디지털 워터마킹 공격 탐지를 위한 계층적 워터마킹 기법

김도은[1], 박소현[2], 이일구[3]
[1]성신여자대학교 융합보안공학과 학부생
[2]성신여자대학교 미래융합기술공학과 박사과정
[3]성신여자대학교 융합보안공학과/미래융합기술공학과 교수
20221080@sungshin.ac.kr, 220227022@sungshin.ac.kr, iglee@sungshin.ac.kr

Hierarchical watermarking technique for detecting digital watermarking attacks

Do-Eun Kim[1], So-Hyun Park[2], Il-Gu Lee[1,2]
[1]Dept. of Convergence Security Engineering, Sungshin Women's University
[2]Dept. of Future Convergence Technology Engineering, Sungshin Women's University

요 약

디지털 워터마킹은 디지털 컨텐츠에 정보를 삽입하는 기술이다. 종래의 디지털 워터마킹 기술은 견고성과 비가시성 사이에 트레이드오프 관계를 가지고, 변형 및 노이즈 공격 등에 취약하다. 본 논문에서는 호스트 이미지의 비가시성을 보장하면서 효율적인 공격 탐지와 소유자 식별이 가능한 워터마킹 기법을 제안한다. 제안한 방식은 주파수 분할 기반의 계층적 워터마킹 및 공격 탐지 시그니처 삽입을 통해 비가시성을 보장하며 용량과 견고성 측면에서 종래의 방법보다 향상된 성능을 보였다. 실험 결과에 따르면 종래의 디지털 워터마크가 무력화되는 왜곡 공격 상황에서 공격 탐지 시그니처 검출이 가능하여 워터마크 공격을 탐지하고 소유자를 식별할 수 있었다.

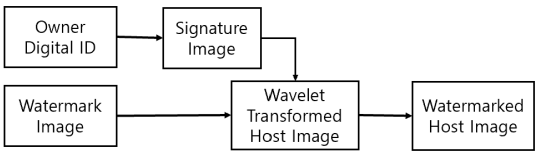
1. 서론

디지털 워터마킹(Digital Watermarking)은 디지털 컨텐츠에 정보를 삽입하는 기술이다. 최근 생성형 인공지능 기술이 널리 활용되면서 디지털 컨텐츠의 저작권 및 소유권을 보호하고 추적하기 위한 디지털 워터마킹이 주목받고 있다 [1]. 대표적인 워터마킹 알고리즘으로는 LSB(Least Significant Bit)를 활용한 공간 도메인 방식과 DCT(Discrete Cosine Transform), DWT(Discrete wavelet transform), SVD(Singular Value Decomposition)와 같은 주파수/변환 도메인 방식이 연구되었다. 그러나, 종래의 디지털 워터마킹은 노이즈, 적대적 공격, 압축, 변형 공격에 취약하다 [2].

본 논문에서는 워터마크가 훼손되어 검출되지 않는 상황에서도 워터마크 공격을 탐지하고 워터마크 존재를 확인하기 위하여, 호스트 이미지의 주파수 대역을 분할하고 워터마크와 공격 탐지 시그니처를 계층적으로 삽입하는 워터마크 위변조 탐지 기법을 제안한다.

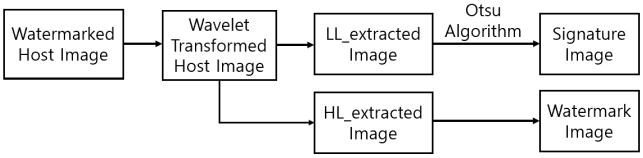
2. 계층적 워터마킹 방법

워터마크 훼손 공격 탐지 및 소유자 식별을 위해 정보주체의 디지털 ID를 이미지화한 공격 탐지 시그니처를 생성하고 워터마크 이미지와 함께 임베딩한다. 워터마크 삽입 알고리즘으로는 DWT를 이용하였다.



