KIPS Trans. Softw. and Data Eng.
Vol.8, No.1 pp.37~44 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.1.37

Printer Identification Methods Using Global and Local Feature-Based Deep Learning

Soo-Hyeon Lee† · Hae-Yeoun Lee††

ABSTRACT

With the advance of digital IT technology, the performance of the printing and scanning devices is improved and their price becomes cheaper. As a result, the public can easily access these devices for crimes such as forgery of official and private documents. Therefore, if we can identify which printing device is used to print the documents, it would help to narrow the investigation and identify suspects. In this paper, we propose a deep learning model for printer identification. A convolutional neural network model based on local features which is widely used for identification in recent is presented. Then, another model including a step to calculate global features and hence improving the convergence speed and accuracy is presented. Using 8 printer models, the performance of the presented models was compared with previous feature-based identification methods. Experimental results show that the presented model using local feature and global feature achieved 97.23% and 99.98% accuracy respectively, which is much better than other previous methods in accuracy.

Keywords : Global Feature, Local Feature, Deep Learning, Printer Identification, Convolutional Neural Network

전역 및 지역 특징 기반 딥러닝을 이용한 프린터 장치 판별 기술

이 수 현† · 이 해 연††

요 약

디지털 IT 기술의 발달로 인하여 프린터와 스캐너의 성능이 향상되고 가격이 저렴해지면서 일반인들도 쉽게 접할 수 있게 되었다. 그러나 이에 따른 부작용으로 공문서 및 사문서 위조 등의 범죄들이 쉽게 이루어질 수 있다. 따라서 해당 문서가 어떤 프린터를 사용하여 출력 되었는가를 특정할 수 있다면 수사 범위를 줄이고 용의자를 판별하는데 도움이 된다. 본 논문에서는 프린터 장치 판별을 위하여 딥러닝 모델을 제안한다. 먼저 최근 인식 등에서 범용적으로 활용되는 지역 특징 기반의 컨볼루셔널 뉴널 네트워크를 이용한 프린터 장치 판별 모델을 제안하고, 전역 특징 기반의 처리 과정을 네트워크 모델에 도입함으로 인하여 수렴 속도 및 정확도를 향상한 기법을 제안한다. 제안한 모델의 성능은 8개의 프린터 장치를 활용하여 기존 프린터 판별을 위한 특징 기반 기술과 비교를 수행하였다. 그 결과 제안하는 지역 특징 기반의 모델과 전역 특징 기반의 모델이 각각 97.23% 및 99.98%의 높은 판별 정확도를 달성하였고, 기존 기술들에 비하여 높은 정확도를 갖는 우수성을 보였다.

키워드 : 전역 특징, 지역 특징, 딥러닝, 프린터 장치 판별, 컨볼루셔널 뉴럴 네트워크

1. 서 론

디지털 IT 기술의 발전으로 컴퓨터 및 주변 기기들도 같이 성능이 향상되고 있으며, 구매 비용도 낮아지고 있다. 이로 인해 일반인들도 고성능의 컴퓨터 시스템과 다양한 주변기기를 접하고 사용할 수 있게 되었다. 또한 소프트웨어 측면에서는 오픈 소스화가 진행되고 있어서 고수준의 소프트웨어를 누구나 쉽게 구할 수 있게 되었다. 이러한 발전은 대부분의 정보가 인터넷에서 공유되는 현대 사회에 있어 기술과 정보의 불평등을 해소하는 측면에서 긍정적이지만 이를 이용한 다양한 범죄들이 파생되고 또한 그 수준이 고도화되고 있어서 사회적인 문제로 대두되고 있다.

주변 기기 중에 대표적인 것은 입력을 위한 스캐너와 출력을 위한 프린터가 있다. 스캐너 성능의 향상으로 원본 정보를 완벽히 포함하는 영상을 획득할 수 있고, 프린터 성능의 향상

※ This work was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2017R1D1 A1B03030432).
† 비 회 원 : 금오공과대학교 소프트웨어공학과 석사과정
†† 정 회 원 : 금오공과대학교 컴퓨터소프트웨어공학과 교수
Manuscript Received : October 4, 2018
First Revision : November 19, 2018
Accepted : December 1, 2018
* Corresponding Author : Hae-Yeoun Lee(haeyeoun.lee@kumoh.ac.kr)

으로 일반인이 원본과 차이를 식별할 수 없을 정도의 인쇄물을 대량으로 빠르게 생산할 수 있다. 그 과정에서 영상을 수정하여 왜곡시키는 일도 인터넷에서 공유되고 있는 소프트웨어를 이용하면 어렵지 않다. 이와 같은 위변조는 공문서 및 사문서의 위변조와 직결되기 때문에 이들 위변조를 방지하기 위한 기술은 점점 중요해지고 있다.

위변조 범죄를 수사하기 위해 다양한 기술들이 사용되고 있으며 대표적으로 위변조 여부 판별 기술과 입력 및 출력 장치 판별 기술이 있다. 위변조 여부 판별 기술은 영상이 변형되면서 남긴 흔적을 식별하여 해당 문서의 어느 부분이 위변조 되었는지를 찾아내는 것을 목표로 한다. 입력 및 출력 장치 판별 기술은 특정 사용자가 생산한 같은 범주의 문서는 대부분 같은 장치를 사용하여 생산된다는 점을 사용하여 위변조 문서를 생산하는데 사용한 입력 및 출력 장치의 유사성을 판단하는 것을 목표로 한다.

