위너 필터와 명암도 동시발생 행렬을 통한 컬러 레이저프린터 포렌식 기술

(Color Laser Printer Forensics through Wiener Filter and Gray Level Co-occurrence Matrix)

이 해 연 † 백 지 연 †† 공 승 규 †† 이 흥 수 †† 최 정 호 †††
(Hae-Yeoun Lee) (Ji Yeoun Baek) (Seung Gyu Kong) (Heung-Su Lee) (Jung-Ho Choi)

요 약 고성능 디지털 인쇄기기의 대중화와 손쉬운 이미지 편집 프로그램들의 등장으로 인하여 위·변조 범죄가 증가함에 따라 여러 가지 사회적인 문제를 야기하고 있다. 이를 해결하기 위해서 디지털 포렌식 기술이 활발하게 연구되고 있다. 본 논문에서는 디지털 포렌식 기술의 한 분야인 컬러 레이저 인쇄기기 판별기술을 제안한다. 각 제조사마다 인쇄방법이 다르기 때문에 육안으로 판별할 수 없는 미세한 차이가 출력물에 존재한다는 점을 이용하였다. 출력물의 노이즈를 추정하여 이러한 미세한 차이를 분석하였으며, 제안하는 방법에서는 출력물을 스캔한 이미지에 대해 위너필터를 거쳐 노이즈를 제거한 이미지를 차감하여 노이즈를 추출한다. 계산된 노이즈 대해 명암도 동시발생 행렬을 계산하여 특징값들을 추출한 뒤 이를 서포트 벡터 머신 분류기에 적용하여 인쇄기기를 판별하였다. 제안한 알고리즘의 성능을 분석하기 위하여 7대 프린터에서 각 371장씩 출력된 총 2,597장 이미지로 실험하였다. 제안한 알고리즘은 컬러 디지털 인쇄기기의 제조사를 판별하는데 있어서 97.6%의 정확률을 보였고, 동일 제조사의 모델을 판별하는데 84.5%의 정확률을 나타냈다.

키워드 : 디지털 포렌식, 위너필터, 명암도 동시발생 행렬, 서포트 벡터 머신 분류기

Abstract Color laser printers are nowadays abused to print or forge official documents and bills. Identifying color laser printers will be a step for media forensics. This paper presents a new method to identify color laser printers with printed color images. Since different printer companies use their own printing process, each of printed papers from different printers has a little different invisible noise. After the wiener-filter is used to analyze the invisible noises from each printer, we extract some features from these noises by calculating a gray level co-occurrence matrix. Then, these features are applied to train and classify the support vector machine for identifying the color laser printer. In the experiment, we use total 2,597 images from 7 color laser printers. The results prove that the presented identification method performs well using the noise features of color printed images.

Key words : Digital Forensics, Wiener Filter, Gray Level Co-occurrence Matrix, Support Vector Machine Classifier

· 본 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0007789)

† 정 회 원 : 금오공과대학교 컴퓨터공학부 교수
haeyeoun.lee@kumoh.ac.kr
(Corresponding author임)
†† 학생회원 : 금오공과대학교 컴퓨터공학부
jiyeoun1128@naver.com
lhslife@kumoh.ac.kr
kkongs86@gmail.com
††† 학생회원 : KAIST 전산학과
jhchoi@mmc.kaist.ac.kr

논문접수 : 2010년 3월 29일
심사완료 : 2010년 6월 8일

Copyright©2010 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 소프트웨어 및 응용 제37권 제8호(2010.8)

1. 서 론

컴퓨터의 대중화로 인해 우리 사회에서는 고성능의 컴퓨터와 주변기기들을 저렴한 가격에 구매할 수 있다. 또한 다양한 소프트웨어의 개발로 인해 초보자들도 컴퓨터를 이용하여 복잡한 작업을 쉽게 할 수 있다. 이로 인한 긍정적 측면이 있지만 부정적 측면도 있다. 그 대표적인 예가 위·변조 범죄이다. 그림 1에 나타난 것과 같이 해가 거듭할수록 다양한 종류의 위·변조 범죄가 증가하는 것을 볼 수 있다. 이는 '포토샵'이나 '일러스트레이터'등 이미지 편집 프로그램의 발전으로 인해 손쉽게 위·변조가 가능하기 때문이다. 또한 위·변조 기술이 점점 교묘해지면서 일반인이 위·변조를 판별할 확률은 1% 정도로 매우 희박하다.

이미지 위·변조의 증가로 인해 이미지에 대한 신뢰성이 점차 중요해지고 있다. 이러한 이미지의 신뢰성을 강화하기 위해 디지털 이미지 포렌식 기술이 활발하게 연구되고 있다. 디지털 이미지 포렌식 기술은 영상에 숨겨져 있는 고유의 특성을 찾아내 이미지를 만들어 낸 기기를 밝혀낸다든지 이미지 위조 여부를 판별하는 기술로, 영상 획득 장비 판별, 영상 위·변조 탐지 기술로 크게 나뉘어져 연구되고 있다.

영상 획득 장비 판별기술은 주어진 영상으로부터 영상을 획득하는데 사용된 장비가 무엇인지 판단하는 기술이다. 주어진 영상만을 이용해서 영상을 획득하는데 사용한 장비의 제조사 뿐만 아니라 정확한 모델명 수준까지 판별하는 것을 목표로 한다.

영상 위·변조 탐지 기술은 주어진 영상이 불법적인 위·변조 단계를 거쳤는지 여부를 판별하는 기술이다. 디지털 이미지의 위·변조는 눈에 띄는 단서를 남기지 않지만, 영상을 구성하는 통계적 특성에 변화를 준다. 따라서 변화된 통계적 특성을 찾음으로써 위·변조 여부를 찾을 수 있다.

본 논문에서는 디지털 이미지 포렌식 기술 중에서 영상 획득 장비에 해당하는 컬러 레이저프린터 판별기술에 대하여 설명한다. 지금까지 연구되어진 프린터 판별 방법 기술은 주로 흑백 레이저프린터를 판별하는 기술들이 대부분이다. 하지만 최근 컬러 레이저프린터 보급률이 높아지면서 컬러 레이저프린터에 적합한 판별기술의 필요성이 대두되고 있고, 본 연구에서는 컬러 레이저프린터에 대한 새로운 판별 기술을 제안하였다. 특히, 위너필터와 명암도 동시발생행렬을 이용하여 컬러 레이저프린터의 고유한 특징을 계산하여 활용하였다. 컬러 레이저프린터 판별기술을 통해, 출력 된 디지털 이미지가 어떠한 프린터에서 출력되었는지 알 수 있다면 위·변조 범죄를 일으킨 범인을 찾는데 도움이 될 수 있으며, 나아가 위조지폐 및 위조 공문서를 쉽게 식별할 수 있을 것이다. 이와 관련하여 2절에서 자세히 알아보도록 하겠다.

본 논문의 구성은 다음과 같다. 2절에서는 프린터 판별과 관련된 기술들에 대하여 설명하고, 3절에서는 위너필터와 명암도 동시발생 행렬을 활용한 프린터 판별기술을 제안한다. 4절에서는 다양한 프린터를 사용한 실험결과를 제시하고, 5절에서 결론을 짓도록 하겠다.

