基于图像识别技术的核电文档智能化应用实践

杨 强,胡心宇
(江苏核电有限公司,江苏 连云港 222000)

摘 要:为采用人工智能领域的新一代信息技术来有效提升实际业务开展的效率,达到降本增效的目的。本文对采用图像识别技术来辅助核电企业开展文档智能化应用进行研究。文章阐述了图像识别技术的概况、分析了采用图片识别技术的业务背景、项目实施的主要过程、项目实现的基本原理和典型应用场景解决方案等内容,通过对基于图像识别技术的扫描文件清晰度的自动化检测、基于光学图像文字识别技术的文件自动化拆分和比对这两个应用场景,阐述了所需要解决的问题、解决方案的原理、具体的程序功能设计方案以及最终的应用效果。根据对相关实际应用效果的评估,证明了采用图像识别技术的技术可行性,能够为文档智能化应用发挥重要的辅助作用。通过本课题的研究和实践,为基于人工智能等技术的文档信息化技术提升做出了有益的探索。

关键词:图像识别技术;光学图像文字识别;档案信息技术;文档信息化;信息资源利用
文章编号:2096 - 4633(2019)11 - 0058 - 06 **中图分类号**:C39 **文献标志码**:B

任何核电项目都离不开合格的文档作为工作依据,文档在项目管理中的作用不可替代。为了加强档案的真实性、完整性、可靠性和安全性等质量控制要求,档案管理人员实施了较多前端控制的措施,但却因缺少有效的工具,导致单纯利用人工很难快速完成数百万份文件的质量检查和文件处理工作。这一关键问题未能解决不仅延滞了项目进度,而且导致档案的不完整、不准确、有效性差、不规范,不系统的问题一直存在。

人工智能是一种引发诸多领域产生颠覆性变革的前沿技术,当今的人工智能技术以机器学习,特别是深度学习为核心,在视觉、语音、自然语言等应用领域迅速发展,已经开始像水电煤一样赋能于各个行业。图像识别技术是人工智能的一个重要领域,是指对图像进行对象识别,以识别各种不同模式的目标和对像的技术。通过对图像识别技术的应用能够快速提升人工智能水平,为用户提供个性化、精准化、智能化服务,大幅提升业务体验,与生产生活的各个领域相融合,有效提升各领域的智能化水平,给传统领域带来变革机遇。图像识别技术目前已经处于成熟技术,国内外的众多学者的研究成果显示,图像目标检测、视频内图像的实时分析、图像 OCR 识别技术等均已具有工业化使用的条件[1-6]。

1 核电文档智能化应用的业务背景

文档信息化是核电企业开展信息化建设的一个子领域,相比较其他生产运营为核心的一线业务而言,文档部门的定位是为企业核心业务提供服务。同理,文档信息化也是为了更好地向企业的各级人员提供资源服务。目前的文档信息化主要集中在文件档案的电子化、文档业务流程的管理、文档的全过程管理,这些功能基本上能够满足企业开展文档整体业务过程的需要,可以支撑起核电企业主要的业务运作,对于实现管理精细化和标准化起到了关键作用。但是在日常工作中,大量的重复性工作还是需要人工进行处理,这些重复性工作消耗了大量的人工工时资源。主要体现在:

一方面,根据核电档案验收工作的要求,对于移交到业主文档部门的的纸质文件和电子扫描文件,需要确保两者的一致性,且针对接收到的扫描文件要确保其清晰度,以免模糊不清的文件传入系统,影响到后续档案验收检查以及实际业务工作文件的使用。目前此项工作需要人工来进行检查,每份文件平均耗时需要 10 分钟,而当前已累积 3 万余份文件待检查。如继续这种人工逐页检查的模式,将会严重影响到档案验收工作的完成。

另一方面,为满足文档精细化管理要求,江苏核电对于移交的竣工文件、设备文件等采用按件管理的模式,由于所接收的材料无法提供原始清单类文件,需要用户人工对扫描后的整卷电子文件进行拆解以及针对提供的导入表进行数据核对。当前一份文件的拆解工作需要平均花费半小时,而每月需要处理的文件量达到上千份。在当前档案验收时间紧、工程进度快的节奏下,如继续通过人工进行拆解的方式,将无法按时完成相应的工作任务。

为了保证按期保质保量的完成文档验收工作,文档管理人员积极探索各种信息技术手段来开展工程文档移交工作。发挥技术对于生产运营和经营管理等相关领域的支持,有效提升核电企业的核心竞争力。

