디지털산업정보학회 논문지 제21권 제1호 - 2025년 3월
http://dx.doi.org/10.17662/ksdim.2025.21.1.045

디지털기록관리에서의 AI OCR 데이터 활용 및 확장방안 연구*

오 세 종** · 이 선 아***

A Study on AI OCR Data Utilization and Expansion in Digital Record Management

Oh Sejong · Lee Sun-A

〈Abstract〉

With the amendment of the 'Framework Act on Electronic Documents and Electronic Transactions' digitized documents were granted the same legal validity as paper documents. In this context, OCR technology gained significant attention. We conducted a study on the improvement of AI OCR technology for the digital conversion of paper records. By analyzing the results of projects such as OCR for public administration documents, OCR for educational materials, and multilingual OCR for cultural tourism, we identified key directions for enhancing AI OCR in record digitization processes. The research method was an in-depth expert interview, and records manager and developers participated. The analysis resulted in three recommendations. First, the development of a semantic-based search tool and the provision of summarized information. Second, the enhancement of security and compliance for records. Third, the automation of repetitive tasks through the integration of OCR with RPA(Robotic Process Automation) algorithms. The expansion of OCR technology has resulted in reduced human errors and increased reliability in the digitization process, decreased workload for specialized personnel, and enhanced security in records management.

Key Words : Non-Electronic Records, Records Management, OCR, AI OCR, Semantic Search

Ⅰ. 서론

1.1 연구의 배경 및 방법

'전자문서 및 전자거래 기본법'이 개정(2020년 6월)되면서 전자화대상문서가 폐기되고, 디지털화된 문서가 종이문서와 동일한 법적 효력을 갖게 되었다[1]. '전자문서 및 전자거래 기본법'에 따르면, 법령의 개정은 인공지능 시대에 적합한 전자 문서의 활용을 높이기 위함이며, 종이 문서를 전자화하여 공인 전자 문서 센터에 보관하는 경우 해당 종이 문서는 보관하지 않아도 된다. 법령의 개

* This research was financially supported by Hansung University.
** 한성대학교 AI응용학과 조교수
*** 한국외국어대학교 정보기록학과 박사수료

정 이후로 종이 문서의 생산, 보관에 큰 비용이 발생하는 금융, 의료, 부동산 산업 등에서는 문서의 전자화가 지속적으로 확대되고 있다[2]. 특히, 국내에서는 2021년 디지털 뉴딜 사업의 일환으로 D.N.A(Data, Network, AI) 생태계 강화가 추진되면서 디지털화가 미미한 문서를 대상으로 한 OCR(Optical Character Recognition) 디지털 전환 기술이 주목받았다.

Adobe의 광학문자인식(OCR)은 인쇄된 문서를 디지털 이미지 파일로 변환하고, 카메라 이미지, 이미지 전용 PDF, 스캔문서 등에서 텍스트 데이터를 추출하는 기술이다. 광학문자인식은 주민등록증 인식, 명함 인식, 신용카드 인식, 서식 인식, 진료비계산서, 정보마스킹, 필기체 인식, 문서 인식, 문서복원, 실시간 단어 뜻/번역 등으로 활용되고 있다. OCR 기술은 데이터 입력의 자동화에 기여하여 인력에 투입되는 비용을 절감하고, 데이터 기술(記述)의 정확성을 높이면서 금융, 의료, 법률 등 다양한 산업에서의 기록을 디지털화하고 있다.

본 연구는 과학기술정보통신부와 한국지능정보사회진흥원(NIA)의 지원 사업 중, '기록물 DB 구축 솔루션, 비전자기록물관리 솔루션, 디지털 아카이브 솔루션' 등을 담당하는 디지털 기록관리 전문기업인 쇼우테크(SHOWTECH)에서 수행한 '공공기관의 공공행정문서 OCR, 교육 학습용 OCR, 문화 관광 다중언어 OCR'에서 한계점을 도출하고, 단순 문자 인식에서 구조인식, 문장변환, 요약정보를 제공하는 AI OCR의 향후 확장 방향을 고찰하였다. 연구 방법으로는 분야별 결과 분석과 기록물관리전문요원 및 개발 전문가 인터뷰를 활용하였다.

1.2 선행연구

Taniguchi[3]는 OCR을 활용하여 서지 기록의 중복을 감지하는 방법론을 논의하고 실험을 진행하였다. 서지 데이터의 정확성과 일관성을 높이기 위해 중복 감지 알고리즘을 적용하였고, 기록의 매칭과 일치율의 분석 과정에서 유용함을 증명하였다. 강지홍[4]은 비전자기록물 중 타자기록물에 대하여 OCR 적용 이후의 성능을 평가하고 개선하였다. OCR 결과로 생성된 PDF 파일을 검색엔진에 제공하여 전문 검색, 키워드 추출, 색인 등록에 활용할 수 있음을 보이며, CAMS 내 시스템 기능 명세를 정의하였고, 시스템 개발에 필요한 구체적인 요건을 도출하는 방식으로 연구를 진행하였다. 안세진, 황현호, 임진희[5]는 AI-OCR을 적용하여 공공기관(김포시)의 기록화 사례를 소개하고, 내부에서의 활용 평가 및 개선사항을 제안하였다. 기록 검색 범위 확대, 개인정보 마스킹, 한자 및 주소변환, 검색어에 대한 학습을 통한 데이터 추출 등에서 아쉬움이 있었음을 이야기하였고, OCR 기술 적용의 과정에서 기록물관리전문요원의 역할과 책임에 대하여 소개하였다. 이준, 유민선[6]은 타자체, 필기체에 특화된 OCR 모델로 범위를 특정하여 연구를 진행하였고, OCR을 활용한 비전자기록물의 전자화를 위해서는 담당자간의 활발한 업무 협의, 기록관리자의 체계적인 기록 관리 계획, 이용자 의견 수렴을 통한 전반적인 행정업무의 개선 등이 필요하다고 주장하였다.

기존 선행 연구는 OCR 적용 이후의 데이터의 정확성, 품질 평가, 기록화 사례 분석, 기록관리자의 업무 효율화 등의 비전자기록물의 품질 사례 연구와 행정업무의 개선에 대한 내용에 한정되어있다.