2. 관련 연구

출력문서를 이용하여 프린터를 판별하는 방법은 크게 수동적 방법과 능동적 방법 두 가지가 있다. 수동적 방법은 프린터의 내부에서 물리적 특성으로 인해 필연적으로 발생하는 출력물 상에서의 보이지 않는 특징들을 찾아서 그것을 인쇄한 프린터를 찾는 방식으로 내재적 특성(intrinsic signature)을 이용하여 프린터를 판별한다. 능동적 방법은 프린터의 인쇄 과정 중 매개변수를 직접 수정해서 출력물에 보이지 않는 특정한 정보를 직접 삽입하는 방식으로 외재적 특성(extrinsic signature)을 이용하여 프린터를 판별한다. 능동적 방법을 이용하여 문서에 특정한 정보를 삽입하는 것에 대한 소비자들

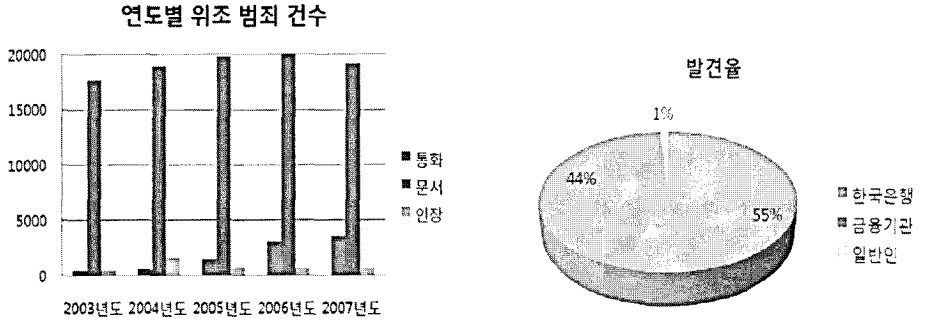


그림 1 연도별 위·변조 범죄 현황 및 발견률

의 야기할 수 있기 때문에 현재 능동적 방법보다는 수동적 방법에 맞추어 프린터 판별기술 연구가 진행되고 있다.

2.1 프린터 스테가노그라피

프린터 스테가노그라피란 능동적인 프린터 판별방법의 일종으로, 제3의 사용자로부터 정보가 불법적으로 사용되거나 변조 되는 것을 막기 위해 인쇄된 종이에 노란 점을 출력하는 기술이다[1,2]. 이 노란 점은 그냥 눈으로 봐서는 잘 보이지 않으며, 고해상도 스캐닝 장비나 특수한 광선을 비추어 암호를 해독할 수 있다. 이 노란 점에는 프린터의 고유한 시리얼 정보 및 인쇄한 날짜 등 다양한 정보가 들어가 있다.

그림 2의 좌측 이미지에서 노란 점들이 출력되는 위치로 출력된 시간, 날짜, 장비번호 등의 정보를 나타는 것을 볼 수 있으며, 그림 2의 우측 이미지에서 프린터 스테가노그라피를 사용하여 출력된 문서의 모습을 볼 수 있다. 이 기술은 레이저프린터로 출력한 문서에 사적인 정보 내용을 노란 점으로 암호화하여 심은 것이 심각한 프라이버시 침해라는 이유로 사용자들에게 반발을 받았다. 또한 눈에 잘 띄지 않는다고 해도 일단 출력물에 인쇄가 되기 때문에 화질 및 인쇄물의 품질 저하가 일어날 수 있다. 따라서 최근에는 기계에 종속적이지 않은 디지털 포렌식 기술 연구를 하는 방법이 연구되고 있다.

2.2 줄무늬 주기를 이용한 프린트 판별기술

레이저프린터의 인쇄 과정 중 광학 광전도체 드럼을 돌리는 기어의 역회전 현상으로 인해 각속도가 일정하지 않게 됨에 따라 발생하는 준 주기적인 변동 때문에 인쇄 방향을 따라서 밝고 어두운 줄무늬가 주기적으로 생긴다. 이렇게 문서에 새겨지는 줄무늬 주기는 프린터의 기계적인 특성을 반영하고 있다고 할 수 있다. Delp et al.는 이와 같은 주기적인 특성을 활용하여 흑백 레이저프린터를 판별하는 알고리즘을 제안하였다[3]. 그림 3은 레이저프린터 인쇄물에서 나타날 수 있는 줄무늬를 나타내고 있다[4].

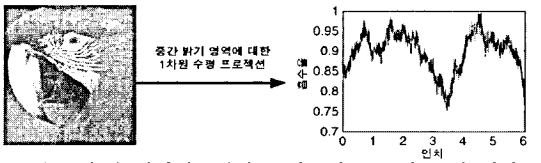


그림 3 흑백 레이저프린터 출력물의 줄무늬 주기 현상

일반적으로 하프톤 이미지 출력물을 이용하여 줄무늬 주기를 구하는 방법은, 프린터의 줄무늬 주기를 구하는 방법처럼 수평 방향으로 투사한 다음 이진푸리에변환을 이용하는 방법을 그대로 이용하면 된다. 프린터 판별방식은 어떠한 프린터로 인쇄했는지 모르는 하프톤 이미지 출력물의 줄무늬 주기를 위의 방법으로 구한 다음 그 줄무늬 주기와 일치하는 줄무늬 주기를 갖는 프린터를 찾는 것이다.

이러한 줄무늬 주기는 하프톤 이미지 출력물에는 적용하기가 쉽고 결과도 좋게 나오지만 하프톤 이미지 출력물이 아닌 일반 텍스트 출력물에는 적용하기 힘들다는 단점이 있다.

2.3 거리 변환을 이용한 프린터 판별방법

레이저프린터로 프린터 된 문서들을 눈으로 볼 때 원거리에서는 동일하게 보이지만, 크게 확대해서 보면 프린터의 품질에 따라 구별되는 특징을 볼 수 있다. 그림 4는 그 예를 보여주고 있다.

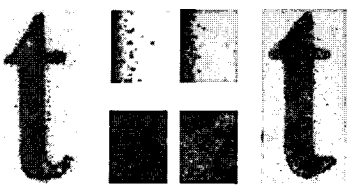


그림 4 프린터 품질에 따른 구별되는 특징

거리 변환을 이용한 프린터 판별방법에서는 문자 이미지 매칭과 프린트된 문서에 존재하는 글자 사이의 거리를 측정하는데 거리 변환 알고리즘을 사용한다. 그림 4

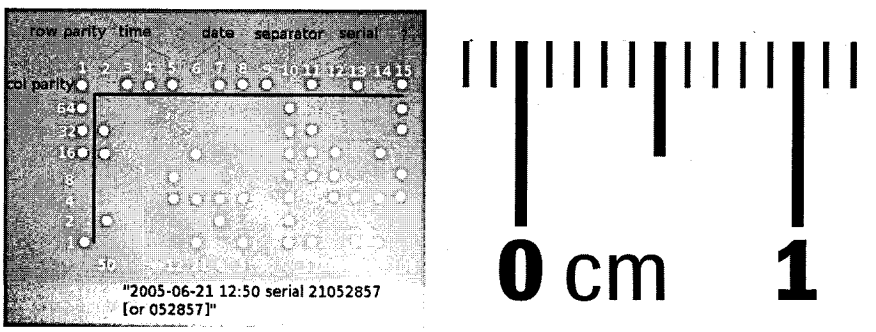


그림 2 프린터 스테가노그라피의 예

와 같이 프린터마다 미세하게 다른 출력 결과의 특징을 분석하여 프린터 판별에 이용한다[5].