2　图像识别技术的概况

图像识别是计算机对图像进行处理、分析和理解,以识别各种不同模式的目标和对像的技术。识别过程包括图像预处理、图像分割、特征提取和判断匹配。简单来说,图像识别就是计算机如何像人一样读懂图片的内容。借助图像识别技术,不仅可以通过图片搜索更快的获取信息,还可以产生一种新的与外部世界交互的方式,甚至会让外部世界更加智能的运行。OCR(optical character recognition,OCR)是指光学设备检查纸上打印的字符,通过检测暗、亮的模式确定其形状,然后用字符识别方法将形状翻译成计算机文字的过程,就是计算机对文字的阅读。语言和文字是我们获取信息最基本、最重要的途径。图像识别技术作为我们的辅助工具,可以为我们自身的人类视觉提供了强有力的辅助和增强,带给了我们一种全新的与外部世界进行交互的方式。

3　图像识别技术在核电文档智能化应用的思路分析

核电文档的主要管理对象就是各种各样的信函、纪要、技术文件,移交归档的案卷是扫描的文件,可以认为是由多份图像进行组成的文件。为解决在文档管理实际业务过程中遇到的难点和人工投入较大的问题,江苏核电文档技术人员对业务过程进行分解和逐一突破,采用图像识别技术,进行辅助性功能开发和应用,有力地提升文档信息化的技术水平。

针对需要人工对文件的清晰度进行肉眼检查的问题,考虑将待检查扫描文件转换为高清无损图片,再借助计算机视觉技术对图片的清晰度进行评价,根据评价的结果由用户按照清晰度的取值大小进行自主检查,如系统给出的最不清晰的结果数值的实际文件可以满足人工阅读需要,则不再对剩下的文件进行人工检查。这样原本需要肉眼一一查看的文件,就转化为检查系统认为不清晰的文件的问题。图像清晰度评价算法有很多种,在空域中,主要思路是考察图像的领域对比度,即相邻像素间的灰度特征的梯度差;在频域中,主要思路是考察图像的频率分量,对焦清晰的图像高频分量较多,对焦模糊的图像低频分量较多。图像梯度算法是考虑图像的每个像素的某个邻域内的灰度变化,利用边缘临近的一阶或二阶导数变化规律,对原始图像中像素某个邻域设置梯度算子,通常我们用小区域模板进行卷积来计算。利用梯度的本质目的,是要找到某像素与其相邻像素的灰度差值,并放大这种差值,从而用于图像增强。Laplacian 算子是最简单的各向同性微分算子,它具有旋转不变性。一个二维图像函数的拉普拉斯变换是各向同性的二阶导。Laplacian 算子利用二阶导数信息,具有各向同性,即与坐标轴方向无关,坐标轴旋转后梯度结果不变。使得图像经过二阶微分后,在边缘处产生一个陡峭的零交叉点,根据这个对零交叉点判断边缘。将原始图像通过拉普拉斯变换后增强了图像中灰度突变处的对比度,使图像中小的细节部分得到增强并保留了图像的背景色调,使图像的细节比原始图像更加清晰[7-8]。

针对当前需要人工对扫描 PDF 文件进行拆分和核对的实际工作难点,分析该问题由两部分组成,一个是由于提供的目录是扫描格式,需要人工根据目录里面的文件的页码起止范围进行文件的拆分形成新的文件,该动作在 PC 端上操作耗费时间较长。另一个是需要人工对该扫描文件以及用做文件著录的可编辑表格进行比照,该动作也需要在 PC 上操作。因此,拟对传统人工拆解文件的过程进行优化,采用光学图像文字识别技术,通过自主研发自动化拆分工具软件,

识别数据并根据数据来处理待拆分文件,将人工的工作由软件自动化完成,在拆分的同时把数据比对的工作完成。

OCR 对文本资料的图像文件进行分析识别处理,获取文字及版面信息。文字检测即检测文本的所在位置和范围及其布局。通常也包括版面分析和文字行检测等。文字检测主要解决的问题是哪里有文字,文字的范围有多大。由于核电文件的目录格式基本一致,且均为表格,因此需要解决的问题是表格文本识别。所以采用采购成熟的表格识别解决方案。经过前期的技术评估,拟采用阿里云的 OCR 服务。该服务在识别的准确率、服务响应速度、售后支持力度等方面相比较其他云服务产品提供商均较为满意。在安全性上,在阿里云提供的云市场服务协议中,存在"OCR 识别服务接收的用户图片数据,基于服务提供的需要会暂时缓存用于识别程序计算,计算完成后将会及时释放缓存资源,不做留存。"这一条款。且供其识别的内容仅为卷内文件目录信息,不包括具体正文内容。安全性上可视为可控。