그 외에도 Xujun Peng, Huaigu Cao, Srirangaraj Setlur, Venu Govindaraju, Prem Natarajan[7]은 수십 년간 OCR 관련 연구는 단일 언어 분석이 대부분이기에 다국어 OCR 연구가 필요함을 주장하였다. 다국어 OCR 분석 연구를 위해서는 복잡한 문자 세트, 구음 부호, 다국어 컨텍스트, 낮은 입력 품질의 개선이 필요하다. 오현호, 윤준석, 배유석, 김형일[8]은 영어 및 중국어 기반 다중언어 학습 문자인식에 대한 성능을 제시하고 있다. 단일 언어에 대한 딥러닝 기반 문자인식 기술을 적용한 결과에서는 다중 언어가 혼재하는 경우 오류가 발생함을 확인하였다. 또한, 옥외광고나 야외 촬영 이미지를 분석에서는 다중 언어 혼재시 언어 컨텍스트 변화를 통한 문자인식 기술 고도화가 필요하다고 조언하였다. 이창엽, 김동

주, 서영주, 황도경[9]은 OCR을 수행하기 위한 최적의 파이프라인을 제안하였다. 각 언어의 특성을 고려한 언어 분류 모델을 기존 모델에 통합하고, 이를 기반으로 언어별 독립 인식 모델로 구성한 다중언어 OCR이 가능한 방법론을 제안하였다.

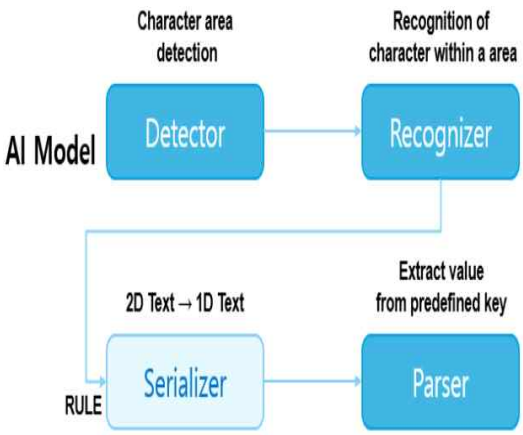
이처럼 선행연구에서는 OCR을 활용하여 종이기록의 데이터화, 데이터셋의 인식률 평가, 행정업무의 개선, 산업 분야별 적용 사례를 주요 연구의 소재로 다루고 있다. OCR 분석은 단일 언어 분석에서 다중언어 OCR 분석으로 확대되고 있지만, 다중언어 OCR에 대한 실험 및 연구는 드물다. 또한, 기록물관리전문요원 및 개발 전문가는 비전자기록물에서 디지털 전환을 수행하면서 다양한 문제가 유발하고 있음을 한계로 언급하였다. 그 문제점을 분류하고, 기술적으로 해결할 수 있는 개선 연구가 필요하다. 본 연구는 공공기관, 교육, 문화관광 다중언어의 실제 활용 사례를 조사하여 시사점을 분석하고, 자가 학습이 가능한 AI OCR 기술 개발과 활용을 위한 추진 방안을 제시한다는 점에서 선행연구와 차이가 있다.

Ⅱ. 이론적 배경

2.1 광학문자인식 정의 및 주요 OCR 현황

광학문자인식은 인쇄된 문서를 디지털 이미지 파일로 변환하는 기술이다. 업스테이지의 OCR 개념은 과거에는 이미지에서 글자를 인식하는 기술만을 의미하였으나 최근에는 이 기술을 활용하고 있는 기업들의 필요한 정보 추출 수요가 높아지면서 OCR 기술들은 인식된 글자에서 정보를 추출하는 파싱(Parsing)까지 포함하여 'OCR'라 정의하고 있다.

예를 들면, upstage[10] 명함 이미지를 추출하는 과정에는 세 가지 OCR 기술이 적용된다. 첫 번째, 검출을 위해 감지기(Detector)가 입력된 이미지를 받아 글자 영역이 어디에 있는지 위치를 알아낸다. 두 번째, 인식(Recognizer)을 통해 영역별로 어떤 글자가 있는지 문자열까지 인식한다. 가끔 편의를 위해 시리얼라이저(Serializer)를 사용하여 글자의 위치와 문자열을 일렬로 정렬하게 되는데, 보통 OCR API를 사용하게 될 경우 여기까지의 결과를 얻게 된다. 세 번째, 파싱(Parsing)이란 'Company' = '한국외대', 'Phone Number' = '010-1234-5678'과 같이 앞서 뽑아낸 문자들을 key&value 쌍의 형식으로 구조화하는 방식이다. 재가공된 광학문자인식의 추출 프로세스는 아래와 같다.



〈그림 1〉 업스테이지의 광학문자인식(OCR) 및 추출 프로세스

네이버에서 제공하는 CLOVA OCR은 읽는 순서와 방향을 추정해 이미지 속 문자를 인식하며, 한국어 지원에 특화되어있어 방언까지 인식한 데이터를 추출하고 있다[11]. 구글은 Google Cloud AI OCR을 서비스하며, 객체 인식, 감정분석 등의 기술이 접목되어 활용도가 다양한 편이다[12]. CLOVA OCR보다 개발자의 의도대로 API 수정 및 커스터마이징이 용이하지만, 소규모로 OCR을 진행하려는 국내의 개인 사용자들은 복잡한 기능에 어려움을 느낄 수 있다.

〈표 1〉 네이버 Clova OCR과 구글 Cloud OCR 비교

Category	Naver Clova OCR	Google Cloud OCR
Business Size	Rationalize pricing policies for individual users and small businesses	High-speed image processing (Global Biz)
Multination Al language	Specialized in Korean language support, it is strong in recognizing special characters and dialects	Supports many languages and is easy to use globally
Scalability	It is difficult to expect functions beyond simple text recognition due to the lack of advanced functions available compared to Google	Present more usability through AI technologies such as object recognition and emotion analysis of Google
Customization	Easy to access, but difficult to modify and customize APIs to fit the developer's intentions	Can modify and customize API directly to meet developer intentions
Interface	user-friendly interface	The interface is a bit difficult for first-time users
Customer Response	Ease of customer support in Korea	Difficulty in supporting from headquarters in case of a problem
Template	Easy to identify layouts and forms within a document through a particular template	Customizable but somewhat difficult for light users
Infrastructure	Suitable for users who want fast, small-scale OCR processing in Korea	Easy to work with other Google services, great for global target businesses with vast global infrastructure

2.2 기록관리분야에서의 AI OCR

광학문자인식을 통한 이미지 파일의 텍스트 추출을 기록관리 분야에 접목한 사례로 정부지원 사업을 수행한 쇼우테크의 과제를 분석하였다. OCR 추출 프로세스는 비전자기록물 상태에서 목록 위주의 관리를 진행하는 과정에 활용하였으며, 필요에 따라 물리적으로 기록을 검색하고 소장된 기록을 활용하여 새로운 콘텐츠를 개발하는 작업에도 접목하였다. AI OCR을 통한 전문 인식은 방대한 양의 문서 내용을 신속하고, 효율적으로 찾을 수 있는 환경을 제공하기에 기록을 활용하고자 하는 이용자의 편의에 기여한다.