문자 이미지 매칭은 일반적으로 이진 이미지를 이용한 거리 변환에 기반하고 있다. 이진 이미지 $I(x,y)$는 오직 0과 1의 두 가지 값을 가지고 있으며, 0은 글자의 픽셀, 1은 배경 픽셀을 의미한다. $I_d(x,y)$는 이미지의 거리 변환을 나타내며, 이것은 글자의 픽셀 (x,y)과 이 픽셀과 가장 가까운 배경 픽셀 간의 거리를 의미한다. 거리 변환의 계산은 two-pass 알고리즘을 사용하여 수행된다. 거리 변환 이후 두 이진 이미지 $f(x,y)$, $g(x,y)$의 거리 차이는 아래 식을 통해 계산할 수 있다.

$$D_{fg} = \frac{1}{N_f}\sum_{y=1}^{H}\sum_{x=1}^{W} f(x,y)g_d(x,y)$$

H, W는 각각 이미지의 높이와 폭을 나타내며, N_f는 이미지 $f(x,y)$ 안의 검정 픽셀의 수이다. $g_d(x,y)$는 $g(x,y)$ 이미지의 거리 변환된 것이다. 매칭 방향에 따라서 다른 매칭 거리를 얻을 수 있기 때문에, 두 매칭 방향 D_{fg}, D_{gf}를 계산하여 조합하여 사용한다.

사전에 알려진 프린터들로 인쇄한 글자를 OCR 기능을 이용하여 문자를 인식하고, 인식된 문자를 데이터베이스에 등록해놓는다. 그 후에 프린터 판별을 위해 이미지가 들어오면 문자를 인식하고, 데이터베이스와 비교하여 거리 변환 방법을 통하여 거리 차이값을 계산한다. 이렇게 계산한 거리 차이값에 대하여, 최소 거리차를 갖는 경우의 문자들을 포함한 프린터를 해당 문서를 인쇄하는데 사용한 프린터로 판별하게 된다. 이와 같은 프린터 식별 시스템의 구조는 그림 5에서 보여주고 있다.

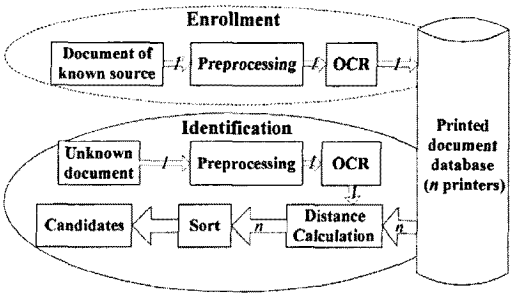


그림 5 거리 변환을 이용한 프린터 식별 방법

2.4 이산 웨이블릿 변환을 이용한 프린터 판별기술

이산 웨이블릿 변환을 이용한 프린터 판별기술은 컬러 레이저프린터마다 디지털 이미지를 실제 문서로 출력할 때 색을 재구성하는 방식이 조금씩 다르다는 점에 착안한 프린터 판별방법으로 Choi가 제안하였다[6]. 이는 기존에 있던 흑백 레이저프린터 판별기술과는 달리, 컬러 문서의 컬러 정보를 이용하여 컬러 레이저프린터를 판별하는 기술이다.

이산 웨이블릿 변환을 이용한 프린터 판별과정은 그림 6에 도시하였다. 여러 대의 컬러 프린터로 인쇄한 출력물들을 하나의 스캐너로 스캔한 다음, 기본 컬러 영역인 RGB영역에서 R, G, B 각각의 영역에 대해서 이산 웨이블릿 변환을 수행한다.

이산 웨이블릿 변환을 거친 후 만들어지는 4개의 영역 중 그림 7에서의 HH는 가로와 세로방향으로 고주파 필터를 거쳐서 만들어진 이미지로 원본 이미지의 노이즈가 남아 있다. 이 노이즈를 통계적으로 분석하여 특징점을 추출하게 된다.

이산 웨이블릿 변환으로 변환된 각각의 R, G, B 영역 중에서 노이즈 성분이 있는 HH 영역에서의 표준편차, 왜도, 첨도, R과 G, G와 B 및 R과 B 밴드의 공분산과 상관계수를 구하여 그것들을 하나의 특징점 벡터로 정한다. RGB영역에서 표준편차, 왜도, 첨도, 공분산, 상관계수 등 총 15개의 특징점을 추출할 수 있다. 각각의 특징점에 대한 식은 표 1에 정리되어 있다. 이 과정은 YCbCr 영역에서도 똑같이 이루어진다. 그리하여 두 영역에서 추출한 총 30개의 특징점을 하나의 벡터로 합친

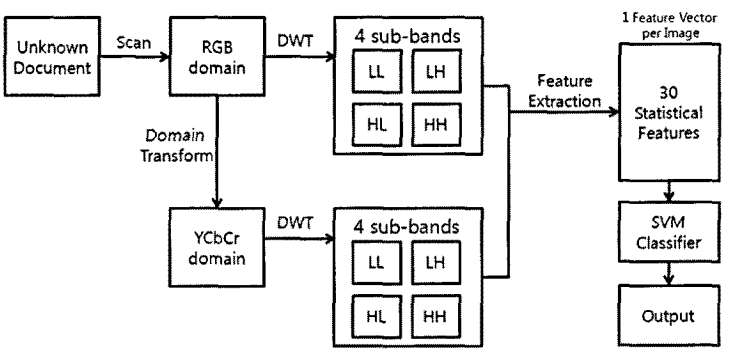


그림 6 이산 웨이블릿 변환을 이용한 컬러 프린터 판별과정

표 1 이산 웨이블릿 변환을 이용한 컬러 레이저프린터 판별방법의 특징점

특징점	식
표준편차	$s=\left(\frac{1}{N-1}\sum_{i,j=1}^{N}(I(x,y)-\overline{I(x,y)})^2\right)^{\frac{1}{2}} where\ \overline{I(x,y)}=\frac{1}{N}\sum_{i,j=1}^{N}I(x,y)$
첨도	$skewness=\frac{1}{N\cdot s^3}\sum_{x,y=1}^{N}\left((I(x,y)-\overline{I(x,y)})^3\right)$
왜도	$kurtosis=\frac{1}{N\cdot s^4}\sum_{x,y=1}^{N}\left((I(x,y)-\overline{I(x,y)})^4\right)$
공분산	$Cov(I_1,I_2)=\frac{1}{N}\sum_{x,y=1}^{N}I_1(x,y)\cdot I_2(x,y)-\frac{1}{N_1}\sum_{x,y=1}^{N_1}I_1(x,y)\cdot\frac{1}{N_2}\sum_{x,y=1}^{N_2}I_2(x,y)$
상관계수	$Corr(I_1,I_2)=\frac{Cov(I_1,I_2)}{\sqrt{Var(I_1)\cdot Var(I_2)}}$

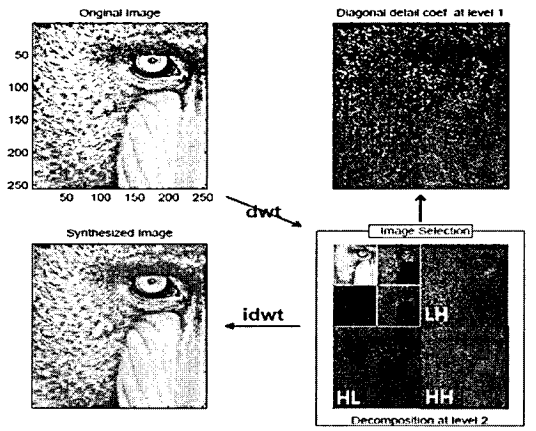


그림 7 이산 웨이블릿 변환을 거친 후 만들어지는 HH 영역

다. 각 이미지마다 하나의 특징점 벡터를 생성한 다음, 서포트 벡터 머신을 이용하여 생성한 특징점 벡터들 중 절반은 학습에 이용하고 나머지 절반은 문서를 출력한 컬러 레이저프린터를 판별하는데 이용한다.

