웨이블릿 변환 기반 영상 시퀀스 표현과 ViT 모델을 이용한 얼굴 프레젠테이션 공격 탐지

최문성, 서우진, 김용강*

국립공주대학교

cbhoo3@smail.kongju.ac.kr, ca1364@smail.konju.ac.kr, ygkim@kongju.ac.kr*

Face Presentation Attack Detection using Wavelet Transform-based Video Sequence Representation and Vision Transformer (ViT)

Moonseung Choi, Woojin Seo, and Yonggang Kim*
Kongju National University

요 약

본 논문은 기존 얼굴 프레젠테이션 공격 탐지(PAD) 기술이 갖는 정적 RGB 정보 의존성과 고정된 필터 구조의 한계를 극복하기 위해, 웨이블릿 변환과 시공간 모델링을 결합한 새로운 탐지 프레임워크를 제안한다. 입력 영상 시퀀스의 각 프레임에 이산 웨이블릿 변환(DWT)을 적용하여, 조명 및 색상 변화에 강건하면서도 공격 매체에서 발생하는 물리적 왜곡을 효과적으로 강조하는 주파수 특징을 추출한다. 이후 Deformable CNN을 통해 기하학적 변형에 유연한 공간 특징을 학습하고, Transformer 기반 시계열 모델링을 통해 프레임 간의 시간적 동적 패턴을 통합적으로 학습함으로써, 실제 운영 환경에서도 다양한 공격 시나리오에 강건한 얼굴 프레젠테이션 공격 탐지 모델을 제안한다.

Ⅰ. 서 론

최근 보안 서비스와 사용자 인증 등 다양한 분야에서 얼굴 인식 기술이 보편화됨에 따라, 사진과 디스플레이 장치를 통한 동영상 재생 등을 통해 시스템의 보안성을 위협하는 얼굴 프레젠테이션 공격(Presentation Attack, PA)에 대한 대응이 필수적인 요소로 대두되고 있다. 이러한 위변조 공격을 사전에 식별하고 방어하는 기술인 얼굴 프레젠테이션 공격 탐지(Presentation Attack Detection, PAD)는 생체 인식 시스템의 신뢰성을 보장하기 위한 핵심 기술로써 그 중요성이 강조되고 있다. 그러나 대부분의 선행 연구는 정적인 RGB 정보에 의존하여 조명 변화나 미지의 공격 매체에 취약할 뿐만 아니라, 고정된 필터 구조로 인해 실제 운영 환경의 기하학적 변화를 효과적으로 반영하지 못한다. 특히 정지 영상 기반의 접근은 프레임 간의 동적인 부자연스러움을 포착하지 못하여, 실제 사람의 미세한 생체 움직임을 구별하는 데 구조적인 한계를 갖는다[1]. 따라서 단순한 이미지가 아닌, 영상 시퀀스를 입력으로 사용하여 프레임 간의 연속적인 정보를 학습함으로써 다양한 공격 유형에 대한 일반화된 탐지 성능을 확보한다. 이후 전처리된 시퀀스에 적용되는 웨이블릿 변환은 육안으로 식별하기 어려운 얼굴의 미세한 질감을 주파수로 강조하여 RGB 정보 의존의 한계를 극복한다. 여기에 결합된 변형 가능한 합성곱은 얼굴의 기하학적 변형에 유연하게 적응하며, 어텐션 메커니즘은 시퀀스 내의 장기적인 의존성 모델링하여 동적인 위조 흔적을 포착하는 데 결정적인 역할을 수행한다. 이에 본 논문에서는 주파수 도메인 분석과 변형 가능한 CNN과 어텐션 메커니즘을 결합하고 웨이블릿 변환 기반 영상 시퀀스 표현과 Deformable ViT를 이용한 PAD 모델을 제안한다.

Ⅱ. 본론

본 논문에서 제안하는 모델은 입력 얼굴 영상 시퀀스를 웨이블릿 변환을 통해 다중 주파수 서브밴드로 분해한 뒤, Deformable CNN과 Transformer Encoder를 순차적으로 적용하여 공간적·시간적 단서를 공동으로 학습한다. 이를 통해 고정된 수용 영역이나 단일 프레임 기반 접근의 한계를 극복하고, 다양한 공격 매체 및 환경 변화에 대해 보다 강건한 판별 성능을 확보하고자 한다. 전체적인 모델 구조는 그림 1에 제시한다.