4 图像识别技术在企业新一代技术架构中的定位

在设计文档相关的新一代信息技术应用的整体技术框架时,主要按照分层架构的形式来设计。一共分为基础设施层、文档业务管理和物联网技术应用层、大数据中心层、实际应用层,这几层主要从角色作用上进行划分。基础设施层作为实现技术落地的基础,位于设计的最底层,它主要由数据库、操作系统、服务器、网络设备以及在云端的基础设施组成,在该层面除了传统技术外,还会使用到云计算技术。文档业务管理和物联网技术应用层面向的是文档管理人员,一方面解决在日常工作要处理的文件管理、档案管理、流程管理以及文档服务等具体业务需求,另一方面解决库房实体管理的问题。这里会使用到物联网技术来实现智慧档案馆,例如可视化库房、温湿度控制、智能密集柜、监控与门禁集成等功能。大数据中心层由文档中心、长期保存平台、知识中心这几个部分组成,将下层的文档资源数据按照大数据分析利用的维度进行收集、整理、清洗和加工,汇总形成具有数据分析价值的平台。本部分也会使用到一些云服务,通过这些市场化的服务能够大大降低大数据分析的难度。在业务应用层,涉及到各类与最终用户直接交互的媒介,例如各类业务系统、人工智能应用、移动互联网应用、定制工具等,当然,大数据也是形成这些业务应用的基础技术之一[9-10]。

图像识别技术作为一种核心的基础技术资源,涉及到云计算、深度学习等技术的应用,主要涵盖了上述的基础设施层、应用层。在基础设施层上,通过对云服务形式提供的图像识别相关服务进行调用,可以认为企业的基础设施延伸到更为广泛的领域;在应用层,图像识别技术可以广泛的应用在文档领域的方方面面,甚至可以将该技术用于核电的其他业务领域,例如人脸识别、实时监控预警等[11-17]。

5 大数据技术在设备缺陷分析的应用探索

5.1 基于图像识别技术的扫描文件清晰度检测

本项目采用的技术方案是 Python + Flask + OpenCV,Python 是一个高层次的结合了解释性、编译性、互动性和面向对象的脚本语言。Flask 是一个基于 Werkzeug WSGI 工具箱和 Jinja2 模板引擎,使用 Python 编写的轻量级 Web 应用框架。OpenCV 是一个用于图像处理、分析、机器视觉方面的开源函数库。OpenCV-Python 是 OpenCV 的 Python API,结合了 OpenCV C + + API 和 Python 语言的最佳特性。OpenCV-Python 是一个 Python 绑定库,旨在解决计算机视觉问题。项目采用 Python 开发,并通过 Flask 提供 Web 访问,用户在该网页上选择 PDF 文件并上传,系统在接收到文件后会自动转换成图片,通过 OpenCV 来检测并给出最终结果。主要的方法就是调用 OpenCV 的 cv2. Laplacian (image, cv2. CV _ 64F) 然后通过 numpy 数学计算库的 var 方法给出分值。结果的示例如图 1 所示,右侧的文件图像在清晰度上属于较低,其分值也较低。

可以看出系统提供的结果可以快捷地对文件的清晰度进行评价,使得用户可以快速发现存在问题得内容,无需全部核实一遍。这种采用智能化图像识别技术对扫描文档的文件进行自动化识别,进行广泛测试和对清晰度阈值范围分析,发现海量扫描

数据中的异常图像并给出可能存在问题的页面结果,辅助文档管理人员快速处理。

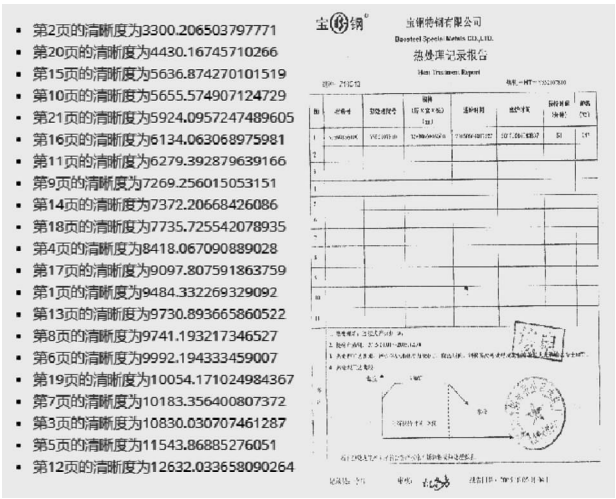

图 1　检测结果的示意图

Fig. 1　Schematic diagram of testing result

5.2　基于光学图像文字识别技术的自动化文件拆分与比对

本项目使用云端 OCR 技术服务,自主研发自动拆分及比对工具,根据每个案卷提供卷内文件目录清单,将 1 个案卷级 PDF,拆分成多个文件级 PDF,分别上传到 ECM 系统对应条目中,便于后续精细化利用。避免承包商提供纸质档案卷内文件目录和电子文件详细目录导入表出现不一致,通过 OCR 技术,进行数据自动比对,若不一致,提醒文档人员校核修订,保障信息一致性,提高文件检索准确性。总体逻辑示意如图 2 所示。