3. 제안하는 컬러 레이저프린터 판별기술

본 논문에서 제안하는 컬러 레이저프린터 판별기술은 그림 8과 같이 크게 훈련과정과 판별과정 두 가지로 나뉜다. 훈련 과정에서는 판별하고자 하는 프린터로 인쇄한 이미지를 사용하여 각 프린터의 고유한 특징을 추출하고, 학습기반 데이터 분류기의 파라미터로 입력하여 훈련을 수행한다. 판별과정에서는 임의의 데이터가 들어왔을 때, 특징을 추출하고, 학습기반 데이터 분류기에 입력하여, 어떠한 인쇄기기로 출력을 했는지 판별을 수행한다.

컬러 레이저프린터 판별 알고리즘의 성능은 프린터를 고유하게 판별할 수 있는 특징 추출에 영향을 크게 받으므로, 본 연구에서는 프린터 이미지에 존재하는 고유한 특징을 추출하기 위하여 다양한 특징점에 대한 연구를 수행하였다. 최적의 결과를 나타내고 각 프린터의 고유한 특징을 정확하게 판별하는 방법은 그림 8의 특징점 추출기에서 특징값을 추출하는 방법과 같다.

3.1절에서는 본 연구에서 제안하는 위너 필터를 이용한 잡음을 추출하여, 이에 대해 명암도 동시발생 행렬을 이용한 특징의 추출 방법을 설명하고, 3.2절에서는 학습기반 데이터 분류기에 대해서 설명한다.

3.1 위너 필터와 명암도 동시발생 행렬을 활용한 특징 추출

컬러 레이저프린터에 존재하는 고유한 특징을 추출하기 위하여, 먼저 컬러 레이저프린터로 인쇄된 디지털 이미지를 스캐너로 스캔한 후 RGB 영역에서 CMY 영역으로 변환한 후 위너 필터를 거쳐 노이즈가 제거된 영상을 생성한다. 원본 영상과 노이즈가 제거된 영상의 차이를 통해 노이즈만을 추출하였다. 분류 정확도를 향상시키기 위해 추출한 노이즈에 대해 0°, 45°, 90°, 135° 4가지 방향에 대해 명암도 동시발생 행렬을 계산 후, 이에 대하여 질감분석을 위한 대조, 상관관계, 합계, 유사성, 공분산을 계산하여 이를 특징점으로 사용하였다.

사용하는 도메인 영역에 대하여 CMY 영역을 사용하였다. 프린터가 문서를 출력할 때는 컴퓨터 상에서 문서를 표현할 때 이용하는 RGB 영역에서, 인쇄 과정에 이용하는 CMY 영역으로 변환한 뒤 출력한다. 그렇기 때문에 다시 스캔을 한 RGB 영역에서 CMY 영역으로 변환시키면 프린터로 출력될 때 생기는 잡음에 대한 훼손이 적기 때문에 다른 도메인에 비해 프린터마다의 차이가 잘 드러난다. CMY는 Cyan, Magenta, Yellow의 3가지 색상을 나타내며. CMY 영역으로 표현한 예제는 그림 9와 같다.

위너필터는 비정상성 잡음을 제거하는 데에 유용하다. 그림 10은 CMY 영역으로 변형된 이미지가 위너 필터를 통해 잡음이 제거된 모습이다. 그림 10(b)는 그림 10(a)에 비해 잡음이 제거되어 이미지가 부드러워진 것

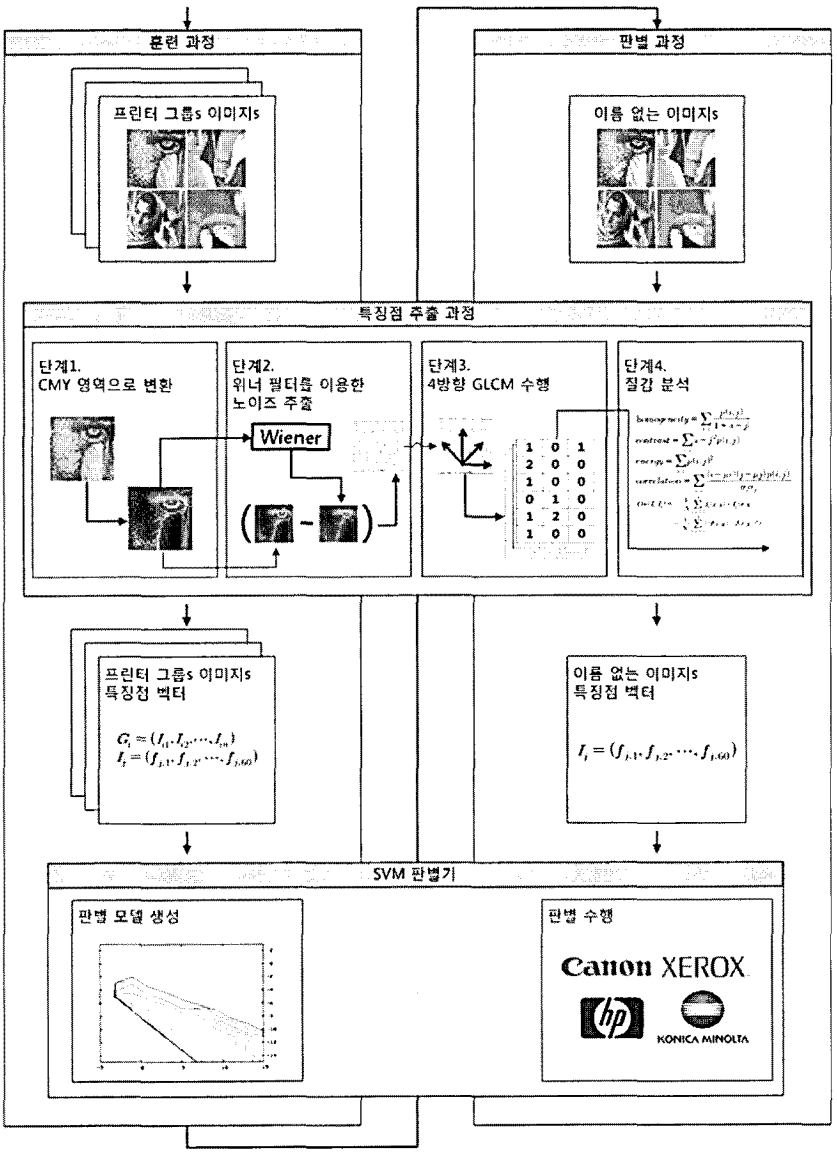
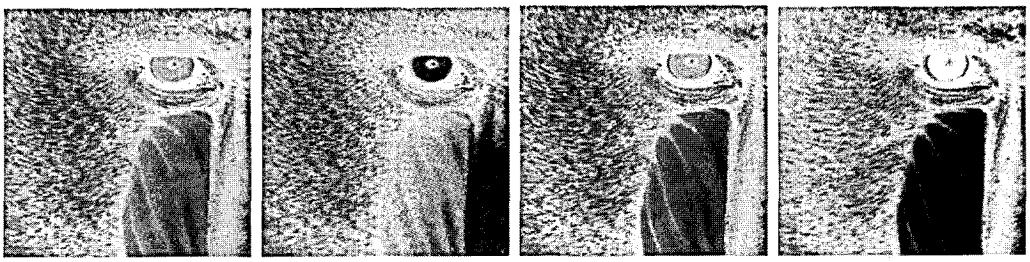


