DOI:10.20033/j.1003-7241.(2026)04-0110-06

基于 PaddleOCR 的电力系统继电保护定值单信息抽取方法

周陈斌, 孟屹华, 沈蛟骁, 谢夏寅, 薛 峰, 童勤毅

(国网江苏省电力有限公司苏州供电分公司, 江苏 苏州 215004)

摘 要:电力系统继电保护定值单在继电保护设备运行维护中具有关键作用,在定值单信息化过程中,由于大量定值单以扫描件或照片形式保存,不同厂商设备定值单布局与字段名称不统一,难以通过传统光学字符识别(optical character recognition,OCR)等方法实现信息抽取。为此,提出一种基于 PaddleOCR 的定值单信息抽取方法,首先利用可微分二值化网络(differentiable binarization network,DBNet)完成文本区域检测,通过 CRNN-CTC 模型提取文本内容,获得定值名称、参数、单位等基础字段,在此过程中,针对定值单语义专业性强、结构复杂的特点,引入多模态预训练模型 LayoutXLM 实现语义实体识别(semantic entity recognition,SER)。随后采用 PaddleOCR 中的图神经网络(graph neural network,GNN)执行关系抽取(relation extraction,RE),推断实体间的逻辑关联,生成包含实体 ID、标签及文本内容的结构化关联数据。为保证 SER 与 RE 训练效果,使用 PPOCRLabel 完成文本块标注,并通过编号方式结合代码自动生成实体 linking 关系,构建用于训练的完整数据集。实验结果表明,该方法在复杂版式及低质量图像条件下仍能稳定识别关键字段,并准确重建定值单内部的语义结构,实现非结构化文档向结构化数据的高效转换。

关键词:PaddleOCR;定值单;文本检测;文本识别;语义实体识别;关系抽取

中图分类号: TP391 **文献标志码**: A **文章编号**: 1003-7241(2026)04-0110-06

An extraction information method for relay protection settings sheets in power systems based on PaddleOCR

ZHOU Chenbin, MENG Yihua, SHEN Jiaoxiao, XIE Xiayin, XUE Feng, TONG Qinyi

(Suzhou Power Supply Branch State Grid Jiangsu Electric Power Co., Ltd., Suzhou 215004, Jiangsu, China)

Abstract: The relay protection setting sheets of power system play a crucial role in the operation and maintenance of protection devices in power systems. During the digitization of these documents, however, a large number of setting sheets are stored as scanned copies or photos, and the layout structures and field names vary across manufacturers. This diversity makes it difficult for traditional optical character recognition (OCR) methods to effectively extract information. To address this issue, this paper proposes a setting-sheet information extraction method based on PaddleOCR. First, a differentiable binarization network (DBNet) is employed to detect text regions, and a CRNN-CTC model is used to extract textual content, obtaining basic fields such as setting names, parameters, and units. Considering the highly specialized semantics and complex structure of setting sheets, the multimodal pre-trained model LayoutXLM is introduced to perform semantic entity recognition (SER). Subsequently, PaddleOCR's graph neural network (GNN) module is applied for relation extraction (RE) to infer logical associations between entities and generate structured relational data that include entity IDs, labels, and textual content. To ensure the effectiveness of SER and RE training, PPOCRLabel is used for text-block annotation, and entity linking relationships are automatically generated using an indexing strategy to construct a complete training dataset. Experimental results show that the proposed method can reliably recognize key fields and accurately reconstruct the internal semantic structure of setting sheets, even under complex layouts and low-quality image conditions, enabling efficient conversion of unstructured documents into structured data.

Keywords: PaddleOCR; setting sheet; text detection; text recognition; semantic entity recognition; relationship extraction

电力系统继电保护定值单是记录设备保护定值和操作参数的重要技术文档,其信息在设备的运行维护中发挥着关键作用。目前,国内调度部门因信息化程度不一,新旧设备及信息系统的混杂使用,导致需维护的定值单形式多样,且表单内容不规则,基本以表格的形式存在。大量的定值单均为经纸版拍照或扫描后的文件格式,这些定值单因设备厂商不同,其表格布局也分散多样,对此类文件的信息抽取较为困难。以图片或 PDF 文件的形式存在的定值单,会丢失易于被计算机理解的原结构信息。因此,如何让程序从文件中识别表格并抽取信息,成为此类文档电子化的重要研究问题。

