基于 CRNN 的质量文档识别系统

曹叶欢 [1]
CAO Yehuan

摘 要 在海洋工程类项目质量管理过程中，为便于后期竣工资料管理如文件信息追溯、查找，同时形成企业历史项目数据库，时常需要将纸质文件转录为电子数据。传统处理方法中，很大一部分纸版报告受限于不同专业内容，格式复杂多样，只能由人工手动转录完成，其效率、时效性和准确度越来越难以满足当下企业生产需求。因此，一种可直接由纸版文件转为电子数据的方法将节省大量人工录入成本，进一步丰富完善企业完工项目数据库。基于机器学习中的神经网络模型 CRNN 进行文字识别，完成了输入图像优化处理、文本定位分割、生产环境数据集训练等工作，最终在实际应用场景数据测试中达到了 93% 的准确度，能满足企业实际生产需要。

关键词 CRNN；文本识别；深度学习；质量文档

doi：10.3969/j.issn.1672-9528.2024.10.012

0 引言

光学字符识别（optical character recognition，OCR）是指通过电子设备获取文字或字符的数字影像信息，再经由一系列数据处理，最终将图像信息转换为文本信息的技术手段。早期 OCR 识别仅能识别印刷体字符，且准确率较低，随着深度学习技术的发展，OCR 识别准确度也获得了极大提高。CRNN 是深度学习算法的一种，当前深度学习在图片识别、语音识别等领域已得到十分广泛的应用。其核心中的多层卷积神经网络由若干个权重矩阵和偏置矩阵组成，通过预训练数据自动进行网络参数调整，以达到像人一样的智能决策。CRNN 神经网络模型由 CNN 卷积神经网络和 RNN 循环神经网络组成，CNN 网络可将输入图像转换为高度为 1 的任意长

1. 中海福陆重工有限公司 广东珠海 519000

[8]GOH A S, PREISS M, STACY N J S, et al.Bistatic SAR experiment with the ingara tmaging radar[J].IET radar, sonar & navigation, 2010,4:426-437.

[9]GUO Y, YU Z, LI J, et al.Focusing spotlight-mode bistatic GEO SAR with a stationary receiver using time-doppler resampling[J].IEEE sensors journal,2020,20:10766-10778.

[10]WANG Y, TAN W, HONG W, et al.Focusing bistatic circular SAR data using polar format algorithm[C]//2009 2nd Asian-Pacific Conference on Synthetic Aperture Radar (APSAR 2009). Piscataway: IEEE,2009:989-992.

[11]ZHANG J, LIAO G, XU J, et al.Study on performance of bistatic circular synthetic aperture radar imaging using geometric diversity[J].IET Radar, Sonar & Navigation, 2018, 12(4): 458-465.

[12]XIE H, SHI S, LI F, et al.Comparison of imaging properties between monostatic and bistatic circular SAR[C]//2017 2nd International Conference on Image, Vision and Computing (ICIVC). Piscataway, NJ: IEEE Computer Society,2017:601-605.

[13]LI T, CHEN K, JIN M J.Analysis and simulation on imaging performance of backward and forward bistatic synthetic aperture radar[J].Remote sensing,2018,10:1676.

[14] 笪敏 . 主动式近距离毫米波成像的距离徙动算法研究 [D]. 合肥 : 合肥工业大学 ,2018.

