Computer Science and Application 计算机科学与应用, 2019, 9(9), 1696-1702
Published Online September 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.99189

Hans汉斯

Research on the Automatic Segmentation Recognition and Data Checking Method in Nuclear Electronic File Archiving

Jing Ma, He Bai

China Nuclear Power Design Company Ltd., Shenzhen Guangdong
Email: 116389207@qq.com

Received: Aug. 27th, 2019; accepted: Sep. 11th, 2019; published: Sep. 18th, 2019

Abstract

In order to effectively regulate and control the quality of mass design documents, CGN developed a process management system, which not only controlled the whole life cycle of documents, but advanced the document quality control to the stage of file generation as well, and realized the integration and automation of nuclear power document management. The system established a new project document front-end control mode to ensure the effective control and accurate transmission of the project documents. However, the electronic document review process in the system, which is to verify the consistency between structured document metadata and electronic document entities, still requires a large amount of manpower. To solve the problem, we will introduce AI technology into the system for secondary development. This paper starts from the enterprise knowledge management and the intersection of software and document management informatization, and elaborates how AI technology innovates in the field of document management and improves the standardization efficiency of document flow.

Keywords

Artificial Intelligence, Image Segmentation Recognition, Project Document Quality Control

自动化分割识别与数据校验方法在核电电子文件归档审查中的应用研究

马　菁，白　鹤

中广核工程有限公司设计院，广东　深圳
Email: 116389207@qq.com

收稿日期：2019年8月27日；录用日期：2019年9月11日；发布日期：2019年9月18日

文章引用: 马菁, 白鹤. 自动化分割识别与数据校验方法在核电电子文件归档审查中的应用研究[J]. 计算机科学与应用, 2019, 9(9): 1696-1702. DOI: 10.12677/csa.2019.99189

摘　要

为有效规范管控海量设计文件的质量，中广核开发了过程管理系统，对文件全生命周期通过系统进行管控，同时将文件质量控制提前到文件生成阶段，实现核电文档管理一体化、电子化和自动化，该系统建立起全新的项目文档前端控制模式，确保项目文档有效受控与准确传递。但该系统的“移交归档”电子文件审查环节，即核查结构化文档元数据与电子文件实体一致性，仍需投入大量人力。为解决该问题，笔者将人工智能(AI)技术引入系统，并依托系统进行二次开发，成功解决这一难题。本文将从企业知识管理、软件及文档管理信息化的交叉技术领域出发，阐述人工智能(AI)技术如何在文档管理领域创新应用，提高文档流转的规范性与效率。

关键词

人工智能，图片分割识别，项目文档质量控制

1. 引言

在核电建设项目中，工程总成本的约 3%~5%是由核电工程设计文件质量的问题导致工程变更和工程实施的错误所造成的。核电工程企业内容信息数据复杂，工程设计文件数量庞大，达到百万级别。在项目管理中，项目文档控制与合同管理、费用控制、进度控制、材料控制、质量控制同等重要，是不可或缺的重要管理模块。为有效规范管控海量设计文件的质量，笔者所在企业开发了过程管理系统(Design Process Management System，简称 DPMS)，将文件质量控制提前到文件生成阶段，通过系统完成文件起草、校审、移交归档、分发、存储、利用，降低文件全生命周期管理过程中的人因误差。该系统建立起全新的项目文档前端控制模式，确保项目文档有效受控与准确传递。但“移交归档”电子文件审查环节，核查结构化文档元数据与电子文件实体一致性，仍需投入大量人力，如何减少人力投入，提高该环节自动化水平，是企业亟需解决的问题。

本文将从企业知识管理、软件及文档管理信息化的交叉技术领域出发，通过对半结构化海量核电设计文件图片的分割识别与技术处理，以自动化的方式实现核电文件检查及元数据匹配的规范要求，并应用在过程管理系统(DPMS)中，提高文档流转的规范性与效率，降低人工成本达 90%以上。

