所示。

1.1　客户端

1）C/S 客户端：提供系统所有的基础工具，例如数据建模工具、动态展现定置工具、流程定义工具、数据导入工具、数据展现工具等。这些工具基于 Java、C＃以及 C＋＋进行开发，用到的技术框架包括 Qt、EMF（eclipse modeling framework）、GEF（graphical editing framework）、SWT（standard widget toolkit）、JFace、Dom4J、Castor、XFire HTTP Client 等。

2）B/S 客户端：作为网页应用功能的集合，包括数据查询、管理以及相关维护工作[8]。通常使用 Javascript 开发语言，基于 js 语言所涉及的技术框架包括 VUE、JSTL、JSP、Ajax（Ext）等。

1.2　服务器端

1）模型层：创建数据模型，提供各模型之间的数据交

互操作。

2）控制层：主要负责前端页面与用户的交互操作，其原理是从通过与数据模型的读写操作实现用户的输入或输出。

3）服务层：主要负责对客户端开发工具的交互，采用 WebService 技术实现前后端的交互。

4）业务逻辑层：主要负责创立规则、实现涉及系统流程的设计，作为相对关键的一层，其效果是承上启下，加强上下之间的耦合程度。

5）持久层：主要负责与数据库之间的交互，其原理是整合数据库增删改查等操作指令，通过自定义方式实现系统与数据库的交互操作，不再使用数据库查询语言。

6）基础层：作为系统运转的主要呈现方式，采用 Spring 容器搭建整体系统的软硬件环境。

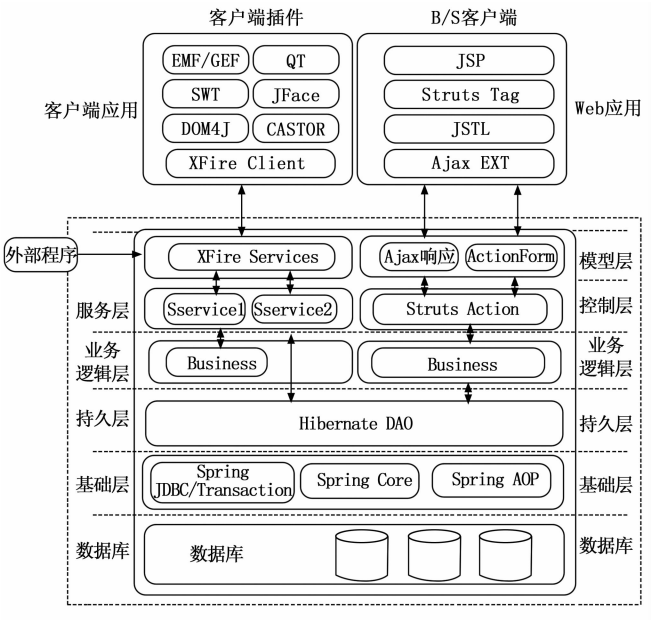

图 1 系统技术架构

整个系统的核心是基于 OCR 的表格检测、表格识别和数据提取功能，将采用 Python 语言进行开发，可以实现对 PDF 文档的转换，自动筛选表格，提取大量数据，以达到所需设计要求。

2 系统软硬件设计

2.1 版面分析

PDF 文档本质上不包含任何段落、句子和单词内容，只包含字符及其位置，因而导致很难区分组成段落的字符与组成表格的字符、页脚、图形描述字符之间的不同，相较于其他格式文档缺少文本流。因而本系统采用 PDF-Plumber 工具库来实现对 PDF 文档的解析。

PDFPlumber 工具库采用 Python 语言开发，基于 pdfminer. six 构建，主要功能包含获取 PDF 中的文本字符、矩形和行的详细信息、表格提取。该工具库由 PDF 类和 Page 类组成，PDF 类的功能是打开 PDF 文档获取整体信息；Page 类作为工具库的核心，获取 PDF 文档的详细信息，包含页面的高宽、页码数量、每个页面的数据对象。

