Decision Tree 기반의 인물 이미지 조작 탐지 모델 설계*

최효경**, 최은정**
**서울여자대학교 정보보호학과
e-mail : hyokyung@swu.ac.kr

Image Manipulation Detection Model with Binary Feature based-on Decision Tree

Hyo-Kyung Choi**, Eun-Jung Choi**
**Department of Information Security, Seoul Women's University

요 약

본 논문에서는 특정 인물의 이미지를 합성하여 타인의 명예를 실추함과 동시에 사회적 혼란을 야기하는 이미지 조작 기술의 악용을 방지하기 위한 인물 이미지 조작 탐지 기법을 제안한다. 제안하는 탐지 모델은 딥러닝 이미지 인식 기법으로 이미지 조작을 탐지하던 기존 기술의 한계를 해결하기 위한 병행 모델로써, Decision Tree 알고리즘을 기반으로 인물 이미지의 원시 데이터를 추출해 text화한 attribute를 학습하여 이미지 조작을 탐지한다. 원시 데이터 증거로 인물 이미지 조작을 탐지해내는 실험의 예측 정확도는 98.75%를 기록하였다.

1. 서론

정보는 자극적일 수록 이목을 집중시킨다. 정책과 경제, 도움이 되는 상식, 새로운 기술과 같은 정보는 정확하고 유용한 공급이지만 수요의 대상은 제한적인 반면, 무용한 정보임에도 불구하고 유명 인사를 대상으로 한 이슈는 진위 여부가 명확하지 않더라고 네트워크를 통해 쉽고 빠르게 퍼져나간다. 이러한 정보의 특성을 악용하여 특정 인물에 대한 가짜 정보를 배포하여 그들의 명예를 실추하고 개인정보를 도용하는 사건들이 증가하는 추세이다. 조사[1]에 따르면 최근 등장하고 있는 가짜 뉴스나 가짜 사진과 같은 정보들은 정상적인 인지 능력을 갖춘 사람이라도 진위를 구별할 수 없을 정도로 논리적이고 정교하다. 특히나 가짜 뉴스의 경우는 배포하는 주체가 사람인지, 소프트웨어인지 식별하거나 딥러닝을 활용하여 뉴스 기사에 숨어 있는 조작을 식별하는 기술들이 개발되고 있지만 악의적인 인물 합성 사진의 경우는 연구 단위에 머물러 있어 마땅한 해결책이 없다는 우려를 낳고 있다.

매일 발전하고 있는 유명인을 타깃으로 한 이미지를 조작하고 합성하는 프로그램들에 대처하기 위해서는 정확한 인물 이미지 합성 및 조작 탐지 기법이 제공되어야 한다. 이를 위해 연구되고 있는 방법 중 가장 좋은 식별 성능을 기대할 수 있는 것은 이미지 인식 분야의 인공지능 알고리즘을 사용하여 탐지하는 것이다. 하지만 이 방법의 경우 합성 인물 이미지의 정교함이 올라갈수록 오탐율이 비례하여 증가한다는 단점이 존재한다. 따라서 탐지율을 높이기 위해선 딥러닝 이미지 인식 탐지와 병행하여 조작 여부 분별을 수행할 별도의 모델이 필요하다.

본 논문에서는 기존 기술과 차별화 되어 병행 탐지 작업에 사용할 수 있는 인물 이미지 조작 탐지 모델을 제안한다. 제안하는 병행 모델은 Decision Tree를 기반으로 하며, 조작되었거나 합성되었다는 증거로 활용할 text 데이터를 학습 데이터셋으로 사용하여 조작 여부를 탐지하는 방법이다.

본 논문의 구성은 2장에서 이미지 조작 탐지 기법의 관련 연구에 관해 서술하고 3장에서는 제안하는 모델의 구조와 구성 요소에 대해 상세 기술한다. 4장에서는 제안하는 모델을 이용한 실험에 대해 기술하고 5장에서는 실험 결과 Tree에 대해 분석하며 마지막으로 6장에서 결론과 향후 연구 방향을 제시한다.

2. 관련 연구

이미지 조작을 탐지하는 기법과 인물 이미지를 이미지 인식 기술로 분별하는 방안은 이전부터 끊임없이 연구되어왔다.