2. 核电项目文档控制

2.1. EPC 模式下项目文档控制

国际上，工程公司已有近百年的历史，作为一种专营工程项目的智力密集型企业，依靠为业主提供工程建设服务生存。根据美国设计－建造学会 2005 年报告，发达国家近一半工程项目采用工程公司总承包的形式[1]。总承包模式即设计－采购－施工(Engineer-Procure-Construct, EPC)交钥匙(Turnkey)项目管理模式。近几年来，我国专业化的核电工程建设项目管理公司与总承包模式已经成为核电工程项目建设的主流，项目文档移交作为交钥匙项目的信息知识资产，是影响项目质量的重要因素。

项目文档控制管理就是通过建立工程项目文档体系，依据文件管理的计划、程序、标准规范，做到工程项目技术文件的有效控制和及时收发，从文档、信息的角度保证项目的顺利实施[2]。核电厂质量保证安全规定(HAF003)，对文件控制要求进行了明确规定，主要包括文件编审批控制、文件发布和分发、文件变更控制三个方面[3]。核电文档审查工作是文档质量控制的基础，文档遵循 HAF003 及项目质量保障体系要求，通过对设计文件编审批、发布分发及变更三方面进行监管，同时，审核设计文件 24 项元数据与半结构化电子文件实体是否一致，避免设计成品出现低级错误，确保设计成品满足项目质保要求，这是文档介入文件前端控制的重要内容。

2.2. 现有技术解决方案与缺陷

2.2.1. 海量非结构化电子文件的结构化处理方法及系统

为了批量检查海量数据。核电行业开发了过程管理系统，这是一种海量非结构化电子文件的结构化处理方法及系统。该系统包括元数据形式化约束配置模块，用于根据核电技术资料的编码规范及匹配规则制定元数据形式化约束条件；结构化处理模块，用于根据元数据形式化约束条件对海量非结构化电子文件进行结构化处理，得到满足核电企业内容管理系统结构要求的海量结构化数据；以及内容管理系统集成模块，用于将该海量结构化数据导入核电企业内容管理系统。

该系统仅从技术资料的实体电子文件的属性(如文件名称、大小、目录、哈希码等信息)进行了分析和提取，并未对非结构化文档的具体内容，尤其是图像内的数据信息进行进一步处理。因此，该问题解决需要从技术层面来完成，即对非结构化的数据进行分析和处理，通过相应的计算机技术将非结构化的数据转换为结构化数据，而成功转化的关键在于对数据文件的整合以及对数据内容的分析并最终建立能够高效运行的索引库的过程[4]。

2.2.2. 图片中文档定位和拆切方法

该技术公开了一种图片中文档定位和拆切方法，应用在文档照片处理的软件中，在用户用手机拍摄文档后，可以将图像内的文档进行快速裁切和扶正，能够为后继的文档识别模块排除干扰，提高文档内文字的识别率[5]。

但该项技术仅从文档定位的角度进行图像的处理和切割，并未定义标准的模板，也未结合图像切割技术与定位识别技术，无法准确的获取信息的位置。

3. 人工智能(AI)技术在归档审查中的开发与应用

综上可知，核电文档管理已实现了“文件全生命周期”的一体化管理，通过 DPMS 系统形成的海量信息数据，大部分是以半结构化形式(结构化元数据和非结构化电子文件实体)存储在企业内容管理系统(Enterprise Content Management System，ECMS)。DPMS 和 ECMS 仅提供核电企业内容管理的基本方法，对于结构化内部文件的处理有明显的作用。而对于文档规范化审查，尤其是半结构化海量数据的自动化处理未能提供高效的方法和系统。

本文将利用人工智能(AI)技术，开发一种半结构化海量核电文档的自动化分割识别与数据校验处理系统，实现核电企业半结构化海量文档的自动化分割识别与数据校验处理，包括基于结构化元数据与非结构化文档模板的可配置自动校验算法；核电图纸灰度化、二值化预处理算法，基于滤波平滑降噪技术及投影技术的图像校正方法，基于模板的图片的自动化分割提取方法；图纸信息特征提取技术；以及基于人工神经网络的分类识别技术。通过以上技术方法满足企业内部海量内容的结构化数据提取，实现自动化校验审查功能，保证核电文档质量与数据完整性，并降低人力成本，提高工作效率。