본 논문은 위변조를 위한 프린터 장치 판별하는 것을 대상으로 하고 있으며 기존에 인쇄물에 내재된 프린터의 특징을 추출하여 판별하는 다양한 연구들이 이루어졌다. 본 논문에서는 프린터 장치 판별을 위하여 딥러닝 모델을 제안한다. 최근 영상 인식 등을 위해서 딥러닝 기술이 범용적으로 사용되고 있으며, 대부분 지역 특징 기반으로 인식을 수행하는 컨볼루셔널 뉴럴 네트워크(Convolutional Neural Network, CNN)를 활용하고 있다. 먼저 본 논문에서는 지역 특징 기반 컨볼루셔널 뉴럴 네트워크를 활용한 프린터 판별 모델을 제안한다. 그 후에 전역 특징 기반의 처리 과정을 네트워크 모델에 도입함으로 인하여 수렴 속도 및 정확도의 향상을 모색한 모델을 제안한다. 딥러닝 연구는 대부분 특정한 응용을 위한 최적의 모델을 수립하는 것에 기여도를 두고 있으며, 본 논문은 프린터 판별을 위한 모델 수립과 전역적 특징을 활용하는 기법의 제안에 기여도가 있다.

제안한 지역 특징 기반 및 전역 특징 기반의 딥러닝 모델 성능은 8개의 프린터 장치를 활용하여 분석하였고, 기존 프린터 판별을 위한 특징 기반 기술들과 비교를 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련된 연구들을 제시하고, 범용 CNN 기반 기술을 설명한다. 3장에서는 전역 특징 기반 딥러닝을 이용한 프린터 판별 모델을 제시하고, 4장에서는 제안한 모델의 성능을 분석한다. 마지막으로 5장에서는 결론을 맺는다.

2. 관련 연구 및 기반 기술

2.1 프린터 장치 판별 관련 연구

문서를 인쇄한 기기를 판별하기 위한 연구는 국내외에서 진행이 되어 왔다. 문서의 인쇄 방식에 따라 다양한 연구들이 있었고, 새로운 인쇄 방식이 등장하면 그에 맞는 기술을 위한 추가적인 연구들이 진행되었다. 분석 방식은 크게 문자 특성 분석[1-4], 영상의 통계적 특성 분석[5-10] 및 딥러닝을 이용한 분석[11]으로 분류해볼 수 있다.

Mikkilineni et al.은 인쇄 영상의 텍스쳐 특징 분석을 통해 EP에서 생산된 10개의 프린터를 판별하는 방법을 제안하였다[1]. 문서 내에 포함되어 있는 특정 문자를 선택해 분산과 엔트로피 기반 특징과 명암도 동시 발생 행렬(Gray-Level Co-occurrence Matrix, GLCM) 기반 특징을 추출하고 5NN Classifier를 사용하여 프린터를 판별한다. 분석을 위해 500개의 “e”가 있는 문서를 활용하였고 성능 검사를 위해 300개의 “e”가 있는 문서를 이용해서 약 57%의 성능을 달성하였다. 또한, 이를 확장하여 문자의 크기. 글꼴, 용지 종류, 인쇄 시점 등의 요소를 포함하여 같은 방법으로 특징을 추출한 후 SVM 분류기(Support Vector Machine Classifier)을 사용하여 프린터를 식별하는 연구를 진행하였다[2]. 5가지 글자 크기 및 글꼴, 3가지 용지 종류를 분석하였고 같은 학습 및 테스트셋을 구성하였다. 실험 결과 글자 크기 변인의 경우, 2pt 이하의 차이일 경우 90%의 성능으로 프린터를 식별하였다. 글꼴 변인의 경우, 글꼴이 같으면 90%의 성능을 보이고 글꼴이 다르면 70%의 성능을 벼였다. 용지 변인의 경우, PP-0001 용지와 PP-0006 용지는 각각 최소 90%의 성능을 보여주고 PP-0008 용지는 100%의 성능을 보여주었고 인쇄 시점 변인은 70%의 성능을 보여주었다.

Deng et. al은 상기 방법과 유사하게 특정 문자를 분석에 초점을 두었지만 텍스쳐 특징이 아닌 문자 간의 거리를 이용하여 45개의 프린터를 판별하는 연구를 진행하였다[3]. 판별을 위해 각 문서에서 특정 문자의 영상을 획득한 후, 이진화를 통해 깨끗한 문자 영상을 생성하였다. 그 후 거리변환을 통해 문자 픽셀과 배경 픽셀의 거리를 구하였고 입력한 영상의 거리 값과 입력해놓은 프린터 중 가장 근사한 거리 값을 보이는 프린터로 판별하였다. 영문자 3가지 글꼴과 중국어 1가지 글꼴을 대상으로 실험을 하였으며, Top 1 식별률은 평균 25%, Top 5 식별률은 평균 81.7%의 정확도를 보였다. 이전 연구들과 성능은 비슷하면서도 인쇄 DPI와 인쇄 재료에 강인한 결과를 보였다. 하지만 인쇄 부품의 상태에 따라 성능이 변하고 인쇄 시점에 따른 변화에 취약한 결과를 보였다.

Elkasrawi et. al은 프린터별 노이즈 및 인쇄 패턴을 분석하여 20개의 프린터를 식별하는 연구를 진행하였다[4]. 문서에 포함된 문장의 높이를 비교하는 방법, 영상에 이진화를 적용 후 중간값 필터링을 사용하고 이와 원본 영상 간의 차영상을 통해 잡음을 뽑아내 비교하는 방법 등을 통해 식별을 시도하였다. 전체 프린터 데이터셋에 대한 정확도는 76.75%를 달성하였고 잉크젯 프린터에 한하여 93.57%의 성능을 나타내었다.

Choi et al.은 이산 웨이블릿 변환(Discrete Wavelet Transform, DWT)을 이용한 프린터 판별을 제안하였다[5]. 컬러 레이저 프린터로 인쇄한 영상을 RGB와 CMYK 도메인으로 나누고 DWT 연산을 수행하여 각각의 HH 밴드를 특징으로 활용한다. 특징값을 SVM 분류기를 이용하여 분류하였고 프린터 브랜드 판별 및 프린터 장치 판별에 있어서 각각 97.89% 및 80.24%의 정확도를 달성하였다.

