韓國IT서비스學會誌
第21卷 第6號
2022年 12月, pp.39-51

Journal of Information Technology Services
https://doi.org/10.9716/KITS.2022.21.6.039

특허와 논문정보를 활용한 OCR 기술발전 동향예측에 관한 연구

김원준* · 이상곤** · 표성국***

A Study on the Prediction for the OCR Technology Development Trajectory based on the Patent and Article Information

Won Jun Kim* · Sang Kon Lee** · Sung Kuk Pyo***

■ Abstract ■

As the 4th Industrial Revolution emerged as a key to improving national competitiveness, OCR technology, one of the major technologies in the 4th industry is in the spotlight. Since characters in various images contain a lot of information, OCR technology for recognizing these characters has evolved into technology used in many industries. In this paper, trends in OCR technology were identified and predicted using thesis data published in 'RISS' and patent data by International patent classification (IPC) under the theme of Optical character recognition (OCR). For patent data 20,000 patents related to OCR technology from 2002 to 2020 were used as data, and 432 papers from 2012 to 2022 were used as data. Through time-series analysis, each patent data and thesis data were investigated since when OCR technology has developed, and various keyword analysis predicted which technology will be used in the future. Finally, the direction of future OCR technology development was presented through network association analysis with patent data and thesis data.

Keyword : OCR Technology, Patent Data, Thesis Data, Data Analysis

Submitted : November 8, 2022 1st Revision : November 15, 2022 Accepted : November 23, 2022

* 한국기술교육대학교 기술경영학과 박사과정
** 한국기술교육대학교 산업경영학과 교수, 교신저자
*** ㈜이투온 연구원

1. 서 론

서류 보관에 중요성이 커짐에 따라 문서 전자화시스템을 도입하는 산업이 증가하고 있다. 전통적인 문서처리 과정은 불편하고 비효율적이다. 디지털화되지 않은 문서로부터 데이터를 추출하는 과정에서 많은 산업분야에서 오류나 문제로 어려움을 겪는다. 하지만 인공지능의 발전으로 인해 수작업 위주의 입력 저장방식을 자동화 방식으로 변화시킨 4차 산업의 기술로 OCR이란 광학문자 인식기술이 개발되었다. 초기 OCR은 1928년 독일의 G.Taushek가 미리 준비된 몇 개의 표준패턴 문자와 입력문자를 광학적으로 비교하여 가장 표준패턴과 유사한 패턴을 대응시켜 입력문자를 인식하는 패턴매칭법[1]을 이용한 문자인식의 원리적 특허를 등록함으로써 시작되었다. 이후 많은 연구를 통해 AI 기술과 병합한 다양한 OCR 기술이 개발되었고 물류와 금융 산업을 필두로 여러 분야에서 상용화되었다. [그림 1]과 같이 기존의 OCR 기술로는 인식이 불가능했던 근대서적의 옛 문자나 복잡한 레이아웃에도 대응이 가능하다. 1860년대 이후에 제작된 서적, 잡지를 대상으로 실험결과 기존의 OCR보다 향상된 정밀도를 보였다.

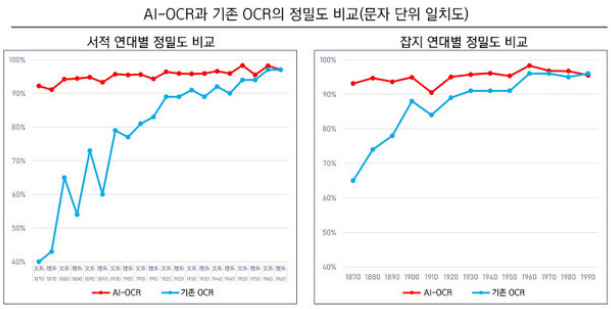
최근 코로나 사태로 인하여 비대면 업무처리가 많아지면서 각종 금융서비스, 보험, 제조, 인사, 공공기관, 의료, 교육 등의 서비스에서도 OCR 기술을 도입하였다.[2,3] 글로벌시장조사업체 그랜드뷰 리서치에 따르면 21년 OCR 산업시장은 연평균 13.7% 성장율을 보였으며[그림 4], 앞으로도 산업경제에 많은 영향을 미칠 것으로 예상하였다. 향후 OCR 기술은 많은 데이터의 학습을 통해 비용 절감과 업무의 효율성 증대, 인공지능 시장발전에 긍정적인 효과를 가지고 올 영역으로 판단하여 본 논문에서는 특허데이터와 논문데이터를 통한 OCR 기술 미래 동향을 예측하였다. 본 논문의 제2장에서는 OCR의 기술적, 산업적 배경을 설명하며, 제3장에서는 OCR의 기술적 분석을 진행하였다. 제4장에서는 특허데이터와 논문데이터를 활용하여 OCR 기술에 대한 각종 데이터 분석을 통해 미래의 OCR 기술산업 동향을 예측하였다.

2. 관련 연구

2.1 OCR 기술적 발전 배경

OCR기술의 실용화 연구는 1920년대 후반부터 시작되었다. 이후 많은 연구를 통해 다양한 OCR기술이 개발되었고, 그 결과 우편번호 추출을 통한 우편물관리, 자동차 번호판 인식[4], 명함인식[5] 등 다양한 산업분야로 적용범위가 확대되었다.

일반적인 OCR 기술은 문서 전체를 읽어 문자를 인식하는 방법으로 불필요한 영역까지 인식이 되는 단점이 존재하였다. 이런 경우 사람이 직접 수작업으로 분류해야 하는 불편함이 존재한다. 그러나 최근 IT 기술이 발달하면서 OCR 기술과 AI를 결합한



출처: 몰포AI솔루션즈.