近年来应用 OCR 光学字符识技术从图像数据中识别数据并抽取信息为研究热点之一。如文献[1]提出了基于 OCR 的身份证识别系统,对身份证等图像进行了识别。

收稿日期:2024-11-01;**录用日期**:2025-05-25

作者简介:周陈斌(1983—),男,硕士研究生,高级工程师,从事电力系统、继电保护及二次系统的专业管理与技术研究的工作。

引用本文:周陈斌, 孟屹华, 沈蛟骁,等. 基于 PaddleOCR 的电力系统继电保护定值单信息抽取方法[J]. 自动化技术与应用, 2026,45(4):110-115. (ZHOU Chenbin, MENG Yihua, SHEN Jiaoxiao, et al. Method for extracting single information of relay protection settings in power systems based on PaddleOCR[J]. Techniques of Automation and Applications, 2026,45(4):110-115.)

文献[2]利用 OCR 技术与 ResNet 残差网络,提出了人物敏感广告识别算法,先构建敏感人名库,再进行规则匹配审核,最后实现基于图文结合的人物敏感广告图片识别。文献[3]提出了基于 OCR 技术的票据识别算法,对火车票和发票等票据文字进行识别,并将非结构化数据转化为结构化数据。此类研究侧重对布局及格式相对固定的图片信息进行数据识别。针对非结构文本尤其是具有不规则表格特性对象的布局及文本识别方面,文献[4]利用手机摄像方式采集定值单图像信息,设计算法对原始图像进行灰度增强、二值化以及倾斜校正等预处理,从固定模板中提取出清晰的定值单表格部分。近些年来,深度学习算法的发展成熟显著地推动了 OCR 技术的普及应用,同时也大幅提升了 OCR 技术的识别率,使得 OCR 在交通、金融、医疗等重要领域得到实际应用[5-7]。针对表格的识别又复杂得多,原因一是表格结构的多样性导致通用识别框架难以构建,二是表格内容的专业性和语义复杂性导致标注困难、模型迁移能力差。传统图像处理方法主要有表格边缘检测、框线检测聚类、投影等[8-12],这些方法有效解决了固定格式表格边界定位识别的问题,但不能实现不规则模板的表格识别。

文献[13]基于形态学检测原理和在 Tesseract-OCR 字符识别的基础上,设计一套原料制式表单识别系统,采用形态学检测对表格框架进行提取,通过动态掩膜及角点检测实现单元格分割,取得不错的效果。文献[14]介绍一种深度学习的表格结构识别方案,能处理简单表格,但随结构复杂度上升效果下降。结合人工智能深度学习和目标检测算法,一些学者采用卷积循环神经网络(convolutional recurrent neural network, CRNN)[15]、快速卷积神经网络[16](Faster region-based convolutional neural networks,Faster R-CNN),提升了检测准确率,但前提是需要大量的数据集进行训练,受限于工业特定领域的数据量和数据采集难度,且对领域专有名词的识别率不理想。此类研究针对表单数据进行图片的结构和文本进行识别,但信息抽取除识别出内容外还需提取信息的逻辑关系及语义关系,且针对特定领域的专业语义的提取是一项有挑战性的工作。本文基于 PaddleOCR 技术,在进行数据预处理及标注的基础上,利用卷积神经网络(CNN)进行图像特征的提取和文本检测。重点在于采用基于 CRNN 和连接时序分类(connectionist temporal classification,CTC)进行文本识别模型对文本区域中的内容进行提取,应用多模态预训练模型 LayoutXLM 完成了表格类定值单中特定语义实体识别(SER)对文本内容进行语义实体的提取。最后采用 PaddleOCR 中图神经网络(GNN)对语义实体之间的关系进行建模,推断这些语义实体之间的关系,生成结构化关联数据,其格式包括实体 ID、实体标签和对应的文本内容。

1 定值单信息抽取方法

PaddleOCR 是由百度公司开源的超轻量 OCR 系统,主要由 DB 文本检测[17]、检测框矫正[18]和 CRNN 文本识别[19]三部分组成。本文通过引入 PaddleOCR 中的检测模型对目标图像进行处理,识别出包含文本的区域,并获取文本框及其位置信息。

