DOI:10.20033/j.1003-7241.(2025)06-0061-06

基于Tesseract-O的抽水蓄能电站工程项目电子档案管理数字化研究

范纪琨[1]，王 艳[2]，钱向清[1]，周保宗[1]，刘 芳[1]

(1.浙江缙云抽水蓄能有限公司,浙江 丽水 321400

2.国网新源控股有限公司,北京 100052)

摘 要: 针对抽水蓄能电站工程项目中,传统的纸质档案管理方式效率低下,易丢失、难查找等问题,研究系统将数据采集、图像预处理、文字识别等多个核心模块集成在一个统一的框架中,设计一种新的基于Tesseract-O的抽水蓄能电站工程项目电子档案数字化管理系统。实验结果表明,该系统的识别准确率为81.2%,成本效益得分为63.4,易用性得分为85.9。综合来看,所提系统在提升档案管理效率、降低丢失风险、增强信息检索便捷性等方面展现出了显著优势。

关键词: 文字识别;管理系统;双边滤波器

中图分类号:TP391.43 文献标识码:A 文章编号:1003-7241(2025)06-0061-06

Research on Digitalization of Electronic Archive Management for Pumped Storage Power Station Engineering Projects Based on Tesseract-O

FAN Ji-kun[1], WANG Yan[2], QIAN Xiang-qing[1], ZHOU Bao-zong[1], LIU Fang[1]

(1. Zhejiang Jinyun Pumped Storage Co., Ltd., Lishui 321400, China;

2. Guowang Xinyuan Holdings Limited, Beijing 100052, China)

Abstract: In response to the problems of low efficiency, easy loss, and difficulty in searching in traditional paper-based archive management in pumped storage power station engineering projects, the research system integrates multiple core modules such as data acquisition, image preprocessing, and text recognition into a unified framework to design a new Tesseract-O based digital archive management system for pumped storage power station engineering projects. The experimental results show that the recognition accuracy of the system is 81.2%, the cost-effectiveness score is 63.4, and the usability score is 85.9. Overall, the electronic archive digital management system for pumped storage power station engineering projects based on Tesseract-O shows significant advantages in improving archive management efficiency, reducing loss risks, and enhancing information retrieval convenience.

Keywords: text recognition; management system; bilateral filter

0 引言

在抽水蓄能电站工程项目中,由于项目周期长、涉及文件多且复杂,传统的纸质档案管理方式不仅效率低下,还存在易丢失、难查找等问题。因此越来越多的研究人员开始关注档案管理的数字化问题,鲜娅静为了提高海量档案高维特征的超大数据集管理及分析效率,结合了下一代网络(next generation network,NGN)和5G网络以及多核支持向量机(support vector machine,SVM)等技术,设计了一种智能档案管理系统,实验结果表明,与传统SVM方法相比,该方法能在不明显降低准确率情况下,大大降低了训练样本的需求量[1]。Gupta S为探究数字化时代知识管理模型的问题,采用半结构化方法调查了37名印度不同行业高管,结果显示,通过企业数字化管理可理解当前和潜在业务状况,实现更好规划和执行,确保业务连续性[2]。在集团公司"数字化转型"的战略部署指导下,研究旨在通过信息化、数字化、智能化技术手段,解决抽水蓄能电站工程项目文件收、管、用的问题,推动工程档案业务从传统工作方式向现代数字化管理模式转型发展[3-4]。研究将数据采集、图像预处理、光学字符识别(optical character recognition,OCR)识别等多个核心模块集成在一个统一的框架中,采用双边滤波器等先进的图像处理技术,对原始图像进行预处理,优化图像

*基金项目:国网科技项目(SGXYKJ-2021-099)

收稿日期:2024-01-31

质量，减少噪声和干扰因素，利用基于长短期记忆(long short-term memory，LSTM)的模型和语言模型进行文本识别，期望这项研究能进一步优化管理流程，健全管理系统，从而加快推动“数字转型”工作机制及任务目标全面落地。