[그림 1] OCR과 AI-OCR 기술차이점

AI-OCR 방법이 여러 산업방면에서 사용된다. 현재는 인공신경망인 CNN (Convolutional Neural Network) 기술[6]과 더불어 AI 학습데이터를 기반으로 이미지의 문자를 인식하는 규칙을 스스로 만들어 필요한 영역만 인식하는 기술로 발전되었다. 향후 AI-OCR은 많은 데이터의 학습을 통해 비용의 절감과 업무 효율성의 증대, 인공지능 시장 효과 증대 등의 산업경제의 긍정적인 효과를 가져올 것으로 예상된다.

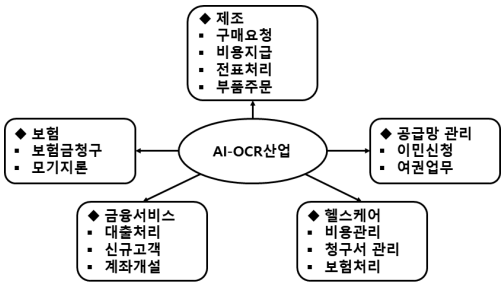
연도	발전과정
1928년	독일의 G.Taushek에 의해 pattern matching 기법으로 시작
1960년	60~760년대 초반 국외 영어 위주 개발 진행
1990년	1990년대 중반 스캐너 보급 및 컴퓨터 처리능력 향상으로 국내 본격 개발 진행
현재	AI 인공신경망을 결합하여 정확하고 정교한기술로 발전

[그림 2] 연도별 OCR 기술 발전과정

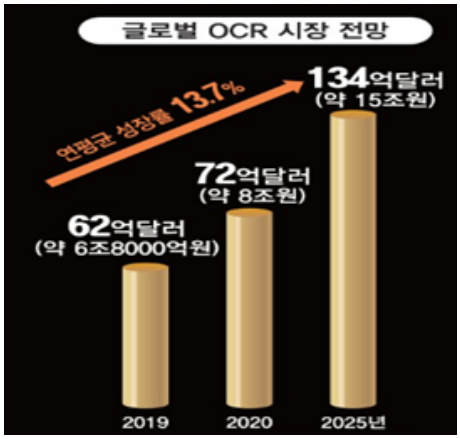
AI-OCR 문자인식 기술의 중요한 부분은 '인식률'이다. 사람의 수작업 과정 없이 AI 자동화 인식산업이 자리 잡으려면 인식률을 높여야 하는 과제를 해결해야 한다. 현재 대부분의 문자와 글꼴을 높은 수준의 정확도로 인식할 수 있도록 개발 되어졌다. 하지만 여전히 오류의 가능성이 존재한다. 조명이나 인쇄상태, 종이의 배경색이나 패턴 등의 다양한 방해요소로 인해 문자 인식률이 떨어지는 경우가 존재하여 산업분야에서는 사람의 수작업 과정이 포함 될 수밖에 없다. 이는 광학 기술의 발달만으로는 해결하기 어려워 소프트웨어의 발전이 동반되어야 한다. 광학 기술영역에서는 전처리 과정에서 불필요한 영역을 제거하고 문자 영역을 추출하는 기술 발전과 더불어 왜곡된 사례들의 학습데이터를 구축하여 학습시켜 문자인식 정확도를 높이는 방법이 연구가 진행되어야 한다.

2.2 OCR 산업적 발전 배경

현재 OCR 기술은 딥러닝과 접목하여 많은 산업분야에서 사용하고 있다. 각종 산업분야에서는 OCR 기술 도입으로 업무 프로세스를 더 간단하고 빠르고 효율적으로 만들었다. [그림 3]과 같이 금융서비스, 보험, 인사, 의료뿐만 아니라 법률서비스, 교육 분야에서도 OCR 기술을 사용하고 있다. 금융서비스는 OCR 자동화 기술을 처음으로 적용한 산업분야 중 하나였으며 컴퓨터가 계좌번호, 서명 및 달러 금액 차이를 읽고 정확하게 인식하여 빠르고 효율적인 서비스를 제공하였다. 보험 분야에서는 수많은 서류 작업을 처리한다. 보험 제안, 신규계정, 정책 갱신 및 청구처리에 모두 서류 작업이 필요하다. 이러한 문서를 수동으로 디지털화 하려면 인력과 비용이 많이 들어간다. OCR 기술로 자동화된 데이터 추출을 통해 정보를 '시스템'에 저장시켜 빠르고 간편하게 업무를 진행 시킬 수 있었다.



[그림 3] AI-OCR 산업 분야



출처: 그랜드뷰리서치.

[그림 4] 글로벌 OCR 시장 전망

2.3 특허데이터 및 논문데이터 활용

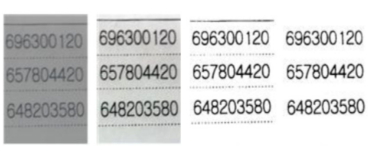
데이터 기반의 동향 분석은 과학기술 분야를 비롯해 다양한 분야에서 적용되고 있다. 기술 동향을 파악하고 예측하기 위해 특허데이터[7]와 논문데이터[8]의 시계열분석 및 키워드 분석, 네트워크 분석을 활용하여 기업 간 기술지식을 분석하고, 분석결과를 다양한 방안에서 활용하는 것이 중요하다고 판단된다. 이에 특허데이터와 논문데이터를 활용한 데이터 분석을 통해 OCR 기술 연구 발전 방향 및 미래산업 동향을 예측하는 연구를 진행하였다. 그러나 아직까지 특허와 논문데이터 분석을 통한 미래 동향예측에 있어 활용성 및 효용성 확대를 위해서는 더욱 다양한 사례에 대한 다양한 연구 분석이 필요할 것이다.

