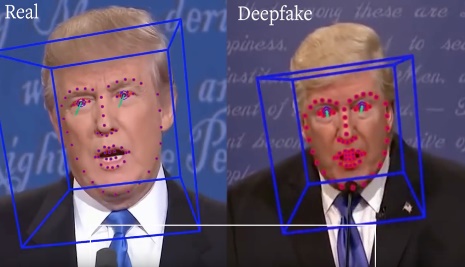
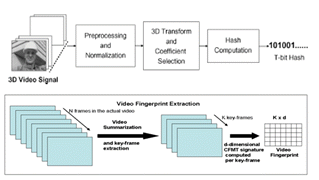
**적대적 노이즈를 활용한 딥페이크에 강인한 영상 생성**

**Generating Video Robust Against Deep-fake Using Adversarial Noise**

**1. 서론**

기존 anti-deepfake 기술로는 아래 그림과 같이 crytographic, tamper-proof, deepfake detection program이 있다. 하지만 crytographic 기술은 불특정 수 많은 사람들을 타겟으로 하는 비디오 매체 특성 상 적합하지 않으며 tamper-proof 와 deepfake detection program을 이용하는 방법은 동영상을 접한 사람들이 추가적인 작업으로 동영상의 무결성을 확인해야 한다는 단점이 있다.

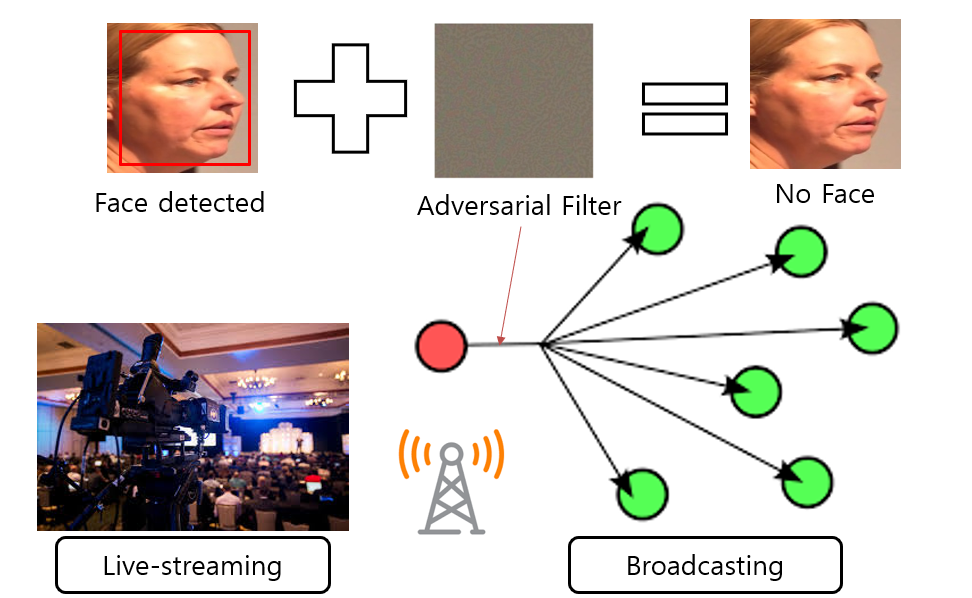


[그림 1] Tamper-proof

[그림 1] 딥페이크 탐지 프로그램

[그림 1] 동영상 암호화

따라서 딥페이크의 제작을 방해하고자 어드버셜 필터를 동영상에 삽입하여 얼굴 탐지를 불가능하게 하는 방법을 제안한다. 아래 그림은 사용할 수 있는 시나리오의 예이다.



