科技创新
Technical Innovation

# 基于 OCR 识别的民生档案数字化管理研究

文◆海南省疾病预防控制中心　陈　丽　任　颖

## 引言

OCR（Optical Character Recognition，光学字符识别）属于一种高效的文字输入方式，亦可称之为文字识别。OCR 技术的运用过程通常涉及将纸张上的文字、图像信息转化为计算机能识别的格式[1]。在档案工作“存量数字化、增量电子化”的要求下，研究 OCR 识别在民生档案数字化管理中的应用，设计基于 OCR 识别的档案数字化管理方案，有助于解决纸质档案在扫描、识别、分类等环节容易出错且耗费大量人力的问题，提升民生服务效率，推动信息化建设再上新台阶。

## 1 OCR 识别的工作原理

OCR 的主要原理是利用光学字符识别技术将图像中的文本转换成计算机可检索、编辑的格式，以提高数据输入及处理的速度和效率[2]。硬件部分主要包括扫描仪、摄像机、光源、光学器件、图像处理器等。扫描仪、摄像机捕捉原始图像，光源为图像提供清晰的照明，光学器件则负责将图像中的文字映射到图像处理器上。而图像处理器则负责处理图像，以供后续识别。在捕捉到原始图像后，会对图像进行预处理，以提高图像质量，便于后续识别。首先，通过灰度处理、二值化处理、噪声去除等消除噪声、去除干扰。其次，进行文字检测和定位。应用边缘检测算法、霍夫变换算法等，分析图像中的字符大小、形状、布局等文字特征，确定文字区域的位置和大小，以便准确检测定位文字区域。再次，进行字符识别。此为 OCR 识别的核心部分，利用深度学习算法，进行大量训练数据，以训练出准确识别各种字符的模型，再识别和分类每个字符。最后，将识别出的文本转换成可检索、编辑的 TXT、XML、PDF 等格式[3]。

## 2 OCR 识别在民生档案数字化管理中的应用

### 2.1 提取档案内容

OCR 在民生档案数字化管理中的应用主要表现在快速检索、内容数字化和自动化归档 3 个方面。OCR 技术在档案内容提取中极具应用前景，在图书馆、博物馆、档案馆、历史档案、法律文档识别、企事业档案管理、家庭档案管理等领域广泛应用，利用 OCR 技术将大量民生档案资料转化为数字格式，为国民生活、工作、文化传承提供了宝贵的资源和更加便捷的服务。运用 OCR 技术，可快速搜索档案中的特定信息，提高档案管理效率和保存安全性，延长档案寿命。

### 2.2 制作电子档案

电子档案现已成为政府部门、学校、医院、服务企业等民生机构的重要数据来源。OCR 技术作为制作电子档案的关键技术，通过将纸质文档中的文字转化为电子文档，有效提升了文档管理效率。首先，收集需要转化为电子档案的纸质文档，确保文档清晰、完整。其次，根据实际需求，确定需要转换的文档范围及类型，如文本、表格等。为确保 OCR 软件正常工作，应配置高分辨率屏幕、高速硬盘等计算机硬件。再次，

【作者简介】陈丽（1977—），女，海南海口人，本科，档案副研究馆员，研究方向：档案管理、档案文献编撰等。

将处理后的图像导入软件中开始识别，观察自动识别结果，如识别有误，则手动修正，确保结果准确、可靠。最后，对于数据庞大的文档，为提高工作效率，可采用批量转换的方式，校核完成后将电子档案导出保存即可。

### 2.3 建立 OCR 文本数据库

在民生档案数字化管理中，建立文本数据库对纸质档案中的文字进行自动识别（OCR）处理非常重要。应选择合适的 OCR 技术，如 Tesseract、ABBYY FineReader 等，具体可根据需求以及预算进行选择，且所选 OCR 技术应兼容各种字体、字号和行距。建立文本数据库的关键步骤是收集数据。在收集过程中，应确保数据多样性，包括纸质文档、扫描图片和其他电子文件等，以实现全面的训练与测试。完成数据收集后，进行 OCR 处理，即将纸质文档和扫描图片中的文字转换为计算机可读的格式。随后选择 MySQL、MongoDB 等合适的数据库管理系统，用于存储和管理文本数据。为确保数据安全，还应定期备份数据库、加密数据库和限制访问权限，防止数据未经授权而被篡改。