3. AI-OCR 분석

3.1 AI-OCR 기술적 분석

3.1.1 전처리(Pre-processing)

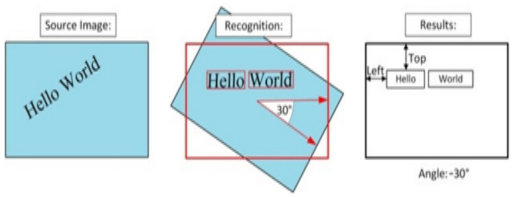
컴퓨터는 사람처럼 직관적으로 문자를 구분하지 못하고, 이미지에서 색상 분석을 통해 비슷한 밝기를 가진 픽셀들을 하나의 덩어리처럼 인식하기 때문에 색상 차이를 분명하게 하여 인식률을 높여야 효과적인 문자인식이 가능하다. 전 처리 단계는 이미지로부터 텍스트 영역을 컴퓨터가 보다 쉽게 인식할 수 있도록 이미지를 보정 하는 역할이다. 전 처리 과정에 사용되는 여러 기술 중에 대표적인 기술은, 컬러 이미지를 회색조로 변환하고, 픽셀 값을 분석하여 밝기와 명암 대비를 크게 변환 후, 픽셀 값을 두 범위로 나누어 0과1로 분류하는 이진화 작업을 수행하는 것이다. 이외에도 얼룩 제거, 라인 제거, 레이아웃 분석 등 여러 기술이 사용된다.



[그림 5] OCR 전처리 과정

3.1.2 문자검출(Text-detection)

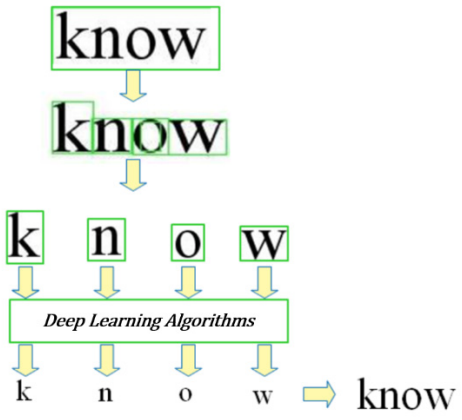
딥러닝 시스템을 이용하여 전체 이미지에서 텍스트 영역을 골라내는 작업을 수행하는 단계이다. 텍스트 영역과 그 영역의 회전 각도를 구하고, 텍스트를 수평 형태로 만든다. 텍스트 영역을 수평으로 만들어야 다음 단계에서 컴퓨터가 더욱 정확하게 문자를 인식할 수 있다.



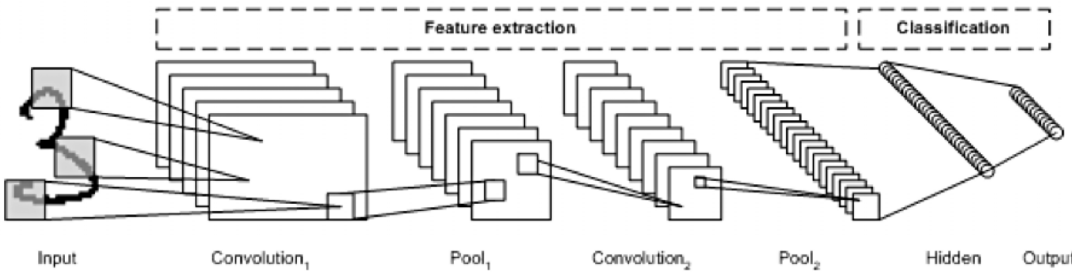
[그림 6] OCR 문자 검출 과정

3.1.3 문자인식(Text-recognition)

AI-OCR의 문자인식은 딥러닝 모델의 학습데이터를 기반으로 스스로 학습하는 인공신경망 CNN을 사용하여 인식하게 된다. CNN의 기본 개념은 필터를 행렬로 표현하여 필터의 각 요소가 데이터 처리에 적합하도록 자동으로 학습되게 하는 것이다. OCR 기술에 딥러닝 도입으로 인해 인식률이 증가하고 응답속도 또한 빨라져 업무 효율이 증가하였다. [그림 8]은 AI-OCR에서 가장 많이 사용되고 있는 CNN 알고리즘이다.



[그림 7] OCR 문자인식 과정



[그림 8] CNN 알고리즘 파이프라인

4. 연구 방법

4.1 분석 대상

본 연구에서는 OCR 기술 주제를 탐색하기 위해 한국에서 2002년부터 2020년까지 출원, 등록되어 특허정보넷 KIPRIS에 공개된 특허데이터 20,000여 건과 RISS에 등록된 OCR 관련 논문 432개의 데이터를 활용하여 분석을 진행하였다. OCR 기술이 본격적으로 사용되던 2002년을 기준으로 특허데이터를 수집, 논문은 게재가 시작되는 2012년도부터 수집하였다. KIPRIS는 특허청이 보유한 국내외 지식재산권 관련 모든 정보를 데이터베이스 기반으로 만들어진 대국민 특허정보검색 서비스이다. OCR 기술특허 검색 결과 중 국제특허분류 IPC 기준으로 광학문자인식과 관련 없는 C, D, E, F섹션을 제외한 A, B, G, H 섹션에 대한 OCR 기술을 분석대상으로 선정하였다. 각 데이터에 대한 노이즈를 제거하고 필요한 정보인 명칭, 초록, 연도, 키워드를 사용하여 시계열분석, 빈도 수 분석, 네트워크 연관성 분석을 진행하였다.

섹션	분야	등록 특허 개수	데이터 사용 여부
A 섹션	생활필수품	4,182개	사용
B 섹션	처리조작, 차량	1,846개	사용
C 섹션	화학, 야금	3,556개	미사용
D 섹션	섬유, 지류	0개	미사용
E 섹션	고정 구조물	87개	미사용
F 섹션	기계공학, 조명	355개	미사용
G 섹션	물리학	15,761개	사용
H 섹션	전기	10,492개	사용

[그림 9] 특허 섹션별 OCR기술

4.2 분석 방법

본 논문에서 [그림 10]과 같이 특허 및 논문데이터를 추출과 정제과정을 거친 후 각각 연도별 시계열분석과 빈도별 키워드 분석을 진행하였다. 그리고 두 데이터의 네트워크 연관성 분석 후 OCR 기술의 동향예측 단계로 진행하였다.