主体流程如图 1 所示,首先通过图片数据预处理对原始图片数据进行整理、拆分、归一化处理,为后续处理提供一致输入。随后对定值单进行数据标注,生成分类标签或位置框等标记信息,为模型训练提供监督支持。后续即是信息抽取的核心步骤,分别为:1)文本检测。通过 PaddleOCR 中的 DBNet 模型,识别出定值单中所有文本区域的边界框;2)文本识别。使用 PaddleOCR 中的 CRNN 模型,将边界框内的文本识别为可读的字符串;3)语义实体识别(SER)。通过 LayoutXLM 模型,结合视觉和文本信息,对文本进行分类,识别出具体的实体类型(如“定值名称”、“定值”、“作用”等);4)关系抽取(RE)。通过图神经网络,推断不同语义实体之间的关系,生成结构化的关联数据。

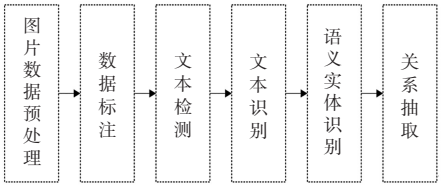

图 1 定值单信息抽取的主体流程

Fig. 1 Main workflow of setting sheet information extraction

1.1 数据预处理

首先,通过将 PDF 格式的定值单转化为单张 jpg 图片,生成模型训练所需的数据集。统一文件命名,并将图片分成训练集和验证集,得到未标注的数据集。根据定值单中的特点进行自定义标签,将所有元素分为以下几类标签。HEADER, QUESTION, ANSWER, INDEX, DATE, OTHER。这些标签可能用于定值单中标识:头部信息、问题、答案、序号、日期等文本区域。将该几个标签保存成 class_ list_ xfun. txt 放在数据集根目录下。

1.2 数据标注

本文使用飞桨提供的 PPOCRLabel 进行数据标注为模型准备训练所需的带有标签和关系信息的数据集,并通过代码为语义实体建立相应的链接关系。导入数据后,打开 PPOCRLabel 工具进行数据标注。在使用时需注意,由于此工具本身不支持关系标注(linking),因此本文在标注阶段通过为不同的关键文本块分配特定的序号(如 INDEX1,QUESTION1,ANSWER1),然后在导出的标注数据中通过代码根据相同序号来自动添加 id 和 linking 字段,建立文本块之间的关系。例如,针对某个问题及其对应的答案和索引项,给它们加上相同的编号,如:

INDEX1:表示这个文本块是序号 1 的索引;

QUESTION1:表示这个文本块是序号 1 的问题;

ANSWER1:表示这个文本块是序号 1 的答案。

以此类推,针对每一组相关文本块,用相同的编号(如 1,2 等)来标记。

PPOCRLabel 在生成的标注文件时不会包含关系(linking),只会导出每个文本块的基本信息,如文本内容、位置坐标和标签等。数据格式如下。

```
[
{
"transcription":"问题 1 的文本",
"points":[[x1,y1],[x2,y2],[x3,y3],[x4,y4]],
"label":"QUESTION1"
},
{...}
]
```

导出标注数据后,需要通过代码为这些文本块生成唯一的 id,并基于相同的编号来建立 linking 关系。处理完所有数据后,将结果保存回文件中,并将其用于训练 RE 模型。在这个过程中,通过手动为文本块标注。相同的编号,再通过代码自动生成 id 和 linking,可以有效地实现 RE 任务所需的关系标注。最后,通过标注以及代码的处理,得到完整的数据集结构如下。

```
[
{
"transcription":"问题 1 的文本",
"points":[[100,100],[200,100],[200,200],[100,200]],
"label":"QUESTION1"
},
{...}
]
```

为每个文本块生成唯一的 id:遍历数据集中的每个文本块,并按顺序为它们分配一个唯一的 id。根据标签中的编号生成 linking:通过遍历数据集中的文本块,找到相同编号(如 QUESTION1,ANSWER1,INDEX1)的文本块。将这些文本块的 id 互相连接,形成 linking 关系。处理完所有数据后,将结果保存回文件中,并将其用于训练 RE 模型。在这个过程中,通过手动为文本块标注相同的编号,再通过代码自动生成 id 和 linking,可以有效地实现 RE 任务所需的关系标注。