面对 PDF 文档每一页无法区分详细类别的数据对象，通过采用 pdfminer. six 的版面分析算法实现对文档内容的分类。版面分析算法通过对字符定位的启发式方法来重建其中的一些结构，尤其是对于语句和段落。版面分析包含有 3 个阶段，将字符分组为单词和行，其次将行分组为框，最后分层分组文本框，其结果是 PDF 页面上布局对象的有序层次结构。

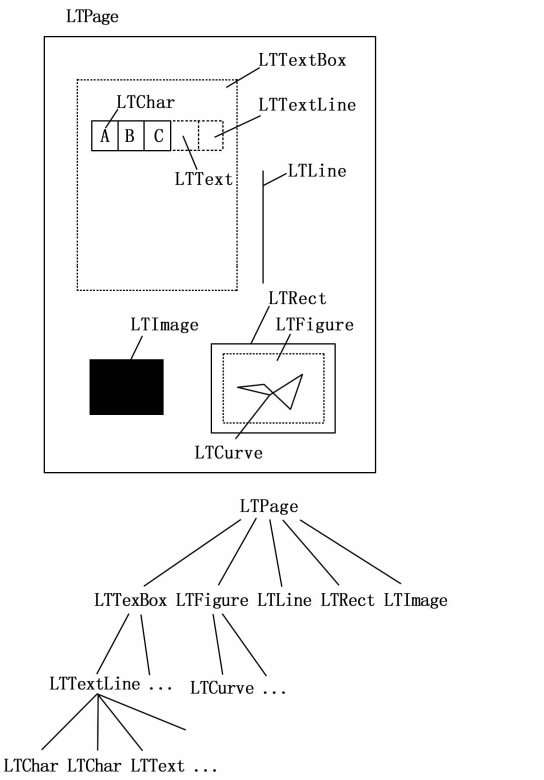

图 2 版面分析结果和层次结构

2.1.1 字符分组

字符转换为文本的第一步是对字符进行分组，每个字符的左下角和右上角都有横、纵坐标，也就是边界框，布局分析算法就根据这些边界框判断字符是否在一起。

水平和垂直接近的字符被分组到一行上。它们之间的距离应该由字符间距（图 3 中的 M）和线条重叠参数决定。两个字符的边界框之间的水平距离应小于字符边距，垂直边界框之间的重叠应小于线条重叠。

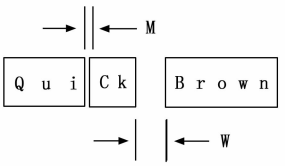

图 3 版面分析结果和层次结构

字符间距和线条重叠的值与字符边界框的大小有关。

字符间距相对于任意一个边界框的最大宽度，而线条重叠相对于任意一个边界框的最小高度。

字符之间需要插入空格，因为 PDF 格式没有空格字符的概念。如果两个字符之间的距离比字距（图 3 中的 W）更远，则插入一个空格。字距是相对于新字符的最大宽度或高度的。更小的字距会创建更小的单词。

通过上述字符分组操作得到的结果是一个行列表。每行包含一个字符列表。这些字符要么是源自 PDF 文件的原始 LTChar 字符，要么是表示单词之间的空格或每行末尾的换行符的插入的 LTAnno 字符。

2.1.2　行分组

字符转换为文本的第二步是分组行。每一行都有一个边界框，该边界框由其所包含字符的边界框决定。通过同样的方法，布局分析算法使用边框对行进行分组。

水平重叠和垂直接近的线被分组。每行的垂直距离由线条边距决定。此边距是相对于边界框的高度指定的。如果边界框的顶部（见图 4 L_1）和底部（见图 4 L_2）之间的距离比绝对线距（即线条边距乘以边界框的高度）更近，则线是接近的。