[그림 10] AI-OCR 연구분석

데이터 수집 단계는 OCR 분석 및 동향예측에 필요한 데이터를 수집 하는 과정으로 특허데이터는 2002년부터 2020년까지의 OCR 기술에 대한 특허를 수집하였으며, 논문데이터는 2012년부터 2021년까지의 OCR 기술 관련 논문을 수집하였다.

데이터 정제 단계는 데이터를 가공하는 단계로써 특허 부분에서는 필요한 데이터 항목만 남기고 노이즈를 제거하는 과정이 진행되었다. 특허데이터에서 사용된 항목은 '명칭', '출원연도', 'IPC'에서 추출한 데이터를 사용하였다. 논문 부분에서는 논문 초록과 키워드를 중심으로 텍스트 마이닝을 통해 중요 키워드를 추출하였다. 추출한 중요 키워드 중 빈도수가 높은 OCR와 같은 불필요한 데이터는 제외하여 분석을 진행하였다. 추출된 키워드별 노출빈도를 계산하는 프로그래밍을 통하여 특허와 논문자료 각각 빈

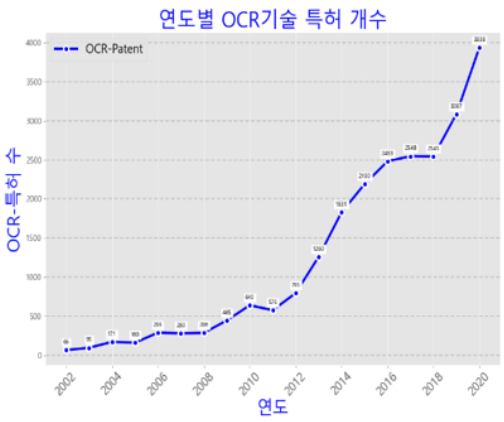
도수 상위 1~40위의 키워드를 추출하였다. 시계열 분석은 연도별 특허 등록 수와 논문게재 수를 분석하였으며, 각 연도별 빈도수 상위 1~20위 키워드를 도출하고, 해당 키워드가 어떤 시계열을 보이는지 확인하여 세부기술 분야별 거시적인 측면에서 동향을 살펴보기 위하여 사용하였다.

특허데이터와 논문데이터의 키워드 분석을 통해 얻은 데이터를 이용하여 네트워크 연관성 분석을 진행하였다. 연도별로 사용되는 단어의 빈도수를 분석하여 OCR 기술이 어느 방향성을 가지고 있는지 어떤 분야에서 많이 사용되고 있는지 기술 동향 예측을 진행하였다.

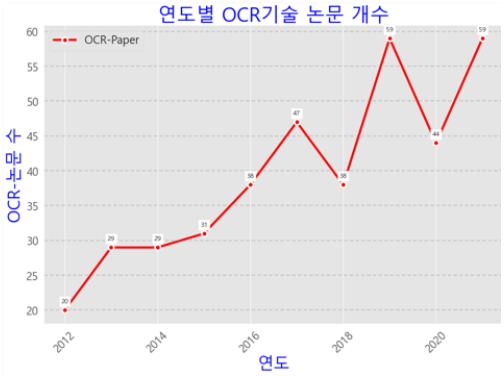
5. 분석 결과

5.1 시계열 분석 결과

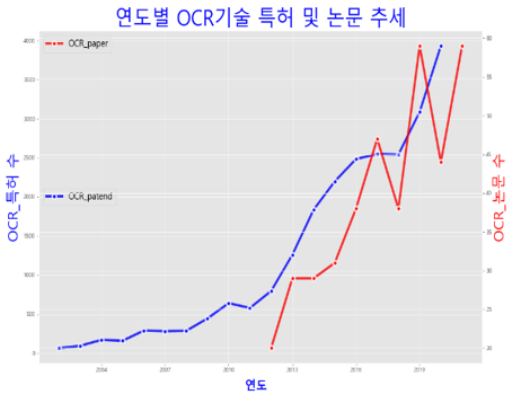
시계열분석은 2002년부터 2020년까지 등록된 OCR 기술특허와 2012년부터 2021년까지 게재된 논문을 연도별로 분석한 결과이다. [그림 11], [그림 12], [그림 13]은 특허와 논문 시계열분석 결과이다. 분석결과에 따르면 OCR 기술특허는 2012년을 기점으로 급격한 증가추세를 보이며 그 후 꾸준하게 증가 현상을 보이고 있다. OCR 특허가 급격하게 증가하는 2012년을 기점으로 OCR 기술관련 논문 또한 게재되기 시작하였다. 이는 IT기술이 발달하면서 각종 물류, 금융, 제조 업무에서 OCR 기술이 사용되는 영향을 받은 것으로 추측되었다. 또한 OCR 기술특허 및 논문이 꾸준히 증가하는 것으로 보아 앞으로의 OCR 기술은 더욱 더 발전된 기술로써 산업에 많은 영향을 끼칠 것으로 예측되었다.



[그림 11] OCR 특허데이터 시계열분석 결과



[그림 12] OCR 논문데이터 시계열분석 결과

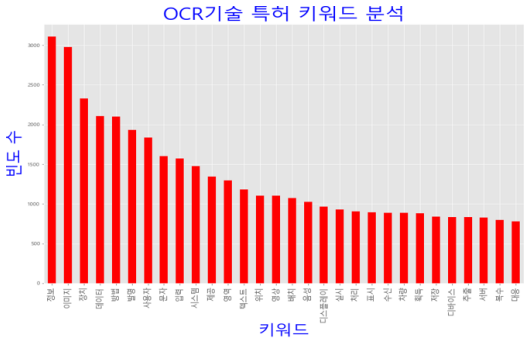


[그림 13] OCR 특허와 논문 추세

5.2 키워드 분석결과

키워드 분석은 특허데이터와 논문데이터에 사용되는 단어 빈도수를 파악하여 자주 사용되는 키워드를 분석한 결과이다. [그림 14]는 특허데이터의 2002년~2020년 OCR 기술특허 중 초록의 키워드 분석을 통해 추출한 단어 30개이다. OCR 기술특허

키워드 분석결과 정보, 이미지, 장치, 데이터 등의 단어를 포함한 특허가 상당 수 존재하였다. 텍스트와 디스플레이 영상 등의 키워드와 사용자, 장치 제공 등의 키워드가 존재하는 것으로 보아 OCR 기술 특허가 많은 사용자들에게 제공되는 서비스 기술로 존재한다는 것을 알 수 있었다.

