국방품질 4.0 시대의 군수품 수리부속 품질관리를 위한 시험성적서(PDF) 데이터 관리 방안

박지환, 강도희
국방기술품질원 품질연구본부 지휘정찰센터
e-mail:pjh1400@dtaq.re.kr

Defense quality 4.0 management of munitions repair parts data extraction method for test report(PDF)

Ji-Hwan Park, Do-hee Kang
C4ISR Systems Center , Defense Agency for Technology and Quality(DTaQ)

요 약

본 논문에서는 국방기술품질원에서 정부 품질보증 활동 중인 무기체계 수리부속 품질관리의 현황과 문제점, 품질 4.0 시대에 맞게 정부 품질관리 활동 진화의 필요성을 제시하였다. 인공지능, 빅데이터에 기반을 둔 품질관리를 위해 기존 계약업체에서 제출한 수리부속 시험성적서를 육안으로 합치성을 확인하는 전통적인 품질관리에서 벗어나 기존 민간에서 활발하게 연구 중인 PDF 성적서로부터 유의미한 정보를 추출하고 해당 데이터를 축적하는 방법을 적용하기 위한 연구를 진행하였고 딥러닝 기반 광학 문자 인식 기술(OCR)을 적용하여 그 가능성을 살펴보았다.

1. 서론

4차산업혁명 시대에 진입하면서 머신러닝(ML), 인공지능(AI) 등의 기술은 기업의 품질 활동에 큰 영향을 미치고 있는데, 이 시대에 필요한 품질 개념으로서 품질 4.0(Quality 4.0)이 등장하게 되었다[1]. 국방 분야도 이를 따라가고 있으며 2022년 국방기술품질원에서 국방품질 종합학술대회를 개최하여 선행적/프로세스 중심의 관리를 통해 품질을 향상시키는 것, 인공지능과 빅데이터 등 4차 산업혁명 기술을 적극적으로 도입한 스마트 방산 체계를 구축하는 것 등의 방향을 제시하였다[2]. 따라서 국방품질 4.0의 핵심은 “디지털화” 이며 방위산업 분야의 품질 데이터 축척이 우선되어야 한다.

하지만 양산단계 무기체계 수리부속은 대부분 영세한 협력업체들이 정부와 계약을 체결하여 품질보증활동을 진행하고 있으며 제조품질에 큰 영향을 미치는 4M(사람(Man), 재료(Material), 설비(Machine), 방법(Method))이 취약하다. 특히 열악한 영세업체 특성상 잦은 인력의 유출, 품질 담당자의 고령화 등에 따라 업무의 연속성이 낮아 일관된 품질 확보가 어려운 상황이다. 또한 일반 민간에서 소품종 다량의 품질관리와는 달리 '20년 지상전술전자전장비 부품 구매계약 품목 단가 및 수량 조사결과에 따른 분류[3]와 같이 군수품 수리부속 계약 대부분이 다품종 소량 품목이기 때문에 통계적인 품질관리가 이루어지기 어려운 실정이다. 따라서 품질보증기관에서는 계약 내 품질관리의 개념을 넘어 계약 간 동일품목에 대한 품질 검사성적서의 데이터(통계량(표준, 표준편차), 샘플링 방법, 제조사, 측정자 등)를 축적하고 관리할 필요성이 있다.

본 연구에서는 광학 문자 인식 기술(OCR)을 이용하여 군수품 시험성적서 PDF(Portable Document Format) 문서에서 유의미한 정보를 추출하는 여러 라이브러리를 조사하고 관련 모델을 적용하고 변환하여 군수품 수리부속 성적서의 디지털화의 가능성을 연구하였다. 이는 향후 국방품질 4.0의 “디지털화” 연구에 기여할 것으로 판단된다.

2. 본론

2.1 관련 연구

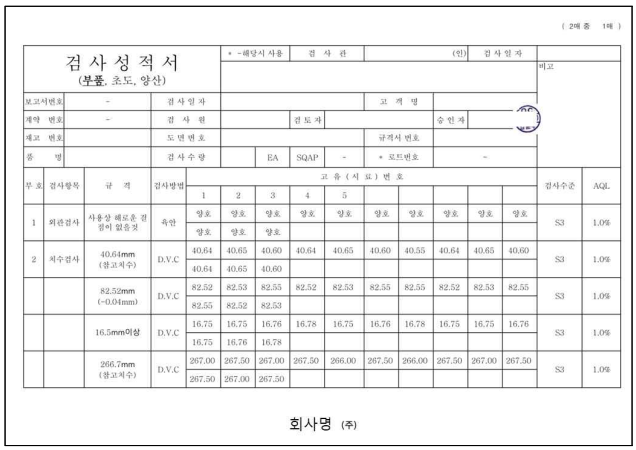
미국에서는 국방부 산하 국방계약관리본부(Defense Contract Management Agency; DCMA)에서 정부품질보증계획(Government Contract Quality Assurance Surveillance Planning)에 따라 데이터 수집과 분석을 실시하고 있으며 국내에서도 이를 벤치마킹하여 무기체계 별 데이터 수집 및 분석 방법의 차별성, 데이터가 자동으로 수집될 수 있는 시스템 구축의 필요성을 제시하고 있다[4].