광학문자인식을 통한 텍스트 추출을 기록관리에 접목하고 있는 주요 기관으로는 해양수산과학기술원, 국립경상대학교, 부산항보안공사 등이 있다. 해양수산과학기술원에서의 AI OCR 처리는 해양수산 연구 보고서(평가보고서, 결과보고서), 연구계획서, 기술문서의 검색, 활용 서비스와 주로 연결된다. AI OCR에 기반한 디지털 전환으로 기록 보관의 체계, 소요 시간의 단축, 기록물관리전문요원의 피로가 감소하는 효과를 확인하였다. 국립경상대학교에서는 오래된 학교사 기록물의 이중 보존 및 검색 서비스에 AI OCR을 적용하였다. 대학신문, 교강사 연구과제, 학사기록 등 오래된 문서류를 경상국립대학교 기록관의 디지털 기록콘텐츠로 전환하여 교육 콘텐츠 기획에 활용하였다. 부산항보안공사에서는 AI OCR을 활용하여 검역 및 관리 기술 기록(수출입 화물 검역, 물품 검사 기록), 내부 안건 및 보고서(항만 보안 계획, 안전관리 지침) 등을 디지털화하면서 자동 분류 기능을 함께 설계하여 내부 자료로 활용할 수 있도록 제공하였다는 점에서 의의를 가진다.

해양수산과학기술원, 국립경상대학교, 부산항보안공사 등의 기관에서 수행한 광학문자인식을 통한 이미지 파일의 텍스트 추출 과정에서 아래와 같은 개선사항도 함께 확인할 수 있다. 첫째, 다양한 서체와 배경 노이즈는 인식의 정확도에 영향을 미친다. 손글씨, 필기체 등의

인식 과정에서 오류가 발생하였고, 손상된 고문서 일수록 정확도가 더 낮았으며, 한·중·일 등의 다국어 인식 사례에서는 언어적 특성과 문법의 차이에서 비롯된 인식 오류가 발생하였다. 특히 다중언어로 표기된 기록의 경우, 추출된 텍스트의 후처리 과정에서 맥락을 고려하지 않으면 의미가 왜곡될 위험이 높기 때문에 더욱 정교한 알고리즘과 데이터셋의 개선이 필요하다. 또한, 한국어 문화권의 외국어 표기법 특성과 해석의 한계에서 비롯되는 문제를 해결하기 위한 지속적 연구와 기술 개발도 요구된다. 둘째, 복잡한 수식이나 도형을 식별 과정에서의 신뢰성이 낮다. 교육기관의 기록에서 다수 확인할 수 있는 수식 및 도형 기록물을 잘못 인식하는 사례가 다수 발생하였다. 수식과 도형의 학습 데이터셋을 확대하여 훈련을 강화하려는 노력이 필요하다. 셋째, 공공문서의 경우 시대적 흐름에 따라 변경되는 서식 구조와 수기체, 타자체 등의 서체 변화 부분에서 문자인식의 정확성이 다소 감소하였다. 데이터 편중화 해소를 위하여 특정 비율을 맞추어 학습용 데이터셋을 구성하여야 하지만, 보존기간이 종료되어 이미 폐기된 기록이 많은 연도, 부서 등의 기록물 그룹에서는 비율을 맞추어 학습을 진행하기에 어려움이 있다. 즉, 공문서 디지털화 과정에서의 높은 정확성 확보를 위하여 생산 연도별, 기관별, 부서별 수집 기준의 범위를 유연하게 조합하는 과정이 필요하다.

본 연구에서는 비전자기록물의 디지털 전환을 위한 단순 문자 인식부터 구조인식, 문장변환, 요약정보 제공 등의 범위까지 확대하여 AI OCR 기술을 기록관리에 접목할 수 있는 방향과 한계점 개선의 방법을 고찰한다.

Ⅲ. 분야별 실증분석 및 활용 사례

3.1 공공행정문서 AI OCR 데이터 활용

공공기관의 공공행정문서는 특화된 문자 인식 AI 모델을 개발하여 학습용 데이터셋을 구축하였다. 국민생활과 밀접한 관계가 있는 지방자치단체에서 생산된 공공행정문서를 대상으로 진행하였으며, 건설, 경제, 농림수산, 복지, 문화, 환경의 6가지 범주를 설정하여 원시데이터를 수집하였다. 수집과정에서 카테고리 분류에 따른 파일링 작업을 수행하였고, 민감정보는 개인정보보호법과 정보공개에 관한 법률에 기반하여 비식별화하였다.

데이터 가공의 단계에서는 저작도구를 사용하여 이미지에 포함된 단어에 대하여 세그멘테이션 및 라벨링 텍스트 등록을 진행하였고, 이미지 품질검사, 세그멘테이션 정확도, 라벨링텍스트 정확도 등 3단계에 걸쳐 검증을 완료하였다. 검수된 데이터는 원천데이터&라벨링데이터 set으로 제공한다.

공공기관 OCR 기록은 문서 내의 문자 인식 검색 서비스로 실생활에서 확인할 수 있는 공공행정문서 내의 행정용어나 법률용어 등의 검색 편의를 위해 활용되고 있다. 행정용어가 많이 포함되어 있는 공공문서를 카메라로 촬영하여 문자를 인식하고 식별된 문자를 웹사이트에서 검색하는 서비스도 제공한다(표 2 참조).