그림 8 제안하는 컬러 레이저프린터 판별과정



(a) CMY 이미지 (b) C (c) M (d) Y

그림 9 CMY 영역으로 표현된 이미지

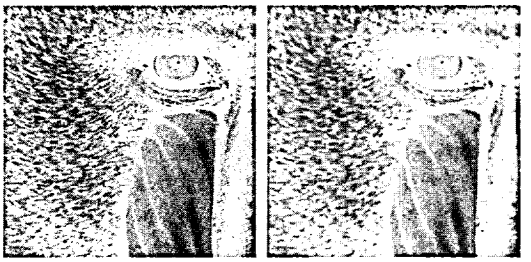
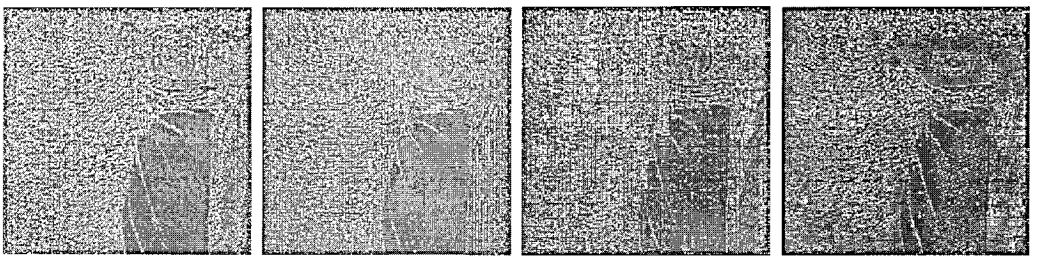


그림 10 CMY 이미지와 위너필터를 거친 잡음이 제거된 이미지

을 볼 수 있다. CMY 영역으로 변형한 이미지를 위너 필터를 사용하여 의도치 않게 들어간 값이나 튀는 값들을 제거하여 새로운 이미지를 만든다. 원본 이미지에서 잡음이 제거된 이미지의 차이를 계산함으로써 원래 이미지에 의도치 않게 들어간 잡음들을 구할 수 있다. 이 잡음들은 그림 11과 같이 각 제조회사별로 차이가 나기 때문에 프린터를 판별할 수 있는 기준이 될 수 있다.

화소 기반의 통계적 텍스쳐 영상 생성을 위한 기본 개념은 Haralick et al.에 의해 제안된 명암도 동시발생 행렬(Gray Level Co-occurrence Matrix)을 통해 체계화 되었다[7,8]. 텍스쳐 영상의 응용 연구에서는 영상 분류 단계에서 텍스쳐 영상을 추가적으로 적용하는 경우, 정성적인 분류결과와 정량적인 분류 정확도의 향상에 텍스쳐 영상이 유용한 자료임이 여러 사례 연구에서 밝혀졌다. 그러나 텍스쳐 영상이 유용함에도 불구하고, 텍스쳐 영상의 선택 기준에 대한 연구 결과는 실제적으로 제시된 바가 거의 없다.

영상의 텍스쳐 특성을 잘 드러나게 하는데 명암도 동시발생 행렬을 이용할 수 있다. 명암도 동시발생 행렬은 영상 내에서 픽셀간의 밝기 값의 관계를 나타내는 방법으로 원본 이미지 그림 12과 같이 방향과 두 점 사이의 거리로 인접한 두 픽셀의 값을 좌표 값으로 이용하여 그러한 좌표 값이 얼마나 자주 나오는지 세어 놓은 행



(a) Xerox (b) HP (c) Canon (d) Konica

그림 11 위너필터를 사용한 각 제조회사별 잡음

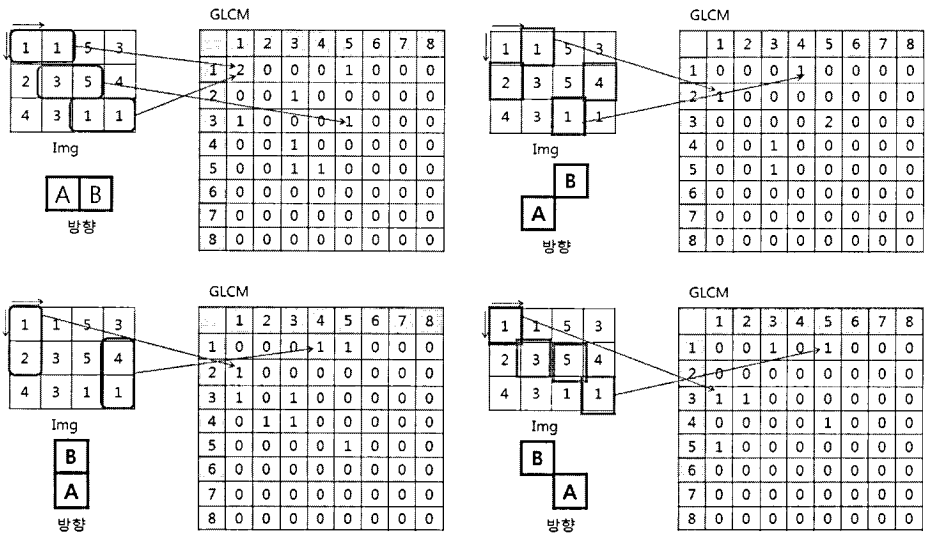


그림 12 명암도 동시발생 행렬 방향성의 예제

렬이다. 이 행렬은 영상의 텍스쳐 특성을 하나의 행렬로 표현한 것으로, 원본 영상의 크기에 상관없이 항상 고정된 크기를 갖는다. 따라서 영상의 텍스쳐 특성을 통계적으로 분석하는데 있어서 유용하고, 각 프린터의 잡음 특성을 잘 표현할 수 있기 때문에 본 연구에서 도입하였다. 특히, 분류 정확도의 향상시켜주는 명암도 동시발생 행렬의 장점을 활용하기 위해 위너 필터를 통해 구한 노이즈를 0°, 45°, 90°, 135° 4가지 방향에 대해 명암도 동시발생 행렬을 계산하였다.