【作者简介】

刘露（1991—），男，安徽舒城人，硕士，工程师，研究方向：雷达总体、雷达成像。

戴一鸣（1990—），男，安徽蚌埠人，博士，高级工程师，研究方向：雷达成像。

郑智坤（2001—），男，福建南平人，硕士，助理工程师，研究方向：雷达成像。

高 敏（1989—），男，安徽含山人，博士，高级工程师，研究方向：雷达总体、雷达成像。

杨 毅（1969—），男，安徽合肥人，本科，研究员，研究方向：雷达总体、雷达成像。

（收稿日期：2024-06-27）

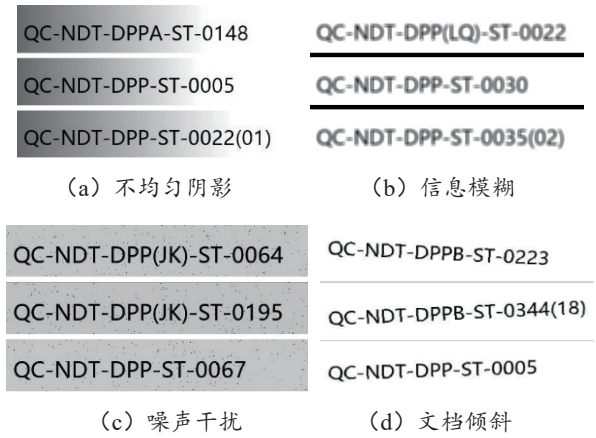
度的图像特征序列，最终经由 RNN 标注输出，广泛应用于文本识别场景中 [1-3]。

在海洋工程项目的质量管理过程中，会涉及大量纸质文件，诸如各类证书、检查报告的审核、信息追溯等工作。在企业数字化转型背景下，质量管理数字化程度的逐步加深，各类配套信息化软件系统逐渐完善，但因为受限于各类文档格式多样、来源不一等实际情况，底层基础数据获取仍有大量数据需要人工转录，一个高速、高精度的 OCR 文件识别系统可以节省大量人力，进一步提高企业办公自动化程度，同时提高数据的准确度 [4-5]。目前基于深度学习神经网络的 OCR 文字识别系统已在银行金融等领域被广泛应用 [6-7]。刘乐等人 [8] 基于改进 SVTR 算法实现了工件坯号的识别，在板坯号识别中取得了良好的检测效果，缓解了实际场景中背景光照变化对图像信息识别的干扰。王秀光等人 [9] 设计并实现了企业通用文档的 OCR 识别平台，针对企业中各类文件版式不同，通过平台内置自定义模板功能，配合内置优化预训练 OCR 模型，以达到较高识别准确度。现阶段文本识别领域正飞速发展，识别模型精度逐渐提高，识别场景也逐渐复杂化，但多为通用领域扫描文档 OCR 识别系统，或特定工况下工件编号识别系统，受限于训练数据等原因，在实际应用场景中识别准确度不佳，无法在企业部署应用。

基于以上原因，本文基于 CRNN 模型，针对企业质量文档，设计并实现了一套文档图片识别系统，探索了一种深度学习在企业落地的方式，为后续其他人工智能模型部署提供指导。

1 数据集准备

当前并无公开质量领域报告数据集，需要自行构建数据集。本文采用反向生成图片和人工标注两种方法相结合来构建训练集和验证集，以兼顾数据集质量及构建速度。为反向生成图像质量更贴近实际扫描情形，在图像生成时加入随机干扰，包括：-2° ～ 2° 字体倾斜、图像模糊、干扰黑点、灰度值变化。人工标注数据来源于以往项目文档扫描存档，按反向生成与人工标注数据 20:1 的比例，随机加入生成数据中，最终形成了 6000 份训练集、800 份验证集。同时，为避免数据场景单一导致模型过拟合，同时引入 ICDAR2015 通用英文识别数据集，进行混合训练。部分训练数据样本如图 1 所示。



（a）不均匀阴影　（b）信息模糊　（c）噪声干扰　（d）文档倾斜

图 1 训练数据示意

2 系统整体组成

系统由四个大的模块组成，分别为图像预处理、文本定位分割、基于 CRNN 神经网络的文本识别模型以及数据格式化输出，如图 2 所示，下面就前 3 个核心功能模块做详细介绍。

