HarassWatch: 소셜 VR 플랫폼에서의 피해자 관점 괴롭힘 행위 탐지*

이준희[1], 김진우[2†]

[1,2]광운대학교 (대학원생, 교수)

HarassWatch: Detecting Harassment from the Victim's Perspective in Social VR Platforms*

Jun-Hee Lee[1], Jin-Woo Kim[2†]

[1,2]Kwangwoon University(Graduate Student, Professor)

요 약

현재 소셜 VR 플랫폼은 메타버스에 대한 몰입감 있는 경험을 제공하지만, 사용자들은 성희롱 등 다양한 형태의 괴롭힘에 노출되고 있다. 실제로 VR 장비를 통한 신체 추적 기능은 아바타에 반영되어 심각한 정신적 피해를 유발할 수 있으며, 특히 10대 사용자를 대상으로 한 성희롱 문제가 자주 보고된다. 그러나 현재 메타버스 환경에서의 괴롭힘 방지를 위한 기술적 대응은 미흡한 실정이다. 본 논문은 이러한 문제를 해결하기 위한 방안으로 피해자 관점에서의 괴롭힘 탐지 시스템을 제안하며, 제안된 시스템은 괴롭힘 데이터로 학습된 YOLOv7-tiny 모델을 사용하여 괴롭힘 상황을 높은 정밀도와 재현율(각각 0.93, 0.92)로 탐지할 수 있음을 입증하였다.

I. 서론

오늘날 메타버스 환경의 발전은 새로운 형태의 소통과 상호작용을 가능하게 하며, 소셜 VR 플랫폼이 대표적인 예로 주목받고 있다. 이러한 플랫폼들은 VRChat, Rec Room, Mozilla Hubs와 같은 서비스에서 사용자들이 VR 장비와 아바타를 통해 몰입감 있는 3D 상호작용을 경험하도록 한다. 기존 2D 기반의 소셜 플랫폼을 넘어서는 이 몰입감은 사용자 경험을 풍부하게 하지만, 동시에 새로운 유형의 프라이버시 침해 및 괴롭힘 문제를 야기하고 있다.

실제 사례로, VRChat은 약 40,000명의 동시 접속자를 기록하며 높은 인기를 끌고 있으나[1], 미성년자의 아바타를 성희롱한 사건으로 용의자가 체포되거나[2], 소셜 VR 플랫폼이 국제 수사기관의 조사 대상이 되는 등[3] 다양한 문제들이 발생하고 있다. 소셜 VR의 특징인 전신 또는 반신 추적 기술(full/half-body tracking)은 사용자의 물리적 움직임을 아바타에 반영하여 상호작용을 극대화하지만, 이로 인해 발생하는 성희롱과 괴롭힘 행위는 사용자에게 심각한 정신적 피해를 유발할 수 있다. 이러한 문제는 인간-컴퓨터 상호작용 및 보안 연구 분야에서 지속적으로 논의되고 있으며, 실제 피해자 인터뷰를 통해 실증적인 증거로 뒷받침되고 있다. 특히, 10대 사용자가 주된 이용층인 소셜 VR 플랫폼에서 성희롱을 포함한 괴롭힘 문제가 빈번하게 보고되고 있어[4], 이에 대한 해결책 마련이 시급한 실정이다.

VR 기술의 발전 역시 현실과 가상의 경계를 허물며 새로운 위협을 초래하고 있다. 실제 공간 또는 가상 객체(예: 아바타)를 통해 이루어지는 폭력과 성희롱 문제는 온라인 괴롭힘을 넘어서는 영향을 미친다. 그러나 현재 메타버스 환경에서 발생하는 괴롭힘 행위를 효과적으로 탐지하고 예방할 수 있는 기술적 대응은 미흡하다. 기존 연구에 따르면, 일부 소셜 VR 플랫폼은 사용자가 겪는 문제에 대해 충분한 안전

* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00457937)

† 교신저자(jinwookim@kw.ac.kr)

조치와 대응책을 마련하지 못하고 있으며, 개발자들은 기술적 한계와 인력 부족을 주요 문제로 지적하고 있다[5]. 이러한 문제로 3인칭 시점의 괴롭힘을 감지하는 HardenVR이 연구된 적 있으나[6], 이는 VR 기기의 1인칭 시점을 다루지 않았을 뿐더러 신고를 위한 증거 수집 기능도 제공하지 않는다.

