OCR을 이용한 AI기반 항만서류 자동인식에 관한 연구

-KNN 알고리즘 적용을 중심으로-

김종은
동덕여자대학교 경영학과(빅데이터 연계전공)
e-mail : jongeun0122@gmail.com

A Study on the Automatic Recognition of AI-based Port Documents Using OCR

- Based on the application of KNN algorithm-

Jong-Eun Kim
Dept. of Business Administration (with BigData),
Dongduk Womens University

요 약

우리나라의 수출입 화물 물량의 대부분은 항만을 통해 처리되고 있으며, 취급화물의 다양성과 선박의 대형화로 주변 국가들의 항만간 경쟁으로 심화되면서 항만 비용의 증가로 발생되고 있다. 이는 항만업무의 효율화와 생산성의 증가로 비용 감소효과를 바라볼 수 있는데, 4차 산업혁명의 주요 기술인 인공지능(OCR, AI알고리즘, 머신러닝, RPA등)의 기술 적용으로 개선할 수 있다. 본 연구에서는 이와 관련된 실질적 항만업무와 관련된 기술을 적용하여 업무의 효율화와 생산성 증가의 기술적 검증을 통해 항만의 경쟁력 강화와 국가 물류발전의 기술적 향상을 도모하고자 한다.

Ⅰ. 서론

우리나라의 수출입 화물량의 96.3%가 항만을 통하여 처리되고, 중국, 일본, 싱가폴, 홍콩 등 주변 국가의 항만간 경쟁에서 항만업무의 효율화와 생산성 확보를 위한 기술적 적용은 매우 필요한 시기이다. 특히, 국내 3위 규모의 최대 액체물류 처리항만인 '울산항만'은 유류 등 액체화물에 대한 항만시설과 관리, 행정적 처리에 많은 투자를 시행하고 있다. 그러나, 한국 컨테이너항만 생산성은 점차 하락하고 있으며, 국내 최대 항만인 부산항의 경우2018년 세계순위는 3위에서 5위로 하락하였으며 가장 큰 원인으로 항만 터미널들의 선석 생산성의 정체 또는 감소가 나타나고 있다. 이는 시설이 지속적인 물동량 증가를 따라가지 못해 생산성 향상의 한계와 컨테이너 물동량 공간의 부족으로 인한 하역능력 저하로 이어지고, 이는 선적 생산성 저하의 원인으로 작용하고 있으며, 울산항만도 예외가 아니다. 물동량증가에 따른 공간부족을 관점에서 바라보면, 로테르담 항, 함부르크 항 등 항만생산성이 높은 항만들은 이미 4차 산업혁명 주요기술을 적용한 '항만 스마트화 전략' 따라 스마트항만을 구현하고 개선하고 있다는 점이다.

따라서, 본 연구에서는 최근 인공지능(AI)을 활용한 연구를 컴퓨터 비전 분야 또는 머신 비전 시스템(Machine Vision System)의 개념을 도입하여 OCR기술적용, AI기반의 머신 비전 적용, KNN알고리즘을 적용한 항만서류 인식율 개선을 통해 항만업무의 효율화와 생산성 증가를 모색하고자 한다. 무엇보다도 여러 이미지의 집합 속에서 유의미한 특징(feature)을 추출해야 하는 자동인식 분야에서 인공지능은 처리 속도를 개선하고 모델을 생성해 항만서류의 직접 작성한 손 글씨 인식(MNIST), 차량 인식과 이미지 자동 인식을 머신러닝 알고리즘인 KNN(K-Nearest Neighbor) 알고리즘을 통해 스마트 항만 서류 자동인식에서 활용한 KNN 알고리즘에 대해 제시하고자 한다.