例如左侧的目录表格里面记录了每个文件的编号、责任者、文件题名、日期、页号,拆分工作里面的每个文件都有相应的页号范围,那么可以按照页号范围来进行 PDF 拆分,问题的关键在于识别这个表格,为此,首先将 PDF 的目录页进行转换成图片,然后将图片进行识别,识别完成后编写代码解析返回的结果,根据结果里面的页号范围对文件进行拆分,形成自动化拆分工具软件。软件制作时需要考虑一些特殊情况,例如档案移交时存在补充页的情况,这些特殊页面的页码是 5.1 之类,需要在程序中进行判断。

图 2　自动化拆分的逻辑示意图

Fig. 2　Schematic diagram of automated splitting

在拆分的同时,需要对导入表的内容进行核对,利用 Python 解析 excel 并与文字识别的结果进行比对,如果发现不一致,提示给用户,用户不在需要逐一对数据进行核对,而只需要关注软件提示的不正常项目,因为文字识别难免有错误,而如果软件解析出来数据是正确的,那么大概率其一定是正确的。在进行内容对比时,需要检查表格的样式以及日期内容的正确性,由于日期存在格式的问题,如不能解析正确会对用户的判断产生误导,因此在程序中需要考虑这种特殊性。

本项工作大幅提升了文件处理效率,通过识别数据并根据数据来处理待拆分文件,大幅提升了文件处理效率,原本平均 30 分钟/件的效率提升到 2 分钟/件。

6　结论

江苏核电采用图像识别技术,对文档管理过程中遇到的难点进行分解和逐一突破,综合采用各种先进技术提升文档信息化的技术水平。采用智能化图像识别技术对扫描文档的文件进行自动化识别,广泛测试和分析合理的阈值范围,自动发现海量扫描数据中的异常图像并辅助文档管理人员快速处理;采用光学图像文字识别技术对传统人工拆解文件的过程进行优化,大幅提升文件处理效率。通过先进技术的应用,效率提高且准确性、规范性大大提高,取得了较好的管理效益。所采用的技术方案完

全免费,在应用层面完全自主开发,拥有自主知识产权。方案在业内属于首创,尚无类似案例。项目的成功实施起到了较好的示范作用,具有一定的推广价值,取得了较好的社会效益。

参考文献:

[1] CHEN CAO, QIMING HOU, KUN ZHOU. Displaced dynamic expression regression for real-time facial tracking and animation [J]. ACM Transactions on Graphics, 2014, 33(04), 1 – 10.

[2] YUNCHAO GONG, SVETLANA LAZEBNIK, ALBERT GORDO, et al. Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2916 – 2929.

[3] CHAO DONG, CHEN CHANGE LOY, KAIMING HE, et al. Image super-resolution using deep convolutional networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(02): 295 – 307.

[4] JAN FUNKE, FABIAN TSCHOPP, WILLIAM GRISAITIS, et al. Large scale image segmentation with structured loss based deep learning for connectome reconstruction [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019, 41(07): 1669 – 1680.

[5] PEDRO F. FELZENSZWALB, ROSS B. GIRSHICK, DAVID MCALLESTER, et al. Object detection with discriminatively trained part-based models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010, 32(9): 1627 – 1645.

[6] BAOGUANG SHI, XIANG BAI, CONG YAO. 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. 2017, 39(11): 2298 – 2304.

[7] KAIMING HE, JIAN SUN, XIAOOU TANG. Single image haze removal using dark channel prior [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011, 33(12): 2341 – 2353.

[8] 崔光茫,张克奇,毛磊,等. 结合多尺度分解和梯度绝对值算子的显微图像清晰度评价方法[J]. 光电工程,2019,46(06): 59 – 69.
CUI Guangmang, ZHANG Keqi, MAO Lei, et al. Micro-image definition evaluation using multi-scale decomposition and gradient absolute value[J]. Opto-Electronic Engineering, 2019, 46(06): 59 – 69.

[9] 邓应松,段秦刚,宋小松. 基于图像识别的保护压板投退状态辨识方法[J]. 陕西电力,2015,43(10): 49 – 53 + 57.
DENG Yingsong, DUAN Qingang, SONG Xiaosong. State identification of relaying plate based on image recognition [J]. Shaanxi Electric Power, 2015, 43(10): 49 – 53 + 57.