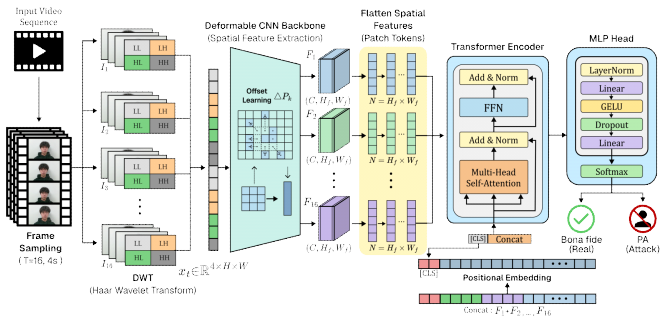

그림 1. Deformable ViT 모델 프레임워크

2.1. 프레임 샘플링 및 웨이블릿 변환 기반 입력 구성

입력 비디오는 효율적인 학습 및 추론을 위해 전 구간을 모두 사용하지 않고, 고정 길이 구간 4초를 기준으로 프레임을 디코딩한 뒤 일정 개수의 프레임을 균일 간격으로 샘플링한다. 구현에서는 FFmpeg 기반 사전 디코딩을 통해 비디오를 PNG 프레임으로 캐싱하며, 학습 시에는 구간 내에서 랜덤 시작점을 사용하고 평가 시에는 중앙 구간을 사용하여 시간 구간 선택에 따른 편향을 완화한다. 또한 최종적으로 각 비디오에서 고정된 개수의 프레임 16장이 입력으로 사용되도록 구성한다.

샘플링된 각 프레임은 얼굴 영역을 중심으로 전처리된 후, 이산 웨이블릿 변환(DWT)을 적용한다[2]. RGB 색상 정보는 조명/화이트밸런스 변화에 민감하고 피부색·인종 등에 따른 편향 가능성이 있는 반면, 주파수 기반 표현은 공격 매체(인쇄물, 디스플레이, 재생 영상 등)의 물리적 특성을 상대적으로 더 직접적으로 반영한다. 본 연구에서는 Haar 웨이블릿을 사용하여 프레임을 LL, LH, HL, HH의 4개 서브밴드로 분해하고, 이를 채널

축으로 결합하여 (4, H, W) 형태의 4채널 텐서로 변환한다. 이는 수식 1과 같다.

$$x_t = Concat(DWT(I_t)) = [LL_t, LH_t, HL_t, HH_t] \in R^{4\times H\times W}$$

수식 1. Haar 웨이블릿 변환

고주파 성분(LH/HL/HH)은 종이 질감, 잉크 번짐, 모아레 및 픽셀 격자, 가장자리 블러 등 왜곡을 강조하고, 저주파 성분(LL)은 얼굴의 전역 구조를 보존한다.

2.2. Deformable CNN을 이용한 공간 특징 추출

구성된 텐서는 Deformable CNN Backbone을 통과하며 공간적 특징이 추출된다.[3] 일반적인 합성곱 연산에서 커널은 고정된 격자 형태를 가지므로, 카메라 각도나 피사체의 거리 변화 등 기하학적 변형에 유연하게 대응하지 못한다. 이를 해결하기 위해 본 모델은 특징 맵의 각 위치마다 학습 가능한 오프셋을 추정하는 변형 가능한 합성곱을 도입하였다. 입력 특징 맵 x와 K개의 샘플링 위치 p_k에 대해, Deformable CNN의 출력 $y(p_0)$는 다음과 수식 2와 같이 정의된다.

$$y(p_0) = \sum_{k=1}^{N} w_k \cdot x_t(p_0 + p_k + \Delta p_k)$$

수식 2. Deformable CNN

여기서는 w_k가중치이며, Δp_k는 별도의 합성곱 레이어를 통해 학습되는 오프셋이다. Δp_k는 소수점 단위의 값을 가질 수 있으므로 이중 선형 보간법을 통해 값을 계산한다. 이 메커니즘을 통해 모델은 얼굴의 포즈가 변하거나 마스크 공격의 경계면이 불규칙하더라도, 공격 판별에 유효한 영역에 적응적으로 집중할 수 있다.

