Journal of KIIT. Vol. 23, No. 10, pp. 55-64, Oct. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.10.55

원자력 품질증빙서류의 화학조성 데이터 추출을 위한 자동화 파이프라인

이상훈*, 문성빈**[1], 오영진**[2], 박상원***, 이해연****

Automated Pipeline for Chemical Composition Data Extraction in Quality Verification Documents of Nuclear Power Plants

Sang-Hoon Lee*, Seong-Bin Mun**[1], Young-Jin Oh**[2], Sang-Won Park***, and Hae-Yeoun Lee****

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다(No. RS-2022-KP002852)

요 약

과거 원자력 품질증빙서류의 화학조성 정보 추출은 페이지 분류, 테이블 탐색, 데이터 입력 전 과정이 수작업으로 이루어져 비효율이 컸다. 선행 연구에서 딥러닝을 통해 화학조성 테이블 영역까지 검출하며 정보 위치 탐색의 자동화를 달성했으나, 최종적 데이터 추출은 여전히 수작업으로 남아있었다. 본 연구는 이를 해결하기 위해, 검출된 테이블 이미지로부터 전후처리 과정을 추가한 OCR을 통해 정형 데이터를 추출하는 자동화 파이프라인을 제안함으로써 화학조성 정보 추출의 전체 워크플로우를 완성하였다. 제안하는 방법의 정보 추출 정확도는 90% 수준으로 일부 수작업 검토가 필요할 수 있으나, 화학조성 정보 추출의 전 과정을 연결하는 자동화된 파이프라인을 최초로 제안하고 그 실효성을 검증했다는 점에서 핵심적인 의의를 갖는다.

Abstract

The extraction of chemical composition data from nuclear Quality Verification Documents was inefficient, as processes including page classification, table localization, and data transcription were carried out manually. While a preceding study automated information localization by detecting table regions using deep learning, the final data extraction remained a manual bottleneck. To address this, the present study proposes an automated pipeline that completes the workflow of chemical composition extraction by applying OCR with additional pre- and post-processing to detected table images, thereby generating structured data. The proposed method achieved an extraction accuracy of 90%, which may still require partial manual review, but its significance lies in being the first to propose an automated pipeline that connects all stages of chemical composition extraction and verifies its effectiveness.

Keywords

nuclear QVD, OCR, data extraction, automated pipeline

* 국립금오공과대학교 디지털융합공학과 박사과정
한국전력기술(주) 디지털솔루션연구소
- ORCID: https://orcid.org/0009-0000-9828-1063
** 한국전력기술(주) 디지털솔루션연구소
- ORCID[1]: http://orcid.org/0009-0009-1048-2473
- ORCID[2]: http://orcid.org/0000-0001-7187-9997
*** 국립금오공과대학교 컴퓨터공학전공 학사과정
- ORCID: http://orcid.org/0009-0007-2039-7067
**** 국립금오공과대학교 컴퓨터소프트웨어공학과 교수(교신저자)
- ORCID: http://orcid.org/0000-0002-6081-1492

· Received: Sep. 22, 2025, Revised: Oct. 13, 2025, Accepted: Oct. 16, 2025
· Corresponding Author: Hae-Yeoun Lee
Dept. of Computer Software Engineering, Kumoh National Institute of Technology, Korea
Tel.: +82-54-458-7548, Email: haeyeoun.lee@kumoh.ac.kr

Ⅰ. 서 론

스캔된 비정형 문서로부터 구조화된 정보를 자동 추출하는 것은 산업 현장의 디지털 전환(DX, Digital Transformation)을 위한 핵심 과제이며, 광학 문자 인식(OCR, Optical Character Recognition)은 이를 위한 기반 기술이다[1]. OCR은 이미지 내의 문자를 인식하여 기계가 읽고 편집할 수 있는 디지털 텍스트로 변환하는 기술로, 현재 많은 분야에서 아날로그 기록 문서의 디지털화를 위해 주목받고 있다. 그러나 최신 OCR 기술은 인쇄체와 같이 정형화된 문서에서는 높은 인식률을 보이지만, 오래되고 복잡한 산업 문서가 가진 고유의 문제 앞에서는 여전히 명확한 한계를 보인다. 특히, 문서 위에 날인된 스탬프나 물리적 훼손, 불균일한 조명으로 인한 음영 노이즈, 불규칙한 패턴의 테이블선 등은 OCR의 문자 인식률을 저하시키는 주요 원인으로 작용한다[2].

이러한 문제는 원자력 발전 분야 문서들에서 더욱 두드러진다. 원자력발전소의 안정적인 장기 운영을 위해서는 배관감육(Pipe wall thinning) 현상에 대한 정밀한 모니터링이 필수적이며, 이러한 해석은 배관 컴포넌트의 정확한 화학조성 데이터에 크게 의존한다. 이 핵심 정보는 설계 및 건설 단계에서 방대하게 생산되는 문서들 중 하나인 품질증빙서류(QVD, Quality Verification Documents) 내 재료시험성적서(CMTR, Certified Material Test Report)에 기록되어 있다. 품질증빙서류는 재료시험성적서 외에도, 열처리 검사 성적서(Heat treatment inspection report), 용접 품질 기록(Welding quality records), 비파괴검사 기록(NDE report) 등을 포함하는 문서 집합체로, 가변적인 테이블 레이아웃, 다양한 폰트의 혼재와 같은 비정형적 특징과 스캔 과정에서 발생하는 이미지 열화 문제를 동시에 안고 있다.