민간에서는 방대한 학술 PDF 논문을 효율적으로 관리하

기 위해서 변환 도구를 활용하여 PDF 문헌에서 의미 정보를 추출하여 데이터 교환의 표준 형식으로 인정되어 실제 많은 시스템에서 이용하고 있는 XML 형식으로 제공하는 연구를 진행하였으며[5], PDF 논문을 세부적으로 검색하기 위해서 논문을 JSON 형식으로 저장하여 구조분석 기반의 PDF 논문 검색 시스템을 연구한 바 있다[6].

2.2 시험성적서 PDF 포맷

일반적으로 군수품 수리부속 시험성적서 PDF 양식은 규정으로 정해진 바는 없으나 계약업체마다 저마다 고유의 양식을 사용하고 있으며 대표적인 예는 [그림 1]과 같다. 문장이나 글보단 표로 한 눈에 알아볼 수 있게 작성되며 품목을 식별할 수 있는 재고번호, 품명, 규격번호, 계약업체, 검사원, 검사수량, 로트번호, 검사수준(샘플링), 검사항목, 검사방법 등으로 구성되어 있다. 계약업체마다 다른 양식으로 제출되고 있으며 High-level의 부품일수록 성적서 양이 방대해지고 기존의 성적서에 소급적용도 어렵기 때문에 상용 PDF 변환 프로그램을 사용하기 곤란한 문제가 있다. 또한 일반적으로 성적서를 스캔하여 PDF 형식으로 제출하기 때문에 서명과 같은 필기체, 불명확한 표 구분, 흐릿한 글씨체 등 정형화된 PDF 문서와는 달리 인식이 어려운 수준이다. 따라서 군수품 수리부속 시험성적서 고유의 시스템 모델이 필요한 시점이다.


검 사 성 적 서
(부품, 소도, 양산)

회사명 (주)

[그림 1] 군수품 수리부속 시험성적서 PDF 양식(예)

2.3 배경기술

PDF에서 이미지 형식의 표를 추출하는 것은 일반적으로 어려운 작업이지만 최근 사람이 쓰거나 인쇄한 문서, 촬영된 사진이나 스캔된 이미지 내의 문자를 인식하여 기계가 읽고 편집할 수 있는 디지털 텍스트로 변환하는 기술인 딥러닝 기반의 광학 문자 인식(OCR: Optical Character Recognition)의 기술 동향이 조사되었다[7]. 군수품 수리부속 성적서 PDF를 먼저 이미지 파일로 변환하여 OCR 라이브러리를 이용해서 다양한 양식의 PDF 성적서를 표준화된 데이터로 추출하는 것이 방법이 필요하다.

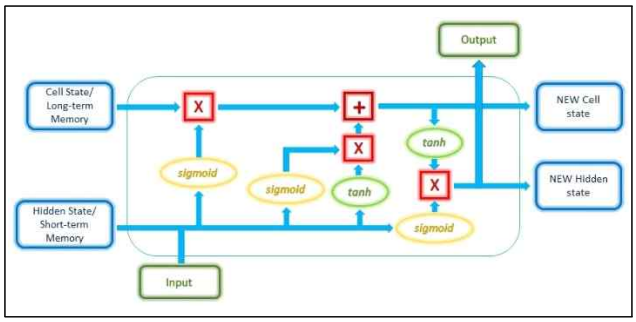
2.4 문자 인식 모델

문자 인식(Text Recognition) 모델은 검출된 문자가 어떤 글자인지를 판별 후 디지털 텍스트 포맷으로 변환하는 인공지능 모델로서, 개별 글자(Character)를 인식하는 방법과 단어(Word) 단위로 인식하는 방법이 있으며 문자 검출 방법에 따른 딥러닝 모델은 [표 1]과 같다[7].

[표 1] 문자 검출 방법에 따른 딥러닝 모델[7]

Method	Models
Bounding box regression	TextBoxes, TextBoxes++, DMPNet, SSTD, RRD, EAST, DeRPN
Part-based methods	SegLink, SegLink++
Segmentation-based methods	Mask TextSpotter, PSENet, TextSnake, Pixellink
Fast scene text detection method	TextBoxes, TextBoxes++, SegLink, RRD, EAST, DBNet, DENet++, CentripetalText(CT)

