웹 기반 도구를 활용한 딥페이크 이미지 탐지 프레임워크*

김건우* 이정인** 민무홍***

*성균관대학교 컴퓨터교육과 (학부생)

**성균관대학교 과학수사학과 (대학원생)

***성균관대학교 컴퓨터교육과 (조교수)

Web-Based Framework for Detecting Deepfake Images

Geon Woo Kim* JeongIn Lee** Moohong Min***

*Department of Computer Education, Sungkyunkwan University
(Undergraduate student)

**Department of Forensics, Sungkyunkwan University
(Graduate student)

***Department of Computer Education, Sungkyunkwan University
(Assistant professor)

요 약

최근 웹 기반 도구를 이용하여 딥페이크 이미지를 손쉽게 생성할 수 있게 되면서, 그로 인한 피해 사례도 증가하고 있다. 이에 대응하기 위해 다양한 딥페이크 탐지 도구가 공개되어 있으나, 각각의 도구는 고유의 탐지 방식의 특성이 있어 한계와 해석의 어려움이 존재한다. 본 논문에서는 자동화된 수치 기반 탐지와 시각 기반 해석 도구를 결합한 이중 구조의 딥페이크 이미지 탐지 프레임워크를 제안한다. 이를 위해 FaceForensics++ 및 FFHQ 기반 이미지와 카카오톡과 텔레그램 전송을 통한 화질 저하 이미지를 활용하여 실험을 수행하였으며, 전문가가 아닌 일반 사용자 관점에서 프레임워크의 유효성을 분석하였다.

I. 서론

딥페이크 기술은 고도화되고 있으며, 웹 기반 이미지 생성 도구만으로도 고화질 조작 이미지 생성이 가능하므로 사기, 사칭 등 여러 범죄가 발생하고 있다[1]. 이런 피해를 막기 위해 딥페이크 사진 여부를 구분하는 능력이 필요하지만, 일반인들이 이를 판별하는 것은 쉽지 않다.

인터넷에는 누구나 이용 할 수 있는 오픈 포렌식 툴들이 존재하지만, 실제 탐지에 얼마나 효과적인지에 대한 검증은 부족하다. 또한, 정량적 판단을 제공하는 자동화된 툴과 달리, 대부분의 포렌식 도구는 시각적 해석에 의존한다.

본 연구에서는 일반인이 접근 가능한 도구를 활용하여 딥페이크 탐지 프레임워크를 제안하고자 한다. 자동화된 탐지 툴을 통해 1차 필터링을 수행하고, 그 결과를 Forensically와 같은 시각화 기반 도구를 통해 추가 판단하는 이중 구조를 통해 탐지 가능성을 높이고자 한다.

이를 위해 FaceForensics++와 FFHQ를 기반으로 다양하게 생성된 합성 이미지를 활용하여 실제 사진 유포 환경을 만들고 카카오톡과 텔레그램 이미지 전송에 따른 화질 저하도 실험을 진행하였다.

본 연구에서 제안하는 프레임워크는 수치 기반 탐지와 시각적 판단을 결합한 이중 구조를 기반으로 한다. 1단계에서는 Sightengine을 활용하여 이미지의 딥페이크 가능성을 수치로 필

* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과(No.RS-2024-00439139, 사이버 위기 대응 능력 및 복원력 시험평가 도구 개발)이며, 과학기술정보통신부 및 정보통신기획평가원의 메타버스융합대학원의 연구 결과로 수행되었음 (IITP-2025-RS-2023-00254129)

터링한다. 이후 탐지가 모호한 경우, Forensically와 같은 시각화 도구를 통해 ELA, Noise, Luminance Gradient 기반 분석을 수행한다. 이를 통해 사용자는 조작 가능성이 있는 부위를 직관적으로 확인할 수 있으며, 전문가가 아니더라도 탐지 및 판단에 도움을 받을 수 있다. 이 프레임워크는 단일 탐지 방식의 한계를 보완하고, 실제 조작 탐지 성능을 높이는 데에 목적이 있다.