1.3 模型训练

模型训练是信息抽取的主体工作,涉及到文本检测模型、文本识别模型、语义实体识别(SER)、关系抽取模型(RE)及部分内容。从预训练模型库中加载适合当前任务的模型,为后续优化提供基础,根据具体任务需求调整训练参数,如学习率和批量大小,以提升模型的训练效果。在调整后的参数下对模型进行训练,使其逐步收敛并优化在特定任务上的表现。最后利用训练好的模型对新数据进行推理,实现对未知数据的预测或分类,该流程为构建高效的机器学习系统提供了标准化的操作步骤。

1.3.1 文本检测模型

文本检测是 OCR 系统中的第一步,用于识别图像中的文本区域。PaddleOCR 通过卷积神经网络(CNN)进行图像特征的提取,并利用特定的检测算法识别文字所在的位置。本文通过 PaddleOCR 中的 DBNet 模型进行文本检测,该模型能够处理复杂背景和不同尺寸的文本区域,尤其适合处理 A4 尺寸的大型文档。检测结果输出带有文本区域的边界框,供后续文本识别步骤使用。

文本检测模型通过分析图像中的视觉特征来定位这些文本区域,并返回它们的位置信息(通常以矩形框或四边形的形式表示),训练好的文本检测模型可以在推理阶段用于检测新图像中的文本区域。模型的输出为每个文本区域的坐标,这些坐标将用于下一步的文本识别。

1.3.2 文本识别模型

文本识别的任务是将检测出的文本区域中的内容提取出来,转换为机器可读的文字。在文本检测之后,CRNN 与 CTC 的结合模型用于将文本区域的内容转化为可读字符串。CRNN 能够有效应对定值单中字符间距不均的情况,适用于不同语言和复杂字体的文本识别。

文本识别模型的主要作用是从文本检测模型输出的文本区域(即文本框)中识别出具体的文字内容。简单来说,文本识别模型的任务是将这些图像中的文字转化为机器可理解的文本信息。它能够处理各种类型的文字,包括印刷体、手写体、不同字体的文本,甚至处理旋转、倾斜或扭曲的文本区域。在 OCR 流程中,文本识别模型是继文本检测之后的关键步骤,直接影响到最终的文字识别准确率。通过识别模型的输出,文档中的文字信息可以被提取并用于进一步的处理和分析。

文本识别模型包括加载 PaddleOCR 预训练模型并在特定数据集上微调以提升性能,使用 CTC 损失函数以处理可变长度的输出序列,并通过设置优化器的学习率和正则化参数来稳定训练和加速收敛等。模型训练完成后,可用于新图像的文本识别,包括输入待识别的图像、输出文字序列及后处理步骤,如无效字符删除和字符校正,以获得最终识别结果。

为了从定值单中提取特定的字段,本文采用 LayoutXLM 模型,该模型结合了视觉与文本特征,能够识别出诸如“序号”、“定值名称”等关键实体,并且对于带有表格和复杂布局的文档也有较高的识别精度。

1.3.3 语义实体模型(SER)

语义实体(semantic entity,SE)模型用于从文档中识别出特定类型的语义实体。在定值单 OCR 任务中,SE 模型可以识别出文档中的关键内容,例如 header、index、question、answer 等实体。这些实体往往具有明确的语义,可以通过 SE 模型进行分类和提取。SE 模型的作用在于对已经通过 OCR 识别出的文字进一步分析和分类,使得模型能够理解文档的结构与内容,为后续的关系抽取和信息提取提供基础。

在模型评估过程中,将训练中模型参数默认保存在 Global.save_model_dir 目录下。在评估指标时,需要置

Architecture. Backbone. checkpoints 指向保存的参数文件。评估数据集可以通过 configs/kie/vi _layoutxlm /ser _vi _layoutxlm_xfund_zh. yml 修改 Eval 中的 label_file_path 设置。

预测结果如图 2 所示。从中可见模型已经正确识别出实体关系,对不同的文本设置了不同的标签(HEADER、INDEX、QUESTION、ANSWER)。

图 2　定值单语义实体模型评估与预测

Fig. 2　Evaluation and prediction of the semantic entity model for setting sheets

1.3.4　关系抽取模型(RE)

关系抽取(relation extraction,RE)是从文本中识别和提取实体之间的语义关系的任务。在定值单 OCR 任务中,RE 的主要作用是将文本块之间的关系明确地提取出来,构建一个结构化的语义网络。例如,将 Header(标题)与 Index(序号)、Question(问题)、Answer(答案)等标注之间的关联关系识别出来,从而能够将多个独立的文本块组合成语义上的整体信息。

