왜곡 공격에 강인한 디지털 워터마킹 기법

김수경[1], 전유란[2], 류정화[3], 이일구[4]
[1]성신여자대학교 융합보안공학과 학부생
[2]성신여자대학교 미래융합기술공학과 박사과정
[3]성신여자대학교 미래융합기술공학과 석사과정
[4]성신여자대학교 융합보안공학과, 미래융합기술공학과 교수
20220115@sungshin.ac.kr, 220247016@sungshin.ac.kr, 220236039@sungshin.ac.kr,
iglee19@sungshin.ac.kr

Digital Watermarking Techniques Robust to Distortion Attacks

Su-Kyoung Kim[1], Yu-ran Jeon[2], Jung-Hwa Ryu[2], Il-Gu Lee[1,2]
[1]Dept. of Convergence Security Engineering, Sungshin University
[2]Dept. of Future Convergence Technology Engineering, Sungshin University

요 약

디지털 기술과 정보통신 기술이 발전하면서 디지털 콘텐츠의 불법복제 및 유통으로 인한 저작권 침해 피해가 증가하고 있다. 저작권 침해 문제를 예방하기 위해 다양한 디지털 워터마킹 기술이 제안되었지만, 디지털 이미지 워터마킹은 이미지에 기하학적 변형을 가하면 삽입된 워터마크가 훼손되어 탐지가 어렵다는 문제가 있다. 본 연구에서는 왜곡 공격에 강인한 상관관계 측정 기반 워터마킹 기법을 제안한다. 제안한 방식은 교차 상관 기법을 이용해 이미지와 워터마크의 상관관계를 계산하고 임계값과 비교하여 공간 영역에서의 비가시성 워터마크의 존재 여부를 검증할 수 있는 디지털 워터마킹 방법이다. 실험 결과에 따르면 표준편차 120의 가우시안 노이즈 공격을 가해도 원본 워터마크와 0.1 이상의 상관관계를 보이며, 종래의 방식보다 높은 탐지 성능을 나타냈다.

1. 서론

디지털 기술과 정보통신 기술이 발전하면서 다양한 유형의 수많은 디지털 콘텐츠들이 인터넷을 통해 실시간으로 유통되고 있다. 그러나 디지털 콘텐츠는 저작자가 아닌 제3자가 편집하기 쉽고, 빠르게 배포할 수 있어 콘텐츠 저작자의 저작권 보호가 중요하다. 최근 저작자의 고유 콘텐츠가 불법으로 사용되는 것을 막기 위해 콘텐츠에 개인의 고유 식별 정보를 삽입하는 디지털 워터마킹 기술이 연구되고 있다. 이 기술을 이용하면 저작자의 소유권을 확인할 수 있지만, 워터마킹된 이미지에 기하학적 변형을 가하면 삽입된 워터마크가 훼손되어 탐지가 어렵다는 문제가 있다.

본 연구에서는 워터마크가 훼손된 경우에도 워터마크의 존재를 입증하기 위해 상관관계 측정 기반 워터마킹 기법을 제안한다. 제안 방식은 교차 상관 기법을 이용해 워터마크의 존재를 확인하므로, 저작물의 위변조 여부를 입증할 수 있다.

2. 관련 연구

최근 기하학적 변형에 강인한 워터마킹에 대한 연구가 활발히 진행되고 있다. Fernandex 등 [1]은 잠재 확산 모델(LDM, Latent Diffusion Models) 기반 이미지 워터마킹을 제안했고, 이미지 자르기, 해상도 변경 등의 공격을 가해도 평균 약 90% 이상의 정확도를 보이며 모델의 강인성을 입증했다. 그러나, 제안 방식은 워터마크를 모델에 구조적으로 삽입하기 때문에 LDM 기반 워터마크만 탐지할 수 있다.

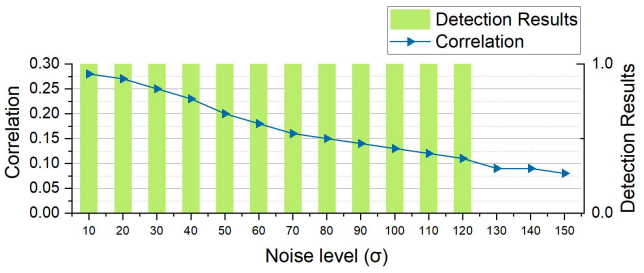
종래의 워터마킹 연구들은 워터마크의 견고성을 평가하기 위해 정규화된 교차 상관 기법을 활용한다. Fierro-Radilla 등 [2]이 제안한 워터마킹 기술은 이미지 압축, 회전, 필터링, 노이즈 공격 환경에서 정규화된 교차 상관 기법 기반 성능 평가를 수행하였고, 원본 워터마크 이미지와 평균 0.95의 상관관계를 보이며 제안한 기술의 강인성을 입증했다. 그러나 제안 기술은 합성곱 신경망 기반 워터마킹이므로 합성곱 신경망 훈련 과정에서 많은 시간이 소요된다.