[그림 1] 제안 방법 사용 예시

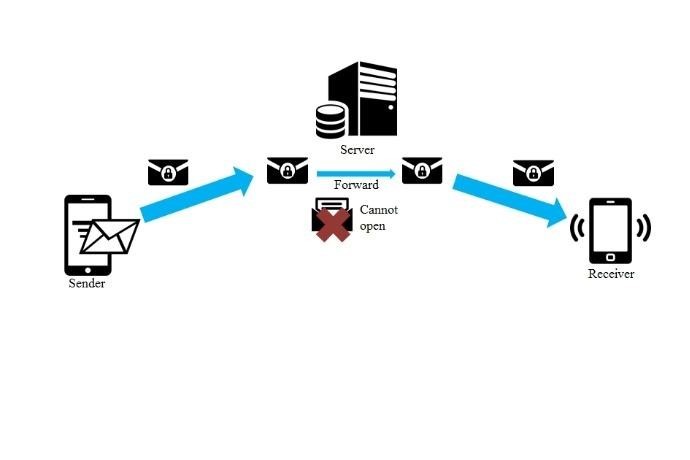
현재 다양한 플랫폼에서 실시간으로 방송되는 동영상을 그대로 가져와 중계하는 방식의 동영상 송출이 급증하고 있다. 이는 TV로 시청하는 수보다 모바일 기기를 사용하는 시청하는 수가 더 많기 때문에 잘못된 정보를 전달하는 큰 문제를 야기할 수 있다. 라이브 스트리밍의 원래 송출지의 동영상을 얻어 딥페이크로 다른 사람이 말과 행동을 한 것같이 속여 해당 동영상을 다시 브로드캐스팅 할 수 있다. 이러한 악용을 막고자 라이브 스트리밍의 최초 송출지에서 어드버셜 어택을 응용한 필터를 만들어 동영상에 삽입시킨다. 이러한 작업은 딥페이크 제작 과정에서 AI가 얼굴을 탐지해야 하는 과정을 불가능하게 만들거나 어렵게 하여 동영상 조작이 불가능하거나 수준이 낮은 조작된 동양상을 유도할 수 있다. 이러한 최초 송출지에서 어드버셜 필터를 적용한 동영상을 송출하게 된다면 그 이후 수많은 플랫폼으로 퍼쳐가는 동영상의 조작 및 변형을 막을 수 있다. 또한 사용자들이 해당 동영상이 진짜인지 가짜인지 판별하기 위해 추가적인 작업을 할 필요가 없게 된다. 하지만 어드버셜 필터를 적용하는 과정에서 AI는 얼굴을 탐지하지 못하지만 사람들은 해당 동영상이 원본과 다른점을 찾을 수 없을만큼 변화가 적은 어드버셜을 적용해야 한다는 점이 있다.

**2. 관련 연구**

**2.1. Anti-fake technology**

**2.1.1. Cryptography**

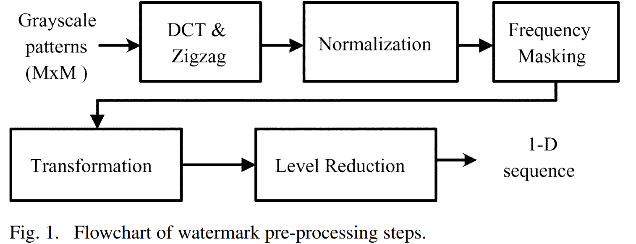
디지털 컨텐츠는 원본 영상이 전송 중 탈취당해 딥페이크 영상이 제작되어 송출될 수 있다. 암호화는 디지털 컨텐츠를 전송하기 전 암호화하여 전달과정에서 발생할 수 있는 제3자에 의해 조작되는 것을 방지할 수 있다. 동영상의 프레임을 블록단위로 나누어 암호화한다. 또한 연산량을 줄이기 위해 동영상의 중요한 부분만 암호화하는 방식이 사용되기도 한다. 동영상을 암호화할 때 암호화된 데이터가 압축과정에서 훼손된다면 디코더에서 복구가 불가능하기 때문에 암호화는 양자화 단계 이후 무손실 압축구간에서 수행된다. 대표적 국제 표준 암호 알고리즘은 AES, DES 등이 있다.