3.2 교육 학습용 AI OCR 데이터 활용

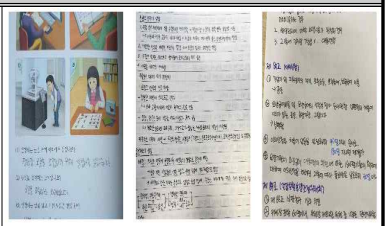
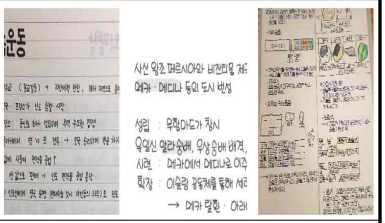
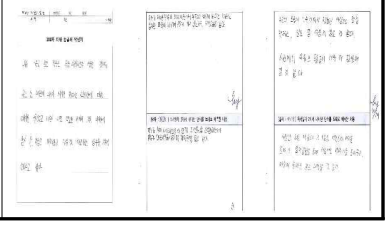
교육 자료 OCR은 수식, 도형, 기호가 포함된 손글씨 문자 인식 모델 개발로 데이터셋을 구축하여 모바일 교육분야 활성화를 위해 활용한다. 학교, 학원에서 학생들이 작성한 노트필기와 답안지를 온·오프라인 플랫폼을 동원하여 PNG형태로 획득한다.

불량, 중복, 혐오표현 등이 많이 포함된 경우를 제외하고 종이로 제출된 자료는 스캐너를 통해 먼저 디지털화하였다. 대상은 초등학교 1학년부터 고등학교 3학년까지 학년별 10,000건씩 총 12만건이며, 수집처는 학교(25%), 교육기관(50%), 자유공모(25%)로 한정하였다(표 3 참조).

〈표 2〉 공공행정문서 AI OCR 서비스 사례

Category	Content
Details of business	Building a Learning Dataset for Developing Character Recognition AI Model Specialized for Public Administrative Documents
Raw data	Collecting public administrative documents produced by local governments Six categories of raw data closely related to people's lives: construction, economy, agriculture, forestry, fisheries, welfare, culture, and environment
OCR Labeling	
OCR Labeling Inspection	
Examples of services	

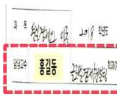
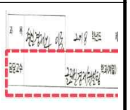
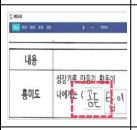
〈표 3〉 원시 형태의 교육 학습용 데이터 수집 과정

Raw data collection	
School	
Educational institution	
Free contest	

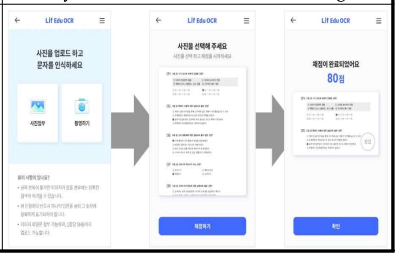
오프라인으로 수집된 자료의 경우 스캐닝을 통해 1차 디지털화를 진행하고, 교육 활동 과정에서 발생한 자료는 모바일 어플리케이션을 이용하여 촬영 후 클라우드 플랫폼으로 업로드하여 원시데이터로 사용하였다. 교육 학습용 데이터는 수집처별 카테고리 분류에 따라 파일링 하였고, 비식별화 도구로 민감정보를 제거한 후 저작도구를 활용하여 라벨링 파일을 생성하였다(표 4 참조).

가공된 교육용 OCR 데이터는 원천데이터와 라벨링데이터 세트 형식으로 적용하였으며, 원천데이터(PNG) 128,771장과 라벨링데이터(JSON) 3,102,747단어를 추출하였다. 구축된 데이터셋은 Training(80%), Validation (10%), Test(10%)로 구분하고 검증절차를 거쳐 유효성을 평가한 뒤 활용하였다. 답안지 채점은 이미지를 업로드하여 문항영역과 정답영역을 인식하고, 사전에 DB화된 정답지를 이용하여 답안 채점 및 결과 표시로 데이터화된 학습 정보를 제공한다.

〈표 4〉 교육 학습용 데이터 정제 과정

Category	Content		
Data collection process using application			
Sensitive information De-identification			
	Before de-identification	De-identification	after de-identification
Data processing (Labeling)			
	Before the Annotation	Register bounding box	Register Labeling Text

〈표 5〉 교육 학습용 AI OCR 서비스 사례

Category	Content
Details of business	Building datasets and activating mobile education by developing handwritten character recognition models that include formulas, figures, and symbols
Raw data	Students' notes and answer sheets written in schools, academies, and contests are classified by grade and collected online and offline De-identify personal information and sensitive information that can identify students and eliminate hate expressions Paper documents are digitized first through scanners, except when there are many defective and redundant targets
Examples of services	

3.3 문화·관광 다중언어 AI OCR 활용

문화, 관광 분야에서 다중언어로 병기되는 이미지를 다중언어 OCR 알고리즘을 통해 다국어 챗봇 서비스에 활용하였다. 간판, 메뉴판, 문화재 안내 팻말, 교통표지판, 동영상 콘텐츠 자막 등을 직접 촬영하거나 스틸컷으로 제작하여 원천데이터를 수집하였다. JPG, PNG, JPEG 등의 이미지 형태로 자료를 획득하였으며, 부산일보, 국제신문, 나단 프로덕션 등의 데이터 제공기관과의 협약을 통해 저작권을 보유한 동영상 번역 자막 이미지 스틸컷을 사용하였다. 병기된 언어별 데이터 수집량은 영어가 65%, 중국어가 20%, 일본어가 15%이며, 수집 전용 어플리케이션으로 데이터 수집 당시의 날씨, 조명, 객체의 색상과 형태 등의 특성 정보까지 함께 취득하였다.