3. 제안 방법

본 연구에서는 교차 상관 기법을 워터마크의 존재 여부를 검증하는 기술로 활용하고자 한다. 제안하는 방식은 교차 상관 기법을 이용하여 공간 영역에서의 비가시성 워터마크를 탐지한다. 이미지와 동일한 크기의 의사난수배열로 구성된 워터마크를 생성한 후 이미지를 32x32 픽셀 단위로 나누어 블록을 생성하고, 각 블록에 워터마크 정보를 삽입한다. 워터마크가 삽입된 이미지에 기하학적 변형을 가한 후 이미지와 원본 워터마크의 상관관계를 계산한다. 특정 임계값 이상의 상관관계를 가지면 워터마크가 존재하는 것으로 간주한다.

4. 실험 및 결과

본 연구에서는 가우시안 노이즈 공격에 의한 이미지 왜곡을 기하학적 변형으로 가정한다. 실험환경을 구성하고, 성능을 평가하기 위해 오픈소스 코드[3]를 참고했다. 워터마크가 삽입된 700x700 픽셀 이미지에 대한 노이즈 공격을 수행하고, 원본 워터마크와의 상관관계가 0.1 이상인 경우 워터마크가 탐지된 것으로 간주한다. 그림 1은 노이즈 공격의 강도에 따른 상관관계와 탐지 여부를 나타낸다. x축은 이미지에 삽입한 가우시안 노이즈의 표준편차 값을 의미한다. 워터마크를 탐지하는 경우 1, 탐지하지 못하는 경우 0으로 나타냈다.



(그림 1) 제안하는 기법의 상관관계 및 탐지 결과

실험 결과에 따르면 이미지에 표준편차 120 이하의 노이즈를 삽입하는 경우 0.1 이상의 상관관계를 보이며 워터마크를 탐지했지만, 표준편차 120을 초과하는 공격을 가하면 0.1 이하의 상관관계를 보이며 워터마크를 탐지하지 못했다. 이를 통해 이미지 왜곡이 큰 상황에서는 워터마크 탐지가 어려움을 확인할 수 있었다.

제안 방식의 성능을 확인하기 위해, 종래 방식의 결과와 비교하였다. 종래 방식은 이미지를 조정해 이미지 내 워터마크 데이터를 추출하는 스냅 태그[4]를 활용하였다. 표 1에 따르면, 제안하는 방식은 표준편차 60 이상의 노이즈 공격을 가하는 경우 종래의 방식보다 높은 탐지 성능을 보인다. 종래 방식은 워터마크 검출 및 복원을 목적으로 워터마크를 탐지하지만, 제안 방식은 워터마크의 존재 여부 입증을 목적으로 워터마크를 탐지하므로 이미지 왜곡 상황에서 종래방식보다 탐지율이 높았다.

<표 1> 탐지 성능 비교

	표준편차 값					
	30	40	50	60	70	80
제안방식	Detected	Detected	Detected	Detected	Detected	Detected
종래방식[4]	Detected	Detected	Detected	Not detected	Not detected	Not detected

5. 결론 및 향후 연구

본 연구에서는 이미지에 워터마크를 삽입하고 노이즈 공격에 의한 이미지 왜곡을 가한 후, 해당 이미지와 원본 워터마크의 상관관계를 계산함으로써 워터마크 존재 여부를 입증하였다. 본 기법은 노이즈 공격 상황에서 종래 방식 대비 높은 성능을 나타냈지만, 워터마크 원본을 알고 있어야 한다는 한계를 갖는다. 향후 연구로는 왜곡 공격에 강인하며, 워터마크에 대한 정보 없이 워터마크 탐지가 가능한 자기상관 기법 기반 탐지 기법을 제안할 계획이다.

사사

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] Fernandez, Pierre, et al. "The stable signature: Rooting watermarks in latent diffusion models.". Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[2] Atoany Fierro-Radilla et al. "A Robust Image Zero-watermarking using Convolutional Neural Networks". 7th International Workshop on Biometrics and Forensics (IWBF). Cancun, Mexico. 2019

[3] BillAivaliot, “Correlation-Based-Watermarking”. 2020.08.25. https://github.com/BillAivaliot/Correlation-Based-Watermarking.git

[4] 김원진 외. “워터마크 데이터의 임베딩 및 추출 방법”. 10-1960290. 2018.07.05. 2019.03.14