对于定值单的场景,识别出文本实体间的关系后,就可以将文本进行结构化,转换成统一的数据格式。定值单中的信息不仅包括单个实体,还涉及到实体间的关系,如“定值名称”与“定值”的对应关系。通过图神经网络(GNN)模型对语义实体间的关系进行建模,重构定值单中的逻辑结构。

在关系抽取(RE)模型的数据准备中,需要文本框的位置信息、文本内容、实体标签(如“Header”、“Index”、“Question”、“Answer”)及其关联关系。由于 PPOCRLabel 工具不支持关系标注,需为每个关联的文本框手动赋予相同序号标记(如“Index1-Question1-Answer1”),以便后续通过代码自动匹配并生成 linking 关系。具体流程包括通过 PPOCRLabel 进行实体和序号标注,随后编写代码读取标注文件并根据序号匹配生成关系数据,以此确保模型能准确抽取关联信息。

如图 3 所示,通过 RE 模型的预测,定值单中的不同实体通过直线连接起来,这些直线代表了实体间的关系已经成功识别。通过这种关系抽取,原本独立的文本框被有序关联,构建出了实体间的语义网络。结合之前通过 SE(语义实体识别)模型得到的实体标签信息,整个定值单中的关键信息,如 Header、Index、Question 和 Answer 等,都被结构化提取并关联在一起。最终,SE 和 RE 模型的结合实现了对定值单所需数据的完整抽取,将其统一转换为一致的格式,便于后续的处理和应用。这种统一的数据格式能够更好地支持自动化分析、信息检索等任务。

图 3　定值单关系抽取模型评估与预测

Fig. 3　Evaluation and prediction of the relation extraction model for setting sheets

2　信息抽取

通过 SE(语义实体识别)和 RE(关系抽取)模型,已经将定值单中的关键信息抽取并识别出来,但为了便于信息抽取,需要将 SE 和 RE 模型的结果进行归一化,转化为统一的数据格式。数据归一化的核心在于将不同的实体和它们的关系转化为结构化的、易于处理的格式。整个流程首先收集 SE 模型输出的文本实体及其类别(如 Header、Index、Question、Answer 等),为关系匹配提供依据。随后收集 RE 模型输出的实体关系(如 Header 与 Index 或 Question 与 Answer 的对应关系)。最后将 SE 和 RE 模型的输出转换为包含实体 ID、实体标签和文本内容的统一结构化数据格式。得到如图 4 结果。

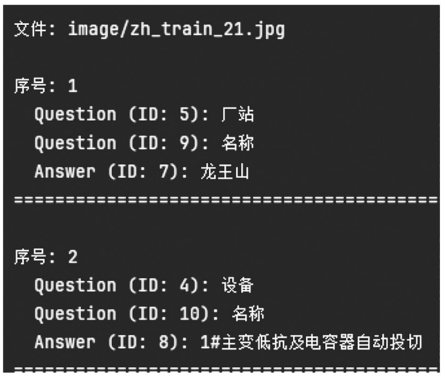

图 4　定值单归一化处理结果

Fig. 4　Normalization processing results of the setting sheet

3　实验验证

实验针对江苏某地区 400 份 PDF 格式的定值单文件进行数据处理。在数据准备阶段,将 PDF 文件拆分为单页图片,每张图片对应一页定值单内容,对图片进行初步筛选,剔除模糊、损坏或无关的图片,以确保数据质量。筛选后的图片被打乱并重新命名,以消除数据顺序对模型训

练的影响。将所有图片按7∶3的比例划分为训练集(70%)和验证集(30%),用于模型训练和性能评估,并确保两者在定值单类型和扫描质量等方面的分布一致,相关数据集如图5所示。

(a)定值单1 (b)定值单2 (c)定值单3 (d)定值单4 (e)定值单5

(f)定值单6 (g)定值单7 (h)定值单8 (i)定值单9 (j)定值单10

(k)定值单11 (l)定值单12 (m)定值单13 (n)定值单14 (o)定值单15

图5 定值单数据集准备及划分

Fig. 5 Preparation and partitioning of the setting sheet dataset

随后对训练集和验证集中的图片进行人工标注,包括文本区域及其内容标注(主要通过自动标注完成,并对特殊符号进行手动标注)以及结构化标注,明确关键字段及其语义对应关系(例如,将“瞬时切低抗电压”标记为“定值名称”)。