## 3 基于 OCR 识别的民生档案数字化管理方案设计

首先，提高传统载体档案的数字化率。积极推动各级单位在档案管理中采用扫描和 OCR 识别等数字技术，实现对纸质档案的数字化管理，使其转化为可检索、复制的数字形式[5]。并在云端或服务器中存储这些数字化档案，方便利用。其次，机器学习、人工智能等先进技术推动档案信息自动化处理，极大地提高了档案管理效率。在新增档案中，应确保其以电子化的形式归档，不仅有利于档案管理员在多个系统中共享和使用档案信息，提高档案的利用效率，还有利于保护珍贵档案资源，降低物理磨损或丢失的风险。基于 OCR 识别的民生档案数字化管理具体方案架构如表 1 所示。

表 1 基于 OCR 识别的民生档案数字化管理方案架构

| | | | | | |
|---|---|---|---|---|---|
| 应用层 | 上层业务 | | | | |
| | 档案服务 | 专项档案 | 档案智库 | 档案展览 | …… |
| 平台层 | 民生档案管理数字化系统 | | | | |
| | 档案接收 | 档案管理 | | 档案分析 | 个人中心 |
| | 档案收集 | 分类 | 编号 | 档案检索 | 权限管理 |
| | 档案著录 | 编目 | 关联关系管理 | 档案统计 | 我的借阅 |
| | 目录管理 | 档案借阅 | 档案销毁 | 档案业务可视化 | 我的审核 |
| | 档案归档 | 档案编研 | 档案审核 | 档案热词 | |
| | | 档案发布 | 档案鉴定 | | |
| 算法层 | 文档预处理算法 | | 文字识别算法 | | 文本理解算法 |
| | 矫正切边、扭曲；去除阴影；锐化增强；去除印章 | | 印刷体识别；手写体识别；中文繁简识别；英文识别；小语种识别；文档识别；印章识别；文档还原；表格识别 | | 实体识别；关系抽取；标签识别 |
| 模型层 | OCR 大模型 | | NLP 大模型 | | |
| | VIMER-StrucText | | ERNIE-Layout | | |
| | VIMER-MaskOCR | | ERNIE-mmLayout | | |

### 3.1 应用层

应用层为上层业务，主要包含以下板块。

（1）档案服务。为用户提供各种档案服务，如在线查询、档案借阅、档案复制等。本方案将使用 OCR 技术识别档案图片中的文字，并将其转化为可编辑的电子文本，以便用户获取档案信息。（2）专项档案。根据不同领域和主题，建立专项档案库，如教育、医疗、人事、财务等。这些档案库将提供详细的档案信息，方便用户查找、使用。（3）档案智库。建立档案智库，为用户提供各种研究资料及数据。（4）档案展览。应用 OCR 技术快速准确地识别和整理展览所需的档案资料，定期举办档案展览。

### 3.2 平台层

平台层是本方案的核心，即民生档案管理数字化系统，由档案接收、档案管理、档案分析、个人中心 4 个子模块构成，负责接收、管理、分析和展示各种档案信息。

档案接收模块负责接收和整理纸质档案、照片档案、音频档案等各种形式的档案，利用 OCR 技术将这些档案识别并转化为数字化形式，以供后续管理。档案接收模块又包括以下内容。（1）档案收集。 OCR 技术自动识别并导入各类纸质档案，提高档案收集效率。（2）档案著录。对导入的档案进行规范化著录，建立档案目录，以便后续管理。（3）目录管理。对档案目录进行分类管理，便于查找、使用。（4）档案归档。将档案按照规定程序归档至数字化档案管理系统，实现档案集中管理。