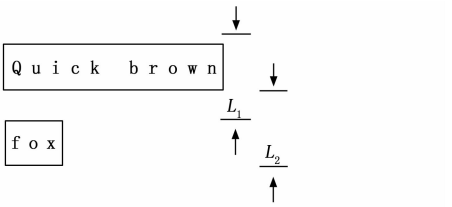

图 4　行分组结果

2.1.3　文本框分层分组

字符转换为文本的最后一步是对文本框进行分组。此步骤重复合并相邻的两个文本框。

边界框的紧密程度计算为两个文本框之间的区域（图中的蓝色区域）。这就是两条线周围的边界框的面积减去每条线的边界框的面积。

通过布局分析算法的三阶段，PDF 文档转换为文本存放在 PDFPlumber. Page 类的性值中，方便后续操作。

2.2　表格检测

通过布局分析算法解析 PDF 文档获取的文本信息包含有多种 PDF 对象，例如文本字符、线条、二维矩阵、图像等。为了进而筛选甄别出表格数据，采用特定的表格检测方法进行表格的提取。

PDFPlumber 的表格检测方法是基于图像形态学的表格线提取，这是一种自底向上的方法，其原理是表格线由水平和垂直方向组成表格且与字符信息在形态上存在较大差别。

2.2.1　候选表格线

在 PDF 文档的每一页中，将可见的线作为候选表格线，适用于抽取线框完整的表格。通过 pdfminer. six 的布局分析算法，已经将文档中的水平、竖直的线解析出来，使用 pdfplumber. container. Container 子类获取线的信息，其中定义了访问 chars、rects、edges 等基本对象的属性。TableFinder 类中的 get _ edges 方法通过 utils 模块中的 filter _ edges 函数对每一个 Page 实例对象中的解析出的 edges 对象进行筛选和过滤，过滤条件包括：方向、最小长度等。

对于线框不完整的表格（包括无线框表格），将根据文本的对齐情况猜测出一些水平和竖直的线，这些线称作为“Text Edge”。TableFinder 类的 get _ edges 方法通过调用同模块中的 words _ to _ edges _ v 和 words _ to _ edges _ h，根据每一页中解析出的 words（word 指的应该是由每一行上彼此间距较小的字符合成的连续字符串）的对齐情况，猜测出竖直方向和水平方向上可能存在的线。

2.2.2　合并边

通过上面的方法将获取到数量不少的线段，其中必然存在冗余，因而需要筛选掉不需要的线段、合并重叠的线段。

1）如果平行线之间的垂直距离非常小，则对他们进行对齐以达到同一条直线上，pdfplumer 使用平均位置进行对齐。

2）对于同一直线上的部分线段出现相互之间近端点的距离非常小的情况，同样需要进行合并。

2.2.3　寻找相交点

因为文档中的表格以及表格单元格基本上都是矩形的，而矩形是可以由其顶点确定的，所以，在找到那些可能是表格或单元格边界的线之后，接下来是找出它们的交点。pdfplumber. table 模块中 edges _ to _ _ intersections 函数，用于找到水平线与竖直线之间的交点，最终的返回的结果是一个字典，以交点坐标作为 key，value 中保存的是相交于该交点的线。

2.2.4　生成单元格

在找到了可能的表格线以及这些线的交点之后，根据这些信息将找到并识别出可能存在的单元格，其步骤主要包含以下几步：

1）首先对所有交点按照自左向右、自上向下的方式排序。

2）找到以每个交点作为左上角的最小的单元格，因为对输入的点进行了排序，所以返回的单元格也是以同样的顺序排序的。

2.2.5　生成单元格

根据前面找到的单元格，把连通的单元格合并到一起生成对应的表格，其步骤主要包含以下几步：

1）对单元格的边框进行处理，生成 4 个角的坐标。

2）根据可用单元格 4 个角的坐标判断单元是否属于当前正在生成的表格。

（1）当单元格与当前正在生成的表格相交时，把该单元格加入到当前表格中，以后该单元格就不再可用了。

（2）当没有单元格可以加入到当前生成的表格的时候，保存该表格，并把当前正在生成的表格设成空表格，判断

剩下可用的单元能够加入到当前表格中。

(3) 当所有单元格都加入到某一表格之后，停止这一过程。

3) 按照表格的左上角坐标进行排序。

4) 过滤掉那些过小的表格。

5) 把剩下的表格封装到 pdfplumber. table. Table 类的实例对象，Table 类中的 extract 方法可以根据表格、单元格以及字符的位置，抽取位于表格及其各个单元格内部的文本，最后以行的形式返回出来。

