한국언론진흥재단
Korea Press Foundation

Media 정책 리포트

2020년 1호

허위정보로서 딥페이크, 어떻게 대처할 것인가

오세욱 | 한국언론진흥재단 선임연구위원
박아란 | 한국언론진흥재단 선임연구위원

Ⅰ. 시작하며

- **지난 2018년 4월 미국의 온라인 매체 버즈피드(Buzzfeed)가 유튜브에 게재한 한 영상[1]이 사람들을 놀라게 했는데, 영상의 내용이 버락 오바마 전 미국 대통령이 도널드 트럼프 현 미국 대통령에게 욕을 하는 것이었기 때문임**
  - 해당 영상 속에서 오바마 전 대통령은 "트럼프 대통령은 진짜 머저리(dipshit)입니다"라고 말했는데, 이는 버즈피드가 딥페이크(Deep Fake) 기술의 위험성에 대해 경고하기 위해 영화감독 조든 필(Jordan Peele)과 함께 만든 조작 영상이었음
  - 영상은 교묘하게 조작돼 일반인들이 봤을 때 조작 여부 판단이 어려웠기 때문에 버즈피드가 스스로 이 사실을 공개하지 않았다면 큰 파장을 불러올 수 있었음

1) BuzzFeedVideo (2018. 4. 17) "You Won't Believe What Obama Says In This Video!" [URL] https://www.youtube.com/watch?v=cQ54GDm1eL0

- **딥페이크 기술은 인공지능의 바탕이 되는 기계학습(machine learning) 기법인 딥러닝(deep learning)을 사용해 원본 이미지나 동영상 위에 다른 이미지를 중첩(superimpose)하거나 결합(combine)해서 원본과는 다른 이미지와 영상을 만들어 주는 이미지/동영상 조작(manipulation) 기술을 말함(최순욱·오세욱·이소은, 2019)**
  - 딥페이크는 사람의 특별한 지시가 없어도 자동으로 최적의 결과물을 산출할 수 있는 기계학습이 적용돼 식별이 굉장히 어려운 허위정보 생성 기술로 활용될 수 있으며, 미디어를 통해 전달되는 것이 실재와는 동떨어진 것일 수 있다는 인식을 심어줄 수 있음

- **우리나라에서는 한류 스타들을 대상으로 한 딥페이크, '지인능욕' 등의 음란물이 이미 사회 문제가 되고 있음**
  - 딥페이크 탐지 기술을 연구하는 딥트레이스(Deeptrace)[2]의 보고서(Ajder, Patrini, Cavalli & Cullen, 2019)에 따르면, 전 세계 딥페이크 기술 관련 음란물 사이트 중 미국이 41%로 가장 많았으며, 한국은 25%로 그 다음이었음
  - 현재 딥페이크 기술은 유명 연예인 등의 얼굴을 음란물에 합성하는 방식으로 주로 사용되고 있지만, 정치적 악용 가능성 등으로 인해 'MIT 테크놀로지리뷰(technology review)'는 가장 조심해야 할 인공지능의 위험 요소 중 하나로 딥페이크 기술의 발전을 지적한 바 있음(Knight & Hao, 2019. 1. 7)

- **실제로 인도 델리 주 의회 선거 하루 전날인 지난 2월 7일 인도 집권당인 인도인민당 (Bharatiya Janata Party)의 의장 마노지 티와리(Manoj Tiwari)는 아르빈드 케지리왈(Arvind Kejriwal) 델리 주 총리를 비판하는 내용을 영어로 된 영상[3]과 힌디어 방언 중 하나인 하르얀비어(Haryanvi)로 된 영상[4]으로 만들어 왓츠앱을 통해 유포했음**
  - 케지리왈 주 총리를 비판하는 내용의 44초짜리 두 영상은 진짜가 아닌 조작 영상으로 티와리 의장이 힌디어로 발언한 51초 길이의 원본 영상[5]에 딥페이크 기술을 적용해 입 모양, 얼굴 표정 및 발언 내용을 조작해 만든 것이었음(Christopher, 2020. 2. 18)
  - 인도인민당은 정치 커뮤니케이션 회사인 아이디즈 팩토리(Ideaz Factory)와 협력하여 이번 작업을 진행하였으며, 인도인민당은 8일 실시된 델리 주 의회 선거에서 결국 패배했지만, 향후 인도 내에서 사용되는 20여 개 언어에도 딥페이크 기술을 적용할 계획임
  - 비록 악의적 조작은 아니었지만 음란물 등에 주로 적용되던 딥페이크 기술이 정치 영역에서 일반적으로 사용되기 시작했다는 점을 보여주는 사례로 오는 4월 총선을 앞둔 우리나라에 주는 함의도 크다고 할 수 있음

- **페이크 뉴스 등 허위정보로 인한 소동이 상징하는 '탈진실(post truth)'을 넘어 딥페이크 기술로 인해 '진실의 종말(end of truth)'로 이어질 수 있다는 사회적 우려가 제기(Horowitz, et al., 2018)되면서 이를 방지하기 위한 법제도적 방안, 기술적 대응 방안, 진위여부 확인 방안 등이 제시되고 있음**
  - 이에 본 보고서는 딥페이크 방지를 위해 제시되고 있는 다양한 대응 방안 등을 검토해 제시하고 대략적이나마 딥페이크 조작 이미지 및 영상의 진위여부를 확인할 수 있는 다양한 도구 및 방안들을 언론인 및 이용자들에게 제시하고자 함

2) https://deeptracelabs.com/
3) https://youtu.be/88GUbuL89bQ
4) https://youtu.be/ZAdrE_wEMM0
5) https://youtu.be/2Tar2O4q0qY

Ⅱ. 딥페이크 관련 법규제 동향

1. 국내 법규제 동향

- **2020년 3월 5일 성폭력범죄의 처벌 등에 관한 특례법(이하 성폭법) 일부개정안이 국회 본회의를 통과함. 이 법안의 주된 목적은 특정 인물의 신체 등을 대상으로 한 영상물 등을 성적 욕망 또는 수치심을 유발할 수 있는 형태로 편집하는 등의 딥페이크를 제작 또는 반포하는 행위를 강력하게 처벌하는 것임**