1 基于Tesseract-O的数字化档案管理系统设计

1.1 工程项目电子档案数字化管理系统总体设计

基于Tesseract-O的抽水蓄能电站工程项目电子档案数字化管理系统主要由三个核心模块组成。首先是数据采集模块，该模块进一步细分为图像采集和文本采集两个子模块。图像采集模块负责捕获和收集与工程项目相关的各种图像数据，而文本采集模块则专注于提取和整理与这些图像相关的文本信息[5-6]。这一模块为后续的数据处理和分析提供了丰富而全面的原始素材。接下来是图像预处理模块，该模块对从数据采集模块获得的图像进行一系列复杂的处理，随后进行空域权重和强度域权重的计算，以进一步精确图像的细节[7-8]。经过这一系列的预处理步骤后，图像数据被输出到OCR算法模块。OCR算法模块是系统的最后一部分，也是最为关键的一环。这个模块利用Tesseract-O的强大OCR技术，将经过预处理的图像转化为可编辑和可搜索的文本信息。这一过程不仅高效地提取了图像中的文字内容，还保证了转化后的文本具有高度的准确性和可读性。通过这种方式，原本难以处理和管理的图像数据变得易于分析和利用，极大地提升了抽水蓄能电站工程项目电子档案的管理效率和便捷性，如图1所示。

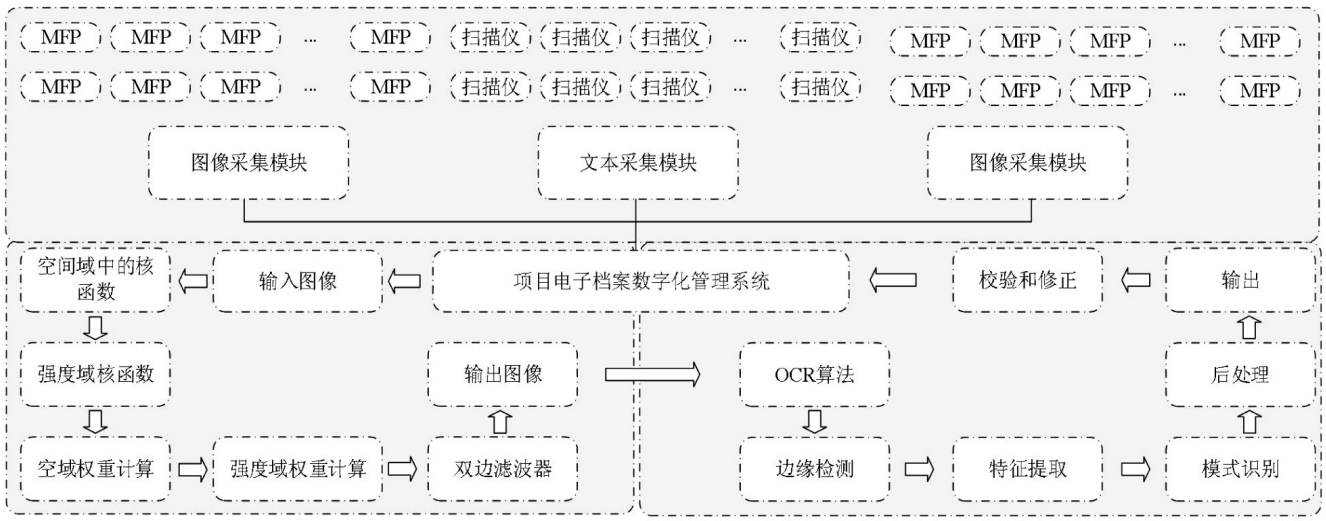


图1 基于Tesseract-O的抽水蓄能电站工程项目电子档案数字化管理系统总体设计

基于Tesseract-O的抽水蓄能电站工程项目电子档案数字化管理系统主要由以下几个实体构成，电子文件、工程项目、往来函、OCR识别结果、用户以及数据字典电子文件实体具有文件ID、文件名、文件类型、文件大小、创建时间、修改时间、存储路径等属性，这些属性详尽地描述了电子文件的各种特性。工程项目实体则包含项目ID、项目名称、项目类型、项目地点、开工时间、竣工时间等属性，这些属性全面地反映了工程项目的相关信息。往来函实体通过函件ID、项目ID、发送者、接收者、发送日期、主题、内容等属性，详细地记录了项目过程中的所有沟通信息。OCR识别结果实体具有识别ID、文件ID、识别文本、识别置信度、识别时间等属性，这些属性准确地呈现了OCR技术对电子文件的识别情况和结果。用户实体由用户ID、用户名、密码、角色等属性组成，角色属性可以区分管理员和普通用户等不同权限的用户。最后，数据字典实体包含词条ID、词条名称、解释说明等属性，为系统中的各种数据提供标准化的定义和解释，其实体-关系(entity-relationship，E-R)图如图2所示。