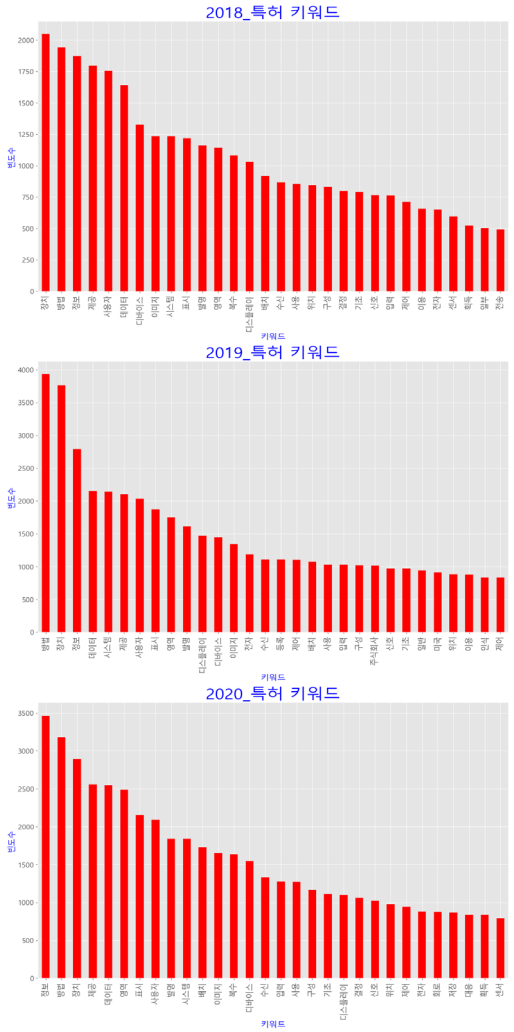


[그림 14] 특허데이터 초록 키워드 분석 결과 (2002년~2020년)

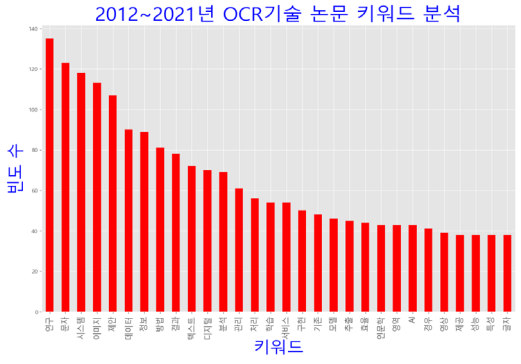
[그림 15]는 특허데이터를 연도별로 키워드 분석을 진행한 결과이다. 실험은 최근 데이터는 2018, 2019, 2020년도 특허데이터를 사용하였다.

실험결과 공통으로 추출된 키워드로써는 장치, 방법, 정보, 제공, 데이터 등이 상위권에 분포되어있었다. 이는 OCR 기술이 다양한 장치로 발명되어 사용자들에게 정보를 제공하며, 데이터를 획득할 수 있는 기술로 발전된 것을 의미하였다. 또한 이미지에서만 사용되던 OCR 기술이 디스플레이, 신호, 음성 등의 영역에서 사용되어가는 것을 볼 수 있었다. 이것은 단순하게 문서에서 문자만 인식하던 OCR 기술이 발전되면서 문서뿐만 아니라 다양한 영역에 접목이 가능해졌다는 것을 확인할 수 있었다.

[그림 16]은 논문데이터 초록 부분 키워드 분석 결과이다. OCR 기술논문에서는 문자나 이미지에서의 정보 추출 및 분석에 관한 키워드가 많이 분포되어있었다. 또한 특허데이터 분석결과랑 비슷하게 사용자에 대한 서비스와 데이터 처리에 관련된 내용이 다수 포함되어있었다. 키워드의 변화를 알아보기 위해 2018~2021년 논문을 분리하여 키워드 분석을

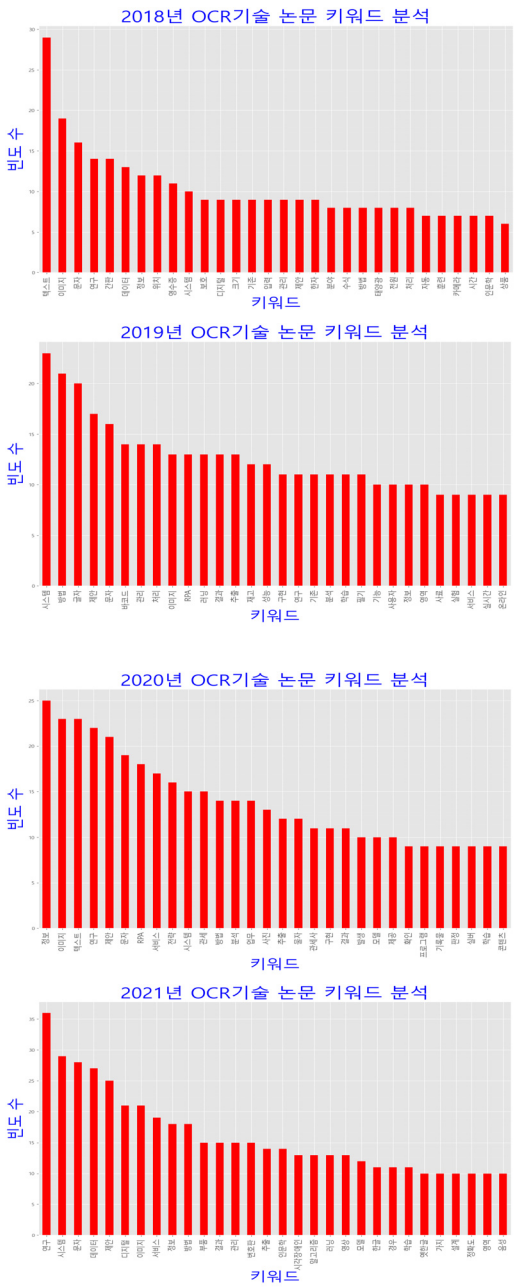


[그림 15] 특허데이터 연도별 키워드 분석 결과 (2018년~2020년)