  - 성폭법 개정안은 다음과 같은 조항을 신설함

> 제14조의2(허위영상물 등의 반포등) ① 반포등을 할 목적으로 사람의 얼굴 · 신체 또는 음성을 대상으로 한 촬영물 · 영상물 또는 음성물(이하 이 조에서 "영상물등"이라 한다)을 영상물등의 대상자의 의사에 반하여 성적 욕망 또는 수치심을 유발할 수 있는 형태로 편집·합성 또는 가공(이하 이 조에서 "편집등"이라 한다)한 자는 5년 이하의 징역 또는 5천만원 이하의 벌금에 처한다.
>
> ② 제1항에 따른 편집물·합성물·가공물(이하 이 항에서 "편집물등"이라 한다) 또는 복제물(복제물의 복제물을 포함한다. 이하 이 항에서 같다)을 반포등을 한 자 또는 제1항의 편집등을 할 당시에는 영상물등의 대상자의 의사에 반하지 아니한 경우에도 사후에 그 편집물등 또는 복제물을 영상물등의 대상자의 의사에 반하여 반포등을 한 자는 5년 이하의 징역 또는 5천만원 이하의 벌금에 처한다.
>
> ③ 영리를 목적으로 영상물등의 대상자의 의사에 반하여 정보통신망을 이용하여 제2항의 죄를 범한 자는 7년 이하의 징역에 처한다.

  - 기존의 성폭법 제14조는 카메라 등으로 성적 수치심을 유발할 수 있는 신체를 촬영하거나 이러한 촬영물 또는 복제물을 반포·제공 등을 한 자를 5년 이하 징역 또는 3천만 원 이하의 벌금에 처하고 있었음
  - 신설된 제14조의2는 이러한 성적 수치심을 유발할 수 있는 영상물 등을 편집·합성·가공하여 제작하거나 반포한 자를 5년 이하 징역 또는 5천만 원 이하의 벌금에 처한다고 규정하여 기존의 음란 영상물 촬영반포죄보다 가중처벌하고 있음. 또한 편집·합성·가공된 음란 영상물을 영리 목적으로 온라인에서 반포할 경우 7년 이하의 징역에 처하고 있음. 따라서 향후 딥페이크 등의 기법을 이용하여 음란물을 만들거나 배포할 경우 가중처벌의 대상이 될 수 있음
  - 개정안이 통과됨으로써 기존에 발의되었던 딥페이크 관련 성폭법 개정안(의안번호 2023469, 의안번호 2023856, 의안번호 2024097)들은 폐기됨

- **딥페이크 규제를 위해 2개의 정보통신망 이용촉진 및 정보보호 등에 관한 법률(이하 정보통신망법) 개정안도 발의되어 있음**

  - 정보통신망법 개정안은 합성영상 유포 방지를 주된 목적으로 하고 있지만, 향후 딥페이크 기술의 사회적 파장이 커질 경우 이를 대비하기에는 부족할 것으로 보임

● **정보통신망 이용촉진 및 정보보호 등에 관한 법률**

- 개정안(의안번호 2024095, 2019. 11. 29.): 인공지능 기술을 이용하여 만든 거짓의 음향·화상·영상 등의 정보를 식별하는 기술 개발을 위한 시책을 과기부장관 또는 방송통신위원회가 마련하도록 규정
- 개정안(의안번호 2024319, 2019. 12. 19.): 사람이 대상인 영상에 다른 사람의 얼굴이나 신체부위 이미지를 중첩·결합하여 만든 합성영상의 유통실태, 피해실태, 기술동향, 예방교육 등의 시책 추진의무를 방송통신위원회가 부담하도록 규정

2. 해외 법규제 동향

● **미국**

- 2018년 상원에서 <악성 딥페이크 금지법(Malicious Deep Fake Prohibition Act)> 법안(S.3805)[6]이 발의됐음. 이 법안에서 딥페이크란 '합리적인 사람들에게 실제로 말한 것이나 행동한 것처럼 보이도록 허위로 만들어지거나 수정된 시청각 기록물'로 정의됨. 이 법안은 '범죄나 불법행위를 유발할 목적으로 딥페이크를 만들어내는 행위'와 '딥페이크임을 알면서 유포하거나 범죄나 불법행위를 유발하기 위해 딥페이크를 유포하는 행위'를 벌금 또는 징역형으로 처벌한다고 규정함
- 2019년 상원에서 <딥페이크 보고법(Deepfake Report Act)>(S.2065)[7] 발의. 이 법안은 딥페이크 등 디지털 기술을 이용한 온라인 위조현황에 대해 국토안전부 장관이 연례보고서를 발행할 의무가 있음을 규정함
- 2019년 6월 하원에서 <딥페이크 책임법(DEEPFAKES Accountability Act: Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act)>(H. R. 3230)[8] 발의. 이 법안은 신분 사칭 등 기술적으로 조작된 온라인 콘텐츠를 만드는 자는 임베디드 워터마크 등을 사용하도록 의무화함.
- 버지니아, 텍사스, 캘리포니아, 뉴욕 등에서도 딥페이크 관련 주(州) 법안들이 발의됨
- 특히 2019년 캘리포니아에서는 1) 타인의 동의 없이 딥페이크 기술로 음란물을 만드는 것을 규제하는 법안(Assembly Bill No. 602)과 2) 선거일전 60일 이내에 선거후보자를 대상으로 악의적인 딥페이크 영상이나 오디오를 만들어 배포하는 것을 금지하는 법안(Assembly Bill No. 703)이 발의됨
- 이러한 미국의 딥페이크 관련 법안들은 딥페이크에 대한 개념정의가 명확하지 않으며 현실적으로 실효성이 뚜렷하지 않다는 점에서 언론계와 학계 등에서 비판을 받고 있음. 특히 수정헌법 제1조 하에서 정치적 표현의 자유는 강력하게 보장되어야 한다는 점에서 딥페이크 관련 법안들은 표현의 자유를 침해할 가능성이 매우 높다고 언론법학자인 미네소타대 제인 커틀리(Jane Kirtley) 교수는 지적함