명암도 동시발생 행렬으로 변형된 영역에 대하여 표 2에 나타난 것과 같이, 유사성, 대조, 합계, 상관관계, 공분산의 통계적인 특성을 사용하여 각 특징점을 추출하였다.

통계적인 특성으로는 4가지 방향성에 대한 명암도 동시발생 행렬 변환을 수행하여 얻어진 C, M, Y 각 영역에 대한 유사성, 대조, 합계, 상관분산과 C와 M, M과 Y, Y와 C 영역간의 공분산이 있다. 이와 같이 추출된 60가지의 특징점들은 하나의 벡터로 합쳐져 학습기반 데이터 분류기에 컬러 레이저프린터를 판별하는데 훈련 및 판별과정 데이터로 이용된다.

3.2 학습기반 데이터 분류기

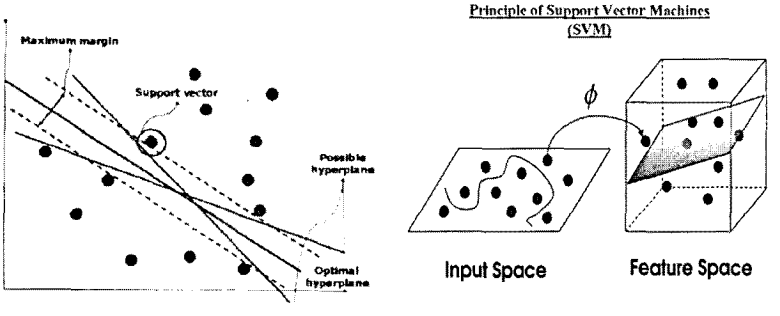
학습기반 데이터 분류기로 Support Vector Machine(SVM)은 1995년 통계학자인 Vladimir Vapnik에 의하여 개발되고 제안된 학습 알고리즘으로 그림 13(a)의 모습과 같이 이진분류를 하기 위한 목적으로 개발되었다[9]. 이진분류란 수집된 훈련 데이터를 이용하여 구조적 위험 최소화의 개념에 기반하여 최적의 선형 결정평면을 찾는 것으로, 현재 SVM은 생물정보학, 문자인식, 필기인식, 얼굴 및 물체 인식 등 다양한 분야에서 성공적으로 적용되고 있다.

SVM은 그림 13(b)와 같이 입력 벡터를 고차원의 특징 공간으로 사상시켜 서로 다른 클래스로 분류할 수 있는 초평면 중에서도 여백을 최대화시키는 특정한 초평면인 최대 여백 초평면을 찾는 것을 목적으로 한다. 최대 여백 초평면에 가장 근접한 훈련 데이터를 Support Vector라고 부르며, 이를 포함하는 초평면 사이의 여백 값이 클수록 분류 성능이 좋아진다.

SVM은 분리 경계가 매우 복잡한 문제를 선형 판별함수의 사용이 가능한 단순한 문제로 변환시키기 때문에 수학적인 분석이 수월하다는 장점을 갖고 있다. 또한

표 2 명암도 동시발생 행렬에 대한 특징점

특징점	식	설명
유사성	$\mathrm{hom}ogeneity = \sum_{i,j} \frac{p(i,j)}{1+\lvert i-j \rvert}$	행렬 내 각 화소들 간의 균일한 정도를 표현
대조	$contrast = \sum_{i,j} \lvert i-j \rvert^2 p(i,j)$	화소간의 명암도 차이를 표현
합계	$energy = \sum_{i,j} p(i,j)^2$	화소들의 합을 표현
상관관계	$correlation = \sum_{i,j} \frac{(i-\mu i)(j-\mu j)\vec{p}(i,j)}{\sigma_i \sigma_j}$	화소간의 상관관계를 표현
공분산	$Cov(I_1, I_2) = \frac{1}{N} \sum_{x,y=1}^{N} I_1(x,y) \times I_2(x,y) - \frac{1}{N_1} \sum_{x,y=1}^{N} ((I(x,y) - \overline{I(x,y)})^3)$	전체 이미지에서 두 픽셀이 연관된 정도를 표현



(a) (b)

그림 13 SVM을 이용한 (a) 이진 분류 및 (b) 사상의 예

적은 양의 학습 자료만으로도 분류를 잘 수행한다. SVM은 학습오류를 최소화하는 경험적 위험 최소화를 기본으로 하는 기존의 원리에 비해, 전체집단을 하위집단으로 세분화 한 뒤 이 집단에 대한 경험적 위험도를 최소화하는 구조적 위험 최소화 개념에 기초하여 일반화 성능이 우수하며 비선형 분리 문제와 같은 복잡한 분류문제를 해결하기 위해 다양한 커널을 제공한다.

SVM light, LIBSVM, SVM Torch 등 다양한 소프트웨어들이 공개적으로 사용 가능하고, 다른 기계학습 방법들보다 상대적으로 사용하기 쉽다. SVM을 사용할 경우 커널의 선택과 그에 따른 커널 변수를 선택하기만 하면 된다. 또한 새로운 패턴이 나타나면 훈련 과정에 모델을 동적으로 갱신할 수 있다.

본 연구에서는 서포트 벡터 머신 중에서도 다중 클래스 분류를 할 수 있게 해 주는 LIBSVM를 사용하였다[10,11]. LIBSVM은 unbalanced 클래스에 대한 penalty 부가 기능을 가지고 있어, 여러 그룹을 확실하게 분류하기 위해 사용자가 초평면을 정하는데 있어서 분류 오류를 최소화 시킬 수 있다.

커널함수는 이론적으로는 가능하나 실제로는 풀기 힘든 매핑함수를 대신하여 원 자료를 고차원 공간으로 이동시켜 특정 공간 내에 선형으로 분리 가능한 입력자료 집합을 만들어주는 역할을 수행한다. 커널함수는 다음과 같이 정의된다.

$$K(\chi_i,\chi_j)=\phi(\chi_i)^T\phi(\chi_j)$$

일반적으로 커널 함수는 다음 표 3과 같이 Linear, Polynomial, Radial basis function, Sigmoid의 네 가지 종류가 있다. 여기서 γ, r, d는 각 커널에서의 파라미터를 의미한다.

표 3 서포트 벡터 머신의 대표적인 커널 함수

커널 함수	식
Linear	$K(\chi_i,\chi_j)=\chi_i^T\chi_j$
Polynomial	$K(\chi_i,\chi_j)=(\gamma\chi_i^T\chi_j+r)^d,\ \gamma>0$
Radial basis function	$K(\chi,\chi_i)=\exp(-\frac{1}{2\sigma^2}\|\chi-\chi_i\|^2)$
Sigmoid	$K(\chi_i,\chi_j)=\tanh(\gamma\chi_i^T\chi_j+r)$

본 연구에서는 Radial Basis Function(RBF) 커널을 사용하였다. RBF 커널은 고차원의 특징공간으로 입력벡터를 비선형적으로 이동시키기 때문에 선형 분리가 불가능한 문제를 해결하는데 매우 유용하다. RBF 커널은 SVM의 예측성과는 라그랑지 승수의 허용오차 상한값인 C와 커널 모수인 γ에 따라 영향을 받게 된다[12,13]. 따라서 두 가지 모수가 적절하게 선택되지 않을 경우 과대적합이나 과소적합 현상이 발생할 수 있다.