[그림 16] 논문데이터 초록 키워드 분석 결과

진행하였다. [그림 17]은 2018~2021년 논문 키워드 분석결과이다. 2018~2019년도에는 텍스트 연구 및 분석 키워드가 상위권에 존재하였다.

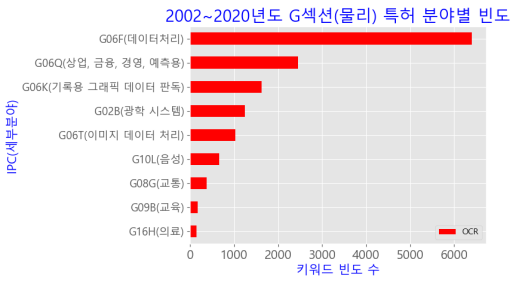


[그림 17] 논문데이터 연도별 키워드 분석결과 (2018년~2021년)

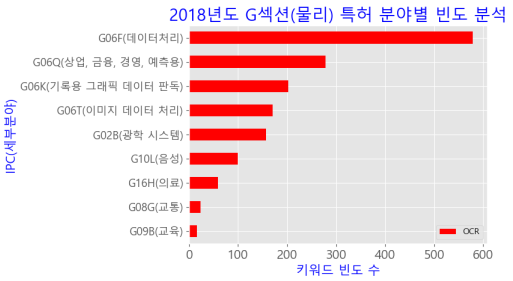
2020년도에는 자동화 기술인 RPA가 언급되었으며 2021년도에는 디지털 기술과 학습데이터 관련 논문이 상위 키워드로 나타났다. 단순하게 텍스트나 이미지에서 정보를 얻는 OCR 기술이 미래에는 자동화 기술 및 데이터처리기술로 발전될 것을 의미하였다

5.3 분야별 분석결과

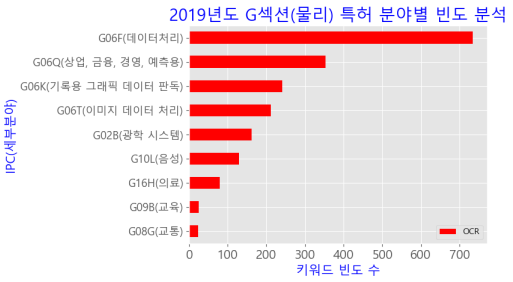
다음은 IPC 섹션 분야별 키워드 분석을 진행하였다. 실험은 G(물리) 섹션 특허 15,000여 개를 진행하였다. [그림 18]은 2002~2020년 G 섹션 특허 분야 분포도 실험결과이다. 실험결과 OCR 기술특허는 데이터 분야에서 많이 사용되었음을 알 수 있었으며 상업, 금융, 경영 분야에서도 OCR 기술특허가 많이 존재하는 것을 알 수 있었다. 그뿐만 아니라 과학, 교통, 의료, 음성, 교육 방면에서도 OCR 기술특허가 존재하여 많은 상업적 서비스에 기여하는 것을 볼 수 있었다. 분야별 특허 수가 어떻게 발전했는지 알아보기 위해 [그림 19]와 같이 2018~2020년도 G 섹션 분야별 분포도를 추출하였다. 실험결과는 다음 [그림 19]와 같이 2018년도에는 600여 개 뿐이던 데이터 관련 특허가 [그림 21]과 같이 2020년도에는 1,600여개 까지 늘어난 것 확인할 수 있었다. 또한 2018연도에 300여개에 불과했던 상업, 경영 분야 OCR 기술특허가 2020년도에는 1,000여개가 넘어가는 것을 보아 앞으로의 OCR 기술은 데이터산업과 다양한 경영, 금융 상업쪽으로 많은 영향을 끼칠 것으로 예측되었다. 뿐만 아니라 그래픽, 교육, 의료, 교통 분야에서도 점차 OCR 기술이 증가하고 있는 추세로 기술적으로 더욱 더 발전될 전망을 보였다.



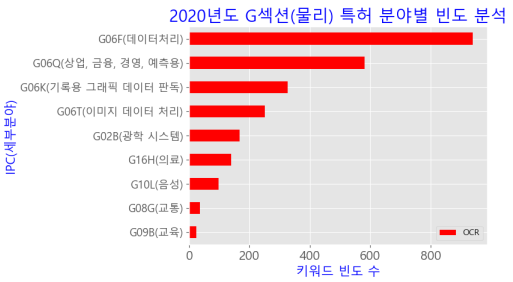
[그림 18] 특허데이터 분야별 분석결과(G섹션)