通过上述五大阶段可实现在 PDF 文档中检测识别出表格信息，通过基于图像形态学的表格线检测方法找寻有效线段，合并重叠的线段，根据线段相交点找寻单元格，最终将单元格合并成表格。

2.3 数据操作

通过 pdfplumber 工具库获取到 PDF 文档中所有的表格数据，系统提供数据展示、数据编辑、数据保存等功能，以满足此系统的实际使用需求。

2.3.1 数据展示

系统采用 B/S 架构搭建，前端采用基于 Javascript 语言的 ExtJS 框架搭建页面，以控件形式展示表格数据。

2.3.2 数据编辑

在前端以控件形式展示提取的表格数据后，系统提供表格数据编辑的功能，包含对表头标题的修改以及表格数据的编辑。

2.3.3 数据保存

在实现对 PDF 文档的解析和表格提取之后，系统提供数据保存功能，主要包含对表格数据、自定义试验信息的保存。

3 系统实现

航天器材料及器件试验数据识别系统将根据上述系统设计进行搭建，本章将详细描述系统实现过程。

图 5 是航天器材料及器件试验数据识别系统的架构设计，采用 B/S 架构实现客户端、服务端的交互操作，基于 OCR 的表格识别功能部署在服务端进行。根据系统设计，客户端界面搭建采用 ExtJS 框架，服务端采用 JAVA 语言开发，其中嵌入运用 Python 语言开发的表格识别功能模块。

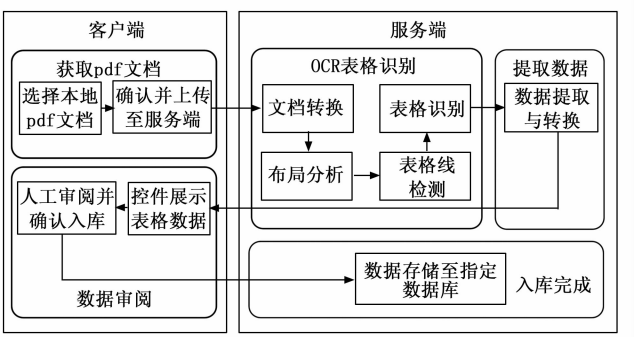

图 5 试验数据识别系统架构

3.1 系统界面

系统界面布局可划分为 3 块，操作区、试验信息区和展示区。操作区提供功能按钮，包含识别前的文档选择按钮、解析完成后的数据编辑按钮、数据保存按钮、原文档预览按钮等。试验信息区显示试验信息表单，提供对试验/仿真名称、数据来源、负责单位、辐照源、类别、时间、人员、国别、子效应类型、测试对象类型、测试对象名称、测试对象型号等。展示区以控件形式展示 PDF 文档识别提取出来的表格数据，包含表头和单元格内容。

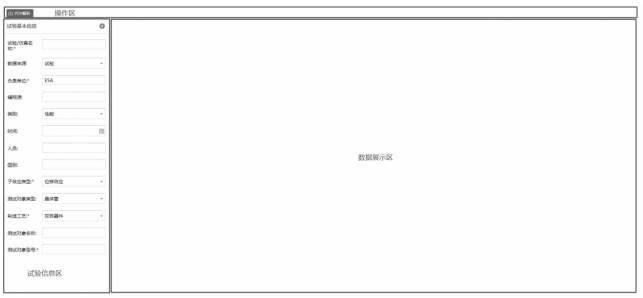

图 6 系统界面

点击 PDF 解析按钮打开文件选择界面，挑选本地 PDF 文档进行识别。

3.2 基于 OCR 的表格识别

3.2.1 PDF 文档读取

在客户端选择 PDF 文档后，其文件路径传输至后台进行解析。首先使用 pdfplumber. PDF 类中的 open 函数打开 PDF 文档，若读取成功则返回一个 pdfplumber. PDF 类实例，该实例中主要包含 metadata、pages 和 len (pdf. pages) 这 3 个属性。metadata 属性是从 PDF 的 Info 中获取元数据键 /值对字典。通常包括 “ CreationDate”，“ ModDate”，“ Producer” 等；pages 属性是一个包含多个 pdfplumber. Page 实例的列表，每一个实例代表 PDF 每一页的信息。