分别训练OCR模型A和信息抽取模型B,以实现从PDF或图片中提取结构化数据的目标。首先,针对OCR模型A,使用内容标注后数据作为训练集,并采用PaddleOCR进行模型训练。在训练过程中,通过验证集评估模型的文本检测与识别精度,并调整超参数以优化模型性能。最终,训练得到的OCR模型A可用于解析PDF或图片,并生成对应的如图6所示文本结果。

{"file_title": "XX继电保护与系统自动装置整定通知书", "table": [{"title": ["苏继第XX号启用", "苏继第XX号作废"], "row": [["站　　名", "XX变", "设 备 名 称", "XX线"], ["互感器变比", "CT: 1200/5　PT: 110/0.1", "变 更 理由", "命名变更，定值调整"], ["最大负荷电流", "600A", "最低运行电压", "90 kV"], ["保 护 装 置 型 号：PSL-621C"], ["距 离 保 护 定 值"], ["软件版本", "4.50", "校验码", "D988"], ["01", "KG1 控制字", "8021", "13", "ZD1 接地距离Ⅰ段", "13.75Ω/3Ω"], ["02", "ZKJ 线路正序阻抗角", "71度", "14", "ZD2 接地距离Ⅱ段", "13.7Ω/3Ω"], ["03", "RRD 距离保护电阻定值", "73.3Ω/16Ω", "15", "ZD3 接地距离Ⅲ段", "16Ω/3.5Ω"], ["04", "104 零序辅助启动门坎", "120A/0.5a", "16", "T1D 接地距离Ⅰ段时间", "0s"], ["05", "K0R 零序电阻补偿系数", "0.6", "17", "T2D 接地距离Ⅱ段时间", "0.3s"], ["06", "K0X 零序电抗补偿系数", "0.6", "18", "T3D 接地距离Ⅲ段时间", "1.1s"], ["07", "ZX1 相间距离Ⅰ段", "13.75Ω/3Ω", "19", "IPT1 过流保护Ⅰ段电流", "2400A/10a"], ["08", "ZX2 相间距离Ⅱ段", "16Ω/3.5Ω", "20", "IPT2 过流保护Ⅱ段电流", "912A/3.8a"], ["09", "ZX3 相间距离Ⅲ段", "73.3Ω/16Ω", "21", "TPT1 过流保护Ⅰ段时间", "0.3S"], ["10", "T1X 相间距离Ⅰ段时间", "0s", "22", "TPT2 过流保护Ⅱ段时间", "2.1 s"], ["11", "T2X 相间距离Ⅱ段时间", "0.3s", "23", "XCJ 测距比例系数", "11.9KM/Ω"], ["12", "T3X 相间距离Ⅲ段时间", "2.1s", ""], ["KG1 保 护 控 制 字 说 明（13-10备用，置0）"], ["位号", "内容", "说　　明"], ["15", "1", "CT、PT自检投入"], ["14", "0", "CT额定电流为5A"], ["13－10", "0", "备用"], ["9", "0", "距离Ⅲ段合闸加速瞬时动作"], ["8", "0", "过流保护退出"], ["7", "0", "距离Ⅲ段偏移特性退出"], ["6", "0", "PT断线时健全相距离保护退出"], ["5", "1", "PT断线时过流保护投入"], ["4", "0", "不对称故障相继速动功能退出"], ["3", "0", "双回线相继加速动功能退出"], ["2", "0", "振荡闭锁功能退出"], ["1", "0", "后加速Ⅲ段退出"], ["0", "1", "后加速Ⅱ段投入"], ["零序保护、重合闸和自动周减载定值"], ["软件版本", "4.51", "校验码", "3465"], ["01", "控制字", "8E41", "13", "重合闸检同期定值", "30度"], ["02", "零序不灵敏Ⅰ段电流", "1800A/7.5a", "14", "重合闸检无压定值", "30 V"], ["03", "零序Ⅰ段电流", "1800A/7.5a", "15", "重合闸时间", "2 s"], ["04", "零序Ⅱ段电流", "960A/4 a", "16", "低周减载频率", "45 Hz"], ["05", "零序Ⅲ段电流", "480A/2 a", "17", "低周减载时间", "20s"], ["06", "零序Ⅳ段电流", "288A/1.2 a", "18", "低周减载闭锁电压", "60 V"], ["07", "零序加速段电流", "720A/3 a", "19", "低周减载闭锁滑差", "5 Hz/S"], ["08", "零序Ⅰ段时间", "0 s", "20", "低压减载电压", "60 V"], ["09", "零序Ⅱ段时间", "0.3s", "21", "低压减载时间", "20s"], ["10", "零序Ⅲ段时间", "0.6s", "22", "闭锁电压变化率", "10 V/S"], ["11", "零序过流Ⅳ段时间", "1.3s", "23", "失灵启动电流", "100a"], ["12", "零序加速段时间", "0.3 秒", ""], ["保 护 控 制 字 说 明（3-4备用，置0）"], ["15", "1", "电流、电压求和自检功能投入。"], ["14", "0", "CT额定电流为5A。"], ["13", "0", "备用"], ["12", "0", "加速段不经二次谐波制动。"], ["11", "1", "零序电流Ⅰ段带方向。"], ["10", "1", "零序电流Ⅱ段带方向。"], ["9", "1", "零序电流Ⅲ段带方向。"], ["8", "0", "零序电流Ⅳ段不带方向。"], ["7", "0", "零序电流加速段不带方向。"], ["6", "1", "零序保护经无3U0突变量闭锁。"], ["5", "0", "PT断线时零序保护不延时动作。"], ["2", "0", "重合闸不检同期。"], ["1", "0", "重合不检无压。"], ["0",