Ryu et. al은 사람의 눈으로 인지할 수 없는 하프톤 패턴을 이용한 프린터 판별 연구를 진행하였다[6]. CMYK 도메인의 하프톤 패턴을 분석하여 9개의 프린터를 분류하였다. 그 결과 브랜드 판별은 91.9%의 정확도를 보여주었고 각 프린터 장치

판별에서는 평균 63%의 정확도를 나타내었다. 하프톤 패턴의 차이는 잘 구분했지만 같은 브랜드의 프린터가 유사한 패턴을 보이기 때문에 프린터에 대해 독립적이지 않아 개별 프린터 분류에는 취약한 정확도를 보였다. Kim and Lee는 이를 확장하여 하프톤 패턴의 텍스쳐 핑거프린터를 이용하여 5개 프린터 장치 판별을 연구하였고 평균 86.14% 정확도를 보였다[7].

Baek et. al은 DWT와 GLCM을 이용한 프린터 판별 기술을 제안하였다[8]. 인쇄물을 HSV 도메인으로 변환을 하고 DWT 연산을 이용하여 컬러 노이즈 정보를 가진 3채널 각각의 HH 밴드를 추출한다. 이를 GLCM으로 변형한 후, 질감분석을 위한 5가지 특성을 추출하여 총 60개의 특징 벡터(HSV 3채널×GLCM 4방향×질감분석특징 5가지)를 구성한다. 특징 벡터를 단일 벡터로 통합 후 SVM을 이용하여 분류를 시행하여 4개 제조사의 프린터에 대해 96.9%의 성능을 보였다. 같은 제조사의 다른 프린터 모델 분류는 평균 79.23%의 정확도를 달성하였다.

Lee et. al은 상기 방법을 확장하여 DWT 대신 위너필터를 사용하여 연구를 진행하였다[9]. 판별 과정은 Fig. 1과 같으며, 색상 도메인을 HSV가 아닌 CMY로 변환하는 점과 DWT의 HH 밴드 대신 위너필터를 이용하여 추출한 잡음을 특징을 활용하는 점을 제외하고는 실험 구성이 동일하다. 이를 통하여 브랜드 판별 및 프린터 장치 판별에서 각각 98.2% 및 84.5%의 정확도를 보였다.

Tsai et al.은 DWT와 특징 선택 알고리즘을 조합하여 프린터 판별 기술을 연구하였다[10]. 5가지의 특징 선택 알고리즘을 이용하여 다양한 특징 그룹을 형성하였다. 일반적으로

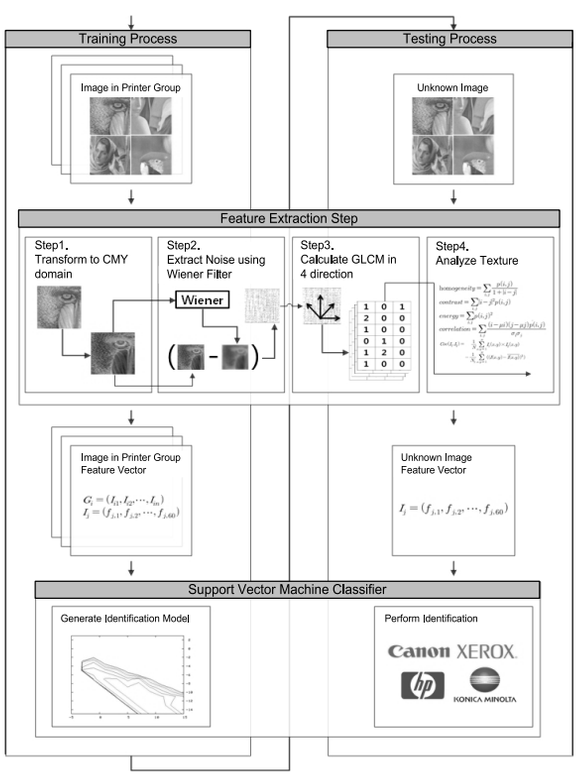

Fig. 1. Printer Identification using Wiener Filter and GLCM

DWT 연산 결과에서 HH 밴드만 주로 사용되지만 LH과 HL 또한 프린터 판별에 있어서 유의미한 특징이라는 결과를 제시하였다. 최적의 특징 그룹을 SVM 분류기를 통하여 분류하였을 때 평균 92.4%의 정확도를 보였다.

Kim et al.은 딥러닝의 CNN을 이용한 프린터 판별 기술을 연구를 진행하였다[11]. 전처리 부분에서 합성곱 연산을 통하여 HCD(Halftone Color channel Decomposition)를 수행하고 이어서 특징 추출을 위한 합성곱 과정을 추가하는 복합적 CNN 모델(Cascaded Learning Framework 구조)을 작성하였다. 8개의 프린터를 대상으로 실험을 하였고 결과적으로 약 96%의 정확도를 보였다.

2.2 딥러닝 기술

딥러닝은 최근 다양한 분야에서 널리 응용되고 있는 기술이다. 인간의 뇌를 모방하기 위한 인공신경망 기술이 비선형적인 문제를 해결할 수 없다는 단점에 의해 사장된 후, 추가적인 연구로 인공신경망을 여러 개를 쌓아서 비선형적인 문제를 해결하는 방법이 제시되면서 각광받기 시작하였다. 일반적인 모델로는 고차원의 데이터 처리에 적합한 심층 신경망(Deep Neural Network, DNN)과 영상 등의 지역적 특징을 지닌 데이터 처리에 적합한 컨볼루셔널 신경망(Convolutional Neural Network, CNN), 그리고 시계열 데이터 처리에 적합한 순환 신경망(Recurrent Neural Network, RNN) 등이 있다.