최근 딥러닝 기반의 문서 이해 기술이 발전하면서, 방대한 문서에서 특정 정보 영역을 자동으로 검출하는 연구가 활발히 진행되고 있다[3]. 본 연구의 선행 연구에서도 이러한 접근법을 활용하여 2단계 파이프라인을 제안하였다[4]. 해당 파이프라인은 CNN(Convolutional Neural Network) 기반 분류기로 대규모 품질증빙서류에서 화학조성 데이터가 포함된 페이지만을 식별하고, 객체 탐지 모델인 YOLOv8을 이용해 해당 페이지 내에서 화학조성 테이블의 위치를 정밀하게 추출한다. 이로써 기존의 수작업 프로세스를 자동화하기 위한 기반을 마련했으나, 해당 연구는 '관심 테이블의 이미지 추출'까지만 다루었고 테이블 내 텍스트를 인식하고 데이터베이스화하는 과정까지는 나아가지 못하는 한계가 있었다.

본 논문은 선행 연구를 확장하여, 추출된 화학조성 테이블 이미지로부터 정형 데이터를 생성하는 자동화된 파이프라인을 제안한다. 제안하는 방법론은 세 단계로 구성된다. 첫째, 테이블 선을 OCR 과정에서 문자와 혼동을 일으키는 주요 방해 요소로 간주하고, 이를 형태학적(Morphological) 전처리를 통해 선제적으로 제거하여 OCR 친화적인 이미지를 생성한다. 화학조성 테이블의 선은 불규칙하게 배치되거나 스캔 품질 저하로 인해 일부가 단선된 형태로 존재하여 문자 인식을 크게 방해하기 때문이다. 둘째, 전처리된 이미지에 최신 OCR 엔진을 적용하여 텍스트를 인식한다. 셋째, OCR 결과의 문자별 바운딩 박스(Bounding box) 좌표를 활용하여 텍스트를 셀 단위로 재구성하고, 최종적으로 데이터베이스에 바로 활용할 수 있는 마크다운(Markdown) 테이블 형식으로 변환한다. 이러한 접근은 선행 연구에서 추출한 이미지로부터 단순히 텍스트를 인식하는 것을 넘어, 인식된 텍스트를 데이터베이스에 직접 연계할 수 있는 정형 데이터로 변환함으로써 원본 문서에서 최종 데이터 추출까지 이어지는 자동화된 파이프라인을 완성한다는 점에서 핵심적 의의를 갖는다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 분석하고, 3장에서 제안하는 프로세스를 기술한다. 4장에서 실험 및 평가 결과를 제시하며, 마지막으로, 5장에서 본 연구의 결론을 요약하고, 한계점과 향후 연구 방향을 논의한다.

Ⅱ. 관련 연구

2.1 OCR 성능 향상을 위한 접근법

현대의 OCR 기술은 딥러닝 기반으로 발전하며 높은 인식률을 달성했지만, 이는 주로 고품질의 정

형화된 문서를 대상으로 한다. 반면, 저품질 스캔 이미지나 비정형 레이아웃을 가진 문서에서는 성능이 급격히 저하되는 문제가 여전히 존재한다. 이러한 한계를 극복하기 위한 파이프라인 접근법은 크게 두 가지 방식을 고려해 볼 수 있다.

첫 번째로, 이미지 전처리(Pre-processing)를 통해 품질을 개선하는 것이다. 이미지 전처리는 OCR 인식률에 직접적 영향을 미치는 핵심 단계이다. 대표적으로, 불균일한 조명 문제는 그림자 제거 등의 기법으로 텍스트와 배경의 대비를 높여 해결하며, 문서 배경의 격자나 밑줄 같은 직선 요소는 허프 변환(Hough transform) 등을 이용해 제거한다[5][6]. 이러한 접근은 OCR 엔진이 문자가 아닌 요소를 문자로 오인하는 것을 방지하고, 실제 텍스트 영역을 더 정확하게 분리할 수 있도록 돕는다. 이는 복잡한 배경과 노이즈가 많은 산업 문서 환경에서 OCR 성능을 확보하기 위해서는 특화된 전처리 과정이 필수적임을 시사한다.

두 번째는 객체 탐지 기술로 필요한 정보 영역만 정밀하게 추출하여 OCR의 정확도와 효율을 동시에 높이는 전략이다. YOLO와 같은 객체 탐지 모델은 차량 번호판, 책 제목, 19세기 역사 문서에서 문단과 제목을 분할, 배관 도면의 심볼 및 텍스트 등을 먼저 탐지하고 해당 영역에만 OCR을 수행하는 데 널리 사용된다[7]-[10]. 이처럼 정보의 위치를 먼저 특정하는 접근법은 OCR 성능을 극대화하는 핵심 전략이며, 선행연구에서도 YOLOv8 모델을 적용하여 화학조성 테이블의 위치를 정밀하게 추출한 바 있다[4].