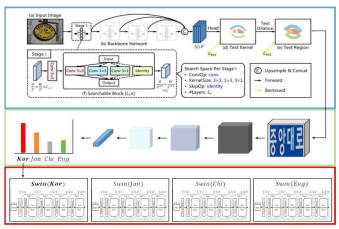
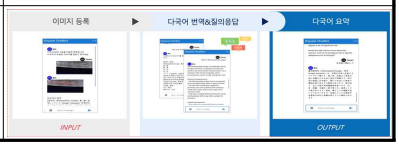
〈표 6〉 네이버 Clova OCR과 구글 Cloud OCR 비교

Category	Number of images	Composition ratio
Korean-English	252,568	63.01%
Korean-Chinese	88,110	21.98%
Korean-Japanese	60,186	15.01%
total	400,864	100%

다중언어 병기 수집 과정에서는 언어 전문 검수단이 3단계에 걸친 데이터 전수 검사를 진행하였다. 특성 정보를 통해 함께 획득된 내용을 바탕으로 이미지 품질, 세그멘테이션, 라벨링 텍스트 등이 정확하게 가공되었는지 확인하고, 정제된 데이터셋의 메타데이터 품질 검사를 통해 다국어 언어정보 영역의 세그멘테이션을 다시 수정하거나 생성하였다. 언어별 특성을 고려하여 음절단위, 어절단위 구분 태깅으로 다중 언어 번역의 품질을 강화하였고, 문자가 쓰여진 진행 방향에 따라 가독정보를 다듬었다. 수집된 다국어 병기 자료로는 외국인을 대상으로한 여행 및 관광분야 챗봇 서비스를 개발하였다. 챗봇 실행 후 다국어가 포함된 이미지를 업로드하면 OCR 기

술을 이용하여 내용을 인식하고 관련 질의응답 및 요약 정보 제공의 기능까지 가능하다.

〈표 7〉 다중언어 OCR의 문화 · 관광분야 서비스 사례

Category	Content
Details of business	Extend multilingual weapon images to multilingual services through multilingual OCR algorithms in culture and tourism
Raw data	Acquire raw data in the form of images with mobile apps and cloud platforms Direct filming of signs, menu boards, cultural heritage information signs, traffic signs, video content subtitles, etc. or still cuts
Learning Model Design	
Examples of services	

3.4 기록물관리전문요원과 IT전문가 FGI

3건의 프로젝트에 참여한 기록물관리전문요원 및 개발 전문가와의 포커스 그룹 인터뷰(FGI)를 진행하였다. 전문가 선정 기준으로는 AI OCR을 활용한 디지털기록관리 프로젝트에 참여한 경험이 있어야 함을 첫 번째로 하였으며, 두 번째는 기록물관리 분야 IT분야 각각의 그룹에서 기록관리와 AI OCR 기술 양쪽에 관한 이해도가 모두 높은 전문가를 2명씩을 선발하였다. FGI는 2024년 5월부터 7월까지 3개월에 걸쳐 진행하였으며, 이메일을 통해 사전 질문을 먼저 진행한 후 1:1 심층 대면 인터뷰를 통해 구체적인 의견을 주고받았다. 프로젝트 설계부터 결과물 활용의 과정까지 AI OCR 기반 기록물 디지털화 프로젝트 전단계를 경험한 전문가들의 의견을 담아 제언을 제시한다는 점에서 선행연구와 차별화하였다.

첫 번째, 기록물관리전문요원 박정빈 차장은 원본의 안전한 보존이 목적이었던 과거와는 달리, 데이터를 다루는 관점에서 기록관리가 이루어질 필요가 있다고 이야기하였다. 기록물 분류기준과 색인 값에 대한 일관성·연속성 유지가 어렵기에 '의미 기반 검색'으로 전환이 필요하며, 데이터 변환 기술의 도입도 고려해야 한다고 조언하였다.

"현재까지 진행되고 있는 기록물 디지털화 사업은 각종 지침과 메뉴얼이 존재하나, 검색 및 활용의 기준이 되는 기록물 분류기준과 색인 값에 대한 일관성 및 연속성 유지에 어려움이 있습니다. 따라서 생산된 데이터의 활용성을 극대화하기 위해서는 기존의 인덱스 기반, 텍스트 기반 검색에서 의미 기반으로 전환되어야 합니다. 생산된 데이터를 보존하는 것에만 만족하지 않고, 공공기록에서 정보를 찾아 데이터 기반 의사결정에 활용할 수 있도록 다양한 IT기술이 기록관리 체계에 도입되어야 할 것입니다."

두 번째, 기록물관리전문요원 차광주 부장은 공공기록물이 생성된 이후 공개 여부를 결정하는데 많은 인력과 시간이 소요됨을 어려움 중 하나로 언급하였다. 기록이 생산되는 시점에서 공개 여부를 결정할 수 있도록 기록의 전문 인식, 분류를 도와줄 기술적 요건과 관련 지식을 갖춘 전문요원이 도입되어야 한다는 의견이다.

"공공기록물은 생산과 동시에 공개여부를 결정하여 대국민 공개 서비스가 원활하게 이루어지도록 생산, 관리되어야 합니다. 하지만 과거에 생산된 기록물은 공개정보가 명확하지 않고 개인정보, 민감정보 등이 다수 포함되어 있을 가능성이 높아 가치 있고 중요한 정보라 하더라도 쉽게 공개하기 어려운 점이 많습니다. 실제 기록물의 공개여부를 결정하기 위해서는 해당 원본 문서를 육안으로 확인하고 검증한 후 공개여부를 결정해야 하므로 많은 인력과 시간이 소요되는 작업입니다. 공공기록

물의 정보공개 업무 효율화 및 대국민 공개서비스의 활성화를 위하여 전문 및 내용에 대한 인식 및 분류 기술이 빠르게 도입되어야 할 것입니다. 그리고 미국, 영국처럼 기록물관리전문요원의 전문성 제고를 위한 계속교육을 확대해야 합니다. 기록과 SW의 융합교육과 시대적인 기술 분석 교육 같은 다양한 계속교육을 계획하고, 기록물관리전문요원 중에서도 디지털기록물관리전문요원 직군을 편성하여 운영하는 것이 필요합니다."

세 번째, 쇼우테크 개발자 조지영 연구원에 의하여 공공기록물의 경우 한국어 받침 인식 부분에서 OCR 기술 적용의 한계가 있음을 확인할 수 있었다. 문서 형식의 다양성과 비표준 글꼴 인식의 문제도 발생하기에 한국어에 특화된 전처리 모델 조정이 요구된다.