4. 실험결과

실험에서는 총 7대 프린터(HP, Canon, Xerox DCC 400, Xerox DCC 450, Xerox DCC 4550, Xerox DCC 6540, Konica)에서 각 371장씩 출력된 총 2,597장의 이미지 셋을 이용하였다. 먼저, 이미지 데이터를 취득하기 위하여 인터넷 검색, 사진 촬영을 통하여 371 장의 영상을 취득하였고, 이를 7대의 컬러 레이저프린터를 사용하여 각각 인쇄하였다. 그 후에 인쇄된 출력물을 EPSON Perfection 2400 스캐너로 300 dpi 해상도로 스캔하여 256×256 픽셀 크기로 분할하여 실험에 사용하였다. 그림 14는 실험에 사용한 이미지 세트의 일부이다.

제안한 알고리즘의 성능을 분석하기 위해 프린터 제조사에 대한 판별과, 동일한 제조사의 모델별 판별 두 가지 실험을 하였다. 컬러 레이저프린터 제조사별 판별을 위해 HP, Canon, Xerox DCC 450, Konica 4대의 프린터로 프린터 된 371장씩 총 1,484장의 이미지를 이용하여 실험을 진행하였으며, 동일 제조사 모델별 판별을 위해 Xerox DCC 400, Xerox DCC 450, Xerox DCC 5540, Xerox DCC 6540 모델의 4대의 프린터로 프린터 된 371장씩 총 1,484장의 이미지를 이용하여 실험하였다.

제안하는 알고리즘과 이산 웨이블릿 변환(DWT)을 이용하는 Choi 알고리즘을 각각 서포트 벡터 머신 분류기를 사용하여 테스트 하였다[6]. 성능 판별의 객관성

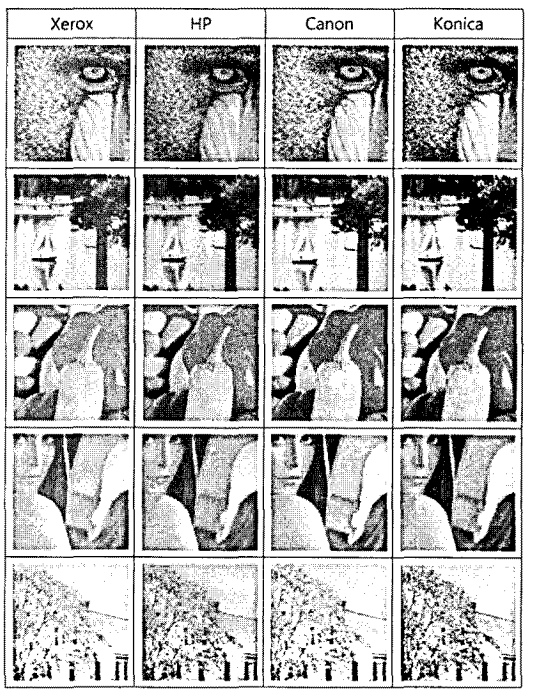


그림 14 실험에 사용된 영상의 예

및 정확도를 위하여 두 가지 실험 모두 각 371장씩 제조회사별로 출력하여 스캔된 1,484장의 이미지를 무작위로 선택하여 742장은 훈련 데이터로, 나머지 742장은 테스트 데이터로 사용하였다.

또한 실험의 오류를 최소화하기 위해 Cross-validation 기법을 사용하였다. 무작위로 선택한 훈련 데이터와 테스트 데이터를 바꾸어 테스트 데이터를 훈련 데이터로 쓰고 훈련 데이터를 테스트 데이터로 사용하여 2번의 실험을 하였다. 2번의 결과 값을 평균 내어 최종 결과를 계산하였다.

4.1 제조사별 판별 성능분석

동일한 실험환경 하에서 제안하는 알고리즘과 DWT를 이용하는 Choi의 알고리즘을 사용하여 제조사별 판별성능을 분석하였다. 실험에 사용된 컬러 레이저프린터의 제조사는 표 4에 정리되어 있다.

표 5에는 DWT를 이용하는 알고리즘과 제안하는 알고리즘을 이용하여 컬러 레이저프린터 제조사를 판별한 판별결과를 나타내었다. 제조사에 대한 판별율의 정확도에 대한 비교는 그림 15에 도시되어 있다.

실험 결과, 제안하는 방법의 프린터별 사이의 분류 결과는 98.2%로 DWT를 이용하는 알고리즘의 분류 결과인 90.82% 보다 7% 정도 높게 나왔다. 또한 Canon 프린터에 대해 DWT를 이용하는 알고리즘 보다 높은 판별률을 보였다.

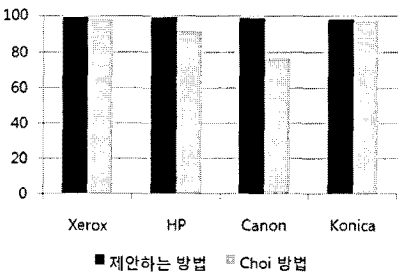


그림 15 컬러 레이저프린터 제조사 판별의 정확도 분석

4.2 동일 제조사 모델별 판별 성능분석

동일한 실험환경 하에서 제안하는 알고리즘과 DWT를 이용하는 Choi의 알고리즘을 사용하여 모델별 판별성능을 분석하였다. 실험에 사용된 컬러 레이저프린터의 모델명은 표 6에 정리되어 있다.

표 7에는 DWT를 이용하는 알고리즘과 제안하는 알고리즘을 이용하여 동일한 컬러 레이저 제조사에 대한 모델명을 판별한 판별결과를 나타내었다. 제조사에 대한 판별율의 정확도에 대한 비교는 그림 16에 도시되어 있다.

실험 결과, 제안하는 방법의 프린터별 사이의 분류 결과는 84.5%로 DWT를 이용하는 알고리즘의 분류 결과인 67.93% 보다 16% 정도 높게 나왔다. 또한 DCC450과 DCC5560, DCC6540 모델에 대해 DWT를 이용하는 알고리즘 보다 높은 판별률을 보였다.

표 4 제조사별 프린터 판별실험에 사용한 컬러 레이저프린터

Label	Printer Model
Xerox	Xerox DCC 400
HP	HP 4650
Canon	Canon iR C2620
Konica	Konica Minolta bizhub C250

표 6 모델별 프린터 판별실험에 사용한 컬러 레이저프린터

Label	Printer Model
DCC400	Xerox DCC 400
DCC450	Xerox DCC 450
DCC5560	Xerox DCC 5560
DCC6540	Xerox DCC 6540

표 5 컬러 레이저프린터 제조사 판별결과

	Xerox		HP		Canon		Konica	
	제안하는 방법	DWT 이용 방법	제안하는 방법	DWT 이용 방법	제안하는 방법	DWT 이용 방법	제안하는 방법	DWT 이용 방법
Xerox	98.9%	98.3%	0.3%	0.7%	0.3%	0.5%	0.5%	0.6%
HP	0.3%	7.0%	98.9%	91.4%	0.3%	1.6%	0.5%	0%
Canon	0%	15.6%	0.8%	1.1%	98.4%	76.3%	0.8%	7.0%
Konica	1.6%	2.2%	1.3%	0.2%	0.5%	0.2%	96.5%	97.3%