(그림 1) 삽입 알고리즘

제안하는 워터마크 및 공격 탐지 시그니처 삽입 알고리즘은 그림 1과 같다. 먼저, 호스트 이미지 소유자의 디지털 ID를 이미지화하여 공격 탐지 시그니처를 생성한다. 본 연구에서는 정보주체의 디지털 ID를 의사난수생성기에 입력하여 생성한 10진수 난수 16자리를 이용하였다. 각 10진수를 2진수로 변환한 0/1 시퀀스를 이미지의 흑/백 픽셀로 치환하고 주기적으로 반복 삽입하는 방식으로 비트를 이미지화하여 공격 탐지 시그니처를 생성한다. 이후, 호스트 이미지를 이산 웨이블릿 변환한 뒤, 위변조 확인을 위한 시그니처 이미지는 저주파(LL, low-low) 대역에 삽입하고 워터마크 이미지는 중주파(HL, high-low) 대역에 삽입한다. 한 번의 웨이블릿 변환으로 분할된 주파수 영역에 계층별로 두 가지 이미지 정보를 임베딩 함으로써 공간이 제한된 호스트 이미지에 더 많은 정보를 삽입할 수 있어서 효율적인 자원 활용이 가능하다.



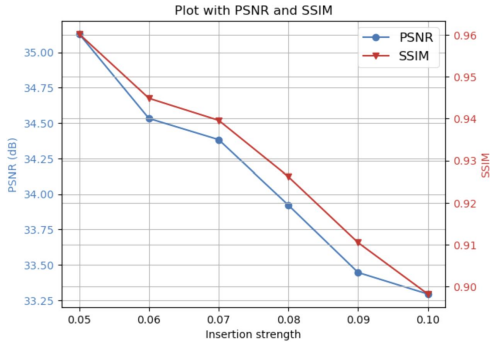
(그림 2) 추출 알고리즘

그림 2는 워터마킹 추출 알고리즘을 보여준다. 워터마킹된 이미지를 이산 웨이블릿 변환 뒤 LL 대역과 HL 대역에서 삽입한 이미지를 추출한다. HL 대역에서는 별도의 처리 없이 NC(Normalized Cross-Correlation) 값이 0.933인 워터마크가 추출된다. LL 대역의 경우, 정확한 비트값 추출을 위해 Otsu 알고리즘을 적용하여 최적의 임계값을 기준으로 이진화한다. Otsu 알고리즘 적용할 때 원본 시

그니처 이미지와의 NC 값이 0.99이고 BER(Bit Error Rate)은 0인 시그니처 이미지를 복원할 수 있다. 그리고 흑/백 픽셀값을 이진수 0/1로 변환하여 소유자의 Digital ID를 획득한다. 이 과정에서 노이즈에 강인하도록 추출된 정보 시퀀스의 최빈값을 출력하도록 설계하였다.

3. 성능 평가

512x512 color 이미지를 이용했으며, 시그니처 이미지는 소유자 시퀀스를 819회 삽입하였다. DWT로 Haar-wavelet을 적용하여 호스트 이미지를 주파수 분할하고, 시그니처 이미지는 LL 대역, 워터마크 이미지는 HL 대역에 삽입하였다. HL 대역의 삽입 강도를 0.03으로 고정하고 LL 대역 삽입 강도를 조절하며 결과 이미지의 PSNR과 SSIM(Structural Similarity Index Map)을 측정한 결과, 0.9 이상의 SSIM을 보장하는 0.09를 LL 대역의 삽입 강도로 설정하였다.



(그림 3) 저주파 대역 삽입 강도에 따른 PSNR과 SSIM

Fig 4. a) Host Image, b) Signature Image, c) Logo Image d) Watermarked Image

제안한 방식으로 생성된 이미지는 원본 이미지의 PSNR이 33.44dB로 시각적으로 구분되지 않으며, SSIM이 0.91이므로 구조적으로도 구분되지 않는다. 또한, 종래의 단일 이미지 워터마킹 방식과 다르게 두 이미지를 계층적 구조의 워터미킹하여 고용량 정보를 삽입할 수 있다.