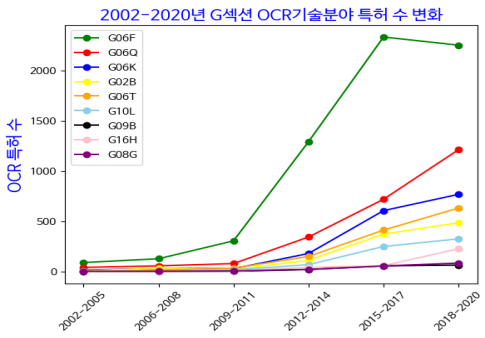
[그림 19] 2018년 G섹션 특허 분야 수



[그림 20] 2019년 G섹션 특허 분야 수



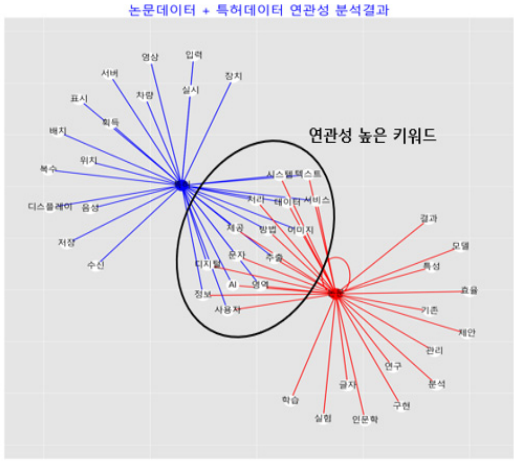
[그림 21] 2020년 G섹션 특허 분야 수



[그림 22] 2002년~2020년 G섹션 특허 분야 수

5.4 네트워크 연관성 분석결과

연도별 추출된 키워드를 통해 네트워크 연관성 분석을 진행할 수 있었다. 연도별로 존재하는 다양한 키워드들이 어떤 연관성을 가지면서 존재하는지 파악할 수 있는 분석이다. 키워드가 중앙에 위치할수록 OCR 기술특허에 많이 사용되는 키워드이다. 논문데이터와 특허데이터의 키워드를 사용하여 네트워크 연관성 분석을 진행하였다. [그림 23]은 각각의 논문과 특허에서 추출된 키워드 중 연관성이 높은 키워드를 추출한 결과이다. 결과에 따르면 OCR 기술은 텍스트, 이미지 정보 추출 기술로 시작하여 현재는 데이터 처리와 사용자에게 시스템을 제공하는 기술로 발전되어왔음을 알 수 있었다. 또한 미래의 서비스 기술로 AI 기술과 접목되어 다양한 디지털 서비스를 제공할 것을 예측할 수 있었다.



[그림 23] 논문데이터 + 특허데이터 네트워크 연관성 분석

OCR 특허와 논문데이터에 대한 분석결과를 다음과 같이 <표 1>, <표 2>로 나타내었다.

공통으로 사용된 의미 있는 키워드의 연도별 사용 증가율과 분야별 증가율을 볼 수 있었다. 특허데이터 관련해서는 다양한 영역에서 사용이 진행되다 보니 영역 키워드가 많은 증가율을 보였고 앞서 분석과 마찬가지로 정보, 데이터, 발명 등의 데이터 정보를

통한 다양한 OCR 기술이 발명되어지고 있는 것을 확인하였다. 논문데이터에서는 OCR 기술의 연구가 활발하게 진행되고 있으며 기술을 시스템화 시키는 논문이 증가한 것을 볼 수 있었다. 또한 사용자에게 다양한 자동화 서비스로 다가오는 OCR 기술로써 RPA와 서비스 키워드 또한 증가하였다. 이를 통해 OCR 기술은 앞으로도 연구가 지속될 것으로 예상되며, 다양한 시스템으로 사용자에게 편의를 제공하는 기술로 더욱 더 발전될 것을 예상한다. 특허데이터 분야별 분석결과 <표 3>에서는 OCR기술이 상업적, 경영적 분야로 많은 특허가 등록된 것을 확인할 수 있었다. 뿐만 아니라 데이터, 의료, 교육 분야에서도 높은 증가률을 보이고 있어 앞으로의 OCR 기술은 이미지, 문자 인식기술을 넘어서 다양한 분야에서 영향을 끼칠 것을 예측할 수 있었다.

〈표 1〉 OCR기술 특허데이터 키워드 분석

키워드	2018	2019	2020	증가율
장치	2,047	3,761	2,894	41%
방법	1,942	3,933	3,181	64%
정보	1,872	2,793	3,465	85%
제공	1,797	2,102	2,557	42%
사용자	1,755	2,034	2,089	19%
데이터	1,640	2,151	2,548	55%
디바이스	1,327	1,442	1,548	17%
이미지	1,236	1,345	1,652	34%
시스템	1,234	2,142	1,840	49%
표시	1,218	1,873	2,157	77%
발명	1,160	1,609	1,841	59%
영역	1,142	1,748	2,489	118%
복수	1,082	1,110	1,637	51%
디스플레이	1,031	1,471	1,100	7%
배치	919	1,074	1,726	88%
수신	865	1,107	1,331	54%
사용	854	1,030	1,274	49%
위치	843	881	975	16%
구성	834	1,021	1,171	40%
결정	799	855	1,060	33%

〈표 2〉 OCR기술 논문데이터 키워드 분석

키워드	2018	2019	2020	2021	증가율
텍스트	48	35	45	32	-33%
이미지	22	14	25	21	-5%
연구	20	15	27	42	110%
데이터	15	-	-	28	87%
디지털	12	-	-	23	92%
정보	12	12	27	22	83%
시스템	11	27	32	39	255%
처리	10	15	-	15	50%
제안	9	16	21	25	178%
방법	9	22	15	20	122%
관리	9	-	-	17	89%
학습	8	13	-	13	63%
구현	-	15	13	-	-13%
RPA	-	14	21	-	50%
추출	-	14	14	16	14%
분석	-	13	15	-	15%
서비스	-	12	19	21	75%
AI	-	12	-	15	25%

〈표 3〉 OCR 특허 분야별 분석

분야	2018	2019	2020	증가율
데이터	574	731	1,651	188%
상업, 경영	303	394	972	221%
이미지	280	311	490	75%
그래픽	204	242	325	59%
과학	30	30	38	27%
음성	99	127	95	-4%
의료	59	78	135	129%
교통	26	25	34	31%
교육	94	139	230	145%

6. 결 론

4차 산업혁명의 흐름 속에서 OCR기술 분야는 폭발적인 성장이 전망되는 분야로 기대되고 있다. 본 연구는 이러한 기대 속에서 국내 OCR기술은 어떻게 발전되어왔는지 어느 분야에 사용되고 있으며, 앞으