표 7 동일 제조사 컬러 레이저프린터 모델명 판별결과

	DCC400		DCC450		DCC5560		DCC6540	
	제안하는 방법	DWT 이용 방법	제안하는 방법	DWT 이용 방법	제안하는 방법	DWT 이용 방법	제안하는 방법	DWT 이용 방법
DCC400	96.5%	92.2%	1.9%	1%	0.5%	3.2%	1%	3.5%
DCC450	0.8%	0.8%	94.6%	72%	1.3%	14.6%	3.2%	12.7%
DCC5560	0.5%	6%	0.5%	4.6%	76.8%	61.5%	22.1%	28%
DCC6540	0.3%	7%	0.5%	11.3%	29.1%	35.6%	70%	46%

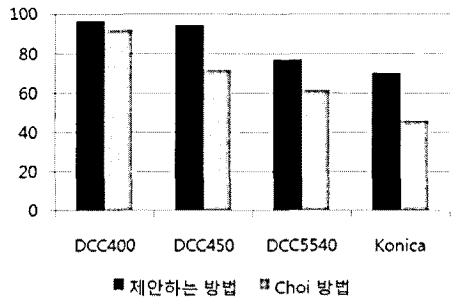


그림 16 동일 제조사 컬러 레이저프린터의 모델별 판별 정확도 분석

5. 결론 및 향후 과제

기존에는 컬러 레이저프린터가 아닌 흑백 프린터를 판별할 수 있는 기술에 대해 많은 연구가 되었다. 하지만 컬러 레이저프린터의 가격이 급격히 떨어지고 그 성능도 점차 발전함에 따라 컬러 레이저프린터의 수요가 늘고 있고, 이러한 점을 악용하여 컬러 레이저프린터를 이용한 오프라인 콘텐츠들의 위변조 범죄도 늘고 있는 추세이다. 따라서 이러한 범죄를 탐지하는데 조금의 기여라도 하고자 본 논문에서는 현재 개발되고 있는 컬러 레이저프린터에 대한 새로운 판별방법을 제안하였고, 실험 결과에 따르면 기존의 연구에 비하여 우수한 판별률을 달성하였다는 것을 알 수 있다.

실험에 따르면 동일 제조사 별 모델 판별에서 제조 공정이 유사할 것으로 판정되는 DCC5540과 DCC6540에 대해서는 판별률이 상대적으로 낮게 나왔다. 반면 서로 다른 제조회사에서 생산된 컬러 레이저 프린터에 대해서는 판별 결과가 높게 나왔다. 이를 통하여 서로 다른 컬러 레이저 프린터 회사는 서로 다른 프린터 제조 기술을 사용하지 때문에 본 논문에서 제안한 특징점인 컬러 노이즈의 통계적인 특징이 다르다는 것을 알 수 있다. 하지만 동일 제조회사의 프린터는 제조 공정이 유사할 가능성이 높기 때문에 본 논문에서 제안한 컬러 노이즈의 통계적인 특징만으로는 높은 판별률을 달성하기 어렵다는 것을 알 수 있다. 이에 대해 프린터 인쇄과정의 미세한 차이를 구분할 수 있는 최적화된 특징점을 찾을 수 있는 지속적인 연구가 필요하다. 또한, 다양한 불법 오프라인 컨텐츠를 추적하기 위해 사용된 토너 종류 판별, 프린터 모델명 판별, 영상을 획득한 디지털 카메라 모델명 판별 등의 연구로 응용할 필요가 있다.

참 고 문 헌

[1] Wikipedia, "Printer steganography," http://en.wikipedia.org/wiki/Printer_stegano graphy.

[2] Electronic Frontier Foundation, "DocuColor Tracking Dot Decoding Guide," http://w2.eff.org/Privacy/printers/docucolor/.

[3] A. K. Mikkilineni, G. N. Ali, P.-J. Chiang, G. T.-C. Chiu, J. P. Allebach and E. J. Delp, "Signature-embedding in printed documents for security and forensic applications," *Proc. of the SPIE Int. Conf. on Security, Steganography and Watermarking of Multimedia Contents*, vol.5306, pp.455-466, 2004.

[4] P.-J. Chiang, N. Khanna, A. K. Mikkilineni, M. V. O. Segovia, S. Suh, J. P. Allebach, G. T.-C. Chiu, and E. J. Delp, "Printer and Scanner Forensics," *IEEE Signal Processing Magazine*, vol.26(2), pp. 72-83, 2009.

[5] W. Deng, Q. Chen, F. Yuan, and Y. Yan, "Printer Identification Based on Distance Transform," *Proc. of Int. Conf. on Intelligent Networks and Intelligent Systems*, pp.565-568, 2008.

[6] J.-H. Choi, "Color laser printer identification using color noise features," MCS 09030, MS thesis, Korea Advanced Institute of Science and Technology, 2009.

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. on Systems, Man, Cybernetics*, vol.3, no.6, pp.610-621, 1973.

[8] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T.-C. Chiu, J. P. Allebach and E. J. Delp, "Printer identification based on graylevel co-occurrence features for security and forensic applications," *Proc. of the SPIE Int. Conf. on Security, Steganography and Watermarking of Multimedia Contents*, vol.5681, pp.430-440, 2005.

[9] Vapnik, V., "Statistical Learning Theory," *Springer*, NewYork, 1998.

[10] S.-J. Ryu, H.-Y. Lee, I.-W. Cho, and H.-K. Lee, "Document Forgery Detection with SVM Classifier and Image Quality Measure," *Lecture Notes in Computer Science*, vol.5353, pp.486-495, 2008.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2001.

[12] Hsu, C.-W., C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification," Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, 2004. Available at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdfs

[13] Tay, F.E.H. and L. Cao, L., Application of support vector machines in financial time series forecasting, Omega, vol.29(2001), pp.309-317.

이 해 연

1997년 성균관대학교 정보공학과(학사) 1999년 한국과학기술원 전산학과(공학석사). 2006년 한국과학기술원 전자전산학과 전산학전공(공학박사). 2001년~2006년 (주)쎄트렉아이 선임연구원. 2006년~2007년 코넬대학교 박사후연구원. 2008년~현재 국립금오공과대학교 컴퓨터공학부 교수. 관심분야는 멀티미디어, 영상처리, 콘텐츠보안, 디지털워터마킹 등

백 지 연

2007년~현재 국립금오공과대학교 컴퓨터공학부 재학 중. 관심분야는 멀티미디어, 디지털포렌식 등

공 승 규

2005년~현재 국립금오공과대학교 컴퓨터공학부 재학 중. 2009년~현재 삼성전자소프트웨어멤버십 회원. 관심분야는 멀티미디어, 디지털포렌식 등

이 홍 수

2004년~현재 국립금오공과대학교 컴퓨터공학부 재학 중. 2009년~현재 삼성전자소프트웨어멤버십 회원. 관심분야는 멀티미디어, 디지털포렌식 등

최 정 호

2007년 한국과학기술원 전산학과(학사) 2009년 한국과학기술원 전산학과(공학석사). 2009년~현재 한국과학기술원 전산학과 박사과정. 관심분야는 멀티미디어, 영상처리, 콘텐츠보안, 디지털포렌식 등