Sevinç Bayram[2]의 연구에서는 변경된 원본 이미지와 원래 이미지를 구분할 수 있는 기술 개발을 위해 이웃 비

* 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음 (2016-0-00022)

트 평면을 기반으로 한 방법을 제안하였다. 해당 연구는 이미지를 '조작'하여 사이버 공격 행위에 사용한 이진 데이터 변형을 탐지하는 데에는 효과적일 수 있으나 포렌식 검출기라는 한계점에 근거한다면 조작 목적에 따른 판단을 하는 데에 탐지 과정이 번거롭다는 단점을 가진다.

Peng Zhou[3]의 연구에서는 RGB 스트림과 노이즈 스트림을 사용하여 이미지 조작 감지를 위한 새로운 네트워크를 제안했다. 해당 연구에서는 변조된 영역과 인증된 영역 사이의 노이즈 불일치를 캡처하여 이미지의 노이즈 특징을 찾았다. 하지만 변조된 이미지의 노이즈 검출이 미비하게 이루어질 경우 탐지에 한계가 있다는 단점이 있다.

Charina G. Marrion[4]의 연구에서는 Facebook에서 사용한 이미지에 높은 압축률이 적용되어 아티팩트가 형성되기 때문에 이미지가 왜곡되는 현상을 연구하였다. 연구 결과는 왜곡 현상이 원본 JPEG 이미지의 품질과 같은 여러 변수에 의존할 수 있음을 보여준다. 해당 연구 역시 이미지 조작을 탐지하는 데 있어 압축 설정에 대한 정보 이외에 대한 탐구의 신뢰성이 떨어진다는 단점이 존재한다.

이미지 인식에 대한 관련 기법에 대해서는 최인규[5], 변영현[6]의 연구에 따르면 CNN 모델으로 이미지에서 얼굴을 인식하여 얼굴의 표정을 추출하고 얼굴 인식을 수행할 수 있음을 보여주었다. 이미지 인식 기술은 이와 같이 뛰어난 성능을 보여주고 있으나, 정교하게 만들어진 GAN 이미지나 노이즈 추출도 불가능할 정도의 수준으로 합성된 이미지의 경우 여전히 탐지 오탐율이 높다는 한계를 가지고 있다.

본 논문에서 제안하는 모델은 원시 데이터인 binary 데이터를 기반으로 하여 인물 이미지의 조작 여부를 탐지한다. 기존의 이미지 인식 기술에서 정교하게 조작된 이미지는 조작되지 않았다고 오인하였던 한계와 이미지 조작을 탐지하는 기법의 효용성에 대한 단점을 보완하기 위해 간단한 절차를 기반으로 판별 과정을 간소화 하고, 조작된 이미지에 대한 특징 추출을 별도로 시행한다.

3. 제안 모델

제안하는 모델은 text 데이터를 attribute와 label로 나누어 학습한 후 의사 결정을 통해 판별하는 데 뛰어난 성능을 보이는 Decision Tree 알고리즘을 사용한다. Decision Tree는 결과를 분석할 때 의사 결정에 사용한 트리 결과를 토대로 관측 값과 목표 값에 대한 관계를 보다 상세히 파악할 수 있다.

성능이 좋은 학습법으로 학습을 진행하는 것도 중요하지만 이미지에 숨겨진 조작 증거를 선택하는 기준도 중요하다. Decision Tree에서 attribute로 사용할 인물 이미지들마다 가지고 있는 조작된 증거를 찾는 방법은 이미지를 raw data인 hex 단위로 분석하는 것이다. 본 모델에서는 조작된 이미지의 헤더 단위에서 이미지 조작에 쓰인 툴에 대한 string이 존재하거나 DATA 단위에서 조작되지 않은 이미지와 구별되는 특이한 code가 삽입되었다는 특징을 이용하여 인물 이미지의 조작 및 합성여부를 탐지한다.

3.1 시스템 구조

시스템은 학습에 사용할 인물 이미지 데이터셋의 구축 후에 실행된다. 인물 이미지 데이터셋의 확장자는 jpg이며 임의로 조작하거나 타인의 얼굴을 합성한 이미지와 조작되지 않은 이미지로 구성된다. 데이터셋이 구축된 후 모든 데이터셋에 대하여 특징 데이터(Hex)를 추출한다. 특징 데이터는 이미지조작 판별에 사용될 string 증거를 의미한다.