图6 信息抽取结果

Fig. 6 Information extraction results

对于信息抽取模型B,使用结构化标注数据作为训练集,并基于PaddleOCR的SE+RE预训练模型进行训练,根据结构化数据的数据特征,自动匹配合适的模板文件,输出如图7所示统一格式的定值单。

对120份定值单文档进行信息提取,通过和相应下发执行的数字化定值数据进行比对,实验统计结果表明,本文所提方法识别准确率达92.08%,误差主要来自于部分图片文档存在少量的信息缺失。在Intel i7 13代CPU主机上单

次信息提取过程平均耗时为 3.34 s,达到可应用的效果。

第一套系统参数定值

序号	定值名称	整定值
01	定值区号	
02	被保护设备	#2主变
03	主变高中压侧额定容量	240MVA
04	主变低压侧额定容量	120MVA
05	中压侧接线方式钟点数	12
06	低压侧接线方式钟点数	11
07	高压侧额定电压	220kV
08	中压侧额定电压	115kV
09	低压侧额定电压	10.5kV
10	高压侧PT一次值	220kV
11	中压侧PT一次值	110kV
12	低压侧PT 一次值	10kV
13	高压1侧CT一次值	2500A

图 7 归一化导出的定制单格式

Fig. 7 Normalized output format of the setting sheet

4 结论

本文提出的基于 PaddleOCR 的定值单信息抽取系统,能够有效识别和分类定值单中的文本区域,提取其中的关键信息,并重构实体之间的关系。实验结果表明,本文方法在复杂文档场景中具有良好的适应性,能够有效处理大量技术文档的自动化处理需求。本系统可以进一步扩展至更多的领域,例如其他类型的技术文件或多语言场景下的 OCR 任务,同时在大规模数据集上进一步优化模型性能。

参考文献

[1]曹佳宇, 陆汝华, 刘宇平, 等. 基于 OCR 的身份证识别系统[J]. 信息技术与信息化, 2021(1):45-47.

[2]杜佳. 基于残差网络和 OCR 技术的人物敏感广告识别[J]. 电子测试, 2022, 36(18): 47-49.

[3]王兴, 郑勇锋, 严永兵,等. 基于 OCR 技术的票据识别算法研究[J]. 智能计算机与应用,2021, 11(11):101-106.

[4]常俊晓, 应宇鹏, 廖小兵, 等. 基于图像处理的继电保护装置定值自动核对方法[J]. 电测与仪表, 2021, 58(11):7.

[5]史建伟. 基于深度学习的车牌识别系统的设计与实现[D]. 南京:南京邮电大学, 2020.

[6]张振宇, 姜贺云, 樊明宇. 一种面向银行票据文字自动化 识别的高效人工智能方法[J]. 温州大学学报(自然科学版), 2020, 41 (3): 47-56.