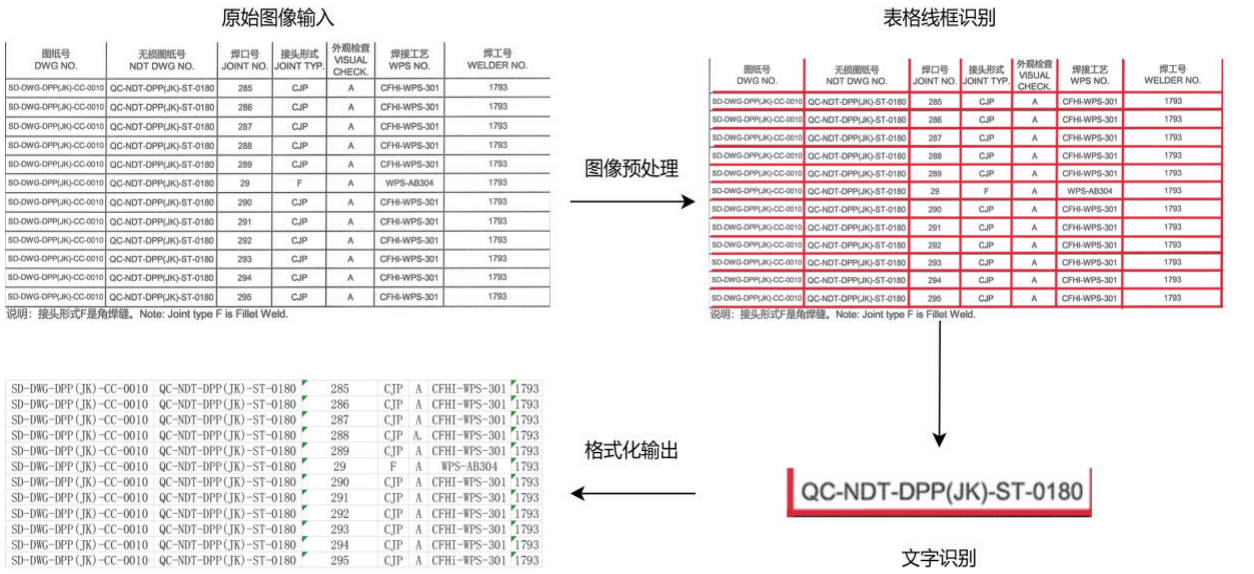


图 2 系统组成

2.1 图像预处理

实际扫描文档中，会出现文档折角、倾斜的现象。为规范数据输入，提高系统特殊场景应对能力。首先通过 Candy 边缘检测获取文档最大外围轮廓，通过比对四点坐标判断输入图像是否完整。同时，基于边缘检测结果和霍夫线变换获取直线角度均值，以实现文档角度修正。为去除扫描过程中可能存在的背景噪声，提高识别的准确度和速度，采用自适应阈值法对输入图像做二值化处理，自适应阈值计算式可表示为：

$$T(i,j)=m(i,j)\times\left[1+k\left(\frac{s(i,j)}{R}-1\right)\right] \quad (1)$$

计算原理为，对于待处理图像，选取 (i,j) 作为中心点，通过选取合适 r 值，计算边长为 r 的圆形区域内的灰度均值与标准差，作为该区域内二值化阈值的选取依据，对比 OSTU 算法，在局部区域内能得到更精细的二值化效果，更好地应对不均匀光照背景。通过设置参数，实现更细小文字与背景的分割，减少文字信息缺失。

由于本文采用 CNN+RNN 的神经网络模型，CNN 输出序列长度受输入图像宽度影响。为提高模型图片感知能力，图像预处理时，保持图像宽度与高度固定比例，同时保证单批次训练输入长度一致要求，长度不足部分做掩码处理，如图 3 所示。

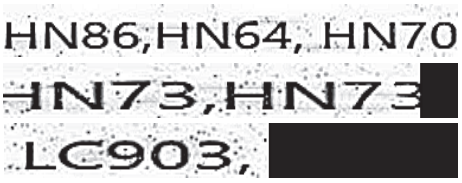


图 3 图像掩码处理

2.2 文字定位

质量文档指在项目全周期质量管理过程中，为确保产品、项目完工质量满足要求各项要求，具备可存储、可追溯属性而制定的一系列规范表格、记录等。为方便后期查找阅读，其关键信息内容往往以表格数据形式呈现，同时因专业领域的复杂与多样性，各类表格样式难以统一。针对一些复杂背景下文本识别任务，熊海朋等人 [10] 提出了一种基于深度学习卷积神经网络的文本定位算法，在对图像预处理后，得到候选文本区域，再进行形态学运算及连通区域限制分析，进一步缩小文本范围，最终结合预先标定好位置信息的数据集训练卷积神经网络，完成最后的细分类。质量报告数据大多排列规范，以表格形式呈现。针对这一特点，本文放弃采用传统文字定位模型，而基于表格特点，分别通过横向和纵向上的膨胀腐蚀操作，来获得表格线框，进而获得交点坐标，从而实现文本信息定位及提取。具体实现方式如下：假定输入灰度图片高度为 H，宽度为 W，创建一个高度为 1，宽度 $W/10$ 的水平方向卷积核，先对图像水平方向进行一次腐蚀操作，再对结果进行一次膨胀操作，可实现表格横向线条框的完整提取，同理可完成纵向方向线条框提取，提取结果如图 4 所示，进一步可得到各表格框位置坐标，完成图像分割。