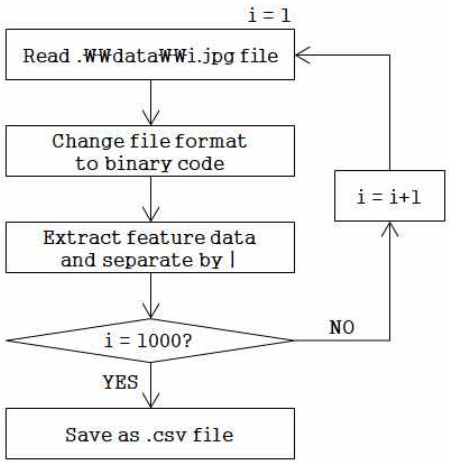

Fig. 1. Feature Extraction Process

이미지를 text화 시켜 추출한 증거 데이터는 Decision Tree 알고리즘으로 학습하기 위해 csv 확장자로 리스트화 한다. 리스트화 한 .csv파일을 읽어 들이기 위한 column separators는 ' | '이며 읽어 오는 정보는 각각 jpg 파일의 이름 정보에 해당하는 Name, 조작에 사용한 도구 정보 등이 담긴 헤더 정보에 해당하는 Header, 특이한 string 정보에 해당하는 Hex, 그리고 마지막으로 조작되었는지 조작되지 않았는지를 구분할 정답에 해당되는 Label이다. 이 때 학습 과정에 불필요한 Name 정보를 제외하고 모두 Import 한다.

	✓	✓	✓
Name	Header	Hex	Label
attribute	attribute	attribute	label
...	...	...	...

Table. 1. Import Configuration

위와 같이 학습 데이터에 대한 Import Configuration을 설정한 후에는 Decision Tree 모델에 따라 학습을 진행하고 탐지가 필요한 이미지를 Tree 값에 따른 의사 결정 방식에 기반 하여 판별함으로써 Performance 결과를 생성한다.

3.2 구성 요소

제안한 모델은 인물 이미지가 조작(합성)되었는지 여부를 판단하기 위해 세 가지 절차를 거친다.

첫 째, 특징 데이터를 학습한다. 본 모델의 실제 동작은 학습 과정이 먼저 선행된다. 수집한 합성 인물 이미지

데이터셋과 조작되지 않은 원본 이미지 데이터셋을 시스템 구조에 따라 처리한 후 지도학습 딥러닝 알고리즘인 Decision Tree로 학습하여 attribute 데이터 값을 기준으로 조작 이미지 특징에 해당하는지 또는 원본 이미지 특징에 해당하는지에 대한 의사 결정을 내릴 수 있는 Tree를 구축한다.

둘 째, 판단 이미지를 전처리한다. 사용자가 탐지의 필요에 따라 입력한 인물 사진의 원시 데이터를 추출하여 조작된 증거로 사용될 값이 포함된 범위를 클리핑한다. 추출한 특징 데이터를 별도로 저장하여 이후 판단 과정에 사용한다.

셋 째, 마지막으로 사용자가 입력한 인물 이미지의 조작 여부를 판단한다. 학습용 이미지 샘플에서 얻어낸 의사 결정도가 그려진 Tree 데이터와 판단이 필요한 인물 이미지의 전처리 데이터가 준비되어 있어야 하며, 앞 과정에서 사용한 Decision Tree 알고리즘을 동일하게 사용하여 의사 결정을 진행한다. 조작 이미지 특징에 더 가까울 경우 조작으로 결론짓고, 그렇지 않을 경우 미조작이라 판단한다.

4. 실험

본 장에서는 제안 모델의 인물 이미지 조작(합성) 여부 판별 성능 검증 실험에 대해 기술한다.

4.1 실험 환경

실험을 위한 환경은 Windows10 pro(64bit) 운영체제에서 Conda 4.4.10과 Python 2.7, Pycharm 2016.3.2., RapidMinerStudio 0.1.1.0, HxD Editor를 이용하여 실험하였다. 전체 상세 실험 환경은 Table 2에 기술되어 있다.

Name	Spec.
OS	Windows10 pro(64bit)
CPU	Intel Core i5-7200U 2.50GHz
RAM	8G
Conda	4.4.10
etc.	Python 2.7, Pycharm 2016.3.2., RapidMinerStudio 0.1.1.0, HxD

Table. 2. Experiments Environments

4.2 실험 방법

탐지 모델 성능 검사를 위해 label이 존재하는 약 3000개의 데이터셋으로 한정한 후 시스템 구조에 기반 하여 데이터 처리 과정을 거친다. 이 때 정답 Label을 조작은 1, 미조작은 0으로 표시한다. 이후 7:3의 비율로 데이터셋을 split하여 학습에 사용할 이미지와 판별에 사용할 이미지를 shuffled sampling한다. 데이터셋을 구분지어 Decision Tree 알고리즘에서 실험을 함으로써 7의 비율로 학습하여 생성된 Tree에 따라 3의 비율의 데이터 판별(이미지 조작 탐지) 결과를 도출한다. Decision Tree에 따라 학습 데이터셋의 조작된 이미지 정보 특징과 유사할 경우 '1'로 판별되며 학습 데이터셋의 조작되지 않은 이미지 정보 특징과 유사할 경우 '0'으로 판별된다.