3.1. 技术方案

本项技术主要包括 10 个模块：内容管理系统集成接口模块、文档数据规范化校验规则管理模块、半结构化数据图像预处理模块、颜色信息提取模块、图像分割提取模块、文件属性提取模块、文字识别抽取模块、清晰度识别模块、综合元数据校验模块以及日志记录模块。

10 个模块关系如图 1 所示。内容管理系统集成接口模块 101 从企业内容管理系统(ECM)中获得文档及相关元数据信息，并将信息整合后传递给半结构化数据图像预处理模块 103，同时根据核电文档内容管理规则与要求通过文档数据规范化校验规则管理模块 102 录入规则数据，作为半结构化数据图像预处理模块 103 的规则输入，经过预处理的图像信息分别作为颜色信息提取模块 104、图像分割提取模块 105、文件属性提取模块 106 的输入，其中图像预处理模块 105 输出的图像区域块发送给文字识别抽取模块 107 及清晰度识别模块 108 进行进一步的处理，并结合颜色信息提取模块 104 和文件属性提取模块 106 的结果，将非结构化数据图像中的提取出的结构化信息与元数据发送给综合元数据校验模块 109 进行校验，将结果输出到自动检查结果显示用户接口，所有的操作过程将记录在日志记录模块 110 中。

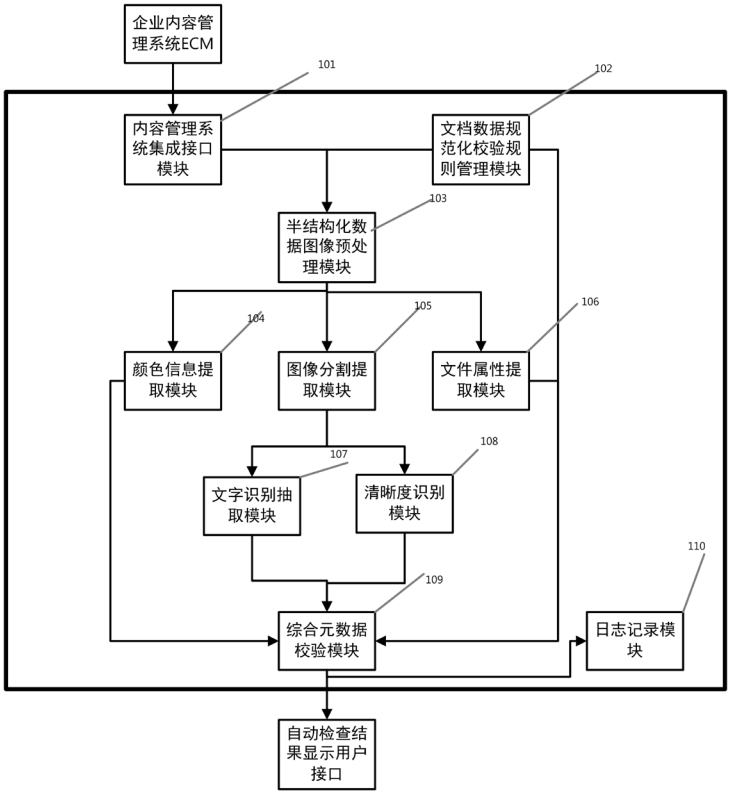

Figure 1. Module relationship diagram
图 1. 模块关系示意图

3.2. 各模块功能说明

下面对主体图 1 模块关系示意图方框内中各个模块的操作进行逐一解释。

3.2.1. 101 内容管理系统集成接口模块

本模块是主体模块与企业内容管理平台的接口模块，负责与核电企业内容管理系统(ECMS)进行数据交互，其中主要包含企业内容分为元数据信息与非结构化文件图像内容。

3.2.2. 102 文档数据规范化校验规则管理模块

本模块是文档数据规范化校验规则数据维护模块，根据核电文档管理规则录入相关数据。数据库主要分为文档分类规则库、校验区域规则库以及元数据校验规则库 3 个库，分别定义结构化元数据、非结构化文件以及二者间的关联关系。

3.2.3. 103 半结构化数据图像预处理模块

本模块负责将半结构化数据图像进行预处理，结合校验规则为下游模块提供输入，预处理主要有 4 个步骤：首先，对图像进行灰度处理，其次，对图像进行二值化处理，接着，利用滤波降噪算法，进行平滑处理，消除噪声，最后，通过方向自动校对模块完成图像调整。