많은 연구들에서 영상 인식을 위해서는 CNN이 유용한 것으로 나타나고 있다. CNN의 기본적인 구조는 Fig. 2와 같이 입력(input) 계층, 컨볼루션(convolution) 계층, 전연결(fully connected) 계층, 출력(output) 계층으로 구성된다.

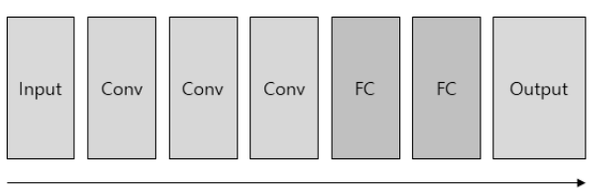

Fig. 2. General Structure of CNN

입력 계층에서는 영상 데이터를 전달받아 다단계의 컨볼루션 계층(Conv)을 거쳐 특징값을 추출하고 전연결 계층(FC)을 통해 특징값 사이의 연관성을 가중치를 통해 조절하고 그 결과를 출력 계층으로 내보낸다. 그 과정에서 컨볼루션 연산[12], 활성화 함수(ReLU[13]), 풀링(Pooling) 연산[12], Drop-out[14], 배치정규화[15], Softmax 등 다양한 기술이 사용된다.

이와 같은 과정을 거쳐 어느 클래스인지 나타내는 값이 출력된다. 학습 중에는 출력값과 정답을 비교하여 컨볼루션 연산의 커널과 전연결 계층의 노드가 가진 가중치를 조정한다. 출력값과 정답의 차이를 최소화하는 방향으로 가중치를 조정하여 출력값이 정답에 가까워지도록 유도한다. 판별 과정에서는 임의의 정보를 넣어 출력값(예측값)의 정답률을 분석하여 정확도를 측정한다.

3. 제안하는 딥러닝 기반 프린터 판별 알고리즘

딥러닝 기반 프린터 판별 과정은 Fig. 3과 같이 딥러닝 모델의 학습과 판별의 과정으로 구성된다. 학습 과정에서는 학습 영상을 입력한 후 모델의 가중치 및 편향 값을 기반으로 입력한 학습 영상의 레이블을 결정함으로서 분류를 수행한다. 모델이 분류한 레이블 값과 학습 시, 레이블한 값이 다를 경우 오류 역전파를 통해 모델의 가중치 및 편향 값에 반영한다. 지정한 횟수의 학습이 이루어지고 난 후, 판별 과정에서는 판별 영상을 입력하여 레이블 값을 결정하여 판별을 수행한다. 판별 영상에 대하여 결정한 레이블 값과 실제 레이블 값을 비교하여 정확도를 측정할 수 있다.

본 논문에서는 딥러닝 모델에 대하여 영상 인식에 대하여 범용적으로 사용되는 지역적 특징을 활용한 컨볼루셔널 뉴럴 네트워크를 사용한 프린터 판별 모델에 대하여 먼저 설명하고, 그 후에 전역 특징의 계산 과정을 포함하는 프린터 판별 모델에 대하여 설명한다.

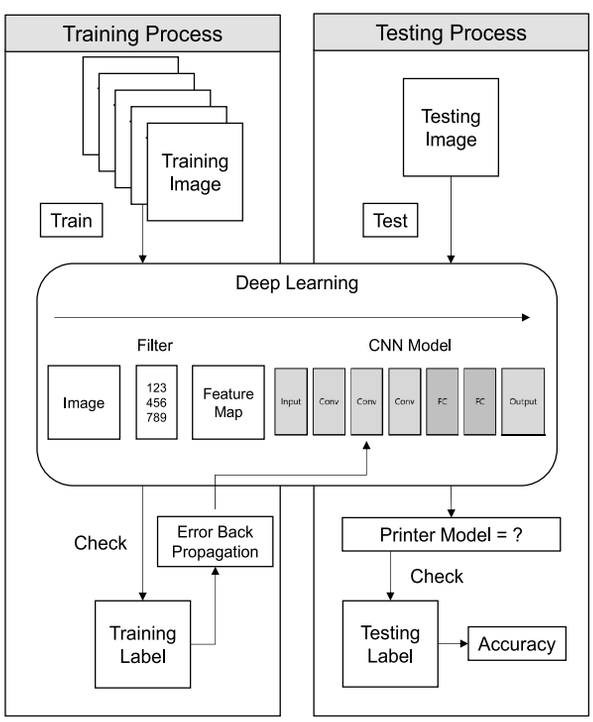

Fig. 3. Training and Testing Process of Proposed Algorithm

3.1 지역 특징 기반 딥러닝 모델

프린터 판별을 위한 지역 특징 기반 딥러닝 모델에서는 영상 인식에 많이 활용되는 컨볼루셔널 뉴럴 네트워크 모델을 도입하였고, 별도의 필터링 처리를 적용하지 않은 순수 영상 데이터만을 이용한다.