로의 동향을 예측하기 위해 특허데이터와 논문데이터 분석 연구를 진행하였다. 기술경쟁력 파악이나 기술예측을 위한 방법으로 델파이 기법, Scenario analysis 등 다양한 방법이 사용되었다. 그러나 주관적인 전문가 판단에 의지하는 정성적 방법은 실험디자인을 위하여 많은 노력이 투입되는 단점이 있었다. 본 연구가 가지는 시사점으로는 일정한 객관적 기준에 근거하여 장시간에 걸쳐서 등록, 게재되는 특허와 논문은 다양한 정보를 내포하고 있다. 특히 특허 정보는 기술정보, 권리정보, 경영정보의 기능성을 갖는다. 이러한 특성으로 성과나 R&D 능력을 평가하고, 경영전력을 분석하기 위한 방법으로 점유율, 키워드 빈도, 분야별 빈도 등의 지표들을 가지고 연구들이 시도되었다.

OCR 기술특허는 2002년부터 특허 등록이 진행되어 2012년부터 급성장을 보였다. 2002년에 66개에 불과했던 OCR 기술특허 수가 2020년에는 3,000여개가 특허기술로 등록되었다. 이와 마찬가지로 OCR 기술 관련 논문 또한 2012년부터 많은 연구 결과가 게재되기 시작하였다. 이로써 20세기 IT 기술의 급격한 발달로 나타난 디지털 혁명으로 인해 OCR 기술이 급격히 발전하였다는 것을 알 수 있었다. 본 논문에서는 30,000여 개의 OCR 기술특허와 432개의 OCR 기술 논문을 분석 데이터로 사용하였다. 각 특허와 논문데이터를 통해 시계열 분석과 키워드 분석을 진행하였다. 연구 결과 문자, 이미지 정보를 추출하는 기술 및 연구가 진행되어 왔으며 현재는 문자를 넘어서 음성 및 데이터 처리에도 사용되고 있으며 다양한 시스템으로 개발되어 사용자에게 제공되고 있다. 또한 기술분야적으로는 데이터처리와 각종산업, 경영 분야에서 사용되는 특허가 많았으며 교육과 교통 및 의료분야에서도 OCR 기술이 사용될 것을 예측할 수 있었다. 마지막으로 특허와 논문데이터의 네트워크 연관성 분석을 통해 앞으로의 OCR 기술은 문자와 이미지뿐만 아니라 데이터 처리를 통해 사용자에게 제공되는 하나의 서비스로 자리를 잡을 것으로 예측되며, 앞으로 발달될 AI 기술과 접목시켜 각종 산업 분야에서 사용자를 위한 서비스를 제공할 것이라고 예측 되었으며, 실무적으로 연구 분석결과는 OCR 기술특허 출원시 참고할 수 있는 유용한 정보로 활용될 수 있고 향후 키워드를 사용하여 다른 주제들과 융합한 새로운 기술개발에 사용될 수 있다. 또한, 본 논문은 국내 한정 특허, 논문 데이터를 사용했지만 전 세계적으로 사용 중인 OCR 기술로서 다양한 나라의 특허와 논문 데이터를 사용하면 세계적인 OCR 기술 동향을 예측할 수 있을 것이다.

참고문헌

[1] 이광로, 정희성, 김명원, “문자인식에 관한 연구”, *전자통신동향분석*, 제4권, 1989, 124-142.

[2] Pedro, M.B.T., “Text recognition for objects identification in the industry”, *International Journal of Mechatronics and Applied Mechanics*, 2017, 81-84.

[3] 국경완, “인공지능 기술 및 산업 분야별 적용 사례”, *정보통신기획평가원*, 2019, 1-13.

[4] Q.M. Aljelawy and T.M. Salman, “Detecting license plate number using ocr technique and raspberry Pi 4 with camera”, *2022 2nd International Conference on Computing and Machine Intelligence (ICMI)*, 2022, 1-5.

[5] Dangiwa, B.A. and S.S. Kumar, “A Business Card Reader Application for iOS devices based on Tesseract”, *2018 International Conference on Signal Processing and Information Security (ICSPIS)*, 2018, 1-4.

[6] Jain, M., M. Mathew, and C.V. Jawahar, “Unconstrained OCR for Urdu Using Deep CNN-RNN Hybrid Networks”, *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, 2017, 747-752.

[7] Ouyang, Y., “Analysis and statistics on patent information of colleges and universities

based on PatSnap: Take wuhan university of technology as an example", *2nd International Conference on Big Data Economy and Information Management (BDEIM)*, 2021, 419-423.

[8] Sakamoto, S. and Y. Okada, "Paper analysis and database of papers of the pelliot collection, dunhuang manuscripts", *2015 International Conference on Culture and Computing (Culture Computing)*, 2015, 203-204.

About the Authors

김 원 준 (sooryutan@koreatech.ac.kr)

현재 ㈜이투온에서 전략기획실장으로 근무 중이며, 기획 및 재무 등 관리분야 전반을 담당하고 있습니다. 고려대학교에서 경영학 석사학위를 받은 뒤 한국기술교육대학교에서 박사학위를 취득하였습니다. 주요 관심 분야는 빅데이터와 AI 서비스 관련 4차 산업입니다.

이 상 곤 (sklee@koreatech.ac.kr)

현재 한국기술교육대학교 산업경영학부 교수로 재직 중입니다. 연세대학교 상경대학 경영학과 학사, 한국과학기술원 경영공학과 석사・박사학위를 취득하였고, 주요 관심분야는 경영정보시스템관리, 기술경영, 그리고 Business Analytics 등입니다.

표 성 국 (skpyo@e2on.com)

현재 빅데이터 및 AI서비스를 제공하는 (주)이투온의 연구원으로, 광운대학교에서 학사 및 석사학위를 취득하였습니다. 주요 관심 분야는 인공지능을 이용한 사물인식에 관한 연구입니다.