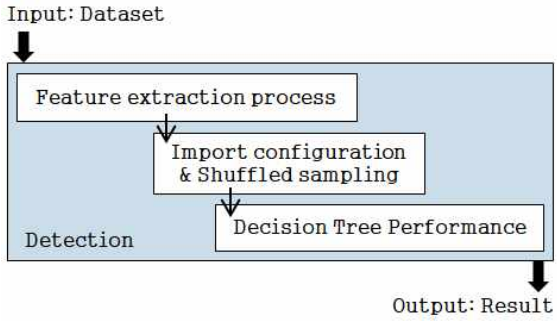

Fig. 2. Architecture of Proposed Model

4.3 실험 결과

제안하는 인물 이미지 조작 탐지 모델의 성능을 실험하기 위해 조작되었거나 조작되지 않은 인물 이미지로 이루어진 샘플 데이터셋을 사용하였다. 이 실험에 사용한 데이터셋의 개수는 3,736개로 학습에 사용할 이미지와 판별할 이미지를 7:3의 비율로 나누어 2,616개의 이미지는 학습에 이용하고, 1,120개의 이미지는 판별에 이용하였다.

이에 따른 실험 결과로는 조작되지 않은 이미지(0)에 대해 조작되지 않은 이미지(0)로 올바른 판단을 한 경우는 573건, 조작된 이미지(1)로 오탐을 한 경우는 0건이었으며, 조작된 이미지(1)에 대해 조작된 이미지(1)로 올바른 판단을 한 경우는 533건, 조작되지 않은 이미지(0)로 오탐을 한 경우는 14건이었다. 결론적으로 해당 모델이 조작된 이미지를 판별할 수 있으며 탐지 정확도는 98.75%가 된다는 것을 알 수 있었다. 실험 결과에 대한 상세 정보는 Table 3에 기술되어 있다.

	true 0	true 1	class precision
pred. 0	573	14	97.61%
pred. 1	0	533	100.00%
class recall	100.00%	97.44%	

Table. 3. The Results of Experiment

5. 분석

이 장에서는 제안하는 모델의 성능 검사 실험에서 사용한 Decision Tree 알고리즘에 따라 도출된 결과를 분석한다.

Decision Tree 알고리즘은 입력된 attribute 값과 label 사이의 연관 관계를 표현하는 데 최적화 되어있다. 이러한 Decision Tree 알고리즘의 특징에 따라 인물 이미지 조작에 대한 attribute로 활용한 header와 hex 데이터 정보를 기준으로 하여 레이블 0(조작되지 않음) 또는 1(조작됨)에 대한 유사도 설명서와 그래프를 확인할 수 있다.

설명서를 확인하면, 인물 이미지에서 추출한 header와 hex 정보 중에서 레이블 0(조작되지 않음)에 가장 근접한

데이터로 분석된 값은 '1010101010101ffdb00430101010101010101010101010101010101010101010101010101010101010101010101010101010101010101010101010101010101010101010101010101010101'로 레이블 0에 대한 score는 1244, 1에 대한 score는 0이었다. 반면 레이블 1(조작됨)에 가장 근접한 데이터로 분석된 값은 '0010000000860101000300000001000000000132000200000014000000069010f0002000000000900000007d000000000466f6f64696500323031383a31303a333120'으로, 레이블 0에 대한 score는 0, 1에 대한 score는 312였다. 이 외에 어떠한 레이블에도 치우치지 않아 판별에 혼란을 주는 정보 값도 있었다. '80a0807070a0d0a0a0b0c0c0c0c07090e0f0d0c0e0b0c0c0cffdb004301020202030303060303060c0807080c0c0c0c0c0c0c0c0c0c0c0c0c0c0c0c0c0c0c0c0c' 해당 헥사 정보는 실제로는 조작되지 않은 이미지의 값이었지만 특이한 특징으로 분류되지 못하여 레이블 0에 대한 score는 55, 1에 대한 score는 49였다.

설명서와 그래프에 따른 결과를 분석해보면, 주로 조작되었는지 조작되지 않았는지를 판별하기 위해서는 이미지에 있는 보편적인 값이 아닌, 조작에 사용된 도구 정보 또는 특이한 string 정보와 같이 다른 값과 구별될 수 있는 특징이 판별 기준으로써 적용되었음을 확인할 수 있다.