2.3. Transformer를 이용한 시계열 학습

Backbone을 통과한 특징맵은 채널C, 높이H, 너비W를 가지며, 이를 1차원으로 평탄화하여 Transformer 인코더의 입력으로 사용한다. 이때, 프레임의 순서 정보를 보존하기 위해 학습 가능한 위치 임베딩을 더해준다. Transformer 인코더는 멀티 헤드 셀프 어텐션을 수행하여 프레임 간의 장기적인 의존성을 모델링한다. 입력 쿼리(Q), 키(K), 밸류(V)에 대한 어텐션 연산은 다음 수식 3과 같다[4].

$$\mathrm{Attention(Q,K,V)} = \mathrm{softmax}\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d^k}}}\right)\mathrm{V}$$

수식 3. Calculation of Self-Attention

이 과정에서 모델은 단일 프레임 내의 공간적 정보뿐만 아니라, 프레임 간의 시간적 변화를 학습한다. 이는 정적인 사진 공격이나 반복적인 패턴을 가진 리플레이 공격의 미세한 부자연스러움을 포착하는데 핵심적인 역할을한다.

2.4. 분류(Classification)

Transformer 인코더의 출력 중 전체 시퀀스의 정보를 요약하는[CLS] 토큰을 추출하여 분류기에 입력한다. 분류기는 LayerNorm과 GELU 활성화 함수를 포함한 MLP(Multi-Layer Perceptron)로 구성되며, 최종적으로 Softmax 함수를 통해 실제 사람과 공격 확률을 출력한다. 모델 학습에는 예측값과 실제 레이블 간의 차이를 최소화하기 위해 이진 교차 엔트로피(Binary Cross Entropy) 손실함수를 사용한다.

Ⅲ. 실험 및 결과 분석

이 실험은 프레젠테이션 공격 감지 데이터 셋 MSU-MFSD를 사용하였다[4]. 평가지표로는 APCER, BPCER, ACER을 사용하였으며, APCER는 공격 샘플을 정상으로 오분류한 비율, BPCER는 정상 샘플을 공격으로 오분류한 비율, ACER는 두 오류율의 평균으로 전체 분류 성능을 종합적으로 평가하는 지표이다. 학습 과정은 동일한 학습 설정에서 초기화 랜덤성의 영향을 확인하기 위해 서로 다른 시드로 총 5회 반복 실험을 수행하였다. 각 실험은 검증 데이터셋에서 ACER가 최소가 되는 임계값을 탐색하여 선택하였고, 선택된 임계값을 테스트셋 평가에 적용하였다.

그 결과는 다음 표 1과 같다. 이는 테스트셋에서 ACER는 평균 13.75±5.44%, APCER는 3.00±0.95%, BPCER는 24.50±11.51%, 정확도(ACC)는 91.63±2.44%를 기록하였다. 이는 모델이 전반적으로 낮은 APCER를 유지하는 경향이 있으나, 시드에 따라 BPCER 변동이 상대적으로 크게 나타날 수 있음을 보여준다.

Seed	APCER(%)	BPCER(%)	ACER(%)	ACC(%)
13	3.33	22.50	12.92	91.88
193	2.50	42.50	22.50	87.50
235	4.17	12.50	8.33	93.75
752	3.33	17.50	10.42	93.13
4687	1.67	27.50	14.58	91.88
Mean±Std	3.00±0.95	24.50±11.51	13.75±5.44	91.63±2.44

표 1. 각 시드 별 평가 성능과 평균·표준편차

Ⅳ. 결론

본 논문에서는 얼굴 프레젠테이션 공격 탐지(PAD)의 일반화 성능 향상을 목표로, 웨이블릿 변환 기반의 주파수 특징과 Deformable CNN 및 Transformer를 결합한 시공간 학습 프레임워크를 제안한다. MSU-MFSD 데이터셋에 대한 실험 결과, 제안 모델은 기존 단일 프레임 및 공간 정보 중심의 PAD 기법 대비 APCER 및 BPCER 등 지표에서 일관되게 우수한 탐지 성능을 보였으며, 주파수 도메인 정보와 시계열 모델링의 결합이 PAD의 일반화 성능 향상에 효과적임을 확인하였다. 향후 연구에서는 다양한 공개 데이터셋에 대한 교차 평가를 통해 제안 기법의 범용성을 추가적으로 검증하고, 실제 엣지 디바이스 환경 적용을 위한 모델 경량화 및 연산 최적화를 수행할 예정이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임
(IITP-2026-RS-2024-00438430).

참 고 문 헌

[1] C. Kong et al., "A survey of deep learning for face presentation attack detection," Neurocomputing, 2025.

[2] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674 - 693, 1989.

[3] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764 - 773.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998 - 6008.

[5] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746 - 761, 2015.