6) https://www.congress.gov/bill/115th-congress/senate-bill/3805/text
7) https://www.congress.gov/116/bills/s2065/BILLS-116s2065es.pdf
8) https://www.congress.gov/bill/116th-congress/house-bill/3230/text

● **중국**

- <온라인 동영상 관리 규정(网络音视频信息服务管理规定)>을 제정하여 2020년 1월 1일부터 시행
- 온라인 동영상 규제 강화를 골자로 하는 이 규정은 딥페이크나 가상현실 등 기술을 활용한 허위영상 전파행위를 불법으로 규정함
- 플랫폼은 이러한 불법콘텐츠를 발견 시 허위정보임을 공시하고 국가인터넷정보판공실, 문화여유부, 국가방송TV총국에 등록하도록 규제를 강화함

Ⅲ. 딥페이크에 대한 기술적 대응 동향

● **기계학습의 한 방법인 딥러닝을 이용해 원본 이미지나 동영상 위에 원본과는 관련이 없는 이미지를 중첩하거나 결합하는 기술인 딥페이크는 영상 속 얼굴 이미지를 조작하는데 주로 사용되고 있음**

- 딥페이크가 영상 속 A의 얼굴을 B로 바꾸는 과정은 크게 추출(extraction)-학습(learning)-병합(mergeing)의 세 단계로 진행됨(최순욱·오세욱·이소은, 2019)
- 추출은 영상 속 얼굴 이미지를 학습하기 위한 자료를 확보하는 것이고 학습은 이렇게 확보된 자료들을 기계가 학습하는 과정이며 이러한 과정이 이루어지면 마지막으로 재구축된 얼굴을 원본 이미지에 병합하는 단계로 이어짐

● **딥페이크 등 기술에 대한 규제는 기술의 발전 속도보다 느릴 수밖에 없기 때문에 기술에 대해 기술로 대응하는 전략도 여러 곳에서 진행되고 있음**

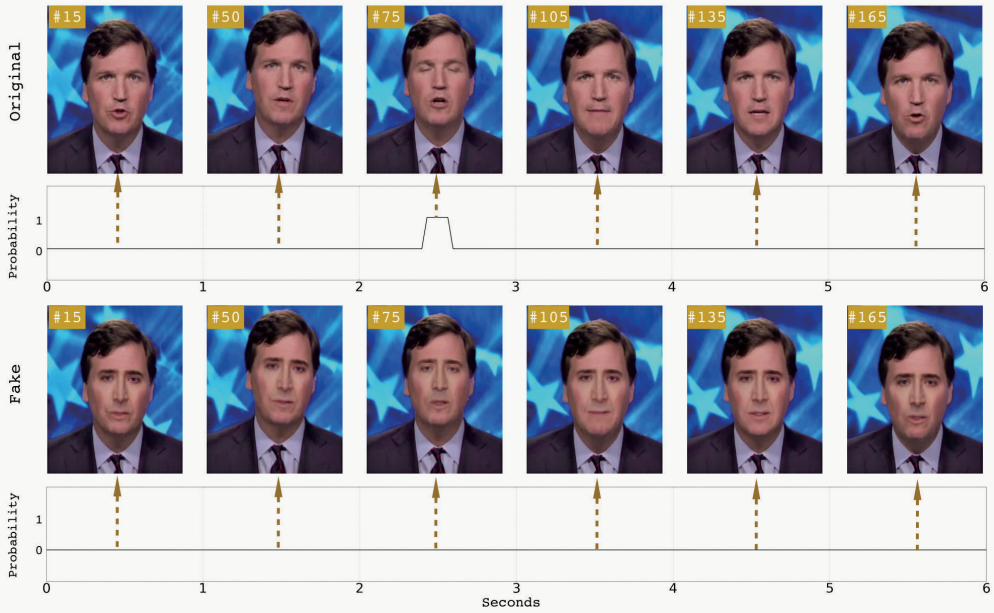
- 추출-학습-병합 과정에서 나타나는 기술적 특성을 파악하여 조작 여부를 판단하는 것임
- 원본과는 다른 이미지의 병합과정이 이루어지기 때문에 인간의 눈으로는 파악하기 어렵지만 조작의 흔적을 남기고 있는데 착안한 기술들로 대표적으로 비정상적 움직임 감지, 기존 이미지 기반 방식, 인공지능 기반 방식 등이 있음(박준·조영호, 2019)

1. 비정상적 움직임 감지

● **딥페이크가 고정된 이미지를 기반으로 사람의 얼굴을 학습하기 때문에 합성 영상에서는 눈 깜빡임 등과 같은 움직임을 적용하기 어렵고 적용하더라도 어색하다는 점에 착안해 영상 속 인물의 눈 깜박임 동작을 분석해 딥페이크 조작 여부를 판단함(Li, Chang, Farid & Lyu, 2018)**

- [그림 1]처럼 정상적인 사람의 경우 대략 6초 이내에 한 번은 눈 깜박임을 하게 되는데, 딥페이크로 합성한 영상에서는 눈 깜박임이 나타나지 않는다는 점을 파악한 것임
- 이와 같은 신호를 자동으로 탐지하는 알고리즘을 적용해 조작여부를 자동 판단함

그림 1 | 그림 내 위 원본 영상에서는 3초가 되기 전에 눈을 깜박이는 모습이 나타나지만, 아래 합성 영상에서는 같은 시간 대에 눈을 깜박이는 모습이 나타나지 않음(Li, Chang, Farid & Lyu, 2018, p. 2)



- **눈 깜박임 외에도 입술의 움직임과 목소리의 일치 여부를 기준으로 딥페이크 적용 여부를 자동으로 판단하기도 함**
  - 서론에서 사례로 제시한 인도인민당의 딥페이크 영상도 발언하는 언어에 따라 입모양과 목소리가 약간씩 다름을 알 수 있는데 이를 상호 비교하여 자동으로 감지하는 방식임
  - 하지만, 딥페이크 기술이 발전함에 따라 입모양과 목소리 일치 여부의 식별율은 그리 높지 않게 나오고 있어 실제 적용이 많이 이루어지고 있지는 않은 상황임(Korshunov & Marce, 2018)