3.2.4. 104 颜色信息提取模块

根据 103 图像预处理模块第 1 步灰度处理前获取的信息，加载对象的颜色信息，并将提取的信息发送给用于校验的综合模块 109。

3.2.5. 105 图像分割提取模块

根据 102 文档数据规范化校验规则管理模块提供的元数据信息，结合数据对象实例生成数据校验清单，清单中记录需要进行校验的各项内容；并根据模板位置信息，提取分割图片。

第一步：边缘识别，利用边缘检测算子对图像进行卷积运算，再采用最大类间方差方法对图像进行二值化处理，然后采用 Hough 算法检测出边缘上的直线段；

第二步：倾斜校正，将上一步得到的直线段按长度从大到小排序，选择最长的几个直线段，计算直线段相对于水平方向的倾斜角度；对这些倾斜角度排序，取中值作为图像的倾斜角度，通过旋转图像对其进行倾斜校正；

第三步：图片分割，保留水平方向和垂直方向上的直线段，去除其他直线段；计算不同直线段端点之间的距离，如果小于设定的阈值，则对直线段进行连接，获得表格的单元格图像；

第四步：定位提取，主要分为 3 个步骤首先根据位置信息提取定位页面位置，其次通过边缘识别算法扣取信息块图片，最后将图片按照规则进行临时保存。

根据 102 模块中给出的位置信息，定位信息块位置，位置信息包括页码、起始点与结束点。

利用边缘算法将扣取的图片保存为 BMP 格式。如图 2 所示。

Figure 2. Extraction results schematic diagram
图 2. 提取结果示意图

3.2.6. 106 文件属性提取模块

根据文件整体信息，获取校验所需要的属性信息，如文件名称、文件大小、文件格式等。

3.2.7. 107 文字识别抽取模块

文字识别抽取模块通过对图片中文字的分割与识别，实现文字信息的提取。主要包扩字符分割、特征提取与分类器三部分组成。

第一步：字符分割：通过对于 105 分割图片的行、字切分，实现字符分割功能。其中行分隔采用二值图像的像素累加方法，如下面公式所示。

$$\left(\sum_{j=1}^{L} F(i,j)\oplus p_1\right)\wedge\left(\sum_{j=1}^{L} F(i+1,j)\oplus p_2\right)\wedge\ldots\wedge\left(\sum_{j=1}^{L} F(i+k,j)\oplus p_{k+1}\right)=0$$

其中 $F(i,j)$ 是文本二值图像，L 是行长，p 是大于零的实验常数，取决于文档的噪点。⊕ 为通配符。当⊕为≥时，表达式若成立，则为行上界；当⊕为≤时则为行下界。上下界之间的可切分为一行。字切分的方法与之类似。

第二步：特征提取：通过对单个文字图片的统计特征进行分析，利用局部灰度算法抽象网格特征，并将特征向量输出提供给分类识别。

第三步：分类识别：基于人工神经网络(ArtificialNeuralNetwork)实现，通过对连接权值的设置，计算非线性激活函数是否大于阈值。进而输出分类信息。

$$y_k = \varphi\left(\sum_{j=1}^{p} \omega_{kj} x_j - \theta_k\right)$$

其中 x_j 是神经元输入信息，ω_{kj} 是神经元 k 连接的权值，θ_k 为阈值，$\varphi(\ )$ 为激活函数，y_k 为神经元 k 的输出，模型如图 3 所示。

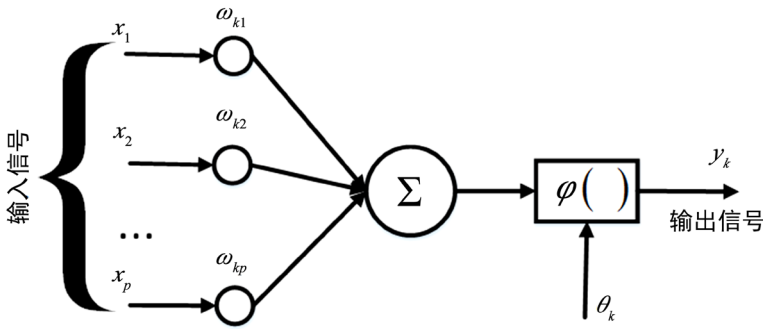

Figure 3. Artificial neural network classification model
图 3. 人工神经网络分类模型

模块 103、105、107 对于图片的核心操作流程如下图所示，该流程是本文的核心流程。通过对非结构化核电文档数据图像的预处理、分割及文字识别，抽取其中的结构化数据信息，作为后续数据校验的基础。如图 4 所示。