pdfplumber. Pages 类提供了最核心的参数属性和方法函数，其中主要包含有 page _ number、width、height 属性以及 extract _ text (x _ tolerance = 0，y _ tolerance = 0)、extract _ tables (table _ settings) 方法。

3.2.2 表格检测

上述内容实现 PDF 文档读取之后，所有页面信息整合返回一个 pdfplumber. PDF 类实例，其中 pdfplumber. Pages 类包含有字符、线段、图像、表格等信息。pdfplumber 把表格抽取的功能封装在 TableFinder 这个类中，在其构造函数 _ init _ 中，清晰的定义了表格抽取的基本流程。接下来将详细描述实现过程。

1) 识别表格线。

TableFinder 类中的 get _ edges 方法包含了 3 种不同的方式用来确定文档中可能存在的表格线。

对于线框完全的表格，TableFinder 类中的 get _ edges 方法通过通过 utils 模块中的 filter _ edges 函数对每一个 Page 实例对象中的解析出的 edges 对象进行筛选和过滤，

过滤条件包括：方向、最小长度等。

对于线框不完全的表格，TableFinder 类的 get _ edges 方法通过调用同模块中的 words _ to _ edges _ v 和 words _ to _ edges _ h，根据每一页中解析出的 words（word 指的应该是由每一行上彼此间距较小的字符合成的连续字符串）的对齐情况，猜测出竖直方向和水平方向上可能存在的线。words _ to _ edges _ v 和 words _ to _ edges _ h 函数的逻辑归纳为根据 words 位置进行聚类，筛选掉 words 少于 word _ threshhold 的文本行，最终把剩下文本行的边缘线作为找到的边返回。

2）合并表格边。

pdfplumber. table. TableFinder 类的 get _ edges 方法会调用同一模块下的 merge _ edges 函数实现上述功能，如果像素内的平行线不再同一水平或垂直位子则进行修正；如果同一直线上的线段不能拼接为单个线段则进行修正。

3）寻找相交点。

pdfplumber. table 模块中 edges _ to _ _ intersections 函数，用于找到水平线与竖直线之间的交点，最终的返回的结果是一个字典，以交点坐标作为 key，value 中保存的是相交于该交点的线。

4）生成单元格。

pdfplumber. table. TableFinder 类调用同一模块下的 intersections _ to _ cells 函数，根据前面找到的线和交点找出可能存在的单元格。

5）生成表格。

pdfplumber. table. TableFinder 类调用同一模块下的 cells _ to _ tables 函数，根据前面找到的单元格，把连通的单元格合并到一起生成对应的表格。

3.3　表格数据提取和展示

在完成上述表格识别操作后，对 PDF 文档中每一页内容进行表格数据提取，提取出来的数据存放至本地新创建的 Excel 文档中，根据 PDF 页码顺序划分 sheet 页进行数据存放，数据提取效果如图 7 所示。

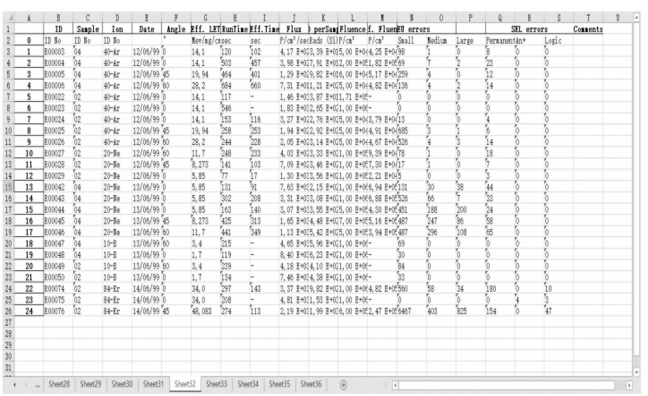

图 7　表格数据提取

试验数据识别系统读取存放表格数据的 Excel 文档，通过 tab 控件形式将每一页提取出的表格数据进行展示，其中根据表头内容是否为空以及特定关键词判定来剔除部分无需展示的表格数据，数据展示如图 8 所示。