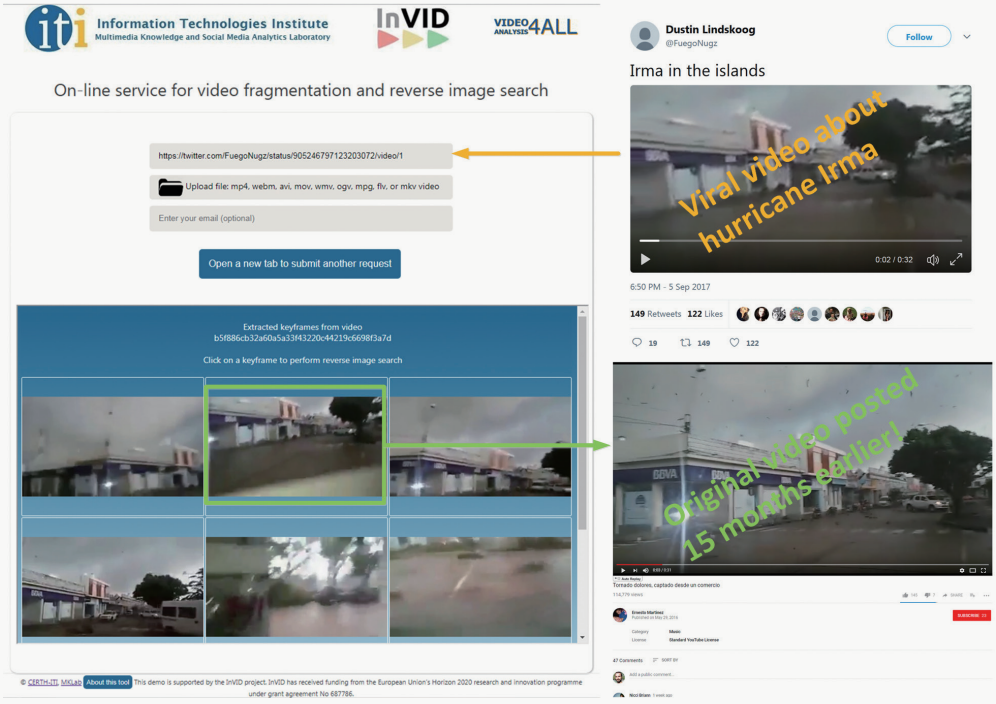
2. 이미지 기반 방식

- **유럽연합(EU)이 2016년 1월부터 EU의 혁신 프로그램인 호라이즌 2020(Horizon 2020)[9]의 자금으로 지원하고 있는 인비드(InVID) 프로젝트[10]는 SNS 등을 통해 폭넓게 유통되고 있는 동영상의 진위 여부를 자동으로 파악할 수 있는 기술적 장치를 개발하고 있음**
  - 인비드가 개발한 도구를 사용하면 방송사, 뉴스 에이전시, 웹 언론사, 신문사 등이 동영상을 포함한 소셜 미디어 콘텐츠의 신뢰성을 [그림 2]처럼 확인할 수 있음

9) 2008년 금융위기 이후 촉발된 경기침체 해소, 경제 시스템 안정화, 경제적 기회 창출 요구에 대응할 수 있도록 R&D 부문 혁신을 강화하고 투자를 확대하기 위해 구축된 EU 최대 규모의 연구기금 지원 프로그램

10) https://www.invid-project.eu

그림 2 | 인비드의 리버스 이미지 검색을 통한 영상 진위 여부 판단



※ 출처 : https://www.invid-project.eu/other-invid-technologies/

- 예를 들어, [그림 2]처럼 허리케인 어마(Irma)와 관련한 영상이 소셜 미디어 등에 게시되면 이 영상을 프레임별로 이미지로 추출한 후 인비드가 보유하고 있는 이미지 데이터 베이스와 비교해서 해당 영상이 15개월 전에 이미 게시된 영상임을 밝혀냄
- 인비드의 검증 소프트웨어(InVID Verification Application)는 입력된 동영상이 과거 뉴스 등에 사용된 전력이 있는지를 확인해 주며, 소셜미디어 기반의 맥락 분석(contextual analysis)을 실시해 해당 영상과 관련된 주변 정보를 제공해 줄 뿐만 아니라, 프레임 단위로 다양한 필터 기반의 동영상 포렌식을 실시, 프레임 수준에서 영상에 특이사항이 있는지를 확인함

● **구글도 이미지 기반의 딥페이크 탐지 기술 개발을 위해 훈련용 이미지 데이터셋(FaceForensics dataset)[11] 을 제공하고 있음**

- 구글의 데이터는 얼굴 이미지로 한정돼 있지만, 그동안 가짜로 밝혀진 이미지들, 원본 얼굴 이미지 등을 폭넓게 제공해 기존 이미지로부터 딥페이크 조작 여부를 판단할 수 있도록 돕고 있음

11) https://github.com/ondyari/FaceForensics/

3. 인공지능 기반 방식

- **미국 국방성의 연구, 개발 부문을 담당하고 있는 방위고등연구계획국(DARPA)의 '미디어 포렌식(Media Forensics: MediFor)[12]' 프로그램은 어떤 이미지나 동영상이 딥페이크와 같이 AI를 이용해 수정되거나 변형되었는지의 여부를 판단할 수 있는 기술을 개발하고 있음**
  - 미디어 포렌식은 이미지 수집 및 분석 등 전통적인 디지털 포렌식 기법을 자동화하기 위해 만들어졌으나 2018년 하반기에 AI가 작성한 위조 이미지/영상 감별로 초점을 전환했음
  - 컴퓨터 비전(computer vision) 분야 전문가인 매튜 튜렉(Matthew Turek) 박사가 이 프로그램을 이끌고 있으며, 그는 이미지와 동영상의 일부가 변형되었음을 알 수 있는 미묘한 신호들을 발견했고 이를 복합적으로 활용해 진위 여부를 판단하고 있다고 밝혔음(Knight, 2018. 8. 7)
  - 눈 깜빡임, 머리 움직임, 눈 색깔 등 기존에 알려진 것뿐만 아니라 인공지능을 활용해 인간이 알 수 없는 변화 등을 복합적으로 판단해 딥페이크 기술에 의한 조작여부를 판단하고 있음