결과 이미지에 가우시안 노이즈 공격, JPEG 압축 공격, 크롭(crop) 공격의 이미지 처리 및 변형 상황에서 추출된 워터마크 품질과 디지털 ID 추출 여부를 실험하였다.

표준편차 0.3의 가우시안 노이즈 공격 시, HL 대역에서 추출된 워터마크 NC는 0.8440으로 극심히 훼손되었다. 하지만 LL 대역에서는 NC 값 0.9895의 공격 탐지 시그니처를 추출하여 786회의 정상 디지털 ID를 검출 가능하였다. 크롭 공격은 원본의 30%만 남긴 상황에서도 잘린 이미지에 삽입된 공격 탐지 시그니처로 디지털 ID를 검출하였다. 이 경우 HL 대역에서 추출된 이미지를 원본과 비교한 NC는 0.0262로 원본 확인이 어렵지만, 30%만 남은 제한적 시그니처 이미지로부터 소유자 증명이 가능하다. JPEG 압축 공격 시 압축 품질을 20까지 저해하면 HL 대역에서 추출한 워터마크 이미지의 NC는 0.5327로 심하게 훼손된다. LL 대역에서 추출한 직후 시그니처 이미지 역시 NC 0.8951로 훼손되지만 Otsu 알고리즘을 적용해 이진화하면 NC 값을 0.9456까지 높여 디지털 ID를 545회 검출했다.

<표 1> 공격 상황별 추출 이미지 상태

공격 유형	추출된 워터마크 NC (종래모델 / 제안모델)	시퀀스 이미지 BER, 시퀀스 검출 수	시퀀스 추출 가능 임계점
Gaussian noise attack	0.8761 / 0.8440	BER : 0.0017, (786/819)	표준편차 0.3
Crop attack	0.0232 / 0.0262	BER : 0.0000, (61/819)	원본의 30% Crop
JPEG Compression attack	0.5636 / 0.5327	BER : 0.0090, (545/819)	압축 품질 20

표 1의 실험 결과에 따르면 단일 이미지만을 삽입한 종래 방식과 제안 방식에서 추출한 워터마크는 모두 훼손되어 원본의 확인이 어렵다. 제안 방식은 공격 탐지 시퀀스 추가 삽입으로 인해 추출 시 NC가 종래 방식 대비 근소히 감소하지만, 워터마크가 손상된 상황에서도 시퀀스 이미지를 통해 워터마크 공격을 탐지하고 디지털 ID를 복원해 정보 주체의 저작권과 소유권을 보장할 수 있다.

4. 결론

본 연구는 디지털 워터마크 위변조 감지 및 소유자 식별 기능을 갖춘 고용량 계층적 워터마킹 방식을 제안하였다. 제안된 알고리즘은 원본 이미지를 이산 웨이블릿 변환한 뒤 HL 대역과 LL 대역에 각각 워터마크 이미지와 소유자의 디지털 ID로 생성한 공격 탐지 시그니처 이미지를 삽입하였다. 실험 결과에 따르면 제안된 방식은 정보의 효율적 삽입뿐 아니라 데이터 위변조 탐지 측면에서 향상된 성능을 보였다. 공격으로 인해 워터마크 이미지가 심하게 훼손되더라도 공격 탐지 시그니처 이미지의 복원 추출을 통해 위변조 탐지 및 소유자 정보 검출이 가능하다.

ACKNOWLEDGMENT

본 논문은 2024년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Shweta Wadhera, Deepa Kamra, Ankit Rajpal, Arun a Jain, & Vishal Jain. (2021). A Comprehensive Review on Digital Image Watermarking.

[2] Evsutin, O., & Dzhanashia, K. (2022). Watermarking schemes for digital images: Robustness overview. Signal Processing: Image Communication, 100, 116523.