Ⅱ. 본론

1. 항만법

4차 산업혁명의 인공지능기술을 적용하기 위한 관련법인 '항만법'은 '법률 제16287호, 2019. 1'은 고시가 되어 있다. '항만물류'에 대한 정의는 항만에서 화물이 공급자로부터 수요자에게 전달될 때까지 이루어지는 운송·보관·하역 및 포장 등 일련의 처리과정을 말한다. 항만과 관련 내용을 정리하면, 첫째, '항만 물류 통합정보체계'란 해양수산부장관 및 항만을 이용하는 자가 항만물류비의 절감 및 각종 정보의 실시간 획득 등을 위하여 항만이용 및 항만물류의 과정에서 발생하는 정보를 정보통신망을 이용하여 상호교환·처리 및 활용하는 체계를 말한다. 둘째, '항만건설통합정보체계'란 해양수산부장관 및 항만건설 관련자가 신속한 행정업무처리와 비용 절감 등을 통하여 항만건설사업의 총체적인 효율을 높이기 위하여 항만건설의 계획·설계·계약·시공·유지 및 관리의 과정에서 발생하는 정보를 정보통신망을

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

이용하여 상호교환·처리 및 활용하는 체계를 말한다. 셋째, '항만운송사업 및 항만운송관련사업'은 항만운송사업, 항만하역사업, 검수사업, 감정사업, 검량사업, 항만운송 관련사업으로 구분할 수 있으며, 항만운송사업법, 항만운송사업법 시행령, 항만운송사업법, 항만운송업무 처리지침으로 적용되고 있다.

<표 1.> 항만운송사업내용

구분	내역
항만용역업	통선업, 급수업, 경비·줄잡이업, 청소업
선박연료공급업	선박용 연료를 공급하는 사
선박수리업	선체, 기관 등 선박시설 및 설비를 수리, 교체 또는 도색하는 사업
컨테이너수리업	컨테이너를 수리하는 사업
선용품공급업	선박(건조 중인 선박 및 해양구조물 등을 포함한다)에 음료, 식품, 소모품, 밧줄, 수리용 예비부분품 및 부속품, 집기, 그 밖에 이와 유사한 선용품을 공급 사업

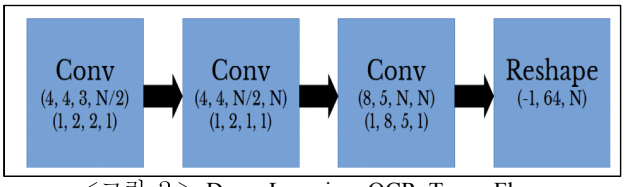
2. 머신 비전 시스템(Machine Vision System)

머신 비전 시스템(Machine Vision System)은 다양한 영상장비(예, 카메라 등)를 사용하여 획득한 영상을 처리하고, 항만관련에 의도하는 결과를 화면과 사용자에 맞게 나타내는 시스템이다. 영상장비, 프레임그래버, 어플리케이션 소프트웨어를 통해 영상장비를 통해 검사물체에 도달한 빛으로 렌즈인 영상장비로 전달되며, 영상장비에 입력된 빛은 전자 신호로 변환된다. 변환을 만든 전자 신호는 프로세서로 입력과 동시에 소프트웨어의 알고리즘에 따라 사전 정의한 프로세스로 결과를 보여준다.

<그림 1.> 항만 머신비젼시스템 구성절차

3. OCR((Optical character recognition; OCR)

관련 항만법령을 시행하기 위한 서류는 OCR기술을 적용한다. 광학 문자 인식(Optical character recognition; OCR)은 사람이 쓰거나 기계로 인쇄한 문자의 영상을 이미지 스캐너로 획득하여 기계가 읽을 수 있는 문자로 변환하는 기술이다. 복잡한 전산 및 수기로 작성하는 항만 관련 서류를 AI기반의 이미지 인식을 통해 자동인식을 하고자 한다. 이미지 스캔으로 얻을 수 있는 문서의 활자 영상을 컴퓨터가 편집 가능한 문자코드 등의 형식으로 변환하는 소프트웨어로써 일반적으로 OCR이라고 하며, OCR은 인공지능이나 기계 시각(machine vision)의 연구분야이다. 항만 관련 서류는 항만법 제30조(항만시설의 사용 및 사용료 등), 항만법시행령 제28조(항만시설 사용료의 종류와 요율 등), 항만공사법 시행령 제12조(항만시설의 사용 또는 임대), 항만공사법 시행규칙 제9조(항만시설의 사용 신청 등) 등 기준에 따라 다양하다. 관련 OCR이미지 기술을 통해서 관련 법규의 컴플라이언스를 제공하고, 오류와 법적 문제, Sanction List점검등을 자동으로 추출하고 검증한다. 해당 본 연구에서는OCR문자인식에서 가장 중요한 변수인 OCR Training을Training의 횟수와 OCR Training의 software인 Deep Learning OCR TensorFlow and Python을 사용하였다.