"AI OCR 기술은 공공, 교육, 다국어 분야에서 다양한 도전 과제를 안고 있습니다. 공공 분야에서는 문서 형식의 다양성과 비표준 글꼴 인식 문제가 발생합니다. 교육 분야에서는 학생들의 필기체 인식을 어렵게 하는 요인이 있으며, 이를 해결하기 위해 다양한 데이터셋을 활용한 모델 학습이 필요합니다. 또한, 다국어 분야에서는 언어마다 서로 다른 문법과 구조가 문제로, 언어별 맞춤형 전처리와 모델 조정이 요구됩니다. 이를 통해 정확성과 효율성을 높일 수 있는 방안이 필요합니다. 그리고 한국어는 다국적 언어인 영어와 다르게 받침글자의 사용이 많아 인식의 어려움도 존재하지만 새로운 신조어의 출현이 빠른 편입니다. 새로운 세대가 나타날수록 줄임말을 이용하는 등 기존의 단어, 문장과는 다른 의미로의 사용이 높은 편입니다. 과거의 문자, 문장위주로 만들어진 문자 인식 기술 기반에서는 새로운 신조어(단어, 문장)가 나타나면 전혀 다른 의미로 해석될 수 있을 것입니다. 빠르고 다양하게 변화하는 한국어의 특성을 감안하여 신조어 등에 대한 지속적인 데이터 확보를 통하여 새로운 환경에 대응하는 문장 및 의미 인식 기술이 개발 및 발전되어야 합니다."

네 번째, 쇼우테크 개발팀장 황강연이사의 의견에 따르면, AI OCR 모델의 문자 탐지 영역에 따른 인식 텍스트를 문장 기반으로 조정하여 자연스러운 문장으로 변환하거나, 문장 기반의 인식을 통해 전문검색이 가능한 방향으로 기술의 발전을 도모하는 것이 기록관리 분야에 실질적으로 적합한 AI OCR의 적용이라 할 수 있다.

"공공기관에서 보존중인 중요기록물의 전문 검색을 위해 OCR 인식 서비스를 도입하는 과정에서 문장 기반의 인식이 꼭 필요함을 확인하였습니다. 문장기반 인식은 전문검색이 가능하게 하고, 요약정보를 제공 할 수도 있으며, 문서정보 내에서의 빅데이터를 활용하는 탐색 기능으로까지 확대할 수 있습니다. 공공기관의 각종 연구보고서, 기술문서의 검색, 활용등을 비롯하여 교육기관에서의 고문서 이중 보존과 바른 문장으로의 자동 문장 변환을 통한 OCR기반 검색 서비스 구현, 요약정보 제공 등은 업무 리소스를 최소화 하면서도 기록의 활용성을 확대하는 중요한 요소입니다. 키워드 검색에서의 한계를 극복하기 위한 방안으로 사용자의 질문을 이해하는 의미기반의 검색 시스템 개발이 기록관리업무에 도움이 될 것입니다."

Ⅳ. 제언 및 확장 방안 연구

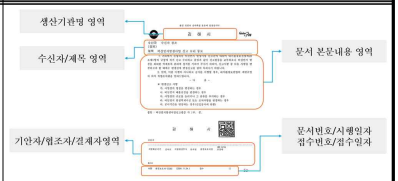
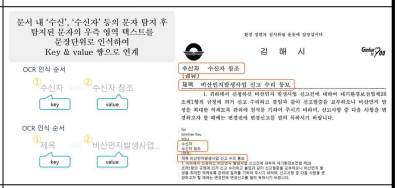
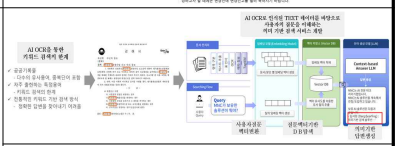
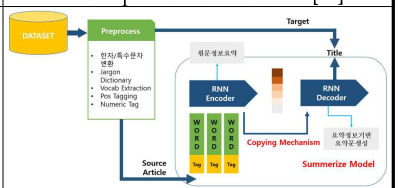
쇼우테크의 과제 수행 사례에 기반하여 기록관리 분야에서의 AI OCR 기술 활용 방안과 관련된 3가지 확장 방안을 제시하고자 한다. 기록관리자의 업무수행에 도움이 되면서도 기록과 기록관리의 본질적 의미를 해치지 않는 범위 내에서 발전 가능한 방향으로 모색해보고자 한다.

4.1 의미 기반 검색도구 개발 및 요약 제공

AI OCR을 활용하여 문서의 내용을 디지털화하는 과정에서는 본문의 기본 텍스트 데이터와 문서의 작성일, 저자, 주제와 같은 속성정보 그리고 문서의 형식, 구조

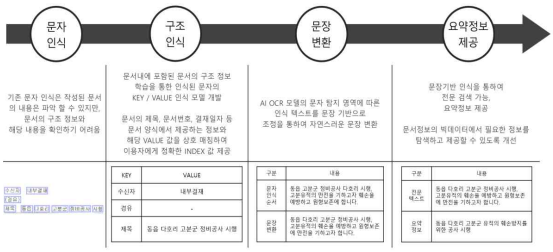
등의 형태 정보가 수집된다. 추출된 텍스트를 분석하여 주요 개념과 주제를 정의한 후 각 개념을 속성 및 값과 함께 분류할 수 있으며, 정의된 값과 개념간의 관계를 설정하면 온톨로지형의 구조가 형성되고, 이는 기록물 간의 관계를 명확하게 하여 정보의 맥락을 이해하는데 도움을 주기에 검색과 분석에 유용하다. 기록 내 정보의 맥락을 이해하여 검색하는 의미 기반의 검색은 정보와 정보를 유기적으로 연결하여 사용자가 원하는 정보의 pool을 확장한다. 키워드 중심의 검색에 비하여 기록물 내 정보탐색 과정에서의 정확성과 효율성을 높일 수 있다.

〈표 8〉 공공기록 관리 분야에서 AI OCR의 데이터 추출 프로세스

Category	Content
Recognition and Extraction	
Key & Value Association	
Semantic-based Search	 RAG Pipeline Reconstruction[13]
Summary Service	 Reconstructing the Document Summary Framework[14]

key와 value의 연계는 데이터의 일관성을 높이고 항목별 의미를 분명히 하는 작업으로, 새로운 개념이 추가되거나 기존의 개념이 변경될 때 key-value 쌍을 여러 상황에서 재사용하여 조정 가능하도록 온톨로지형 유연성을 형성하는데 사용한다. RAG(Retrieval-Augmented Generation) Pipeline은 광범위한 문서 집합에서 정보를 검색하고 검색내용을 활용하여 응답을 생성하는 방식으로, 지식 집약적인 작업환경에서 특정 분야의 지식 도입에 용이하다. LLM(Large Language Model)에 RAG를 적용할 경우, LLM은 사전에 학습한 지식 외에도 최신 정보나 특정 도메인에 대한 세부 정보를 보강하여 답변할 수 있으며, 사용자는 출처를 인용하여 답변의 정확성을 검증할 수 있다[15]. 온톨로지 형태의 데이터셋을 구축하고 RAG Pipeline을 OCR과 융합하여 지식기반의 응답이 가능하도록 기록관리 시스템을 설계한다면, 사용자가 기록을 활용할 때 참고할 수 있는 정보의 범위가 확장될 것이다.