图 4 表格线框识别示意

在具有表格类特色文档的文字位置识别上，该方法比文本检测类网络模型运行速度更快，且在表格信息密集处，分割结果更为准确，无需耗费大量资源进行训练集的位置标注工作。

2.3 文字识别

本文文本识别模块采用 CRNN 网络结构 [11]，共由 7 个卷积模块通过 ReLU 激活函数连接，由若干池化层与归一化层和一个长度为 512 高度为 1 的特征序列输入双向长短时记忆网络 LSTM 层，以及一个全连层组成，具体模型结构如图 5 所示。

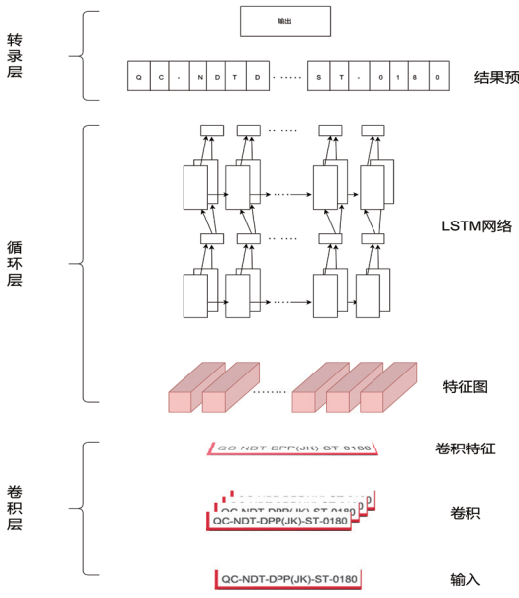


图 5 CRNN 卷积神经网络模型

图片经预处理和 CNN 网络后，输出大小为 [batch_size, seq_length, modle_n] 的张量，最终经 RNN 网络输出为 [seq_length, 1, dict_size]，其中 dict_size 为文字对照字典大小，经由解码函数得到最终所需的文本信息，解码函数采用 Beam Search，相较于 Greedy search 其搜索范围更大，拥有更好的全局解码结果。

3 实验结果分析

本文 CRNN 模型训练配置如下：算法基于 Paddlepaddle-GPU 2.3.2 框架，编程语言为 Python 3.8.19，运行环境 Linux Ubuntu 18.04 操作系统，CPU 型号 Intel(R) Core(TM) i5-10400F，主频 2.90 GHz，运行内存 32 GB DDR4，显卡型号 RTX 2060，显存 6 GB。

模型训练过程中，损失函数采用 CTC 编码方式计算损失函数 [12]，在应对输入与输出长度不一致问题上，具有更好的表现。优化器采用自适应学习率算法 Adam，在文本识别任务中，相较于 RMSprop 算法，拥有更高的识别准确度 [13]，单次读取图像批次大小为 64，学习率大小取 5e-5，损失函数下降曲线如图 6 所示。

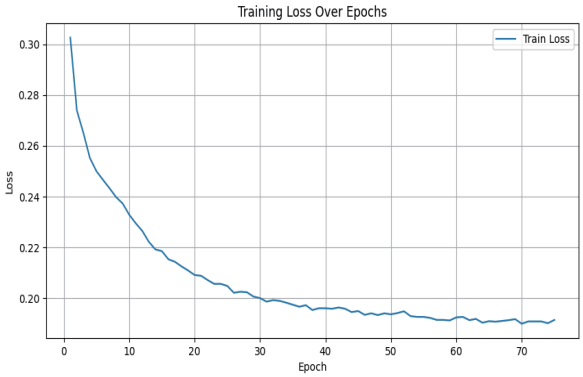


图 6 损失函数下降曲线

模型训练完成后，经测试、验证集评估准确度可达 93%。应用部署于测试用笔记本电脑，主频 3.2 GHz，识别速度为每秒 6 张图片，单张 PDF 文档识别时间为 30 s。扫描文档识别结果如图 7 所示。