[10] 付文龙,谭佳文,吴喜春,等. 基于图像处理与形态特征分析的智能变电站保护压板状态识别[J]. 电力自动化设备,2019,39(07): 203 – 207.
FU Wenlong, TAN Jiawen, WU Xichun, et al. Protection platen status recognition based on image processing and morphological feature analysis for smart substation [J]. Electric Power Automation Equipment, 2019, 39(07): 203 – 207.

[11] 欧家祥,史文彬,张俊玮,等. 基于深度学习的高效电力部件识别[J]. 电力大数据,2018,21(09): 1 – 8.
OU Jiaxiang, SHI Wenbin, ZHANG Junxi, et al. Recognition of efficient electrical components based on deep learning[J]. Power Systems and Big Data, 2018, 21(09): 1 – 8.

[12] 陈良琴,唐海城,肖新华. 基于深度学习的输电线路风险预警识别研究[J]. 电力大数据,2018,21(12): 1 – 5.
CHEN Liangqin, TANG Haicheng, XIAO Xinhua. Research on risk warning and recognition of transmission line based on deep learning[J]. Power Systems and Big Data, 2018, 21(12): 1 – 5.

[13] 林怀德,罗毅初. 一种利用深度图像进行电缆弯曲度测量方法[J]. 电力大数据,2018,21(11): 88 – 92.
LIN Huaide, LUO Yichu. A method for cable curvature measurement based on depth image[J]. Power Systems and Big Data, 2018, 21(11): 88 – 92.

[14] 李黎,王惠刚,刘星. 基于改进暗原色先验和颜色校正的水下图像增强[J]. 光学学报,2017,37(12): 176 – 184.
LI Li, Wang Huigang, Liu Xing. Underwater image enhancement based on improveddark channel prior and color correction [J]. Acta Optica Sinica, 2017, 37(12): 176 – 184.

[15] 黄新波,李菊清,张烨,等. 复杂环境下覆冰绝缘子识别检测技术[J]. 高电压技术,2017,43(03): 891 – 899.
HUANG Xinbo, LI Juqing, ZHANG Ye, et al. Recognition and detection technology of ice-coverd insulators under complex environment[J]. High Voltage Engineering, 2017, 43(03): 891 – 899.

[16] 窦健泰,高志山,马骏,等. 基于图像信息熵的 ptychography 轴向距离误差校正[J]. 物理学报,2017,66(16): 110 – 118.
DOU Jiantai, GAO Zhishan, MA Jun, et al. Correction of axial distance error in ptychography based on image information entropy [J]. Acta Physica Sinica, 2017, 66(16): 110 – 118.

[17] 杨 强,陈超,查风华,等. 核电企业基于"大云物移智"的文档管理创新[J]. 电力大数据 2018,20(09): 35 – 40.
YANG Qiang, CHEN Chao, ZHA Fenghua, et al. Document management innovation of nuclear power enterprises based on "big data, cloud computing, internet of things, mobile technology, artificial intelligence" [J]. Power systems and big data. 2018, 20(09): 35 – 40.

收稿日期:2019 – 09 – 01

作者简介:

杨 强(1988),男,硕士,高级工程师。主要从事核电企业信息化建设、运维及管理工作。

(本文责任编辑:范 斌)

Application practice of nuclear power documents based on image recognition technology

YANG Qiang, Hu Xinyu

(Jiangsu Nuclear Power Co., Ltd., Lianyungang 222000 Jiangsu, China)

Abstract: In order to apply the new generation of information technology in the field of artificial intelligence to effectively improve the efficiency of actual business development, and achieve the purpose of reducing costs and increasing efficiency. This paper studies the use of image recognition technology to assist nuclear power enterprises in the application of document intelligence. The article expounds the general situation of image recognition technology, analyzes the business background of image recognition technology, the main process of project implementation, the basic principle of project realization and the solution of typical application scenarios, Through the automatic detection of scanning file sharpness based on image recognition technology, the automatic resolution of files based on optical image text recognition technology, and the comparison of these two application scenarios, This paper expounds the problems to be solved, the principle of the solution, the specific program function design scheme and the final application effect. According to the evaluation of the relevant practical application results, it is proved that the technical feasibility of adopting image recognition technology can play an important supporting role for the intelligent application of documents. Through the research and practice of this topic, it makes a beneficial exploration for the promotion of document information technology based on artificial intelligence and other technologies.

Key words: image recognition technology; optical image text recognition; archive information technology; document information; information resource utilization