2.2 테이블 정보 추출 연구 및 기술적 한계

테이블에서 정보를 추출하는 것은 단순 텍스트 인식을 넘어, 셀 단위의 구조적 정보를 함께 이해해야 하는 복합적인 과제이다. 일반적으로 이러한 문제는 테이블의 구분선을 구조적 정보로 활용하여 셀을 먼저 정의하고, 분할된 셀 영역 안의 텍스트를 인식하는 접근법이 주를 이룬다. 예를 들어, 테이블 영역 검출 후 구분선을 기준으로 셀을 분리하고 CNN-RNN(Convolutional Neural Network - Recurrent Neural Network) 결합 모델로 텍스트를 인식하거나, GCN(Graph Convolutional Network)를 이용해 의미적으로 연결된 셀 영역을 찾는 연구들이 이에 해당한다[11][12]. 최근에는 테이블 탐지, 구조 인식, 내용 인식을 통합 처리하기 위해 DETR, CascadeTabNet, PP-OCR v2와 같은 최신 딥러닝 모델들을 단일 파이프라인으로 결합하는 End-to-End 접근법이 활발히 연구되고 있다[13].

그러나 이러한 최신 연구들조차 명확한 테이블 구조를 전제로 한다는 점에서, 본 연구가 해결하고자 하는 문제와는 근본적인 차이가 있다. 대부분의 연구는 PubTabNet과 같이 사전에 분석 대상 테이블 이미지만 선별하여 구성한 데이터셋을 활용하는데, 이는 현실의 산업 문서가 마주한 복합적 문제를 충분히 반영하지 못한다[14].

저품질 스캔 문서의 경우 단선, 불완전한 수직선, 불분명한 테이블 선 등이 OCR 인식에 직접적인 방해 요소로 작용한다. 기존 연구들이 테이블 선을 구조 파악을 위한 정보로 활용하는 것과 달리, 본 연구에서는 OCR 성능 향상을 위해 이를 제거 대상으로 보고 전처리 과정에서 제거하는 접근법을 사용한다. 이 과정에서 손실된 테이블 구조 정보는 인식된 텍스트의 위치 좌표 정보를 활용하여 재구성한다.

이와 같은 접근법은 산업 문서가 가진 다층적 노이즈(Multi-level noise) 문제를 단계적으로 해결하는 파이프라인의 일부로 볼 수 있다. 선행 연구에서 YOLOv8을 도입하여 문서 레벨의 노이즈, 즉 방대한 문서 더미 속에서 원하는 테이블을 탐색하는 문제를 해결한 데 이어, 본 연구는 그 후속 단계로 이미지 레벨의 노이즈 문제를 다룬다. 이를 통해 정보 탐색에서부터 데이터 추출에 이르는 전 과정을 유기적으로 연결함으로써, 실제 산업 문서 환경에 적용 가능한 실용적 파이프라인을 제안하고 그 유효성을 확인하였다.

III. 제안하는 파이프라인

본 연구는 선행 연구를 확장하여 원본 문서로부터 정형 데이터를 추출할 수 있는 자동화 파이프라인을 완성하는 데 목적이 있다. 그림 1과 같이 전체 파이프라인은 선행 연구에서 수행된 테이블 이미지 추출 단계와, 이번 단계에서 핵심적으로 다루는 데이터 추출 및 구조화 과정으로 구성된다.

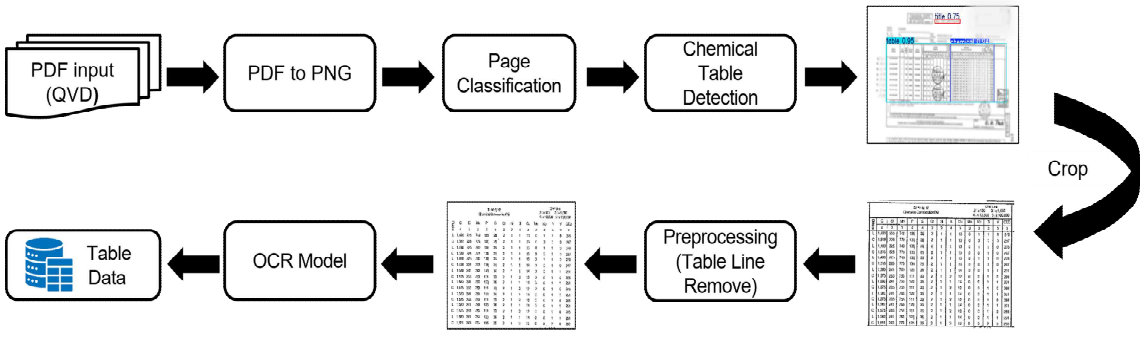

그림 1. 제안하는 파이프라인의 처리 흐름
Fig. 1. Process flow of the proposed pipeline

선행 연구가 PDF 형태의 품질증빙서류에서 목표 테이블 이미지를 성공적으로 추출했다면, 본 연구는 그 이미지를 입력받아 최종적인 정형 데이터로 변환하는 후속 처리 파이프라인을 설계하고 구현한다. 이 과정은 테이블 선 제거를 통한 이미지 전처리를 시작으로, OCR을 이용한 문자 인식을 거쳐, 마지막으로 추출된 텍스트를 구조화하는 순서로 진행된다.