1.2 基于图像处理及OCR的数字化识别模型

在电子档案数字化管理系统中，管理往来函电子文件是系统的重要核心。在处理工程管理往来函电子文件的数字化过程中，首先需要对原始电子文件进行预处理，这包括优化图像质量、转换图像到灰度以及应用图像滤镜[9-10]。双边滤波器(bilateral filter)是一种常用的图像滤波算法，用于平滑图像并保持边缘信息。双边滤波器考虑了像素之间的空间距离和像素值的相似度[10-11]。在双边滤波器中，每个像素的新值由其附近像素的加权平均计算得出。这些权重由空间域权重和强度域(像素值)权重两个因素决定。空间域权重衡量了像素之间的空间距离，使得滤波器在平滑图像时更加保护边缘。通常情况下，距离较近的像素具有较高的权重，而距离较远的像素具有较低的权重。强度域权重则根据像素值的相似程度

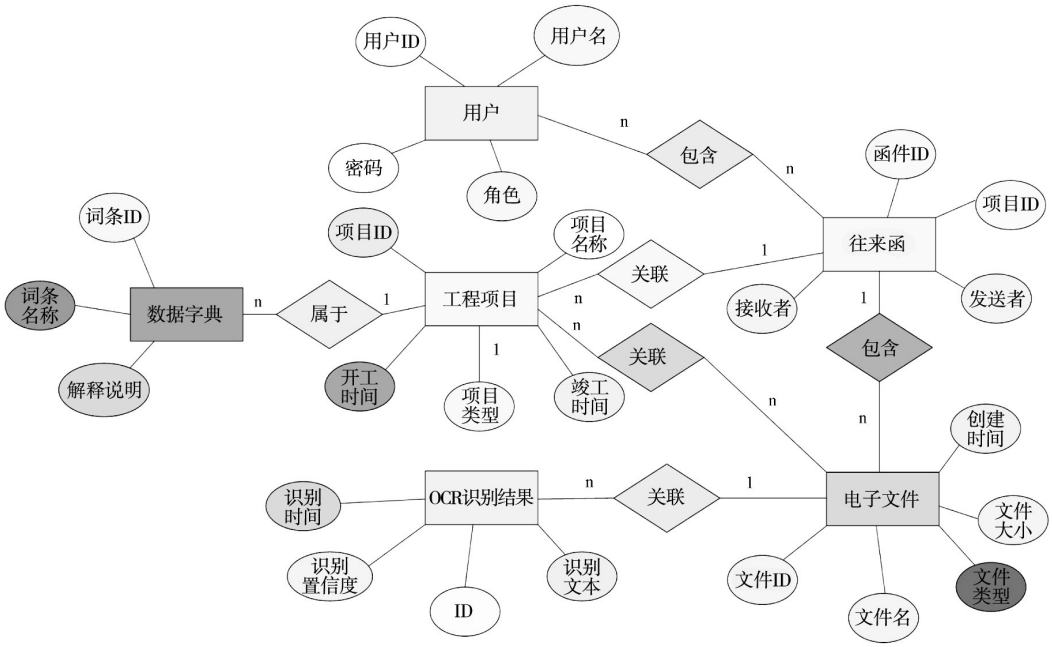


图2　基于Tesseract-O的抽水蓄能电站工程项目电子档案数字化管理系统ER图

进行调整，以确保相似的像素有较高的权重。这样可以避免对边缘和纹理等细节进行过多的平滑处理，其算法流程示意图如图3所示。

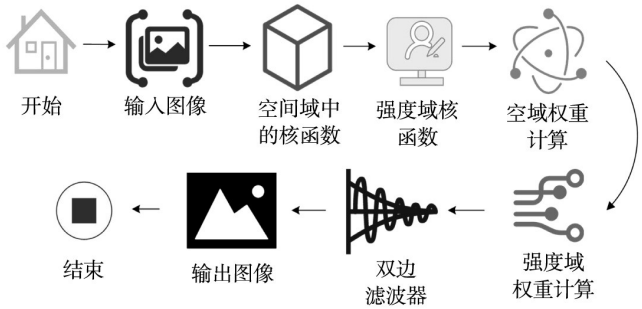


图3　双边滤波算法的主要过程

双边滤波器通过综合考虑空间域和强度域的权重，在降低图像噪声的同时，有效地保持了图像的边缘清晰度，使得图像的平滑效果更为自然。这种滤波器在图像降噪、边缘保留以及纹理增强等多个领域都有广泛的应用。然而，对于细节丰富的图像，双边滤波器可能会导致边缘模糊或不自然的过渡效果。为了解决这个问题，将高斯函数引入到传统的双边滤波中。高斯函数能够构造出与双边滤波权值相似的深度权重，从而改进滤波器的性能。假定ψ表示的是以像素点为中心的邻域范围，$w_r(i,j)$表示的是灰度域权值，则这种改进后的滤波器的表达式如式(1)所示。