档案管理模块负责管理数字

化档案，具体包括以下内容。（1）分类。根据档案类型和属性，将其分类存储在相应目录下。（2）编目。对档案进行标签、描述和关键词提取，便于检索、利用。（3）档案借阅。提供档案借阅申请、审批流程，实现档案共享利用。（4）档案编研。对档案进行深度挖掘和整理，形成有价值的编研成果。（5）编号。为每份档案分配唯一的编号，便于识别和管理。（6）关联。建立档案间的关联关系，便于用户通过关键词或主题查找相关档案。（7）关系管理。对档案之间的关系进行规范化管理，形成关系网络。（8）档案销毁。对超过保存期限或无利用价值的档案进行销毁，节省存储空间并防止信息泄露。（9）档案审核。对新增或修改的档案进行审核，确保档案信息的准确性和完整性。（10）档案鉴定。定期对档案进行鉴定，分别界定长期保存、归档数字化及销毁范围。

档案分析模块提供多形式的档案检索和分析功能，具体包括以下内容。（1）档案检索。提供关键词检索、主题检索、时间检索等多种检索方式，方便用户查找所需档案。（2）档案统计。统计分析各类档案的数量、利用情况等，为决策提供数据支持。（3）档案业务可视化。通过图表、数据可视化等方式，展示档案管理业务的整体情况。（4）档案热词。分析用户在检索过程中常用的关键词，了解用户感兴趣的档案主题类型。

个人中心模块提供用户个人账户管理和权限管理功能，同时提供用户借阅记录和审核记录等功能，方便用户查询和管理档案信息。

### 3.3 算法层

算法层包括文档预处理算法、文字识别算法和文本理解算法。

（1）文档预处理算法。利用图像处理技术，对档案图像进行切边校正，使其符合预期的尺寸和形状。利用图像修复技术，校正因纸张变形等原因导致的扭曲。利用图像分析技术，去除档案图像中的阴影部分。通过增强图像的对比度和清晰度，提高文字识别精度。采用自动化算法去除档案图像中的印章，以免干扰文字识别。（2）文字识别算法。利用机器学习算法，自动识别印刷体文字。采用先进的神经网络模型，识别手写体文字。采用多模型算法识别中国汉字的复杂性和多样性，自动识别常见的英文字符。利用现有语种模型识别其他语种。采用特定算法识别和还原含表格的档案文件。（3）文本理解算法。自动识别文本中的实体，如人名、地名、机构名等。通过分析文本中的关系，如时间、地点、人物等关系，获取更丰富的信息。为档案文件提供标签和分类，便于后续数据分析和利用。

### 3.4 模型层

模型层由 OCR 大模型和 NLP 大模型构成。其中，OCR 大模型包括 VIMER-StrucText 模型、VIMER-MaskOCR 模型。VIMER-StrucText 模型用于识别结构化数据，可准确识别档案标题、时间、地点、事件等信息，为后续的数据挖掘提供支持。VIMER-MaskOCR 模型可识别档案图像中特定区域的文字，便于档案的分类整理和检索。NLP 大模型包括 ERNIE-Layout、ERNIE-mmLayout 等模型，可对数字化档案进行智能分类、摘要生成和关键词提取。ERNIE-Layout 模型通过分析档案的布局和格式，对档案内容进行初步分类，并生成摘要。ERNIE-mmLayout 模型可整体分析多页档案，提取关键信息，提高档案摘要生成效率。

## 结语

档案数字化管理是信息化时代的必然趋势，研究基于 OCR 识别的数字化管理具有重要的现实意义。应用 OCR 技术有利于实现民生档案信息的快速识别、智能分类和高效利用，提高档案管理的效率和水平，为民生幸福提供更加优质的技术支持。

### 引用

[1] 蔡军,陈欣欣.OCR技术在城建档案文件级著录信息自动获取与校核中的应用[J].兰台世界,2022(6):95-97+107.

[2] 刘妍.大数据背景下OCR全文检索对档案著录带来的机遇与挑战研究[J].档案天地,2023(8):37-40.

[3] 郑慧,刘思含.人工智能与档案开发利用:应用、愿景与进路[J].山西档案,2022(5):5-10+28.

[4] 胡欧哲,张欣.基于OCR技术的高校财务档案安全管理系统研究[J].武汉理工大学学报(信息与管理工程版),2023,45(1):160-164.

[5] 毛海帆,李鹏达,傅培超,等.基于数据挖掘技术构建辅助档案开放鉴定模型[J].中国档案,2022(12):29-31.