3.1 테이블 선 제거를 통한 이미지 전처리

스캔된 재료시험성적서의 테이블 이미지는 스캔 품질 저하, 물리적 훼손 등으로 인해 OCR 성능을 저해하는 복합적 노이즈를 포함한다. 다양한 전처리 기법을 검토한 결과, 복잡한 노이즈 환경에서 가장 강건하고 일관되게 성능을 향상시키는 기법은 테이블 선 제거임을 확인하였다. 복잡한 노이즈 환경하에서 테이블 선은 OCR 엔진이 문자의 일부로 오인하도록 하는 핵심 방해 요소이기 때문이다.

테이블 선 제거는 다음과 같은 과정으로 수행된다. 먼저, 이진화 및 색상 반전으로 이미지를 정규화한다. 이후, 수평 및 수직 커널을 이용한 형태학적 연산으로 선 요소만을 정밀하게 분리해낸다. 최종적으로 원본 이미지에서 분리된 선 요소를 제거함으로써, OCR 엔진이 문자 인식에만 집중할 수 있는 최적의 환경을 조성한다.

3.2 OCR을 이용한 문자 인식

전처리가 완료된 이미지는 오픈소스 OCR 엔진을 통해 텍스트 정보로 변환된다. 본 연구에서는 원자력발전소 관련 문서의 보안상 제약으로 외부 API 기반 OCR 도구를 활용하기 어려워, 오픈소스 엔진을 중심으로 검토하였다. 재료시험성적서는 스캔 노이즈와 비정형 테이블 구조 등 복잡한 특성을 가지므로, 이러한 환경에서 성능이 제한적인 Tesseract와 같은 전통적 오픈소스 OCR 도구보다 딥러닝 기반 프레임워크가 더 강점을 보인다[15]. 이에, 산업계와 학계에서 널리 사용되는 최신 딥러닝 기반 OCR 프레임워크인 PaddleOCR과 docTR을 활용하여 인식 성능을 비교 분석하였다[16][17].

PaddleOCR은 중국 바이두(Baidu)가 개발한 PaddlePaddle 딥러닝 프레임워크를 기반으로 하는 오픈소스 OCR 툴킷이다. 경량화된 모델 구조를 통해 빠른 처리 속도를 제공하며, 한국어를 포함한 80개 이상의 다국어 인식을 지원하는 강점이 있다. 특히, 최신 버전에서는 저품질 이미지나 장면 텍스트 인식 성능을 지속적으로 개선해 왔으며, 본 연구의 대상과 같이 스캔 품질이 일정하지 않은 문서에 높은 적용성을 보인다.

docTR은 문서 AI 전문 기업인 Mindee에서 공개한 오픈소스 OCR 엔진으로, PyTorch와 TensorFlow 프레임워크를 모두 지원한다. 텍스트 영역을 찾는 탐지 단계에서는 DBNet, LinkNet 등 다양한 다양한 최신 아키텍처를 활용하고, 텍스트를 인식하는 단계에서는 CRNN, MASTER와 같은 강력한 모델을 선택적으로 조합할 수 있는 유연한 End-to-End 파이프

라인을 제공한다. 이는 복잡한 문서의 레이아웃을 분석하고 특정 문제에 최적화된 모델을 구성하는 데 강점을 가진다[18]-[21].

각 엔진을 통해 추출된 텍스트는 해당 위치를 나타내는 바운딩 박스 정보와 함께 출력되어, 다음 구조화 단계의 입력으로 사용된다.

3.3 추출 텍스트의 구조화

OCR 엔진을 통해 추출된 결과는 순서가 없는 단순 텍스트와 좌표의 리스트 형태이므로, 원본 테이블의 논리적인 행과 열 구조를 복원하기 위한 후처리 과정이 필수적이다. 본 연구에서는 바운딩 박스 좌표를 활용하여 다음과 같은 휴리스틱 기반의 정렬 및 구조화 알고리즘을 적용하였다.

먼저, 모든 텍스트 바운딩 박스를 y 좌표 기준으로 정렬한 뒤, 순차적으로 탐색하며 행을 군집화한다. 각 텍스트 박스의 중심 y 좌표와 높이를 고려하여, 수직 중심 거리와 높이 중첩 비율이라는 두 가지 임계값을 동시에 만족하는 경우에만 같은 행으로 그룹화하여 강건한 행 분리를 수행한다. 이렇게 군집화된 각 행의 텍스트들은 x 좌표 기준으로 좌에서 우로 다시 정렬된다.

다음으로, OCR 과정에서 오인식될 수 있는 노이즈 행을 필터링한다. 모든 행의 요소 개수를 파악하여 가장 빈번하게 나타나는 개수, 즉 최빈값을 기준 열 개수로 설정하고, 이 기준을 초과하는 행에서는 요소 간 간격이 비정상적으로 좁은 끼인 요소를 찾아 제거한다. 또한, 최소 행 길이 기준을 두어 의미 있는 데이터 행만 선택적으로 남긴다.