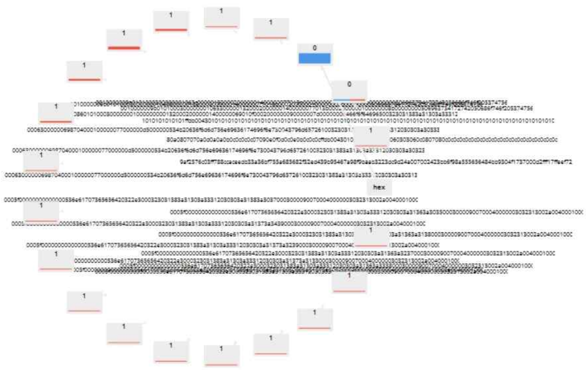

Fig. 3. Decision Tree Graph

6. 결론

본 논문에서는 사회적인 혼란을 야기하는 인물 이미지 조작 및 합성 문제를 해결하기 위한 새로운 인공지능 이미지 조작 탐지 모델을 설계하였다.

제안하는 모델의 조작된 인물 이미지 탐지 성능을 시험하기 위해 조작되었거나 조작되지 않은 데이터셋을 이용하여 학습과 판별을 실행한 결과 판별에 사용한 1,120개의 이미지에 대하여 98.75%의 정확도로 조작과 미조작을 구분해 냈다. 실험 결과를 실험에 사용한 Decision Tree 알고리즘에 따라 분석해 본 결과, 조작된 이미지에 표시되어 있는 조작에 사용한 도구에 대한 string 또는 조작을 하였을 때 손상된 hex 데이터와 같이 특이한 특징을 기준으로 판별을 진행하였으며, 보편적인 데이터의 경우는 판별을 하는데 혼란을 기하는 것을 확인할 수 있었다.

기존 연구에서는 이미지 인식 알고리즘을 이용하여 이미지 자체에 대한 분석을 통해 탐지하는 기법이 존재하였으나 정교화 되어가는 이미지 조작 기술을 대처하지 못하여 오탐율이 높다는 단점이 존재하였다.

제안하는 모델은 기존 연구들과는 달리 이미지에서 추출한 원시데이터를 기반으로 binary 분석을 진행하여 탐지율에 있어 우수한 성능을 보인다는 장점을 가진다.

그렇기 때문에 제안하는 탐지 모델은 합성 및 조작으로 특정인에 대한 명예를 실추하거나 거짓된 정보를 확산시키는 문제에 대한 해결책으로써의 역할을 할 수 있다.

하지만 해당 모델은 병행 모델인 만큼 기존의 이미지 인식 기술과 병합되지 않을 경우, 헥사 값이 알아볼 수 없게 이미지가 지속적으로 변형되었거나 다양한 방식을 사용해서 2차 3차적으로 유포되었을 경우 특징 추출이 어렵다는 한계가 존재한다.

따라서 이에 대해 이미지 인식 기술을 조작 기술에 대처할 수 있도록 고도화 시키는 방안과 이미지 인식 기술로 해결할 수 없는 binary detection에 대한 개발을 기반으로 인물 이미지가 조작되었을 경우 효과적으로 판별할 수 있는 추가적인 연구가 필요하다.

참고문헌

[1] 김준래, "가짜뉴스, 합성사진 꼼짝 마," https://www.msit.go.kr/webzine/posts.do?postIdx=324, 과학기술정보통신부 웹진 2018 4월호 문화&라이프, 2018.

[2] Sevinç Bayram, İsmail Avcıbaş, Bülent Sankur, Nasir Memon, "Image manipulation detection with Binary Similarity Measures," Proc. European Signal Processing Conf., 2005.

[3] Peng Zhou, Xintong Han, Vlad I. Morariu, Larry S. Davis, "Learning Rich Features for Image Manipulation Detection," arXiv:1805.04953, 13 May 2018.

[4] Charina G. Marrion, "Digital Image Manipulation Detection on Facebook Images," University of Colorado at Denver, 2016.

[5] 최인규, 안하은, 송혁, 고민수, 유지상. (2016). CNN 기반의 얼굴 표정 인식. 한국방송미디어공학회 학술발표대회 논문집, , 320-321.

[6] 변영현, 곽근창. (2014). 로봇환경에서 3차원 CNN을 이용한 비디오 기반 얼굴 인식. Proceedings of KIIT Summer Conference, , 26-29.