II. 배경 및 관련 연구

Forensically, FotoForensics 등의 시각 기반 도구에서 제공하는 ELA(Error Level Analysis)는 JPEG 압축 과정에서 발생하는 손실 특성을 활용하여, 재압축된 이미지와 원본 간의 차이를 시각화함으로써 조작된 영역을 부각시키는 방식이다. 이와 함께, 이미지의 잡음(Noise) 패턴 역시 조작 탐지의 핵심 단서로 작용할 수 있다. JPEG 블록 단위의 압축 아티팩트나 잡음 분포의 불연속성은 비정상적인 패턴으로 감지될 수 있으며, 특히 조작된 영역에서는 일반적인 노이즈 분포와 달리 불균형하거나 과도하게 강조된 양상이 나타난다. 또한, 명암이나 색상 변화율을 분석하는 Luminance Gradient 역시 시각적 단서로 활용 가능하다. 일반적으로 자연스러운 밝기 변화는 일관되게 복원되지만, 조작된 영역은 이러한 변화율이 왜곡되거나 sharpening 강도가 다르게 나타나는 경향이 있다[2].

그러나 이러한 웹 기반 도구들은 다양한 생성 방식이나 압축 손상이 포함된 이미지에 대한 탐지 성능이 충분히 검증되지 않았다. 특히 메신저 전송 등으로 품질이 저하된 이미지에 대해서는 탐지 정확도가 낮아질 수 있다.

III. 실험

3.1 실험 준비

본 연구에서는 세 가지 유형의 이미지 데이터를 활용하였다. (1) FaceForensics++ C23[3]에서 추출한 원본-딥페이크 이미지 쌍 15쌍, (2) Face Swapper 및 Remaker AI로 생성한 FFHQ[4] 기반 고화질 합성 이미지 24장, (3) (2)의 이미지에 대해 카카오톡(일반, 저화질)과 텔레그램(압축 화질)을 통해 전송 후 재수집한 3종 화질 버전이다.

3.2 수치 기반 탐지 도구

Sightengine API의 Deepfake 탐지 기능을 활용하여 이미지를 정량적으로 분석하였다. 데이터는 (1)FaceForensics++, (2)Face Swapper 및 Remaker AI 생성 이미지, (3)카카오톡, 텔레그램 전송 압축 이미지로 구성되며, 각 이미지에 대해 딥페이크일 확률값을 수치로 출력하였다.

3.3 시각 기반 탐지 도구

시각 기반 탐지 도구로는 누구나 접근 가능한 대표적인 이미지 포렌식 도구인 Forensically와 FotoForensics를 이용하여 조작 여부를 분석하였다.

Forensically는 웹 기반 오픈 소스 이미지 포렌식 툴로 다양한 시각화 기반의 분석 기능을 제공한다. 본 실험에서는 그 중에서도 ELA, Noise, Luminance Gradient를 활용하여 조작 여부 탐지를 시도하였다.

FotoForensics도 웹 기반으로 이미지 포렌식 툴로 ELA, JPEG %와 같은 분석 기능들을 제공한다. 본 실험에서는 그 중에서 ELA를 활용하여 탐지를 시도하였다.

IV. 결과

4.1 수치 기반 탐지 도구 결과

각 데이터별 Sightengine의 딥페이크 사진인 확률을 나타내는 결과의 평균값은 다음과 같다.

표 1 사진 종류별 딥페이크 확률 결과

FaceForensics ++	
원본 사진	1%
합성 사진	98.25%

FFHQ 원본	
원본 사진 (원본 화질)	7.83% *
원본 사진 (카카오톡 일반화질)	2.5% *
원본 사진 (카카오톡 저화질)	1.33%
원본사진 (텔레그램 압축)	2.66%*

Face Swapper	
합성 사진 (원본 화질)	99%
합성 사진 (카카오톡 일반화질)	99%
합성 사진 (카카오톡 저화질)	99%
합성 사진 (텔레그램 압축)	99%

Remaker AI	
합성 사진 (원본 화질)	99%
합성 사진 (카카오톡 일반화질)	99%
합성 사진 (카카오톡 저화질)	99%
합성 사진 (텔레그램 압축)	99%

사진의 원본 화질, 카카오톡 일반 화질, 텔레그램 압축일 때 값이 7.83%, 2.5%, 2.66%로 확인되었다. 이는 합성하지 않은 사진이 원본 화질일 때 42%, 카카오톡 일반화질일 때 10%, 카카오톡 저화질일 때 3%, 텔레그램 압축일 때 10%로 딥페이크 탐지 확률을 보였기 때문이다.