QC-NDT-DPPA-ST-0001	S22	CJP
QC-NDT-DPP-ST-0002	S33	F
QC-NDT-DPPA-ST-0003	S35	F
QC-NDT-DPP(LQ)-ST-0004	S22	CJP
QC-NDT-DPPA-ST-0005	S25	F
QC-NDT-DPPA-ST-0006	S35	CJP

QC-NDT-DPPA-ST-0001	S22	CJP
QC-NDT-DPP-ST-0002	S33	F
QC-NDT-DPPA-ST-0003	S35	F
QC-NDT-DPP(LQ)-ST-0004	S22	CJP
QC-NDT-DPPA-ST-0005	$25	F
QC-NDT-DPPA-ST-0006	S35	CJP

图 7 节选文档识别结果

4 结语

本文基于 CRNN 神经网络，设计实现了一套针对质量文档的 OCR 识别系统，通过图像预处理、文本分割定位、文字识别三大模块，完成了纸质报告信息获取。在文本分割定位上，针对质量文档中表格线框的特点，提供了一种表格文本定位分割的方法，通过腐蚀膨胀处理、提取线框的方式来获取文本位置信息，对比文字检测模型，在信息密集处有优秀的分割效果，更易于搭建。测试结果表明，本系统训练设备性能要求较低，易部署，具有很强的扩展性，满足企业实际部署应用要求，扩展了企业基础数据获取方式，可加快推动企业质量管理的数字化与智能化。

参考文献：

[1] 沈嘉康 . 基于改进 DBNet 与改进 CRNN 的集装箱箱号识别系统 [J]. 工业控制计算机 ,2024,37(3):54-56.

[2] 徐琦 , 孙顺凯 , 钱杰 , 等 . 基于改进 CRNN 网络的卷烟件烟上行码识别方法研究 [J]. 中国烟草学报 ,2024,30(3):125-131.

[3] 石志强 , 周新辉 , 沈康畅 , 等 . 融合 CTPN 和 CRNN 对识别影像图片文字及应用的研究 [J]. 现代计算机 , 2023, 29(23): 42-46.

[4] 高天 .OCR 识别技术在钢结构制造企业精细化管理的应用 [J]. 中国信息化 ,2023(11):106-107+101.

[5] 曹菁 , 陈康 , 齐宁 , 等 . 基于 OCR 和图像检测的盖章文书图像自动审核方法 [J]. 应用科学学报 ,2023,41(6):1058-1067.

[6] 王阳 , 李振东 , 杨观赐 . 基于深度学习的 OCR 文字识别在银行业的应用研究 [J]. 计算机应用研究 , 2020,37(S2):375-379.

[7] 吴永飞 , 王彦博 , 陈志豪 , 等 . 商业银行基于 DCU 技术的 OCR 应用研究 [J]. 中国信用卡 ,2023(12):47-50.

[8] 刘乐 , 张晓松 , 黄锋 , 等 . 基于改进 DBNet 和 SVTR 算法的连铸板坯号检测与识别 [J]. 电子测量与仪器学报 , 2024, 38(2): 67-75.

[9] 王秀光 , 尹世阁 .OCR 技术在企业文档识别中的研究与实践 [J]. 信息与电脑 (理论版), 2022,34(18):175-178.

[10] 熊海朋 , 陈洋洋 , 陈春玮 . 基于卷积神经网络的场景图像文本定位研究 [J]. 电子科技 ,2018,31(1):50-53.

[11]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, 2018,39(11):2298-2304.

[12] 郭浩 , 宁初明 , 韩寿松 , 等 . 基于 DBNET 与 CRNN-CTC 的自然环境文字识别系统 [J]. 计算机应用与软件 , 2023, 40(9): 132-136.

[13] 李凯鹏 , 刘刚 , 李帅 , 等 . 基于 CNN 的高精度手写体数字识别 [J]. 信息与电脑 (理论版),2022,34(10):67-70+75.

【作者简介】

曹叶欢（1998—），男，湖南郴州人，本科，焊接系统工程师，助理工程师，研究方向：质量管理数字化、智能化。

（收稿日期：2024-07-04）