- **딥트레이스(Deeptrace)는 바이러스 백신 소프트웨어처럼 소셜미디어나 검색 엔진의 백그라운드에서 눈에 띄지 않게 작동하면서 영상 및 이미지를 탐색하는 딥페이크 탐지 소프트웨어를 개발했음**
  - 딥트레이스는 딥페이크를 단순한 얼굴의 교체가 아니라 인체 움직임을 모방하고 사람의 음성을 합성하는 기술로 폭넓게 정의하고 사이버보안에 상당한 위험을 초래하는 기술로 간주함
  - 딥트레이스는 바이러스 백신 프로그램처럼 딥페이크에 항시 대응 가능한 인공지능 시스템을 기존의 사용 환경 위에서 구축하는 것이 중요하다고 강조하면서 2019년에 자신들의 시스템으로 발견한 인터넷 상 딥페이크 영상의 수는 14,678개였으며 이 중 96%가 음란물이었다고 밝힌 바 있음(Ajder, Patrini, Cavalli & Cullen, 2019)

Ⅳ. 딥페이크 검증 확인을 위한 도구 제안

- **이렇듯 기술적 대응 노력들이 이어지고 있지만, DARPA의 기술자들은 "우리는 이 전투에서 지고 있다"라고 말하고 있음(Knight, 2018. 5. 23)**
  - 실제 사진에 합성하거나 실제 사진을 조작하는 경우는 흔적을 남기지만, 기계가 생성한 이미지는 그 흔적조차 알 수 없기 때문이라는 것임
  - 기술의 발전으로 필립 왕(Phillip Wang)이 만든 웹사이트 '이 사람은 존재하지 않는다(This Person Does Not Exist)'[13]처럼 실제로 존재하지 않는 사람들의 이미지도 생성되고 있음

12) https://www.darpa.mil/program/media-forensics
13) https://thispersondoesnotexist.com/

- **딥페이크 기술의 발전은 그나마 우리가 믿을 수 있다고 생각했던 동영상마저도 조작이 가능하다는 생각을 일반화시켜 허위정보 등으로 인한 문제와 더불어 기술적으로 조작가능하지 않은 미디어는 없다는 일반적 인식을 심화시킬 수 있음**
  - 알고리즘적 미디어 합성에 의해 어떠한 사건도 조작될 수 있음을 우리 사회가 모두 알게 되면 어떠한 일이 발생할 것인가
  - 결국 매개된 현실(mediated reality)의 평가자들이 더 중요한 역할을 해야 할 것으로 이는 기존 저널리스트들에게 새로운 기회가 될 수 있음
  - 고도로 발전하는 딥페이크 기술의 조작 여부 판단은 디지털 포렌식 도구로 무장한 저널리스트들보다 더 잘 할 사람들은 없으며, 이를 위해 언론사와 저널리스트들은 디지털 도구에 대한 훈련, 테크놀로지에 대한 이해 함양, 업무 투명성 강화 등의 조치가 필요함(Diakopoulos, 2018. 5. 15)
  - 이러한 관점에서 딥페이크 이미지 및 영상 조작여부를 판단하는데 활용 가능한 다양한 도구들을 소개[14]하고 일상생활에서 딥페이크 조작 여부를 판단할 수 있는 방법을 소개함

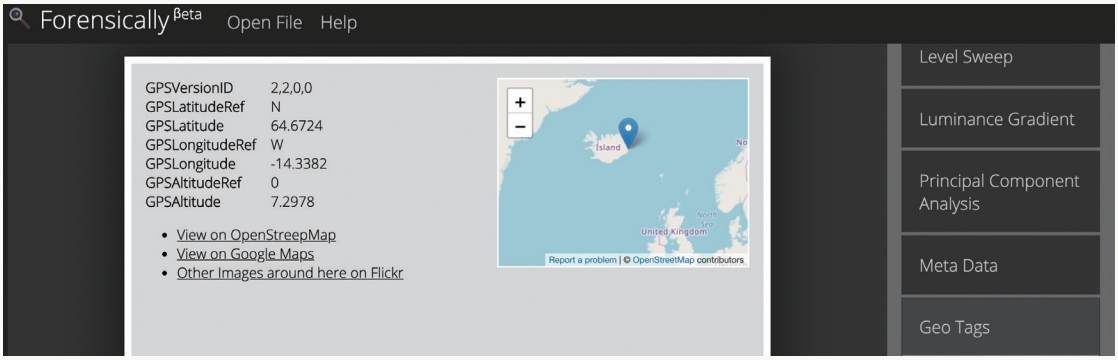
1. 저널리스트들을 위한 검증 도구 제안

① ExifTool : https://www.sno.phy.queensu.ca/~phil/exiftool/

- 다양한 이미지 파일들을 읽어들여 해당 메타데이터들을 확인해 주는 기능을 제공함
- 사진 촬영일시, 장소 등을 확인하여 해당 사진이 언제, 어디서, 어떻게 작성됐는지를 알 수 있게 해 조작 여부 판단을 가능하게 해 줌
- 개인 PC에 별도 설치해야 하며, 버전 업데이트가 잘 돼 최신 내용 반영 중