이후, 필터링된 행들의 모든 텍스트 중심 x 좌표를 수집하여 1차원 클러스터링을 통해 열 구조를 정의한다. x 좌표 간의 거리가 특정 임계값 이내인 것들을 같은 클러스터로 묶고, 각 클러스터의 평균 x 좌표를 해당 열의 대표 중심 좌표로 정의한다. 마지막으로, 이렇게 정의된 열 구조를 기준으로 각 행의 텍스트 요소들을 가장 가까운 열 중심에 매핑하여 2차원 테이블 구조를 완성한다.

최종적으로, 2차원 배열로 완벽히 구조화된 텍스트는 가독성과 데이터베이스 호환성이 우수한 마크다운 테이블 형식으로 변환된다. 이 과정을 통해 비정형 스캔 문서의 이미지로부터 데이터베이스에 즉시 활용 가능한 정형 데이터를 생성할 수 있다. 이처럼 다단계 휴리스틱에 기반한 접근은 완벽한 자동화에는 일부 한계가 있을 수 있으나, 테이블 형태의 이미지 문서에서 발생하는 다양한 예외 상황에 유연하게 대응할 수 있는 실용적인 방법이다. 그림 2는 화학조성 이미지 입력부터 마크다운 출력까지 진행 과정을 보여준다.

IV. 실험 및 결과

본 장에서는 제안하는 후단 파이프라인의 핵심 단계인 테이블 전처리 및 문자 인식의 성능을 정량적으로 평가한다. 이를 위해 먼저 성능 평가 지표를 정의하고, 실제 실험 결과를 통해 제안하는 방법론의 실효성을 분석 및 검증한다.

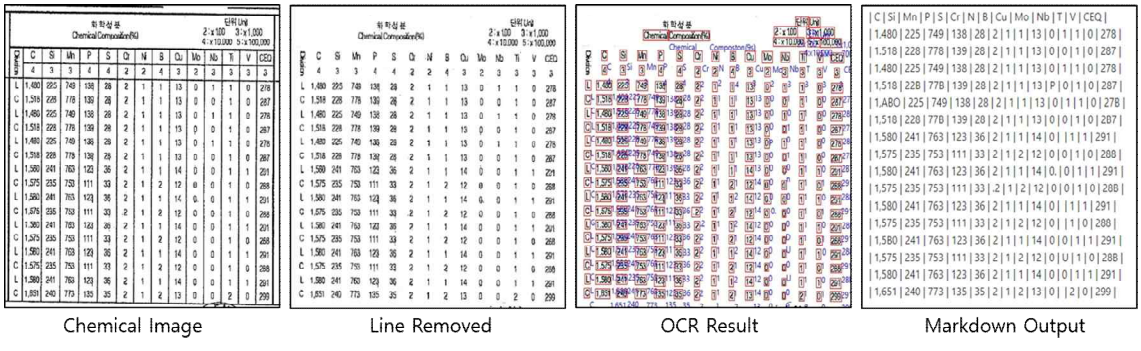

그림 2. 제안하는 파이프라인의 처리 예시
Fig. 2. Example of the proposed pipeline process

4.1 실험 데이터셋

본 연구의 성능 검증에는 실제 원자력 발전소 품질증빙서류 26장을 활용하였다. 전체 데이터셋은 총 4,678개의 셀과 8,703개의 문자를 포함하며, 이에 대한 정답 데이터(Ground truth)를 직접 구축하여 실험에 사용하였다. 해당 데이터셋은 다양한 문서 양식과 구조적 특성을 반영하고 있어, OCR 인식 성능 및 전처리 기법의 효과를 평가하기에 충분한 대표성을 가진다. 데이터셋의 규모는 제한적이지만, 실제 현업 문서에서 구축한 정답 데이터를 기반으로 평가함으로써 제안 기법의 효과를 현실적 조건에서 검증하였다.

4.2 성능평가지표

OCR 성능은 두 가지 관점에서 평가하였다. 첫째는 순수 텍스트 인식의 정확도를 측정하는 문자 정확도(Character accuracy)이며, 둘째는 테이블의 구조적 특성을 반영하여 실제 데이터의 유용성을 평가하는 셀 정확도(Cell accuracy)이다.

문자 정확도는 전체 문자 수 대비 올바르게 인식된 문자의 비율을 나타내며, 기본적인 OCR 성능을 평가하는 지표이다. 1에 가까울수록 개별 문자 인식 성능이 우수함을 의미하며, 식 (1)과 같이 계산된다.

$$character\ accuracy = \frac{correctly\ recognized\ characters}{total\ characters} \quad (1)$$

셀 정확도는 테이블 내 전체 셀 중에서 OCR 결과가 정답 텍스트와 완벽히 일치하는 셀의 비율을 나타낸다. 이 지표는 추출된 데이터가 별도의 후처리 없이 얼마나 실용적으로 사용될 수 있는지를 직접적으로 보여주는 핵심적인 척도이며, 식 (2)와 같이 계산된다.

$$cell\ accuracy = \frac{correctly\ recognized\ cells}{total\ cells} \quad (2)$$

4.3 End-to-End 접근법과의 비교 고찰

앞서 관련 연구에서 언급한 End-to-End 접근법은 문서 구조 분석과 텍스트 인식을 단일 파이프라인으로 처리하는 방식으로, 정제된 공개 데이터셋에서는 우수한 성능을 보여왔다. 반면, 본 연구는 스캔 품질 저하와 비정형 테이블 구조가 혼재된 산업 문서를 대상으로 하며, 맞춤형 데이터셋을 적용하여 객체 탐지 기반 테이블 검출, 이미지 전처리, 그리고 OCR을 단계적으로 결합한 실용적 파이프라인을 제안하였다. 이와 같이 두 접근법은 적용 대상과 전제 조건에서 차이가 크므로, 직접적인 성능 비교는 본 연구의 범위에 포함되지 않으며, 이에 따라 제안한 파이프라인 내에서 전처리 기법과 OCR 엔진의 성능을 검증하는 데 중점을 두었다.