$$f(x,y)=\frac{1}{w_{pd}}\sum_{i,j\in\psi} w_d(i,j)*w_r(i,j)*w_s(i,j)*I(i,j) \tag{1}$$

式中，$w_s(i,j)$表示的是空间域权值，$I(i,j)$表示的是噪声图像$w_d(i,j)$表示的是深度域权值，其表达式如式(2)所示。

$$w_s(i,j)=\exp(-\frac{(i-x)^2+(j-y)^2}{2\sigma_s^2}) \tag{2}$$

式中，σ_s表示的是空间域权值标准差，假定σ_r表示的是灰度域权值标准差，灰度域权值的表达式如式(3)所示。

$$w_r(i,j)=\exp(-\frac{I(x,y)-I(x,y)^2}{2\sigma_r^2}) \tag{3}$$

式中，假定σ_d表示的是深度域权值标准差，则深度域权值的表达式如式(4)所示。

$$w_d(i,j)=\exp(-\frac{\lambda d(ij)-\lambda d(ij)^2}{2(\sigma_d^{`})^2}) \tag{4}$$

式中，λ是未知的，因此还需要利用调控深度标准差来消除，所以$w_d(i,j)$的表达式如式(5)所示。

$$w_d(i,j)=\exp(-\frac{d(ij)-d(ij)^2}{2\sigma_d^2}) \tag{5}$$

接下来，通过图像二值化和去除噪点，可以更清晰地区分文字与背景。文本区域检测和图像分割步骤则有助于识别并分离文本行、单词或字符。正规化文字大小和提取字体、字符特征进一步为OCR算法的应用提供了准备。使用Tesseract-OCR引擎，Tesseract-OCR引擎是系统的核心部分，它结合了基于LSTM的模型和语言模型来进行文本识别。LSTM模型是一种深度学习模型，特别适合于处理序列数据，如文本。它能够学习到文本中的长期依赖关系，从而提高识别的准确性[12]。而语言模型则是利用大量的语料库训练出来的，它能够提供文本的上下文信息，帮助OCR算法更准确地识别出文本内容。通过Tesseract-OCR引擎和这些模型的结合，系统