지역 특징 기반 딥러닝 모델은 CNN 기반으로 수립을 하였고 Fig. 4에 제시하였다. 딥러닝에 대한 개발 경험과 관련 연구 결과[16]를 바탕으로 입력 영상에 대하여 3개의 컨볼루션 계층과 2개의 전연결 계층으로 구성하고 판별을 하여 결과를

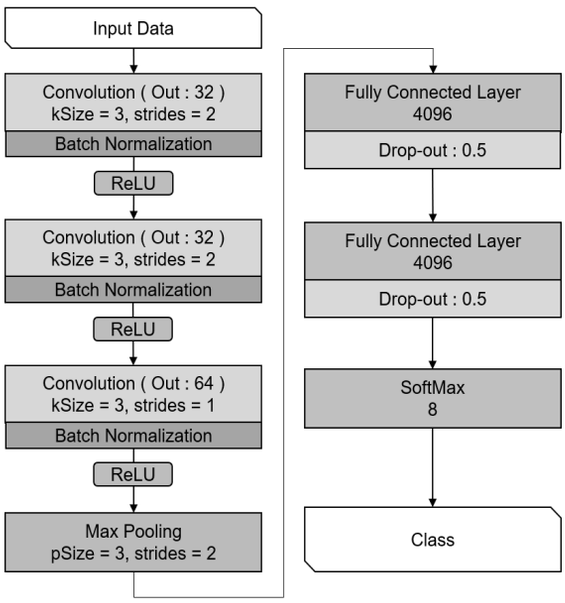

Fig. 4. Local Feature-based Printer Identification Model

출력하도록 하였다. 또한 최적의 성능을 나타내도록 파라미터들에 대한 설정과 처리 과정을 구성하였다.

제안한 모델은 128x128 크기의 3채널 RGB 영상을 입력 데이터로 사용한다. 첫 번째와 두 번째 컨볼루션 계층은 32개의 feature map을 다루고 strides 값을 2로 두어 feature map 크기를 줄인다. 세 번째 컨볼루션 계층은 64개의 feature map을 다루며 strides 값은 1로 두어 feature map 크기를 유지한다. 모든 컨볼루션 계층은 배치 정규화(batch normalization)과 ReLU를 거친다. 그 후 max 풀링을 진행하여 연산량을 줄여 학습 속도를 향상시킨다. 그 후, 4096개의 노드를 갖는 전연결 계층을 2번 통과하며 특징값의 상관관계를 가중치를 통해 계산한다. 마지막으로 softmax 계층에서 8개의 클래스 중 가장 유사한 클래스의 값을 출력한다.

3.2 전역 특징 기반 딥러닝 모델 (GLCM Model)

영상 인식에서 범용적으로 사용하는 컨볼루셔널 뉴럴 네트워크 모델은 지역적 특징을 기반으로 인식을 수행하는 모델이며 3.1절에서 제안을 하였지만, 프린터 장치 판별을 위해서는 인쇄된 문서에 분산되어 나타나는 전역적인 특징을 고려하는 것이 성능 및 속도 측면에서 효율성을 높일 수 있는 것으로 사료된다. 따라서 전역 특징 기반 판별 모델에서는 인쇄된 문서로부터 전역적 특징을 추출하는 과정을 네트워크 모델에 도입을 수행하였다.

전역적 특징으로는 영상 픽셀 데이터를 필터링 없이 그대로 입력 데이터로 사용하는 것이 아니라, 영상에 대하여 명암도 동시 발생 행렬(Gray-Level Co-Occurrence Matrix, GLCM)을 생성한 후에 이를 컨볼루셔널 뉴럴 네트워크 모델의 입력 데이터로 활용한다. GLCM은 영상 내에서 픽셀 단위에서의 밝기값 관계를 추출하는 방법으로 인접 픽셀의 값을 좌표로 이용하여 같은 값이 반복해서 나오는 횟수를 특징으로 활용하

며, 텍스쳐 특성을 잘 나타내는 특징 추출 방법이다[17].

GLCM의 feature map 크기는 영상의 픽셀값 범위에 의존한다. Fig. 5에는 1픽셀을 3bit로 가정하여 GLCM을 이용하여 8x8의 feature map을 추출하는 과정을 도시하였다. 본 논문에서 사용한 영상은 8비트(0~255)이기 때문에 영상의 크기와 상관없이 256×256 크기의 feature map이 생성되고 R, G, B 각 채널에 대해 4가지 방향(0°, 45°, 90°, 135°)에 대해 특징을 추출하여 256×256 크기의 feature map을 12개 생성하였다. 하지만 256×256×4×3의 데이터를 딥러닝 모델에 직접 학습하기에 하드웨어적인 성능의 제약으로 인하여 Fig. 6과 같이 각 R, G, B 채널의 값을 같은 방향끼리 행렬합 연산을 수행하여 256×256×4의 데이터로 만들어 딥러닝 모델의 입력 데이터로 활용하였다. 전역 특징 기반 프린터 판별 모델의 컨볼루션 계층, 전연결 계층 및 출력 계층과 이들에 대한 구성과 파라미터의 설정은 지역 특징 기반 판별 모델과 동일하도록 구성하였다.

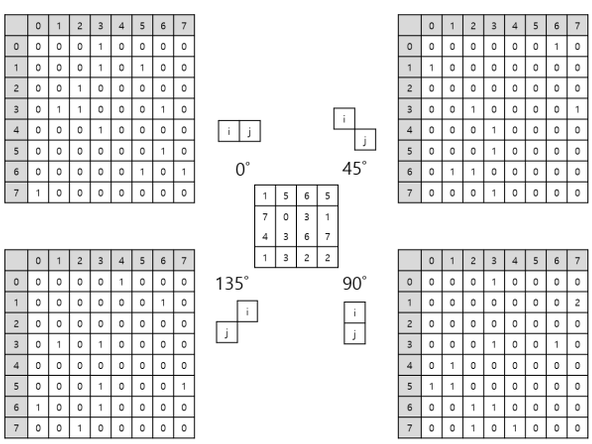

Fig. 5. Gray Level Co-occurrence Matrix Calculation (in case of 3 bits image)