[7]陈德华, 冯洁莹, 乐嘉锦, 等. 中文病理文本的结构化处理方法研究[J]. 计算机科学, 2016, 43(10):272-276.

[8]XIAO D, SUN H, BAO Z. A multi-table image recognition system based on deep learning and edge detection[C]//Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science (AICS 2019). Shenyang: Northeastern University, 2019:200-207.

[9]LIANG Q, PENG J, LI Z, et al. Robust table recognition for printed document images [J]. Mathematical Biosciences and Engineering, 2020, 17(4):3203-3223.

[10]高良才, 李一博, 都林, 等. 表格识别技术研究进展[J]. 中国图象图形学报, 2022, 27(6):1898-1917.

[11]张泽彧. 基于对比学习和 SAB 模块的重叠关系抽取方法研究[J]. 自动化技术与应用,2025,44(11):106-110.

[12]张志刚. 基于深度学习的铁路通信继电保护测试异常诊断技术研究[J]. 自动化技术与应用,2025,44(8):98-101,114.

[13]方浩东, 鲍敏. 工厂检测检验用手写表格的识别及数字化处理方法[J]. 软件工程, 2023, 26(5):20-23.

[14]SCHREIBER S, AGNE S, WOLF I, et al. DeepDeSRT: deep learning for detection and structure recognition of tables in document images[C]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017). Kyoto, Japan: Institute of Electrical and Electronics Engineers, 2017:1162-1167.

[15]SHI B, BAI X, YAO C. An end-to-end trainable neural network for image- based sequence recognition and its application to scene text recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11):2298-2304.

[16]SUN Ningning, ZHU Yuanping, HU Xiaoming. Faster R-CNN based table detection combining corner locating [C]//2019 International Conference on Document Analysis and Recognition (ICDAR 2019). Sydney, Australia.: Institute of Electrical and Electronics Engineers, 2019:1314-1319.

[17]LIAO M, WAN Z , YAO C, et al. Real-time scene text detection with differentiable binarization [J]. Proceedings of the AAAI Conference on Artificial Intelligence,2020,34(7):11474-11481.

[18]YU Deli, LI Xuan, ZHANG Chengquan, et al. Towards accurate scene text recognition with semantic reasoning networks [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: Institute of Electrical and Electronics Engineers, 2020: 12110-12119.

[19]LI Wei, CAO Longbing, ZHAO Dazhe, et al. CRNN: integrating Classification Rules into Neural Network [C]//2013 International joint conference on neural networks (IJCNN2013). Dallas, Texas, USA : Institute of Electrical and Electronics Engineers, 2013:2196-2203.

(上接第 86 页)

[8]周云浩, 杨宝杰, 刘丹, 等. 基于随机森林算法的电力工程数据预测分析建模与仿真[J]. 电子设计工程, 2024, 32(4):103-106, 111.

[9]周鹤, 黄建军. 基于大数据分析的光通信系统关键设备状态识别[J]. 激光杂志, 2023, 44(12):167-172.

[10]谢雨洁, 肖友刚, 王田天, 等. 基于异常检测的轴承退化阶段识别方法[J]. 中南大学学报(自然科学版), 2022, 53(5):1740-1749.

[11]郝颖, 冬雷, 王丽婕, 等. 基于数学形态学去噪的光伏发电限电异常数据识别算法[J]. 中国电机工程学报, 2022, 42(21):7843-7854.

[12]李峰, 张建业, 霍伟伟, 等. 基于光纤光栅的 OPGW 异常振动信号分类识别[J]. 光学与光电技术, 2023, 21(5):38-42.

[13]王立勇, 朱洪伟, 柳小峰, 等. 基于改进人工鱼群算法的智慧电厂多储能优化调度方法[J]. 自动化技术与应用, 2025, 44(11):175-180.

[14]胡炜, 陈传海, 郭劲言, 等. 考虑工况变化的数控刀架运行状态异常检测方法[J]. 吉林大学学报(工学版), 2022, 52(2):329-337.

[15]杨婧, 辛明勇, 宋强. 基于实时数据流特征提取的设备能耗异常识别算法研究[J]. 自动化技术与应用, 2024, 43(3):74-77.

[16]严宇平, 洪雨天, 陈守明, 等. 基于参数估计的配电网载波通信异常信号识别方法[J]. 电测与仪表, 2022, 59(10):123-129.