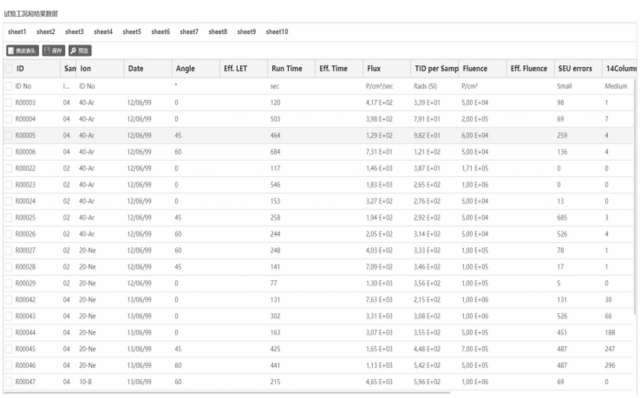

图 8　数据展示

3.4　表格数据操作

3.4.1　数据编辑

1）在数据展示页面，提供表头、表格数据的编辑功能。通过按钮点击打开当前页的表头信息列表，原始表头和修改表头分两列对比展示，如图 9 所示。

图 9　表头信息编辑

2）在 tab 页的表格控件中，通过双击单元格进行数据编辑，如图 10 所示。

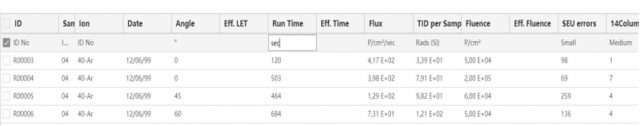

图 10　表格数据编辑

3.4.2　文档预览

数据展示界面提供原始报告预览功能，通过按钮点击实现网页在线预览 PDF 文档，有助于表格数据的编辑操作，如图 11 所示。

4　结果与分析

本文设计了基于 OCR 技术的航天器材料及器件试验数据识别系统，实现了对 PDF 文档的表格识别、数据提取等功能，根据 PDF 文档内表格数量的不同，实现识别速率在

图 11 PDF 文档预览

3 秒到 35 秒不等。在识别规整表格和无旋转的字符数据情况中，识别准确率达到 100%，在识别不规整或存在旋转字符的情况中，识别准确率略有下降，字符识别清晰但字符排列顺序有偏差。经过测试，系统能够正常识别 PDF 文档的表格数据，提取和展示数据正常，对于相对规整的表格不存在表格识别失败或错误的现象，系统运行稳定，对比效果如图 12 所示。