4.2 시각 기반 탐지 도구 결과

표2 사진 종류별 시각 기반 탐지 결과

원본 화질	
ELA	Forensically와 FotoForensics 모두 얼굴 중심부 (눈썹, 눈, 입 등) 영역에서 밝게 부각되거나 무늬가 흐트러진 패턴이 자주 발생함
Noise	원본 이미지들은 얼굴 부위에서 균일한 노이즈 패턴을 보여주는 반면, 딥페이크 이미지들은 얼굴 부위나 특정 윤곽선 주변(눈, 입 등)에 과도한 노이즈 강조가 나타남
Luminance Gradient	원본은 피부 질감이나 명암이 비교적 균일하고 세밀하게 분포되어 있지만, 딥페이크 이미지들은 피부 영역이 매끈하게 나옴
일반화질, 저화질	
ELA	원본 사진은 얼굴이 전체적으로 흐트러져 나오는 반면, 딥페이크 사진은 눈, 코, 입의 영역은 명확하게 표시됨
Noise	원본 화질일 때와 원본 사진은 노이즈가 고르게 분포되어 있지만, 딥페이크 사진은 특정 부위(눈, 코, 입, 턱 등)에 집중되어 있음
Luminance Gradient	원본 화질일 때와 비슷하게, 원본은 피부 질감이나 명암이 비교적 균일하고 세밀하게 분포되어 있지만, 딥페이크 이미지들은 피부 영역이 매끈하게 나옴

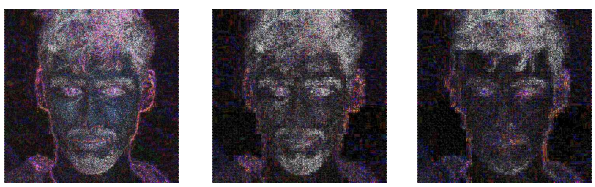

그림1. 원본, 원본 저화질, 딥페이크 저화질 ELA

V. 결론 및 향후 과제

본 연구는 웹 기반의 수치 기반 및 시각 기반 탐지 도구를 결합하여, 일반 사용자들도 활용 가능한 딥페이크 탐지 프레임워크를 제안하였다. Sightengine을 통한 정량적 분석은 대부분의 딥페이크를 높은 정확도로 탐지했지만, 일부 이미지에서 높은 오탐 확률이 나타났다. 이와 같은 이상치는 ELA, Noise, Luminance Gradient 분석을 병행함으로써 시각적으로 구별 가능함을 확인하였다.

또한, 일반 화질 및 저화질 이미지와 같이 압축으로 인해 시각적 정보가 저하된 경우에도, 수치 기반 탐지와 ELA, Noise, Luminance Gradient 분석을 통해 조작 여부를 안정적으로 탐지할 수 있었다.

제안된 프레임워크는 수치 기반 필터링과 시각적 해석을 연계함으로써, 기술적 지식이 없는 일반 사용자도 딥페이크 이미지를 효과적으로 판단할 수 있도록 설계되었다. 이를 이용하여 이중 탐지 구조를 자동화하게 된다면, 일반인들도 보다 쉽게 해당 사진이 조작되었는지, 어느 영역이 조작되었는지 쉽게 판별할 수 있을 것이다.

또한, 추가적으로, 해당 연구는 딥페이크 사진만을 대상으로 했지만, 딥페이크 영상으로 범위를 넓힌다면, 보다 다양한 도구들을 활용한 탐색이 가능해질 것이다.

[참고문헌]

[1] 김홍희, “강남 사는 30대 여성”…딥페이크로 120억 사기친 일당 검거, KBS 뉴스, 2024년 5월 2일. [온라인]. 이용 가능: https://news.kbs.co.kr/news/pc/view/view.do?ncd=8242798&ref=A

[2] Krawetz, N., & Solutions, H. F. (2007). A picture’s worth. *Hacker Factor Solutions*, 6(2), 2.

[3] A. Rössler et al., “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019, pp. 1–11.

[4] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410.