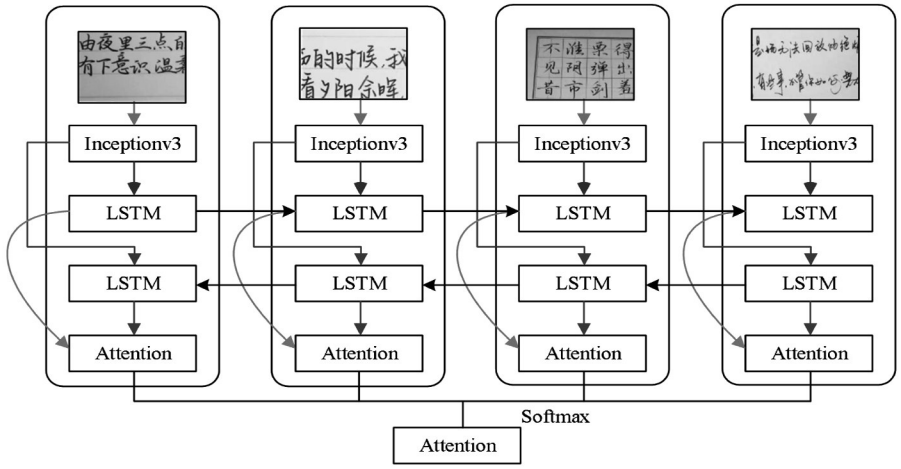


图4 基于LSTM的模型和语言模型的文字识别算法结构

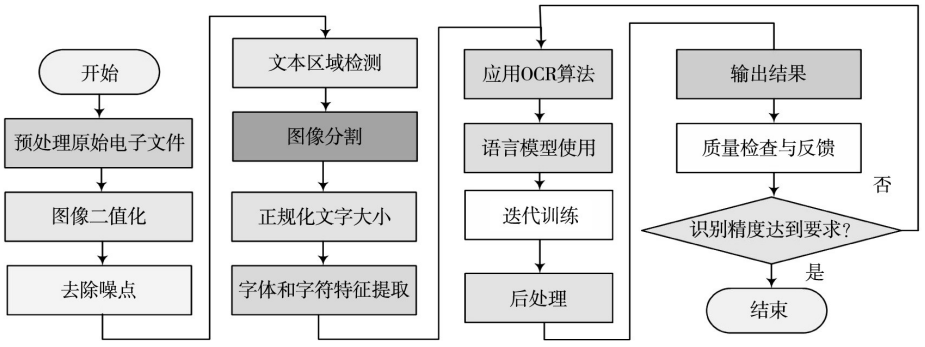


图5 基于Tesseract-OCR模型的文字识别算法流程

能够实现高效且准确的文本识别,其结构如图4所示。

后处理步骤对识别结果进行校正和清洗,确保数据的准确性。最终,识别出的文字内容被输出并整理归档,同时整个OCR过程可以被集成到现有的工程管理系统中,实现自动化处理。此外,质量检查和反馈机制不断优化Tesseract模型,提高识别精度。这一流程确保了工程管理往来函电子文件的高效、准确处理和数据的有效利用。通过这个流程,Tesseract-OCR可以将从电子文件中识别出的文本转换为可编辑、可搜索的格式,这样就为后续的数据分析和归档工作提供了便利。需要注意的是,为了达到最佳的识别效果,可能需要对Tesseract进行适当的训练,特别是在处理具有特殊格式或专业术语的工程文件时。如图5所示。

2 Tesseract-OCR模型性能测试与应用性分析

2.1 Tesseract-OCR模型性能测试

为了对图像进行更加准确的图文转换,研究在处理图像文件的时候首先使用了双边滤波,验证该算法在文档图像识别领域的优越性。实验引入高斯滤波器、灰度变换增强、同态滤波来与之进行对比。以平均梯度作为评价指标,对文字、植物、服装、建筑、美食等图像类型进行处理,实验结果如图6所示。

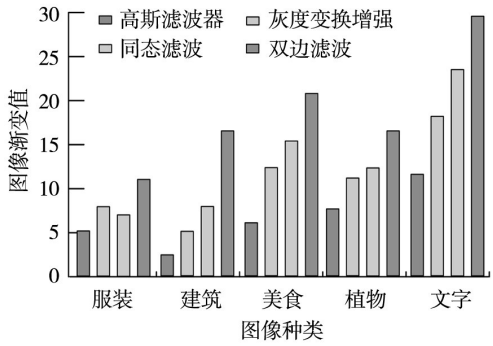
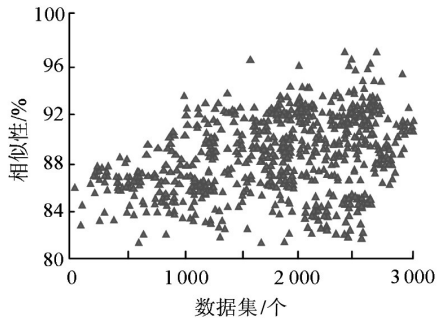