****

**2.1.2. Tamper-proof**

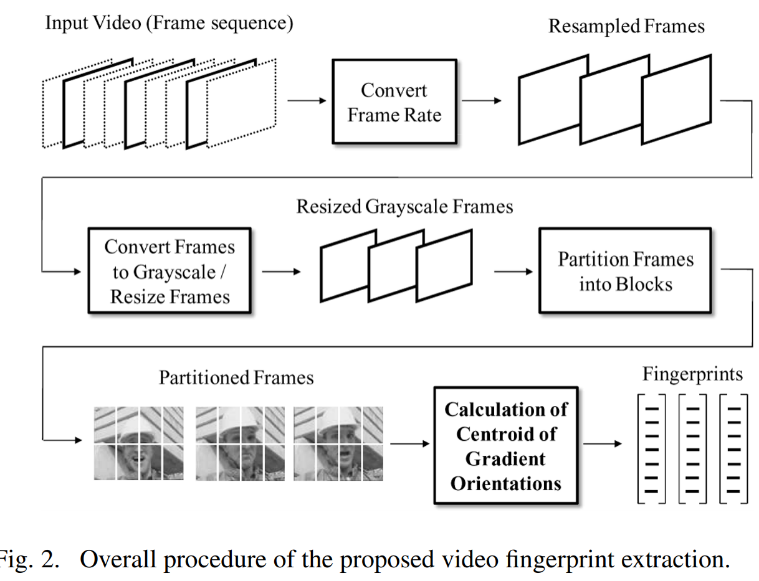
**1. 워터마킹**

워터마킹은 디지털 컨텐츠의 위조나 변조여부를 확인할 수 있는 기술을 뜻한다. 워터마킹은 컨텐츠 보호, 저작권 보호, 불법복제 보호를 위해 사용된다. 소비자는 이용중인 디지털 컨텐츠가 제공자로부터 제공된 원본임을 검증할 수 있다. 워터마크 삽입방법에는 공간영역 삽입방법과 주파수영역 삽입방법이 있다. 공간영역 삽입은 공격과 압축에 약하기 때문에 주파수 영역삽입이 주로 사용된다. 아래 그림은 주파수 영역 워터마크 삽입방법의 흐름을 보인다. 워터마크는 이미지를 블락단위로 잘라 얻은 DCT를 Zigzag 스캐닝해 얻은 시퀀스에서 중요한 정보가 담긴 저주파지역을 중심으로 삽입된다. 워터마킹은 3가지 요소가 잘 설계되어야 한다. 1.Transparency는 워터마킹이 삽입된 데이터가 원본 데이터와 동일한지를 의미한다.2.Robustness는 삽입된 워터마킹이 다운그레이가 되어도 신뢰성 있게 사용할 수 있는지를 의미한다. 3.Capacity는 디지털컨텐츠의 삽입될 수 있는 정보량을 의미한다.



**2. 핑커프린트**

핑거프린트란 원본 데이터에 삽입하여 조작이 되더라도 작성자를 증명할 수 있도록 지문과 같은 데이터를 의미한다. 핑거프린트 추출 과정은 아래 그림과 같다. 동영상을 프레임단위로 나눈다. 프레임을 그레이스케일로 전환한다. 설정한 크기로 그레이스케일 프레임 사이즈가 재설정 된다. 각 프레임을 블록 단위로 분리한다. 각 블록은 그래디언트 방향의 중심으로 계산된다. 이것을 이용하여 핑거프린터 벡터를 생성한다.



**2.2. Deepfake creation**

**2.2.1. Autoencoder**

**2.2.2. GAN**

**3. 제안 시스템**



[그림 1] 제안 시스템 처리 흐름도

[그림 1]은 제안 시스템을 사용해 딥페이크에 강인한 영상을 생성하는 과정을 도식화한 것이다. 제안 시스템은 프레임 추출 모듈, 얼굴 검출 모듈, 적대적 노이즈 삽입 모듈, 영상 생성 모듈로 구성된다. 프레임 추출 모듈은 원본 영상의 프레임을 추출하여 각 프레임을 이미지 파일로 저장한다. 프레임 추출 모듈이 추출한 이미지 파일은 얼굴 검출 모듈의 입력 데이터로 사용한다. 얼굴 검출 모듈은 각 프레임에서 사람의 얼굴을 추출한다. 얼굴 검출을 위해 Haar[1], HOG[2] 등 객체 검출 알고리즘을 사용한다. [그림 2]는 Haar 알고리즘을 사용해 이미지에서 얼굴을 검출한 사진이다. 이미지에서 추출된 얼굴 부분은 적대적 노이즈 삽입 모듈로 전달하여 해당 부분이 얼굴로 인식되지 않는 수준의 노이즈를 삽입한다. [그림 3]은 적대적 노이즈 삽입 모듈을 사용해 JSMA[3] 알고리즘으로 생성된 적대적 노이즈를 삽입한 후, [그림 2]에서 사용한 것과 동일한 알고리즘을 사용해 얼굴을 검출한 사진이다. [그림 2]와 달리 얼굴 부분을 정상적으로 검출하지 못하는 것을 확인할 수 있다. 적대적 노이즈가 삽입된 각 프레임을 영상 생성 모듈이 취합하여 다시 인코딩함으로써 딥페이크에 강인한 영상을 생성할 수 있다.

[그림 3]의 이미지를 생성하기 위해 JSMA 알고리즘을 사용해 수준의 적대적 노이즈를 삽입했다. 이러한 적대적 노이즈 삽입은 프레임에서 얼굴을 검출하지 못하도록 하는데 성공했으나 노이즈로 인해 원본 영상보다 선명도가 떨어져 보이는 것을 확인할 수 있다. 따라서, 육안으로 확인하기 어려운 수준이면서 동시에 얼굴을 검출하지 못하도록 하는 적절한 적대적 노이즈 삽입 수준을 결정할 방법이 필요하다. 3.1장에서 적절한 적대적 노이즈 삽입 수준을 결정하기 위한 처리 과정을 자세히 설명한다.

A person wearing a graduation cap

Description automatically generated with low confidence

[그림 2] Haar 알고리즘을 이용한 얼굴 검출

A person wearing a graduation cap

Description automatically generated with medium confidence

[그림 3] 적대적 노이즈 삽입 후 얼굴 검출