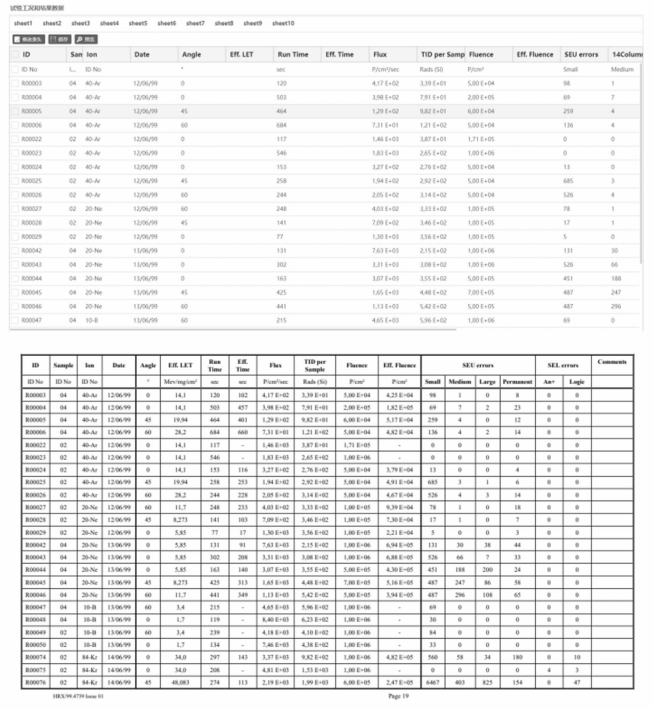

图 12 PDF 文档识别对比

5 结束语

本文介绍了基于 OCR 的航天器材料及器件试验数据识别系统，一个利用 OCR 技术对 PDF 文档格式的航天器材料及器件试验表格数据识别和提取的 B/S 架构系统。系统主要实现了 PDF 文档转文本、表格检测、数据提取、数据编辑等功能。本文结合 EXT、JAVA、Python 等技术，设计并实现了航天器材料及器件试验数据识别系统，具体工作内容如下：

1）简要介绍航天器材料及器件试验数据识别系统的业务架构和数据流程。

2）重点阐述航天器材料及器件试验数据识别系统的关键功能设计。

3）详细展示整体系统的实现过程和功能演示效果。

在系统的实际运行过程中，所设计的 PDF 文档转文本、表格检测、数据提取、数据编辑等功能正常，达到预期目标。针对 pdf 格式文档借助文字识别、图像处理等技术实现自动提取表格数据，解决传统人工提取数据存在的耗时长、出错率高等弊端。对于相关领域的数据识别系统设计具有一定的参考价值和借鉴意义。

该系统仍有改进优化的方面：

1）扩大 PDF 文档识别数据种类和识别复杂度，除了表格，还能包含对曲线、关键词、旋转字符的识别提取。

2）随着深度学习与 OCR 领域的结合，针对 PDF 文档的识别提取能够更为智能和精确，将文档内图像数据也囊括到可识别数据种类中。

参考文献：

[1] 张宁静，袁书培，吴海龙. 基于 OCR 的中文债券图表数据检测和文本识别 [J]. 现代计算机，2021，27（30）：73－81.

[2] 邝 振，崔 喆. 社区选举系统选票中的表格识别算法 [J]. 计算机应用，2017，37（S2）：179－182.

[3] 窦方坤，曹皓伟，徐建良. 基于文本元素的 PDF 表格区域识别方法研究 [J]. 软件导刊，2020，19（1）：113－116.

[4] 黄锦德，郝红卫，张冬霞. 一种新的表格识别特征提取方法 [J]. 计算机工程，2006（6）：215－217.

[5] 包云超，周 全，孔令军，等. 基于行列信息门的表格结构识别网络 [J]. 无线电工程，2022，52（3）：463－469.

[6] 刘长松，潘世言，郑冶枫，等. 一种表格框线检测和字线分离算法 [J]. 电子与信息学报，2002（9）：1190－1196.

[7] 田翠华，张一平，胡志钢，等，李西雨. PDF 文档表格信息的识别与提取 [J]. 厦门理工学院学报，2020，28（3）：70－76.

[8] 曹旭东，曹卫东，朱小宇. 基于 B/S 架构的油田生产数据管理系统应用研究 [J]. 计算机测量与控制，2018，26（8）：142－146.

[9] 梁 鹰，罗伟其. 基于 B/S 的异构数据库信息集成的系统设计与实现 [J]. 计算机工程，2000（12）：23－25，41.

[10] 李霄霄. 基于 OCR 的字符识别的研究与实现 [J]. 科技视界，2017（14）：98－119.

[11] 周长岭. 中文 OCR 中的版面分析技术 [C] //第六届全国汉字识别学术会议论文集，重庆，1996：137－1420.

[12] 王妹华. 文档分析与理解中若干技术的研究 [D]. 南京：南京大学，2001.

[13] 林 强. 基于 OCR 的支票识别系统的研究与实现 [D]. 北京：北京邮电大学，2010.

（下转第 293 页）