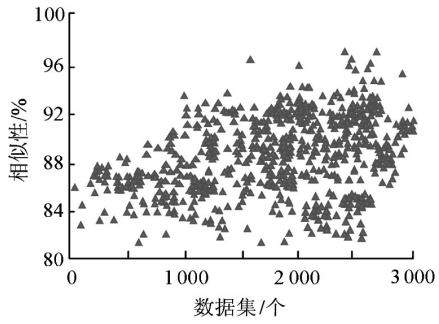
图6 四种算法的图像梯度值比较

从图6中可以看出,高斯滤波器的图像梯度值是最低的。紧接着是灰度变换增强和同态滤波。而在这四者中,研究提出的深度双边滤波器展现了最高的图像梯度值。与高斯滤波器相比,深度双边滤波器的梯度值高出了40%。此外,双边滤波在处理文字类图像后,其梯度图明显更大。这意味着深度双边滤波器在处理图像时能够保留更多的边缘和细节信息,因为它产生的图像梯度值最大。而高斯滤波器可能会使图像更为平滑,导致边缘和细节信息的损失。对于文字类图像,双边滤波似乎能够更好地增强其特征和结构,使得处理后的梯度图更为明显。除此之外,为验证该算法能对不同文字有效的泛化,选取Stanford OCR英语数据集和Devangari汉语数

据集分类泛化能力评估，部分数据样本的平均相似性结果如图7所示。



(a) Stanford OCR英语数据集



(b) Devangari汉语数据集

图7 不同数据集的模型分类能力

图7(a)和图7(b)展示了使用两种不同数据集进行训练的模型的平均相似度随数据集大小的变化情况。从图7(a)可以观察到，当数据集的大小达到1 000之前，Tesseract-O模型在Stanford OCR数据集上的平均相似度主要集中在79.2%至83.4%之间。随着数据集的增加，平均相似度逐渐提高，并最终稳定在83.1%至95.6%之间。从图7(b)可以看出，Devangari数据集在数据集大小达到1 000之前，平均相似度在83.3%至88.1%之间。随着数据集大小增加到2 500，平均相似度稳定在85.1%至97.3%之间。两个数据集的平均相似度差异不大，这表明Tesseract-O方法具有良好的泛化能力。

2.2 Tesseract-OCR模型应用性分析

实验充分验证了研究所提Tesseract-O模型在文字识别领域的优越性，为了进一步验证该模型在实际应用中的表现，实验将研究提出的基于Tesseract-O的抽水蓄能电站工程项目电子档案数字化管理系统应用于集团所属的多个抽水蓄能电站工程项目电子档案管理之中，并对多个文本文字进行了图文转换实验，实验结果如图8所示。

图8 Tesseract-O模型识别工程项目电子文档图文转换结果

图8详细展示了Tesseract-O模型在识别工程项目电子文档图文转换方面的卓越成果。从图中可以清晰地看到，Tesseract-O对于中文的识别准确率达到了令人惊叹的高度。无论是复杂的排版、多样的字体，还是图像质量的差异，该模型都能以出色的稳定性和准确性进行文字提取和转换。这不仅大幅提升了文档处理的效率，也为相关领域的研究和应用提供了有力支持。Tesseract-O的这一表现，无疑证明了其在中文OCR技术方面的领先地位，为未来的智能文档处理打下了坚实基础。除此之外，实验还针对不同语言进行了广泛的测试，包括英文、法文、德文等，以全面评估Tesseract-O模型的多语言处理能力。实验结果通过相似性热力图直观地展示了模型在不同语言间的性能差异，如图9所示。

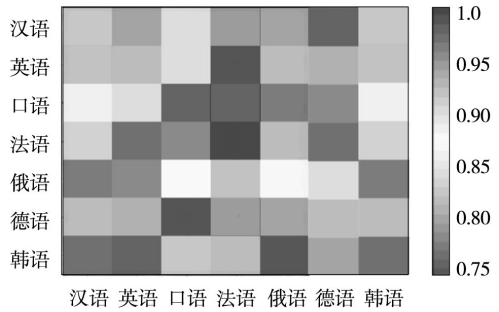


图9 Tesseract-O模型的多语言处理能力相似性热力图

从相似性热力图中，可以清晰地观察到，大多数语言的相似值都稳定地保持在0.80以上。这表明Tesseract-O模型在处理不同语言时，不仅能够准确地识别各种语言的字符，还能有效地处理不同语言之间的语法和语义差异。这种高度的灵活性和适应性充分展现了Tesseract-O在多语言处理方面的卓越能力。这一重要发现为Tesseract-O模型的进一步优化和扩展提供了宝贵的参考依据。实验最后邀请了5名管理系统方面的专家对该系统识别准确率、成本效益、易用性进行了综合评价，评价采用百分制评分体系，实验结果如图10所示。