<그림 2.> Deep Learning OCR TensorFlow

4. 이미지 처리와 KNN(K-Nearest Neighbor)

KNN(K-Nearest Neighbor) 알고리즘은 데이터 내 벡터가 뿌려져있다고 가정할 때, 하나의 벡터를 기준으로 임의로 설정 거리 값 K에 해당하는 가까운 거리에 위치한 벡터끼리 묶는다. 이때 거리 계산 시 사용하는 수식으로는 유클리디언 거리 (Euclidean distance) 와 맨하탄 거리 (Manhattan Distance) 등 다양한 방법이 있다. KNN 알고리즘은 위와 같은 거리 공식을 반영해 묶여진 집합들을 바탕으로 모델을 생성하고, 새로운 벡터가 들어왔을 때 기존 모델을 기반으로 새로운 벡터를 분류 및 예측하는 매커니즘이다. 이러한 원리를 갖고 있는 KNN 알고리즘은 패턴 분류에 활용되는 가장 기초적인 통계 알고리즘이라 할 수 있다. KNN 알고리즘은 모델 생성 및 학습 단계에서 활용되며 전체 단계를 정리해보면, 다음과 같다. KNN 알고리즘은 주변 데이터 값에 기반하여 결정이 된다. 최적의 K 값을 선정하는 것은 데이터에 절대적인 의존적 영향을 미치며, K 값을 크게 줄 경우는 데이터의 구조적 특성을 파악하는데 어려워진다. 또한, 기존에 더 분류된 영역으로 처리되며 편향적 결과를 제시한다. 따라서 K의 값이 작은 경우 K=1 경우 이상치 데이터에 영향을 준다.

(a) Euclidean distance $$D(\underline{x},\underline{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

(b) Manhattan distance $$D(\underline{x},\underline{y}) = \sum_{i=1}^{n}|x_i - y_i|$$

(c) Minkowski distance $$D(\underline{x},\underline{y}) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p}$$

<그림 3.> KNN 유사도 측정방법

5. 이미지 처리과정

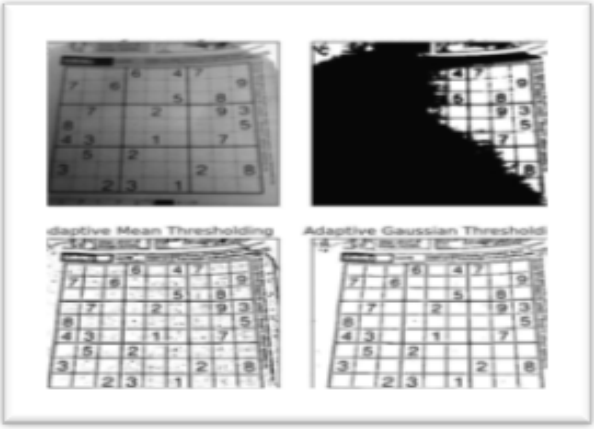
Step 1) & Step 2) 이미지 업로드 및 이진화 방법은 다음과 같다. 항만 서류 이미지는 차량 자동 인식, 얼굴 자동 인식 등과 달리 픽셀 값의 대비가 뚜렷하며 노이즈가 적다.

따라서, 사용하는 필터 또한 노이즈 제거보다는 윤곽선 검출(Edge Detection)에 효과적인 소벨필터(Sobel Filter), 케니필터(Canny Filter)와 임계치(Threshold)를 활용한 윤곽선 검출 필터인 적응임계처리(adaptive Threshold)를 활용해 윤곽선(Edge)을 [그림 4] [그림 5]와 같이 검출했다.