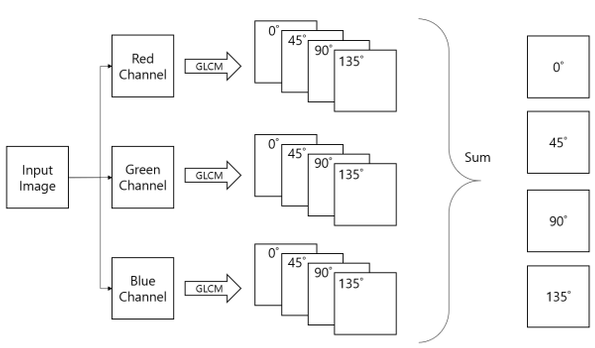

Fig. 6. GLCM Feature Map Processing for Data Reduction

4. 실험 결과

4.1 실험 환경

프린터 장치 판별을 위한 딥러닝 장비 환경을 구축하였다. CPU는 Intel의 7세대 Core-i7을 사용하였고 GPGPU 연산을 담당할 그래픽카드는 호환성이 좋은 Cuda library를 사용하는 Nvidia의 Titan XP(12GB)를 사용하였다. RAM은 16GB로 구성하였다. 구동 환경은 Windows 10에서 구글이 배포한 Tensorflow를 사용하였으며 프로그래밍 언어는 Python을 이용하였다.

4.2 실험 데이터

실험을 위한 데이터를 위하여 총 8개의 프린터 장치에서 동일한 영상을 인쇄하였고 그 영상을 스캐너를 통해 획득하였다. Table 1에는 사용된 프린터 장치 및 영상에 대한 정보를 나타내며 Fig. 7은 획득한 영상의 예시를 프린터 장치별로 제시하였다. 영상의 크기는 256×256이며 모두 RGB 3채널이다.

Table 1. List of Printer Used for Experiment and Their Labels

Label	Brand	Product Name
C1	Canon	iR C2620
C2	Canon	iR C3200N
H1	HP	4650
K1	Konica	C250
X1	Xerox	DC C400
X2	Xerox	DC C450
X3	Xerox	DC C5540
X4	Xerox	DC C6550

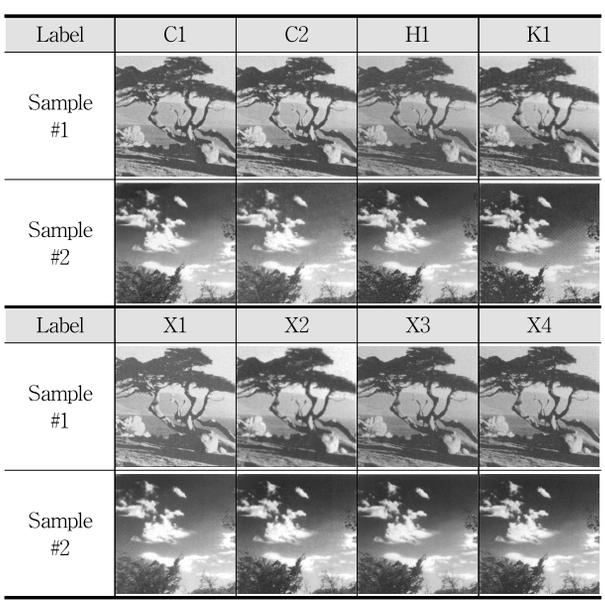

Fig. 7. Samples from Each Printer

딥러닝에 사용되는 하드웨어의 한계로 인해 스캐너를 통해 획득한 영상을 그대로 사용하기에는 무리가 있으며 프린터 별로 371개의 데이터로는 딥러닝을 수행하기에 부족하다. 이에 전체 데이터의 일관성을 유지하고 편향을 방지하기 위

하여 Fig. 8과 같이 128×128 크기로 임의 절단의 과정을 통하여 100배 많은 수의 영상 데이터를 수집하였다.

전체 데이터는 8 (프린터 장치 수) × 37,200 (프린터 장치당 데이터 개수)으로 297,600개이며 학습 데이터와 판별 데이터의 비율은 약 80:20으로 각각 240,000개와 57,600개로 설정하여 성능에 대한 분석을 수행하였다.

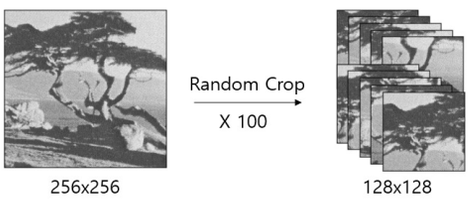

Fig. 8. 128x128 Sample Collection with Random Cropping

4.3 실험 결과 및 분석

본 논문에서 제안하는 지역적 특징(local feature model)과 전역적 특징(global feature model)을 활용하는 모델을 이용한 프린터 장치 판별을 실험한 결과에 대한 정확도 분석을 Table 2와 Fig. 9에 요약하였다. 그림에서 x 축은 전체 데이터셋에 대한 학습 횟수(epoch)를 의미하며, y 축은 판별 정확도를 나타낸다.

Table. 2 Printer Identification Accuracy Per Epoch

Epoch	Local Feature	Global Feature	Epoch	Local Feature	Global Feature
0	12.52	12.52	11	91.48	99.86
1	35.72	97.8	12	95.05	99.92
2	72.45	98.9	13	95.14	99.57
3	81.35	99.82	14	93.49	99.56
4	86.72	99.66	15	92.76	99.98
5	88.6	99.71	16	94.39	99.98
6	90.48	99.87	17	96.53	99.86
7	91.88	99.87	18	96.82	99.92
8	93.35	99.74	19	96.56	99.57
9	93.79	99.1	20	97.23	99.56
10	92.71	99.9	Max	97.23	99.98

지역적 특징을 활용한 모델은 최대 정확도로 97.23%를 보였고, 전역적 특징을 활용한 모델은 최대 정확도로 99.98%를 보였다. 전역적 특징을 활용한 모델이 지역적 특징을 활용한 모델에 비하여 높은 정확도를 보였으며, 학습 속도에 있어서는 더 큰 차이를 보였다.