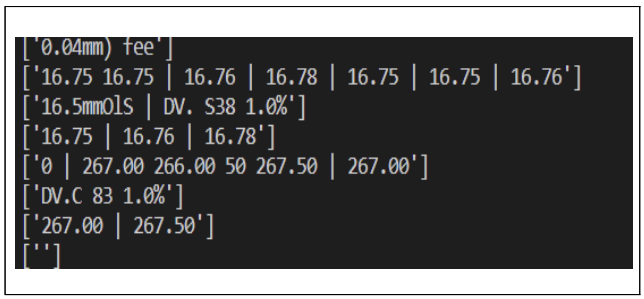
군수품 수리부속 품질관리를 위한 OCR 모델을 선택함에 있어서 먼저 군수품 수리부속 성적서에는 통계량을 표시하는 숫자와 높은 빈도로 쓰이는 정형화된 단어들이 있다. 예를 들어 앞서 말한 재고번호, 검사수준, 검사원, 검사 수량, 로트번호 등이 있으며 반복되는 단어가 사용된다. 따라서 개별 글자를 인식하는 모델을 사용하기보다 단어 단위의 인식 모델을 사용하는 것이 데이터를 집약적으로 관리할 수 있음을 알 수 있다. 본 연구에서는 딥러닝 OCR 라이브러리인 "pytesseract"[8]를 활용하여 성적서를 변환하였다. 해당 라이브러리는 반복적이고 순차적인 데이터 학습에 특화된 RNN(Recurrent Neural Network, 순환신경망)의 단점인 데이터가 길어질수록 앞서 받아들인 데이터 내용이 전달되지 못하는 장기 의존성(Long-term dependency) 문제를 개선한 방법인 LSTM(Long Shot Term Memory Network) 구조를 이용한 모델이며 시간 단위로 입력 노드를 통해 들어오는 데이터를 입력, 저장, 출력할 수 있도록 제어할 수 있으며 원리는 [그림 2]와 같다[9].



[그림 2] LSTM의 구조[9]

2.5 성능 분석

아래 [그림 2]는 [그림 1]의 성적서를 OCR을 이용하여 변환한 결과 일부이다. 결과를 통해 숫자, 영어, 특수문자는 비교적 잘 인식을 한다는 것을 알 수 있었다. 하지만 필기체나 불명확한 선, 특히 한글 인식률이 낮았다. 즉, 현재 OCR 기술은 정형화된 문자(영문), 숫자에 대해서는 높은 인식률을 제공하지만 필기체, 도장, 촬영된 문자 등에 대해서는 더욱 정교한 알고리즘이 필요하다는 것을 뜻한다. 국방품질 분야에서 향후 비슷한 숫자끼리의 평균, 편차 등의 알고리즘을 추가하여 데이터를 축적하는 방식으로 관리한다면 현재 군수품 수리부속 통계적 품질관리 어려움을 극복할 수 있을 것으로 판단된다. 또한 한글 인식율이 높은 OCR 기술이 개발되고 이를 발빠르게 적용한다면 검사원 등의 측정자 등에 따른 품질 위험 식별을 성적서 접수 단계에서 추가 식별할 수 있을 것을 판단되며 나아가 기술자료(도면) 등 과의 규격 일치성 판정까지 활용함으로써 업무 자동화에 도움이 될 것이다.



[그림 2] OCR(Optical Character Recognition) 라이브러리를 통한 성적서 변환 결과 일부

3. 결론

민간에서 다양한 OCR 기술이 개발되고 각종 산업, 공공분야, 개인정보 등록에 사용되고 있지만 국방, 특히 품질관리 분야에서 사용된 이력이 없기 때문에 본 연구에서는 기본적인 OCR 기술을 군수품 수리부속 시험성적서에 응용 가능하다는 것을 보임으로써 품질 4.0 시대에 맞는 군수품 수리부속 품질경영 발전 방향을 제안하였다. 국방품질 분야에서 쓰이는 단어나 형식은 오히려 민간보다 제한적이기 때문에 경량의 딥러닝 OCR 모델을 개발하고 적용한다면 업무 간소화에 큰 보탬이 될 것으로 기대된다.

참고문헌

[1] 서호진 외, “품질 4.0: 개념, 요소, 수준 평가와 전개 방향”, 한국품질경영학회논문지, 제 49 권 4호, pp. 447-466, 12월, 2021년.

[2] 허건영 외, “품질 4.0 시대의 국방품질경영 발전 방향 공유”, 2022 국방품질 종합학술대회, 9월, 2022년.

[3] 박지환, “군수품 구매계약 품질보증활동 효율화 방안 연구: ’20년 지상전술전자전장비 부품계약 중심으로”, 2022년 한국품질경영학회 춘계학술대회 포스터, p116, 5월, 2022년.

[4] 신병철 외, “양산단계 군수품에 대한 정부품질보증활동 실효성 향상 방안”, 한국품질경영학회논문지, 제 44 권 1호, pp. 153-166, 3월, 2016년.

[5] 박민규 외, “PDF 논문으로부터 의미 정보 추출”, 2012년 한국정보과학회 가을 학술발표논문집, 제 39 권 2호, pp. 106-108, 2012년.

[6] 김현준 외, “구조분석 기반의 PDF 논문 검색 시스템에 관한 연구”, 한국컴퓨터종합학술대회 논문집, 제 2020 권 7호, pp. 1780-1782, 2020년.

[7] 민기현 외, “딥러닝 기반 광학 문제 인식 기술 동향”, ETRI 전자통신동향분석, 제37권, 제5호, 10월, 2022년

[8] Https://pypi.org/project/pytesseract

[9] Https://blog.floydhub.com/ong-short-term-memory-from zero-to-hero-with-pytorch/