4.4 실험 결과 및 분석

제안하는 테이블 선 제거 전처리 과정이 OCR 성능에 미치는 영향을 정량적으로 평가하기 위해, 선행 연구에서 추출된 화학조성 테이블 이미지를 대상으로 실험을 수행하였다. 성능 비교를 위해 PaddleOCR과 docTR 두 가지 공개 OCR 엔진을 사용하였으며, 성능은 문자 정확도와 셀 정확도 두 가지 지표로 측정하였다.

먼저, 표 1은 전처리 과정이 개별 문자의 인식 성능을 나타내는 문자 정확도의 변화를 보여준다. 전처리 적용 후 PaddleOCR은 0.92, docTR은 0.94의 문자 정확도를 기록하여, 기존 대비 각각 3.72%, 11.66%의 성능 향상을 보였다.

표 1. 선 제거 전후의 문자 정확도 비교
Table 1. Comparison of character accuracy before and after table line removal

OCR model	Without preprocessing	With preprocessing	Performance improvement
Paddle	0.89	0.92	3.72%
docTR	0.84	0.94	11.66%

더 나아가, 데이터의 실질적인 유용성을 평가하는 셀 정확도의 변화는 표 2에서 확인할 수 있다. 전처리 후 셀 정확도는 PaddleOCR에서 0.87, docTR에서 0.90로, 각각 7.96%, 24.54% 향상되었다. 한 글자만 틀려도 전체 셀이 실패로 처리되는 더 엄격한

셀 정확도 지표가 크게 향상되었다는 점은, 제안하는 전처리 기법이 데이터의 신뢰성과 실용성을 크게 개선했음을 시사한다.

표 2. 선 제거 전후의 셀 정확도 비교
Table 2. Comparison of cell accuracy before and after table line removal

OCR model	Without preprocessing	With preprocessing	Performance improvement
Paddle	0.80	0.87	7.96%
docTR	0.72	0.90	24.54%

이러한 성능 향상은 제안된 전처리 과정이 테이블 선으로 인해 발생하는 복합적 노이즈를 효과적으로 제어했기 때문이다. 테이블 선은 끊어진 일부가 특정 문자로 오인식되는 문제를 유발할 뿐만 아니라, 그림 3의 시각적 비교에서 확인할 수 있듯이 텍스트 영역 자체의 검출을 방해하기도 한다. 본 연구의 전처리 기법은 이러한 방해 요소를 제거함으로써 개별 문자의 오인식 가능성을 줄여 문자 정확도를 향상시켰다. 더 나아가, 누락될 수 있었던 셀 데이터의 정보 완결성을 보존함으로써 셀 정확도를 높여 데이터의 실용성을 강화하였다.

결론적으로, 이러한 결과는 복잡한 테이블 구조를 가진 산업 문서에서 정확한 텍스트를 추출하기 위해 제안된 전처리 과정이 효과적 접근법임을 입증한다. 최종적으로 추출된 텍스트는 후속 단계를 통해 바운딩 박스 정보를 활용하여 마크다운 형식의 정형 데이터로 변환된다.

V. 결론 및 향후 과제

본 연구는 선행 연구에서 제안된 정보 위치 탐색 단계를 확장하여, 추출된 테이블 이미지로부터 문자 정보를 인식하고 정형화하는 데이터 추출 파이프라인을 설계하고 검증하였다. 그 결과, 기존에 전적으로 수작업에 의존하던 데이터 추출 과정을 '페이지 분류 → 관심 영역 추출 → 이미지 전처리 → OCR → 데이터 구조화'의 자동화된 프로세스로 대체할 수 있는 가능성을 실험적으로 입증하였다.

형태학적 연산 기반 전처리 기법은 테이블 선과 같은 비문자적 요소를 효과적으로 제거하여 OCR 인식률을 개선하였으며, 그 결과 셀 정확도가 약 90% 수준으로 향상됨을 확인하였다. 이는 완전한 자동화를 의미하지는 않지만, 기존의 노동 집약적 수작업 입력 과정을 대체할 수 있는 실질적 가능성을 보여주었다.

다만 여전히 일부 셀 단위 오류가 존재하며, 이는 스캔 문서의 노이즈와 비정형성에서 비롯되는 OCR 기술의 본질적 한계로 분석된다. 현재의 OCR은 주로 픽셀 기반 처리에 의존하기 때문에 문서가 가진 구조적·의미론적 맥락을 완벽히 반영하지 못하는 한계가 있다.