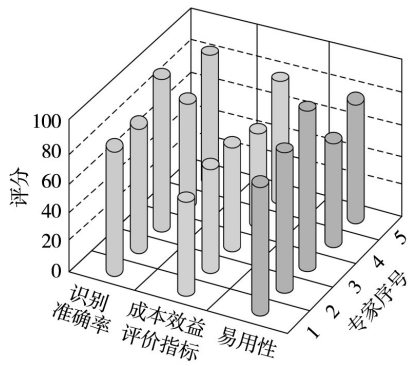


图10 系统个性指标综合得分情况

从图10中可以看出，系统的识别准确率、成本效益、易用性平均得分分别为81.2、63.4、85.9。专家们对系统的识别准确率和易用性给予了较高的评价，认为系统在这两个方面表现优秀。而在成本效益方面，虽然评分略低于其他两个指标，但整体仍处于较高水平，表明系统在经济效益方面也具有较高的竞争力。

3 结束语

为了提高抽水蓄能电站工程项目电子档案管理系统的数字化程度，研究基于Tesseract-O结合双边滤波器ORC识别开发出一种新的档案管理系统。实验结果表明，深度双边滤波器在处理图像时能够保留更多的边缘和细节信息，从而提高图像识别的准确性。与高斯滤波器相比，深度双边滤波器的梯度值高出了40%。Tesseract-O模型在Stanford OCR英语数据集和Devangari汉语数据集上表现良好，平均相似度均稳定在较高水平，具有良好的泛化能力。从相似性热力图中，可以清晰地观察到，大多数语言的相似值都稳定地保持在0.80以上。这表明Tesseract-O模型在处理不同语言时，不仅能够准确地识别各种语言的字符，还能有效地处理不同语言之间的语法和语义差异。在实际应用中，该系统识别准确率和易用性平均得分分别为81.2和85.9，表现出色。成本效益得分为63.4，虽有一定提升空间，但整体经济效益较高。这一研究成果为智能文档处理和多语言OCR技术的发展提供了新的思路和方向。未来可以进一步优化Tesseract-O模型以降低成本，并探索其在更多领域的应用潜力。

参考文献：

[1] 鲜娅静.基于NGN和5G的档案管理系统研究与仿真[J].微型电脑应用，2023，39(2)：161-163，168.

[2] Gupta S，Tuunanen T，Kar A K.Managing digital knowledge for ensuring business efficiency and continuity[J].Journal of Knowledge Management，2023，27(2)：245-263.

[3] 邵旭东，曹志威，樊志杰，等.移动警务跨域消息提醒服务系统的设计与实现[J].计算机测量与控制，2023，31(4)：264-271.

[4] Heng，B.，You，C.，Mei，W.，et al.A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications[J].IEEE Communications Surveys&Tutorials，2022，24(2)：1035-1071.

[5] 吴欣.基于流媒体技术的医学档案信息资源数字化传输[J].微型电脑应用，2023，39(8)：213-216.

[6] 柴丽萍，杜一玮，庄硕，等.数智时代企业智慧文档管理体系构建研究[J].情报科学，2022，40(12)：36-41.

[7] 佟岩，刘柯慧，赵泽与等.企业数字化转型对客户集中度的影响[J].北京理工大学学报(社会科学版)，2024，26(1)：177-194.

[8] 石瑞杰，邹萍，吴夕科，等.电工装备行业云端协同制造及仿真应用研究[J].系统仿真学报，2019，31(4)：771-786.

[9] 邹名璐，罗元.电子文件归档管理系统的核心功能单元设计[J].计算机工程与科学，2019，41(3)：498-504.

[10] 郑剑，刘宁，张毅，等.一种报表自适应推荐算法的设计[J].自动化技术与应用，2023，42(12)：128-130，135.

[11] 李航.基于Web服务器的高校思政教育系统设计[J].自动化技术与应用，2023，42(12)：177-179.

[12] 孙鹏，韩璐，王书源，等.基于数据整合的企业物资采购信息管理系统[J].自动化技术与应用，2023，42(11)：169-173.

作者简介：范纪琨(1986—)，女，本科，高级政工师，研究方向：行政管理、档案管理等。