문서의 요약정보는 <그림2>와 같이 key-value 매칭값에 index를 부여하고 문장 기반으로 내용을 탐지하여 추출한다. 문서 요약의 실질적인 역할을 담당하는 요약모델은 크게 원문의 정보를 요약하는 인코더와 요약한 정보를 바탕으로 요약문을 생성하는 디코더로 나누어지며, 각 부분에는 LSTM(Long Short-Term Memory)과 GRU(Grated Recurrent Unit)를 사용한 RNN(Recurrent Neural Network)셀을 적용할 수 있다. 디코더가 요약문을 생성하는 단계에 있어서는 인코더에서 각 토큰의 출력값을 변환시켜 확률화하고, 그 확률을 최대화할 수 있도록 복사 방법론의 아이디어를 차용하는 것을 제안한다(표8 참조).



〈그림 2〉 AI OCR의 서비스 개발 구조 프로세스 모형

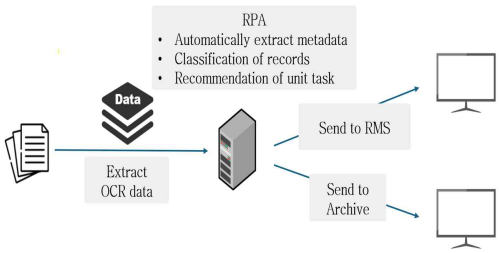
4.2 보안 및 컴플라이언스(compliance) 강화

디지털화된 문서를 저장하고 전송하는 과정에서 특정 사용자나 그룹별 접근 권한 레벨을 지정하거나, 변환하는 과정에서 암호화를 적용하면 데이터가 유출되더라도 인가되지 않은 사용자로부터 내용을 보호할 수 있다. AI OCR 시스템은 기록의 변경 이력이나 접근 이력을 자동으로 남기는 요구사항도 지원 가능하므로, 저장된 로그 기록을 실시간으로 확인하여 중요한 문서가 무분별하게 이용되는 현상을 방지할 수도 있다. 디지털 변환시 조건을 설정하고 자동으로 알림이나 경고를 보내도록 지정하면 관리자가 기록의 무단폐기, 유실, 훼손 등을 방지하기에도 용이하며, 정책 및 규정이 변경되었을 때 기존의 기록관리 형식에 변동내역을 재 적용하기에도 수월하다. 즉, 기록을 보호하고 활용 내역에 투명성을 제공하여 기록관리 과정에서의 책임소재를 명확화하고, 잘못된 사용 및 정보 유출 사고에 신속한 대응이 가능하다. 대부분의 컴플라이언스 요구사항을 충족하기 위해서는 특정 기간 동안 기록을 보존해야 한다. AI OCR을 통한 이력관리로 감사시 필요한 정보를 쉽게 추출할 수 있으며, 실시간 로그로 정기 점검 및 리포트 작성 등 규제 기관의 검토에 대비하기에도 효과적이다.

4.3 OCR과 RPA 통합 활용

OCR과 RPA(Robotic Process Automation)의 통합은 텍스트 추출, 데이터 전송, 정보의 분류 등의 작업 자동화에 실용적이다. OCR은 비정형 데이터를 디지털 형식으로 변환하는 것에 유리하고, RPA는 변환된 데이터를 반복적이고 규칙적인 업무에 반영하여 업무 처리에서의 효율을 높인다. 기록관리의 영역에서 두 기술의 통합은 비전자기록물로부터 데이터 추출과 처리 과정을 자동화하고, 다양한 형식의 문서에서 유의미한 정보를 신속하게 추출하여 종이기록의 디지털화 과정에 신뢰성을 부여하면서 결과물의 품질을 높인다.

OCR-RPA 통합 알고리즘은 다양한 시스템 및 애플리케이션을 통합하여 데이터의 흐름을 원활하게 한다. 예를들어, OCR로 추출한 정보를 이용하여 RPA로 문서구조에 따라 자동으로 생성된 메타데이터를 입력하고, 기록물의 내용적 분류 기능을 통하여 의미 기반의 기록물 분류를 수행한다. 분류된 기록물에 대한 단위과제, 단위업무코드 등의 추천 및 지정도 가능하며, 이를 기록관리 시스템으로 자동 전송하거나 특정 아카이브로 연계 전송할 수 있다(그림4 참조). RPA로 자동화된 OCR 알고리즘은 단순 반복 업무분야에서 사람의 개입을 최소화하여 휴먼에러(Human error) 발생을 낮추고 업무 처리의 일관성과 정확성을 향상시킨다. 즉, 기록관리 업무처리에서 과정은 단조로워지며, 소요 시간이 단축되고, 기록물관리 전문요원들의 업무 피로도는 감소하며, 업무 효율은 높아지는 효과가 있다. 다만, 기록물에서 text를 추출하고 전송하는 과정이 자동으로 이루어지는 만큼, 기록물 내 정보에 대한 유출이나 데이터 분류의 오류가 없도록 관리 감독 및 기술적 보완이 필요할 것이며, 공공기록물의 경우 관련 법령의 점검이 선행되어야 할 것이다.



〈그림 3〉 OCR-RPA 통합 알고리즘 적용 예시

Ⅴ. 요약

본 논문에서는 AI OCR로 기록의 디지털화를 수행하고 디지털화된 기록정보를 신규 서비스로 확장한 사례를

분석하여, 기록관리 분야에서 OCR 확대 적용의 강점과 향후 방향성을 제언하였다. OCR로 디지털화된 공공 기록이 대국민 서비스로 활용되고, 교육 기록이 교육의 품질향상을 위해 재사용되는 등의 사례를 통해 이미 기록관리 산업에서 OCR 기술 활용의 필요성은 확인되었다. 필요성에 대한 인식과 더불어, 앞으로의 디지털 기록관리 방향에 대한 탐구 과정에서 기록물관리전문요원의 역량과 역할에 대한 논의가 필요하다. 기록관리자는 프로그램의 설계·구현·검수의 단계에서 기술적 구축을 담당하는 IT전문가에게 기록학적 관점에서의 분류·선별·보존 프로세스를 제시할 수 있어야 하며, 시스템을 통한 기록물 재사용의 과정에서 어떤 방식으로 데이터를 제공하거나 주고받을 것인지와 관련된 의사결정을 내릴 수 있어야 한다. 기록학 전문성이 인문학 및 정보과학 기술 전문성 등의 분야와 협력·융합하고, 더 나아가 역동적으로 전개되는 4차 산업혁명 시대에서 다양한 형태의 데이터를 포착하여 기록의 생태계를 풍부하게 만들어야 한다[16].