구분	처리프로세스
Step 1	항만 서류 텍스트 이미지를 온라인으로 업로드한 후 .jpg 혹은 .png 이미지 형태로 저장 및 배열화(change to matrix) 시킨다.
Step 2	저장한 이미지를 이진화(binary) 한다
Step 3	이진화한 이미지로 segmentation 및 분류 알고리즘을 만든다.
Step 4	KNN 알고리즘을 활용해 분류

<표 2.> KNN 알고리즘 처리프로세스

<그림 4.> adaptive threshold 예시



<그림5.> adaptive threshold 예시

6. Segmentation 특징 추출

이번 연구에서는 유클리디언 거리를 사용했다. 그 이유는 유클리디언 거리가 이미지의 너비(width) 및 길이(length)만 필요로 하며 거리 계산을 하기 때문이다.

$$d^2_{ij} = \min_{(x,y)\in B}((i-x)^2+(j-y)^2)), \text{ for all } i, j$$

<그림 6.> [수식] 유클리디언 거리

KNN 알고리즘을 활용하여 데이터셋를 학습시킨 후 나온 학습률은 약 90.2%로 좋은 성능을 제시하였다.

Segmentation 특징 추출을 위해 주성분분석(Principal Component Analysis;PCA)과 독립성분분석(Independent Component Analysis;ICA) 을 사용했으며 20,000 개의 한글 서류 데이터 셋을 학습시켰다.

Ⅲ. 결론

본 연구는 기존 텍스트 필터링을 포함하여 그 한계점을 항만 서류에 적용하기 위해 이미지 학습을 통한 필터링 방법을 제시하고자 한다. 초성, 중성, 종성이 결합된 한글의 형태로 한 글자를 여러 줄에 걸쳐 분리·작성과 충분한 의미 전달을 만들 수 있다.

또한, 본 연구는 언어 특성과 서류의 줄에 맞춰 한글의 단어를 변형하여 파이썬(Python)과 텐서플로우(TensorFlow)를 통해 변형시키는 알고리즘을 수정 개발하였다. 또한, 최종 결과를 이미지 파일로 저장하는 프로그램을 통해 인식의 개선을 도출하였다.

향후 연구과제로 서류와 Sanction list인 세계에서 각국의 경제제재를 가장 많이 주도하는 미국의 Sanction 미국 재무부 해외자산통제국의 금칙어 클래스를 많이 적용하게 되면 모델과 검색 성능이 낮아지고, 변형 알고리즘을 적용하는 성능상의 구별에 한계가 있다. 향후 이를 보완하기 위해 주기적인 학습 모델 적용과 업데이트가 필요하고, 클래스를 생성할 때, 금칙어 기준이되는 단어를 구분하는 필요성이 발생한다.

참고문헌

[1] 이동만, “항만시설사용료 감면제도가 울산항 활성화에 미치는 영향 분석.” 석사논문. (2019)

[2] 정경락, “KNN 알고리즘을 이용한 Fingerprint 기반의 UWB 실내측위.” 석사논문. (2019)

[3] 정현, “OCR을 이용한 라벨 문자인식 및 검사 시스템

의 구현에 관한 연구." 석사논문. (2019)
[4] 원치훈, "KNN에서 변수의 중요도에 관한 고찰." 석사논문. (2018)
[5] Smith, Ray. "An overview of the Tesseract OCR engine." Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). Vol. 2. IEEE, 2007.
[6] Deng, Zhenyun, et al. "Efficient kNN classification algorithm for big data." Neurocomputing 195 (2016): 143-148.
[7] Bellino, Gabriel Martin, et al. "Optimization of the KNN Supervised Classification Algorithm as a Support Tool for the Implantation of Deep Brain Stimulators in Patients with Parkinson's Disease." Entropy 21.4 (2019): 346.
[8] Kannan, R., and V. Vasanthi. "Logistic Regression and KNN Algorithm Experimental Diagnosis to Reduce the Impact of Cardiac Arrest." International Journal of Simulation--Systems, Science & Technology 20.1 (2019).
[9] Huang, Shuai, et al. "Analysis of factors influencing rockfall runout distance and prediction model based on an improved KNN algorithm." IEEE Access (2019).