② Image Forensics : https://29a.ch/photo-forensics/

그림 3 | Image Forensics에 이미지 파일 업로드 후 Geo Tag 확인 화면

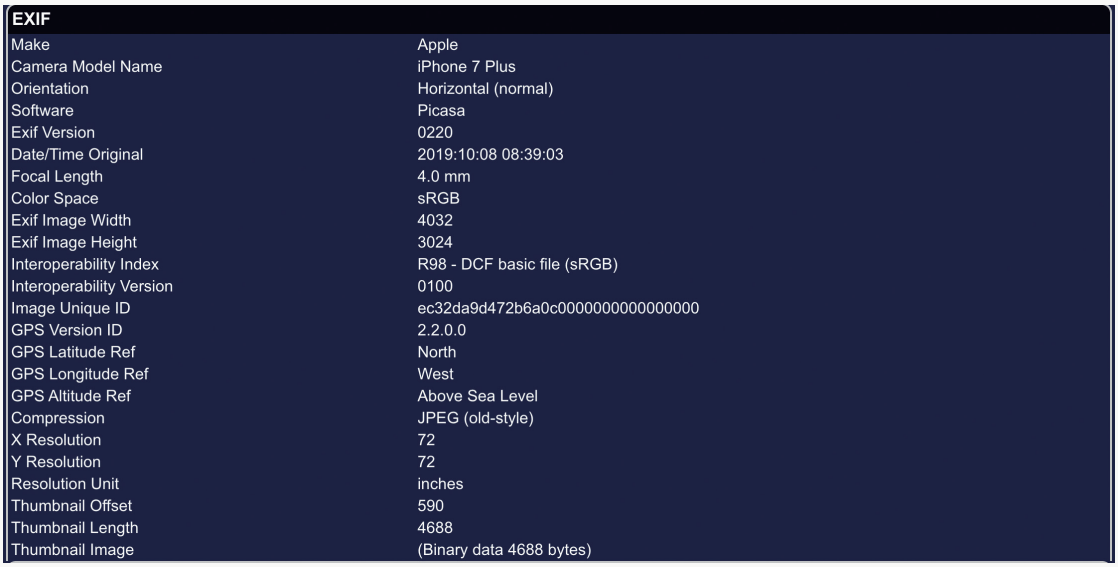


14) 여기서 소개하는 도구들은 'OSINT Essentials'에서 추천한 내용(http://bitly.kr/SYQk62tE)에 바탕을 뒀음

- 사이트에 접속 후 이미지를 업로드하면 해당 이미지가 원본인지 복사 됐는지 여부를 지수로 수치화해 보여줌
- 복사여부 외에도 메타데이터, 이미지 파일의 속성, 수정 여부 등 확인 가능한 다양한 수치들을 제공해 이미지 전문가의 경우 해당 이미지 원본 여부 파악을 쉽게 할 수 있음
- 다만, 아직은 베타여서 지원되는 파일 형식에 제한이 있음

③ FotoForensics : http://fotoforensics.com/

그림 4 | FotoForensics에 이미지 업로드 후 메타데이터 확인 화면



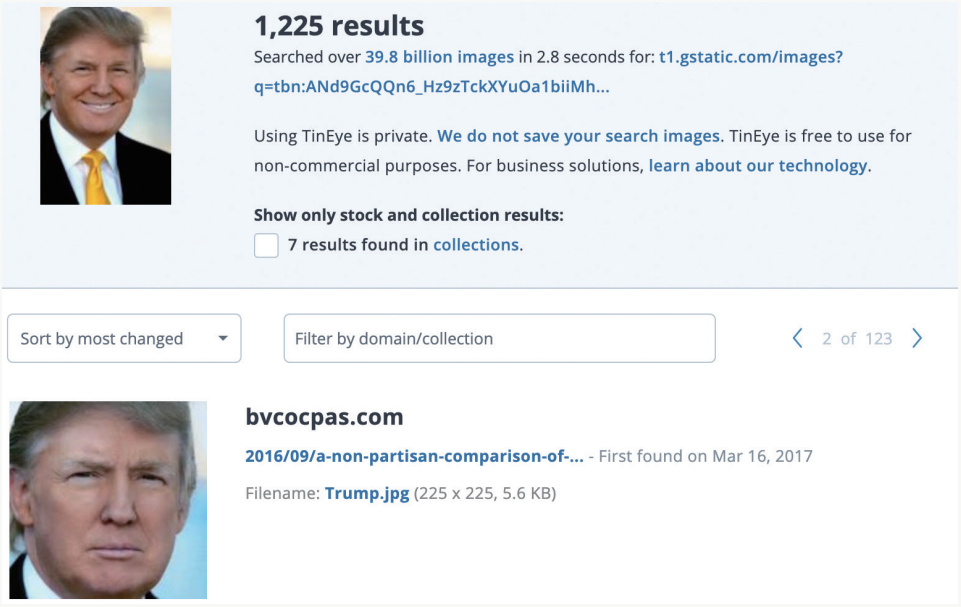
EXIF	
Make	Apple
Camera Model Name	iPhone 7 Plus
Orientation	Horizontal (normal)
Software	Picasa
Exif Version	0220
Date/Time Original	2019:10:08 08:39:03
Focal Length	4.0 mm
Color Space	sRGB
Exif Image Width	4032
Exif Image Height	3024
Interoperability Index	R98 - DCF basic file (sRGB)
Interoperability Version	0100
Image Unique ID	ec32da9d472b6a0c0000000000000000
GPS Version ID	2.2.0.0
GPS Latitude Ref	North
GPS Longitude Ref	West
GPS Altitude Ref	Above Sea Level
Compression	JPEG (old-style)
X Resolution	72
Y Resolution	72
Resolution Unit	inches
Thumbnail Offset	590
Thumbnail Length	4688
Thumbnail Image	(Binary data 4688 bytes)

- ExifTool과 마찬가지로 다양한 이미지 파일들을 읽어들여 메타데이터들을 확인해 줌
- 별도 설치 없이 웹 기반으로 운영되고 있어 취재나 출처 확인 필요시 언제든 쉽게 활용 가능
- ExifTool과 마찬가지로 사진 촬영일시, 장소 등을 확인하여 해당 사진이 언제, 어디서, 어떻게 작성됐는지를 알 수 있게 조작 여부 판단을 가능하게 해 줌
- 특히 소셜 미디어 상에서 확인한 이미지 파일의 검증이 필요할 경우에 유용함

④ 리버스 이미지 검색 : https://tineye.com/

- 선택한 이미지를 업로드하거나 웹 상의 이미지 주소를 입력하면 비슷한 이미지를 찾아 줌
- [그림 5]처럼 자신들이 확인한 이미지들 중에서 비슷한 이미지가 얼마나 있는지를 보여준 후 각 이미지들의 출처와 간략한 정보를 제공함
- 웹 기반으로 사용이 쉬우며, 검색한 이미지를 가장 많이 바뀐 순, 가장 비슷한 순, 용량이 큰 순, 최신 수, 오래된 순으로 정렬해 보여줘 구글 이미지 검색보다 편리함