실제로 NAA, SAA 등 해외의 계속교육 담당 기관에서는 기록물관리전문요원의 역량을 세분화하여 디지털 기록 관리에 특화된 심화교육을 제공하고 있다. 피교육자의 범위를 한정하거나 역량체계에 따라 내용의 난이도를 구분하는 등 교육을 제공하는 방법은 다양하나, 각 기관에서 제공하는 디지털 기록관리와 관련된 교육의 커리큘럼은 기록전문직 협회 차원에서 전문직 집단의 참여를 통한 개발과 관리의 필요성을 보여준다[17]. 역량별로 세분화된 교육과정을 '교육용 기록정보콘텐츠'로 관리하면 융합형 기록관리전문가 양성에서의 효율을 높일 수 있다[18]. 국내에서도 디지털 기록관리 산업 현장의 흐름에 맞추어 IT융합적 시각을 제공하는 구체적인 계속교육 및 융합교육을 계획하고 커리큘럼을 공개하는 등 실무자와 연구자들이 함께 성장할 수 있는 방안이 마련되기를 희망한다.

참고문헌

[1] 전자문서 및 전자거래 기본법(약칭: 전자문서법), 시행 2022. 10. 20, https://law.go.kr/법령/전자문서및전자거래기본법

[2] SAMILi.com, "기업 등의 전자문서 유통・보관 허용 추진," https://www.samili.com/tax/JaryosilOrganView.asp?div_cate=tax&v_seqno=25591&page=1743&v_part=0&searchword=

[3] Shoichi Taniguchi. "Duplicate bibliographic record detection with an OCR-converted source of information," Journal of Information Science, Vol. 39, No. 2, 2012, pp.153‐168.

[4] 강지홍, "소장기록물 특성을 고려한 OCR 인식 성능 개선방안 연구 보고서," 국가기록원, 2020.

[5] 안세진, 황현호, 임진희, "종이기록 데이터화를 위한 AI-OCR 적용 사례연구," 정보관리학회지, 제39권, 제3호, 2022, pp.165-193.

[6] 이준, 유민선, "OCR을 통한 비전자기록물의 전자화 연구: 딥 러닝을 중심으로," 디지털문화아카이브지, 제7권, 제1호, 2024, pp.259-276.

[7] Xujun Peng, Huaigu Cao, Srirangaraj Setlur, Venu Govindaraju and Prem Natarajan, "Multilingual OCR Research and Applications: An Overview," MOCR '13: Proceedings of the 4th International Workshop on Multilingual OCR, Washington DC, USA, No.1, 2013, pp.1-8.

[8] 오현호, 윤준석, 배유석, 김형일, "다중 언어로 학습된 이미지 기반 문자인식 모델의 성능 분석," 대한전자공학회 하계학술대회 논문집, 2023, pp.1058-1061.

[9] 이창엽, 김동주, 서영주, 황도경, "정밀한 다중언어 OCR을 위한 파이프라인," 한국정보기술학회, 제22권, 제8호, 2024, pp.1-7.

[10] upstage.ai, "성공적인 OCR 도입을 위한 최신안내

서," https://www.content.upstage.ai/whitepaper/ocr-adoption-guide-for-finance-industry.
[11] NAVER CLOUD PLATFORM, "Clova OCR," https://www.ncloud.com/product/aiService/ocr
[12] Google Cloud, "세계적 수준의 Google Cloud AI를 활용한 OCR(광학문자인식)," https://cloud.google.com/use-cases/ocr?hl=ko
[13] 마인즈앤컴퍼니, "AI검색솔루션: RAG=Vector Search +LLM 1편," https://blog.mnc.ai/74
[14] 송석민, "태그 정보와 복사 방법론을 활용한 수치 텍스트의 문서 요약," 석사학위논문, 서울대학교 공과대학 산업공학과, p.11.
[15] 최은실, "앙상블 기반 OCR과 RAG 기술을 활용한 ESG 서류 검토 자동화 시스템 개발," 한국컴퓨터정보학회논문지, 제29권, 제9호, 2024, pp.25-37.
[16] 노명환, "데이터 아카이브를 위한 성리학적 구성주의 레코드 컨티뉴엄 이론의 새로운 정립 모색: 지속적인 상호작용·변화·생성에 관한 존재론적 사상들에 기반하여," 기록과 정보·문화 연구, 제13호, 2021, pp.7-64.
[17] 전보배, 설문원, "해외 기록전문직 역량 체계 분석 및 활용 방안," 한국기록관리학회지, 제24권, 제1호, 2024, pp.137-161.
[18] 이선아, 오세종, "대학 실습 교육용 기록정보콘텐츠 가치 연구: 산학연계형 SW실습교육을 중심으로," 문화기술의 융합, 제10권, 제2호, 2024, pp.537-545.

■ 저자소개 ■

오 세 종
(Oh Sejong)

2025년~현재
한성대학교 AI응용학과 조교수
2020년-2024년
한국외국어대학교 AI교육원 교수
2020년
한양대학교 문화콘텐츠학 박사(문화기술)
2016년-2019년 슈피겐코리아
2014년-2016년 NHN AD
2010년-2014년 NHN Search Marketing
2010년 한양대학교 엔터테인먼트콘텐츠학 석사

관심분야 : 인공지능, 소셜미디어 빅데이터, 문화기술, 디지털 마케팅
E-mail : tbells@hansung.kr

이 선 아
(Lee Sun-A)

2021년~현재
한국외국어대학교 정보기록학과 박사과정

관심분야 : 디지털 기록관리, 온톨로지, 시멘틱웹, 데이터분석
E-mail : salee41@hufs.ac.kr

논문접수일 : 2025년 02월 18일
수정접수일 : 2025년 03월 09일
게재확정일 : 2025년 03월 15일