전역적 특징을 이용한 모델은 학습 시작 초반인 1 학습 횟수부터 높은 정확도를 달성한데 비하여 지역적 특징을 이용한 모델은 점진적으로 정확도가 증가하는 형태를 보였다. 20번 학습 횟수를 기준으로 지역적 특징보다 전역적 특징을 활용한 모델의 성능이 2% 이상 높았으며 20번을 넘어서 계속 실험을 진행하였지만 Tensorboard를 통해 확인한 loss 값과 테스트 셋을 통한 성능 값을 확인하였을 때, 지역적 특징을 활용한 모델은 97%를 기준 ± 0.5% 정도로 값이 진동하며 20 학습 횟수과 비교해서 크게 차이가 나지 않았다. 또한 전역적 특징을 활용한 모델의 성능은 100%에 근접하였다. 이는 GLCM에 기반하는 전역적 특성이 프린터 장치들 사이의 차별성을 잘 표현하기에 높은 성능을 보이며 빠른 학습이 된 것으로 판단된다.

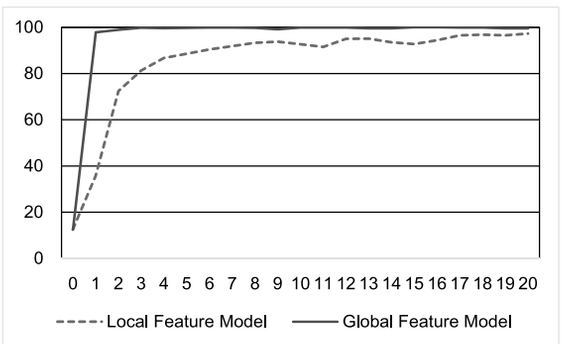

Fig. 9. Printer Identification Accuracy of Two Proposed Models

Table 3에서는 프린터 장치 판별을 위한 성능에 대하여 기존의 특징 기반 연구 중 유사한 스캔 영상 데이터베이스를 이용한 연구들과 성능을 비교하여 분석하였다. 사용한 프린터 장치의 숫자에 대하여 조금의 차이가 있지만, 제안하는 지역 특징 기반 및 전역 특징 기반 프린터 장치 판별 모델이 각각 4% 및 6% 이상의 성능 향상을 보였고, 전역 특징 기반 프린터 장치 판별 모델은 100%에 근사한 것을 확인할 수 있다.

Table 3. Comparison of Printer Identification Algorithms

Method	Printers	Accuracy
Choi [5] - Statistic	9	80.24%
Ryu [6] - Statistic	9	63.00%
Kim [7] - Statistic	5	86.14%
Baek [8] - Statistic	4	79.23%
Baek [9] - Statistic	4	84.5%
Tsai [10] - Statistic	10	92.40%
Kim [11] - Deep Learning	8	96.09%
Our local feature model	8	97.23%
Our global fearue model	8	99.98%

Table 4와 Table 5에 20회 학습을 수행한 후에 최대 정확도를 갖는 경우의 제안하는 지역적 특징 및 전역적 특징을 활용한 프린터 장치 판별 모델이 각 프린터 개별 장치를 판별한 정확도를 정리하였다. 각 프린터 개별 장치 판단하는데 있어서 전역적 특징의 우수성을 확인할 수 있다.

Table 4. Printer identification Accuracy for Each Printer Using Local Feature Model

20 epoch		Prediction								
		C1	C2	H1	K1	X1	X2	X3	X4	Accuracy
Truth	C1	6972	287	9	8	2	0	1	0	95.78%
	C2	167	6890	14	3	1	5	1	0	97.30%
	H1	1	2	7097	1	1	7	1	1	99.80%
	K1	7	2	0	7178	0	1	0	0	99.86%
	X1	11	13	3	2	7263	18	1	0	99.34%
	X2	1	0	9	5	0	7106	0	0	99.79%
	X3	18	23	46	1	1	6	6966	142	96.71%
	X4	2	17	52	4	3	23	670	6535	89.45%

Table 5. Printer Identification Accuracy for Each Printer Using Global Feature Model

15 epoch		Prediction								
		C1	C2	H1	K1	X1	X2	X3	X4	Accuracy
Truth	C1	7259	0	0	0	0	0	0	0	100%
	C2	1	7143	0	0	0	0	0	0	99.99%
	H1	0	0	7210	0	0	0	0	1	99.99%
	K1	0	0	0	7006	0	0	0	0	100%
	X1	0	0	0	0	7272	0	0	0	100%
	X2	0	0	0	0	5	7159	0	0	99.93%
	X3	0	0	0	0	2	1	7292	1	99.95%
	X4	0	0	0	0	0	0	1	7247	99.99%

5. 결 론

컴퓨터 및 프린터와 스캐너 등 다양한 주변 기기들의 성능이 향상되고, 고수준 소프트웨어의 보급으로 인해 일반인의 문서 위변조 가능성이 높아지게 되었다. 공문서 및 사문서 위변조의 범죄는 사회에 끼치는 영향이 크기에 이를 방지하기 위해 많은 기술이 연구되었고 활용되고 있다.

위변조 범죄를 예방하기 위한 기술의 하나로서 본 논문에서는 전역 특징 기반 딥러닝 기술을 이용하여 문서를 인쇄하는데 사용된 프린터를 판별하기 위한 방법을 제안하였다. 최근에 영상 관련 데이터를 처리하는 기술에 많이 사용되는 CNN에서 대부분의 모델이 필터링 등을 통한 지역적 특징을 활용하는데 비하여 본 논문에서 제안하는 모델에서는 텍스쳐 분석에 유용하고 영상 전체의 통계적 특성을 잘 나타내는 GLCM을 이용하여 전역 특징 추출하는 과정을 딥러닝

모델의 계층에 포함하였다. 일반적인 영상처리는 영상 내의 지역적 특성만을 이용하지만, 인쇄물은 생성과정에서 잡음이나 특성이 모든 영역에 걸쳐 발생하기에 전역적인 특징을 추출하여 분석하는 시도를 하였으며 이를 통하여 제안하는 모델의 속도 및 정확성에 있어서 향상을 달성하였다.