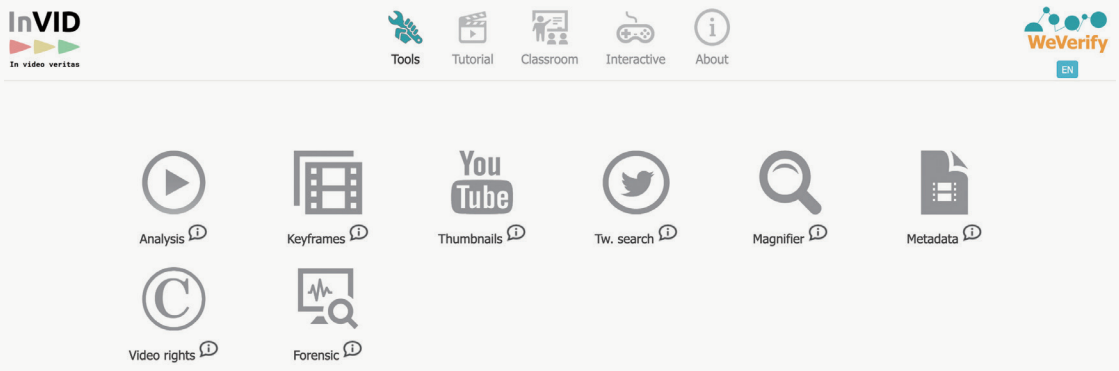
그림 5 | TinEye에서 리버스 이미지 검색 결과



⑤ InVID : https://www.invid-project.eu/tools-and-services/invid-verification-plugin

- 앞서 EU에서 추진 중인 사례로 소개했던 InVID 프로젝트에서 저널리스트들을 위해 만든 웹 브라우저 크롬 및 파이어폭스 플러그인으로 설치하면 페이크 뉴스와 이미지, 비디오를 검증해 줌
- 플러그인을 설치하면 동영상 url, 이미지 url로 검색할 수 있는 기능을 제공하며 [그림 6]과 같이 InVID 페이지에서 통합 검증이 가능한 화면을 제공함
- [그림 6]처럼 통합 검증 페이지를 통해 간략한 분석, 키프레임 분석, 썸네일, 트위터, 페이스북 영상, 메타데이터, 비디오 저작권 등을 확인할 수 있음
- 저널리스트가 이러한 복합적 정보를 활용해 해당 영상 및 이미지 조작 여부를 판단하게 도움

그림 6 | InVID의 통합 검증 화면



⑥ 유튜브 데이터 뷰어 : https://citizenevidence.amnestyusa.org/

- 해당 사이트에 접속 후 검증하고 싶은 유튜브 영상의 주소를 입력하면 해당 영상이 올라온 정확한 날짜 및 시간을 보여주고 비슷한 영상이 올라온 인스타그램 및 페이스북 주소를 제시함
- 해당 영상의 썸네일 및 주요 프레임 이미지를 제시하고 이를 리버스 검색할 수 있도록 제공함
- InVID 플러그인의 통합 검증 결과랑 같이 활용하면 유용함

⑦ Geo Search Tool : http://youtube.github.io/geo-search-tool/search.html

- 특정 도시, 지역을 검색하여 해당 위치의 유튜브 영상을 검색해서 제공함
- 위치뿐만 아니라 특정 시점도 정할 수 있으며 허리케인, 화재 등 특정 사건을 키워드로도 검색 결과를 제공함
- 이를 바탕으로 특정 영상을 일련의 영상들과 비교해 조작여부를 판단할 수 있지만, 기능이 안정화되어 있지는 않음

표 1 | 이미지 및 동영상 검증 도구들

도구 명	특징	작동 방식	URL
ExifTool	입력한 이미지의 메타데이터 확인	PC 설치형	www.sno.phy.queensu.ca/~phil/exiftool
Image Forensics	입력한 이미지의 메타데이터 및 복사 여부 확인	웹 기반	29a.ch/photo-forensics
FotoForensics	소셜 미디어 상 이미지의 메타데이터 확인	웹 기반	fotoforensics.com
TinEye	입력한 이미지의 리버스 검색	웹 기반	tineye.com
InVID	이미지 및 동영상 통합 검증	브라우저 플러그인	www.invid-project.eu
유튜브 데이터 뷰어	유튜브 영상의 메타데이터 제공	웹 기반	citizenevidence.amnestyusa.org
Geo Search Tool	특정 위치 기반으로 유튜브 영상 검색	웹 기반	youtube.github.io/geo-search-tool/search.html

● **이외에도 더 많은 도구들이 있으며 각자에게 맞는 도구들이 있을 것임**

- 지금 소개한 도구들은 각자의 장단점이 있기 때문에 어느 하나만을 고정해 사용하기보다는 여러 개를 복합적으로 활용하면서 검증하고 싶은 영상 및 이미지에 대해 깊이 있게 분석하는 것이 좋음

2. 일상생활에서 검증 방법 제안

● **대부분의 경우 일상생활에서 딥페이크 등이 적용돼 의심이 필요한 내용에 대해서도 쉽게 넘어가는 경우가 있는데 이러한 문제를 해결하기 위해서는 간단한 방법을 통해 일상적으로 보고 있는 내용에 대해 의심하는 습관이 필요함**

- 이러한 관점에서 버즈피드에서 소개한 딥페이크 영상을 구별하기 위한 방법 5가지[15]를 소개함
  1) 바로 누구라고 결론을 내리지 말라(Don't jump to conclusions)
  2) 영상의 출처를 살펴봐라(Consider the source)

15) https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed

3) 다른 곳에도 있는지 검색해 봐라(Check where else it is (and isn't) online)
4) 화자의 입모양을 면밀하게 살펴봐라(Inspect the mouth)
5) 천천히 돌려 봐라(Slow it down)