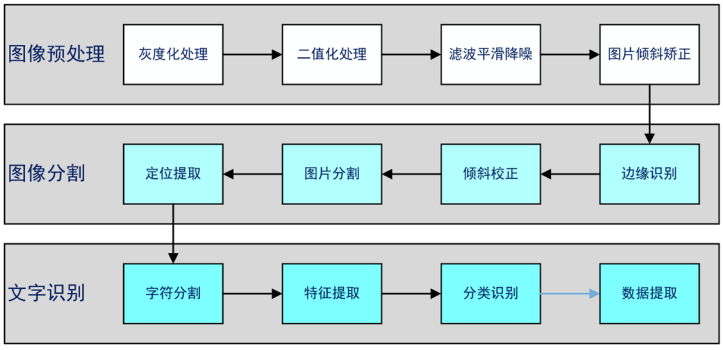

Figure 4. Unstructured document structured information extraction process
图 4. 非结构化核电文档结构化信息提取流程

3.2.8. 108 清晰度识别模块

清晰度识别采用梯度算法，对文档中的图片采用 Sobel 算子提取水平和垂直两个方向的梯度值，基于 TenenGrad 能量梯度函数实现清晰度判断。

$$D(f)=\sum_y\sum_x\left|G(x,y)\right| \quad (G(x,y)>T)$$

$$G(x,y)=\sqrt{G_x^2(x,y)+G_y^2(x,y)}$$

T 是给定的边缘检测阈值，G_x 和 G_y 分别是像素点(x, y)处 Sobel 水平和垂直方向边缘检测算子的卷积。其中 Sobel 算子模板如下。

$$g_x=\frac{1}{4}\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix} \quad g_y=\frac{1}{4}\begin{bmatrix}1 & 2 & 1\\ 0 & 0 & 0\\ -1 & 2 & -1\end{bmatrix}$$

3.2.9. 109 综合元数据校验模块

根据前置模块提供的信息，分别对各项需要校验的文档模板内容进行判断，并导出校验结果给自动检查结果显示用户接口，提供各项校验结果的说明。

3.2.10. 110 日志记录模块

记录操作过程中的日志，如图片分割结果，元数据识别结果等。

4. 结论

本项人工智能技术的应用，填补了核电企业内容自动化图像识别校验的空白，可根据核电企业内容管理要求，针对不同的文件类型和模板实现多样化定制校验规则，适应核电多专业、多机组、多技术路线的发展特点。通过对核电专属信息的自动化图像识别，保证了核电内容的完整性和准确性，极大地提升了生产效率，降低人力成本。并且图像识别技术可应用于各类应用场景，具有一定的推广价值。目前，该项技术已部署到 DPMS 流程系统中，在应用过程中，大大提高了文档移交归档接收的工作效率，实现了海量文档接收的自动化，是新型技术在档案领域应用的重大突破，也为其他新技术在传统档案管理领域的应用尝试拓展了思路。

参考文献

[1] 郭东利. 总承包模式下的核电工程项目管理[J]. 中国核工业, 2008(5): 16-20.

[2] 董宣. EPC 模式下核电工程项目文档控制管理的探索与实践[C]//中国档案学会. 2010 年全国档案工作者年会论文集. 北京: 中国档案出版社, 2010.

[3] HAF003-1991. 核电厂质量保证安全规定[EB/OL]. 2016-11-18.

[4] 张东林. 云计算技术在海量电子病历数据分析中的应用研究[J]. 太原学院学报, 2018, 36(1): 38-42.

[5] 韩智素, 王珏, 刘新科, 谭波. 一种图片中文档定位和拆切方法[P]. 中国, CN201710157232.1, 2017-03-16.