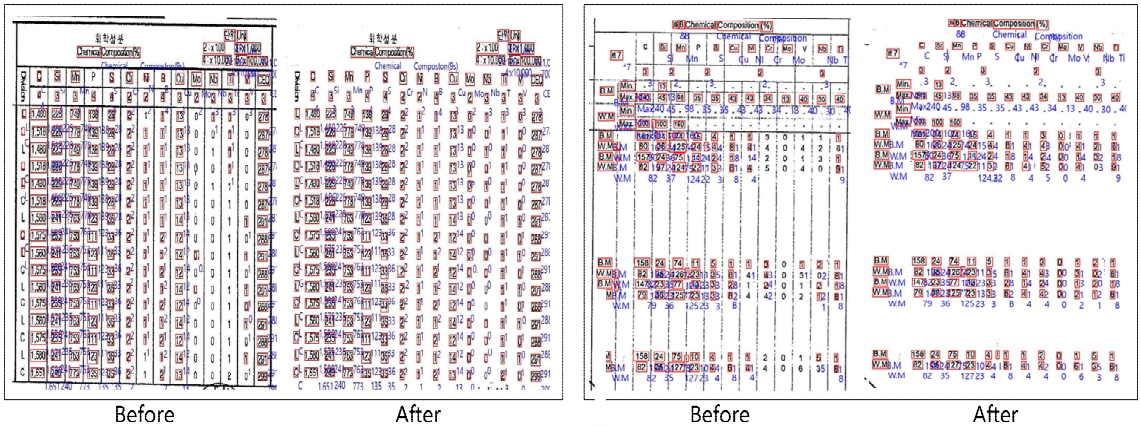

그림 3. 테이블 선 제거 전처리 전과 후의 OCR 결과 시각적 비교
Fig. 3. Visual comparison of OCR results before and after table line removal pre-processing

향후 연구에서는 이러한 한계를 극복하기 위해 멀티모달(Multimodal) 대규모 언어 모델(LLM, Large Language Model)의 도입을 계획하고 있다. 멀티모달 LLM은 이미지의 시각적 정보와 텍스트의 언어적 정보를 통합적으로 이해 및 추론할 수 있어, 저품질 문서에서도 테이블 구조와 문맥을 고려한 정확한 문자 인식 및 데이터 구조화가 가능하다. 이를 본 파이프라인과 결합함으로써, 원자력 분야에서 데이터 기반 안전 관리 체계를 뒷받침하는 완전 자동화 문서 처리 솔루션을 완성하는 것을 목표로 한다.

References

[1] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, and J. Zelek, "Text Detection and Recognition in the Wild: A Review", arXiv preprint arXiv:2006.04305, Jun. 2020. https://doi.org/10.48550/arXiv.2006.04305.

[2] G. Min, A. Lee, K. S. Kim, J. E. Kim, H. S. Kang, and G. H. Lee, "Recent Trends in Deep Learning-Based Optical Character Recognition", Electronics and Telecommunications Trends, Vol. 37, No. 5, pp. 22-32, Oct. 2022. https://doi.org/10.22648/ETRI.2022.J.370503.

[3] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, "Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions", IEEE Access, Vol. 9, pp. 72894-72936, Apr. 2021. https://doi.org/10.1109/ACCESS.2021.3072900.

[4] S. H. Lee, S. B. Moon, Y. J. Oh, S. W. Park, and H. Y. Lee, "Deep Learning-based Automated Detection of Chemical Composition Tables in Design and Construction Documents of Nuclear Power Plants", Journal of KIIT, Vol. 23, No. 9, pp. 75-83, Sep. 2025. https://doi.org/10.14801/jkiit.2025.23.9.75.

[5] I. Christian and G. P. Kusuma, "Improving OCR Performance on Low-Quality Image Using Pre-processing and Post-processing Methods", International Journal of Engineering Trends and Technology, Vol. 71, No. 6, pp. 396-405, Jun. 2023. https://doi.org/10.14445/22315381/IJETT-V71I6P239.

[6] K. Jaiswal, A. Suneja, A. Kumar, A. Ladha, and N. Mishra, "Preprocessing Low Quality Handwritten Documents for OCR Models", International Journal for Research in Applied Science & Engineering Technology, Vol. 11, No. 4, pp. 2980-2985, Apr. 2023. https://doi.org/10.22214/ijraset.2023.50664.

[7] J. J. Kim and C. B. Kim, "Implementation of Robust License Plate Recognition System using YOLO and CNN", Journal of KIIT, Vol. 19, No. 4, pp. 1-9, Apr. 2021. https://doi.org/10.14801/jkiit.2021.19.4.1.

[8] S. Y. Kim, J. W. Park, S. M. Kim, Y. Na, and Y. J. Jang, "Multi-book Label Detection Model using Object Detection and OCR", Journal of KIIT, Vol. 21, No. 2, pp. 1-8, Feb. 2023. https://doi.org/10.14801/jkiit.2023.21.2.1.

[9] D. Fleischhacker, R. Kern, and W. Göderle, "Enhancing OCR in historical documents with complex layouts through machine learning", International Journal on Digital Libraries, Vol. 26, No. 1, pp. 3, Feb. 2025. https://doi.org/10.1007/s00799-025-00413-z.