- **딥페이크를 탐지하는 기술이 발전하고 저널리스트들이 이를 활용하여 조작 영상을 실시간으로 식별하여 제공하더라도 사람들이 조작 내용만을 믿고 기존 생각을 바꾸지 않는다면 딥페이크로 인한 문제는 해결될 수 없음**
  - 온라인 상 허위 이미지에 대한 신뢰 평가에 출처 유무보다 해당 이미지 이용자의 인터넷 이용 능력과 디지털 이미지 편집 경험 및 미디어 이용 정도가 더 큰 영향을 미친다는 연구결과(Shen, Kasra, Pan, Bassett, Malloch & O'Brien, 2019)가 이를 잘 보여줌
  - ASI 데이터 사이언스(ASI Data Science)의 존 깁슨(John Gibson)은 2018년 열린 제15회 얄타 유럽 전략 회의(Yalta European Strategy)에서 영상과 음성 합성 기술의 가능성에 관한 사람들의 인식을 높이는 것이 중요하다며, 딥페이크를 '램프에서 빠져나온 지니'에 비유하면서, 영상 합성 기술이 앞으로 대중들에 의해 더 많이 이용될 것이고 이러한 발전을 정부가 통제할 수는 없을 것으로 전망했음(최순욱·이소은, 2019, 42쪽)
  - 결국 딥페이크 등 허위 정보에 대한 법규제, 기술적 대응, 언론의 사실 확인 등도 중요하지만 이용자들의 인식 제고가 무엇보다 중요하다고 할 수 있음

Ⅴ. 마치며

- **허위정보로서 딥페이크 기술의 위험성에 대해서는 많은 우려의 목소리가 나오고 있지만, 이에 대한 대비는 아직 우리 사회에서 부족한 편임**
  - 현재 한국에서도 법제도 등 여러 대응책이 준비 중이지만 딥페이크는 그 기술적 속성상 정의되기도 어렵고 식별도 해내기가 어려움
  - 특히 선거 등 중요한 민주적 절차에 큰 영향을 미칠 수 있는 딥페이크로 인한 문제를 방지하기 위해 선거기간 동안 딥페이크를 의도적으로 생산하거나 이를 알면서 유포하는 자에 대한 대처방안 마련을 고려할 필요가 있음
  - 딥페이크 식별 기술의 개발 등도 중요하며 현재 개발된 도구들의 활용을 통해 의심스러운 영상을 걸러내고 올바른 정보를 제공하는 것이 더 중요함

- **딥페이크로 인해 믿을 수 있는 미디어 형식으로서의 영상에 대한 신뢰도 저하가 예상됨**
  - 이러한 미디어 형식에 대한 의심은 해당 내용을 전달하는 메신저, 즉 저널리스트와 언론에 대한 신뢰 회복으로 대응하는 것이 가장 현실적일 것임
  - 언론과 저널리스트들은 딥페이크로 인한 새로운 현상에 대비하기 위해 관련 기술에 대한 이해 및 다양한 식별 도구 활용법 등에 익숙해질 필요가 있음
  - 딥페이크가 더욱 심각한 사회적 문제로 부상할수록 신뢰할 수 있는 언론의 필요성에 대한 요구가 더 높아질 것임
  - 따라서 언론도 엄정한 사실 확인을 통해 정확한 기사를 발행하려는 노력을 지속적으로 제고해야 함

참고문헌

박준·조영호 (2019). 딥페이크 영상 탐지 관련 기술 동향 연구. 한국정보과학회 학술발표논문집, 724-726쪽.

최순욱·오세욱·이소은 (2019). 딥페이크의 이미지 조작: 심층적 자동화에 따른 사실의 위기와 푼크툼의 생성. <미디어, 젠더 & 문화> 34권 3호, 339-380쪽.

최순욱·이소은 (2019). 딥페이크와 사실의 위기 : 어떻게 대응할 것인가? <해외미디어동향> 2019-01호.

Ajder, H., Patrini, G., Cavalli, F. & Cullen, L. (2019). The State of Deepfakes: Landscape, Threats, and Impact. Deeptrace.

Christopher, N. (2020. 2. 18). We've Just Seen the First Use of Deepfakes in an Indian Election Campaign. Vice. Retirieved from https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

Diakopoulos, N. (2018. 5. 15). Reporting in a Machine Reality: Deepfakes, misinformation, and what journalists can do about them. Columbia Journalism Review. Retrieved from https://www.cjr.org/tow_center/reporting-machine-reality-deepfakes-diakopoulos-journalism.php

Horowitz, C. M., Allen, C. G., Saravalle, E. Cho, A., Frederick, K. & Scharre, P. (2018). Artificial Intelligence and International Security. the Center for a New American Security. Retrieved from https://www.cnas.org/publications/reports/artificial-intelligence-and-international-security

Knight, W. & Hao, K. (2019. 1. 7). Never mind killer robots—here are six real AI dangers to watch out for in 2019. MIT technology review. Retirieved from https://www.technologyreview.com/s/612689/never-mind-killer-robotshere-are-sixreal-ai-dangers-to-watch-out-for-in-2019/

Knight, W. (2018. 5. 23). The US military is funding an effort to catch deepfakes and other AI trickery. MIT technology review. Retirieved from https://www.technologyreview.com/s/611146/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery/

Knight, W. (2018. 8. 7). The Defense Department has produced the first tools for catching deepfakes. MIT technology review. Retirieved from https://www.technologyreview.com/s/611726/the-defense-department-has-produced-the-first-tools-for-catching-deepfakes/

Li, Y., Chang, M. C., Farid, H., & Lyu, S. (2018). In ictu oculi: Exposing ai generated fake face videos by detecting eyeblinking. arXiv:1806.02877 [cs.CV] Retrieved from https://arxiv.org/abs/1806.02877

Shen, C., Kasra, M., Pan, W., Bassett, G. A., Malloch, Y., & O'Brien, J. F. (2019). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. New Media & Society, 21(2), pp. 438-463.

Media

2020년 1호 정책 리포트

발행인 민병욱
편집인 김철훈
기　획 한국언론진흥재단 미디어연구센터
발행일 2020년 3월 9일

한국언론진흥재단 미디어연구센터
04520 서울특별시 중구 세종대로 124 프레스센터빌딩 13층
전화 (02) 2001-7750 팩스 (02) 2001-7740
www.kpf.or.kr

편집 (주)나눔커뮤니케이션
04034 서울특별시 마포구 잔다리로7길 16 교평빌딩 304호
전화 (02) 333-7136 팩스 (02) 333-7146