또한 기존에 알려진 특징 기반 프린터 판별 기술들보다도 더 높은 정확도를 달성하였다. 영상처리에 지역적인 특징만을 사용하지 않고 적극적으로 전역적인 특징을 이용한 점에서 의의가 있다고 생각한다.

본 논문에서 사용한 데이터는 인쇄물에 있어서 다양한 오염을 적용하지 않은 것들을 대상으로 분석을 수행하였다. 그러나 실제로는 인쇄물에 다양한 변형이 있어서 이후 연구에는 이와 같은 데이터를 구성하여 성능분석을 수행할 필요가 있다. 또한, 프린터 판별 기술의 경우 정확도가 100%가 아니면 법적 증거로 효용성이 없기에 성능의 향상을 모색할 필요가 있을 것으로 판단된다. 딥러닝의 장점은 유의미한 특징값을 사람의 개입 없이 최적화 과정에서 찾아내어 활용한다는 점이다. 하지만 본 논문에서 제안하는 기법은 전역 특징 추출을 위하여 사람이 개입한 특징을 활용하였으므로 차후에 전역적 특징을 추출할 수 있는 딥러닝 모델을 연구하여 적용한다면 좀 더 효율적일 것으로 생각된다.

References

[1] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp, “Printer identification based on texture features,” in *Proceedings of the International Conference on Digital Printing Technologies,* pp.306-311, 2004.

[2] A. K. Mikkilineni, O. Arslan, P.-J. Chiang, R. M. Kumontoy, J. P. Allebach, G. T.-C. Chiu, and E. J. Delp, “Printer forensics using svm techniques,” in *Proceedings of the International Conference on Digital Printing Technologies*, pp.223-226, 2005.

[3] W. Deng, Q. Chen, F. Yuan, and Y. Yan, “Printer identification based on distance transform,” in *Proceedings of the International Conference on Intelligent Networks and Intelligent Systems*, pp.565-568, 2008.

[4] S. Elkasrawi, and F. Shafait, “Printer identification using supervised learning for document forgery detection,” in *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems*, pp.146-150, 2014.

[5] J.-H. Choi, D.-H. Im, H.-Y. Lee, J.-T. W. J.-H. Ryu, and H.-K. Lee, “Color laser printer identification by analyzing statistical features on discrete wavelet transform,” in *Proceedings of the IEEE International Conference on Image Processing*, pp.1505-1508, 2009.

[6] S.-J. Ryu, H.-Y. Lee, D.-H. Im, J.-H. Choi, and H.-K. Lee, “Electrophotographic printer identification by halftone texture analysis,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1846-1849, 2010.

[7] D.-G. Kim and H.-K. Lee, “Colour laser printer identification using halftone texture fingerprint,” *Electronics Letters*, Vol.51, No.13, pp.981-983, 2015.

[8] J.-Y. Baek, H.-S. Lee, S.-G. Kong, J.-H. Choi, Y.-M. Yang, and H.-Y. Lee, “Color Laser Printer Identification through Discrete Wavelet Transform and Gray Level Co-occurrence Matrix,” *KIPS Transactions: PartB*, Vol.17, No.3, pp.197-206, 2010.

[9] H.-Y. Lee, J.-Y. Baek, S.-G. Kong, H.-S. Lee, and J.-H. Choi, “Color Laser Printer Forensics through Wiener Filter and Gray Level Co-occurrence Matrix,” *Journal of KIISE: Software and Applications*, Vol.37, No.8, pp.599-610, 2010.

[10] M.-J. Tsai, J. Liu, C.-S. Wang, and C.-H. Chuang, “Source color laser printer identification using discrete wavelet transform and feature selection algorithms,” in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp.2633-2636, 2011.

[11] D.-G. Kim, J.-U. Hou, and H.-K. Lee, “Learning deep features for source color laser printer identification based on cascaded learning,” arXiv preprint arXiv:1711.00207, 2017.

[12] V. Dumoulin, and F. Visin, “A guide to convolution arithmetic for deep learning,” arXiv preprint arXiv:1603.07285, 2016.

[13] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the International Conference on Machine Learning*, pp.807-814, 2010.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, Vol.15, No.1, pp.1929-1958, 2014.

[15] S. Ioffe, and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the International Conference on Machine Learning*, pp.448-456, 2015.

[16] A. Tuama, F. Comby, and M. Chaumont, “Camera model identification with the use of deep convolutional neural networks,” in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, pp.1-6, 2016.

[17] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.3, No.6, pp.610-621, 1973.

이 수 현

https://orcid.org/0000-0002-3372-5660
e-mail : dark0487@kumoh.ac.kr
2018년 금오공과대학교
컴퓨터소프트웨어공학과(학사)
2018년~현 재 금오공과대학교
소프트웨어공학과 석사과정
관심분야 : Image Processing, Deep Learning

이 해 연

https://orcid.org/0000-0002-6081-1492
e-mail : haeyeoun.lee@kumoh.ac.kr
1997년 성균관대학교 정보공학과(학사)
1999년 한국과학기술원 전산학과(공학석사)
2006년 한국과학기술원 전자전산학과
(공학박사)
2008년~현 재 금오공과대학교 컴퓨터소프트웨어공학과 교수
관심분야 : IoT, Image Processing, Digital Forensics