[10] S. B. Lee and H. Y. Lee, "Deep Learning-Based Object Recognition and Design Information Extraction Method from Piping Isometric Drawings", Korean Journal of Computational Design and Engineering, Vol. 29, No. 4, pp. 400-408, Dec. 2024. https://doi.org/10.7315/CDE.2024.400.

[11] D. S. Lee and S. K. Kwon, "Methods of Classification and Character Recognition for Table Items through Deep Learning", Journal of Korea Multimedia Society, Vol. 24, No. 5, pp. 651-658, May 2021. https://doi.org/10.9717/kmms.2020.24.5.651.

[12] S. Y. Yoo, S. H. Moon, and J. W. Kim, "Computer Vision-based Cell Recognition for Digitization of Equipment Inspection Documents",

Proc. of the 2024 KORMS/KIIE/KSS Spring Joint Conference, Yeosu, Korea, pp. 1-8, May 2024.

[13] A. Anand, R. Jaiswal, P. Bhuyan, M. Gupta, S. Bangar, M. M. Imam, R. R. Shah, and S. Satoh, "TC-OCR: TableCraft OCR for Efficient Detection & Recognition of Table Structure & Content", Proc. of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval (MMIR '23), New York, United States, pp. 11-18, Oct. 2023. https://doi.org/10.1145/3606040.3617444.

[14] X. Zhong, E. ShafieiBavani, and A. J. Yepes, "Image-based table recognition: data, model, and evaluation", Proc. of the European Conference on Computer Vision (ECCV 2020), Glasgow, UK, Vol. 12366, pp. 591-607, Nov. 2020. https://doi.org/10.1007/978-3-030-58589-1_34.

[15] R. Smith, "An Overview of the Tesseract OCR Engine", Proc. of the 9th International Conference on Document Analysis and Recognition (ICDAR), Curitiba, Brazil, pp. 629-633, Sep. 2007. https://doi.org/10.1109/ICDAR.2007.4376991.

[16] C. Li, W. Liu, R. Guo, X. Yin, K. Jiang, Y. Du, Y. Du, L. Zhu, B. Lai, X. Hu, D. Yu, and Y. Ma, "PP-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System", arXiv preprint arXiv:2206.03001, Jun. 2022. https://doi.org/10.48550/arXiv.2206.03001.

[17] Mindee, "docTR: Document Text Recognition", GitHub Repository, https://github.com/mindee/doctr. [accessed: Sep. 01, 2025)]

[18] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-Time Scene Text Detection with Differentiable Binarization", Proc. of the AAAI Conference on Artificial Intelligence, New York, USA, Vol. 34, No. 7, pp. 11475-11482, Apr. 2020. https://doi.org/10.1609/aaai.v34i07.6812.

[19] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation", Proc. of the 2017 IEEE Visual Communications and Image Processing, St. Petersburg, FL, USA, Dec. 2017. https://doi.org/10.1109/VCIP.2017.8305148.

[20] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 11, pp. 2298-2304, Nov. 2017. https://doi.org/10.1109/TPAMI.2016.2646371.

[21] N. Lu, W. Yu, X. Qi, et al., "MASTER: Multi-aspect non-local network for scene text recognition", Pattern Recognition, Vol. 117, 107980, Sep. 2021. https://doi.org/10.1016/j.patcog.2021.107980.

저자소개

이 상 훈 (Sang-Hoon Lee)

2008년 2월 : 인하대학교 기계공학부(공학사)
2020년 8월 : 경북대학교 산업대학원 기계공학전공(공학석사)
2023년 3월 ~ 현재 : 국립금오공과대학교 디지털융합공학과 박사과정
2012년 11월 ~ 현재 : 한국전력기술(주) 디지털솔루션연구소
관심분야 : Computer Vision, Deep Learning

문 성 빈 (Seong-Bin Mun)

2014년 2월 : 부산대학교 기계공학부(공학사)
2024년 2월 : 국립금오공과대학교 디지털융합공학과(공학석사)
2013년 7월 ~ 현재 : 한국전력기술(주) 디지털솔루션연구소
관심분야 : Computer Vision, Deep Learning

오 영 진 (Young-Jin Oh)

1998년 2월 : 서울대학교
원자핵공학과(공학사)
2000년 2월 : 서울대학교
원자핵공학과(공학석사)
2006년 8월 : 서울대학교
원자핵공학과(공학박사)
2006년 12월 ~ 현재 :
한국전력기술(주) 디지털솔루션연구소 연구원
관심분야 : 설비열화 예측, 데이터기반 예측진단

박 상 원 (Sang-Won Park)

2020년 3월 ~ 현재 :
국립금오공과대학교
컴퓨터공학과 학사과정
관심분야 : Computer Vision,
LLMs, Data Mining

이 해 연 (Hae-Yeoun Lee)

1997년 2월 : 성균관대학교
정보공학과(학사)
1999년 2월 : KAIST 전산학과
(공학석사)
2006년 2월 : KAIST
전자전산학과(공학박사)
2008년 3월 ~ 현재 :
국립금오공과대학교 컴퓨터소프트웨어공학과 교수
관심분야 : Digital Forensics, Image Processing, IoT