본 논문은 이러한 VR 환경에서의 괴롭힘 문제를 해결하기 위해 1인칭 관점에서의 소셜 VR 환경에서의 괴롭힘 데이터셋을 제작하고, 비전 모델을 사용하여 괴롭힘 탐지 시 증거 영상을 만드는 괴롭힘 탐지 시스템을 제안하고자 한다.

II. 배경지식

Unity Sentis는 Unity Technologies에서 개발한 AI 추론 엔진으로, Unity 환경에서 신경망 모델을 효율적으로 활용할 수 있도록 설계되었다. Sentis는 낮은 지연시간과 높은 성능을 요구하는 환경에서 AI 모델을 효과적으로 구동할 수 있다. 또한 ONNX (Open Neural Network Exchange) 형식을 지원해 다양한 AI 프레임워크와의 호환성을 유지하며, Unity의 그래픽 엔진과의 통합을 통해 개발자들이 AI를 손쉽게 적용하고 조정할 수 있게 한다[7].

이러한 특징 덕분에 Unity Sentis는 VR이나 AR 환경에서도 복잡한 AI 연산을 실시간으로 수행하는 데 적합하며[7], 특히 사용자 인터페이스(UI) 공격 방지 연구와 같은 보안 시나리오에서도 고성능 AI 모델의 실시간 추론이 필요한 상황에서 중요한 역할을 수행할 수 있다.

III. HarassWatch

3.1 위협 모델

본 논문에서는 서로다른 유저가 상호작용할 수 있는 일반적인 소셜 VR 플랫폼 환경을 기준으로 하였다. 공격자는 유저에게 공격적인 행동 을 할 수 있는 환경을 가정하였다.

3.2 실험 환경

Intel Core i7-13700K CPU, 64GB RAM, GeForce RTX 4080 GPU가 장착된 머신을 사용하였다. Unity는 최근에 공개된 6000.0.23f1 최신 버전을 사용하여 유저간 상호작용이 가능한 소셜 VR 플랫폼 환경을 구성하였다. Unity Library는 각각 Sentis 1.4.0-pre.3, OpenXR plugin 1.12.1 버전을 사용하였으며, VR 기기로는 Meta Quest Pro를 사용하였다.

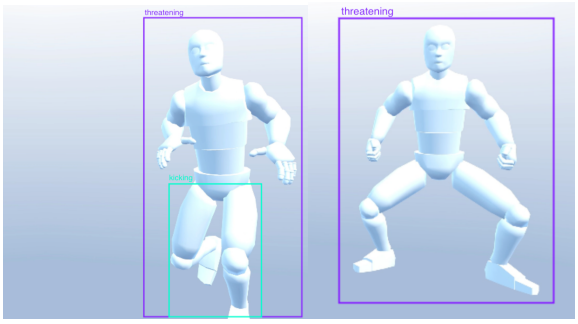

Fig. 1 Images annotated for each category

3.3 괴롭힘 데이터셋

데이터 수집은 직접 구현한 소셜 VR 세계에서 수집하였다. 실제 사람의 행동을 수행하는 아바타를 배치하여, 아바타가 상대방을 괴롭히거나, 폭력적인 행동을 하도록 하고, 이를 VR의 1인칭 화면으로 녹화하였다. 이후 녹화된 화면을 Roboflow에 업로드하여 라벨링하였다. 데이터는 500개의 이미지 데이터로 Punching, Kicking, Threatening 총 3개의 클래스로 이루어져있다. Threatening은 사용자를 향한 위협적인 행동을 말한다. Fig. 1은 Roboflow을 통해 라벨링한 이미지 데이터이다.

3.4 모델 학습 및 탐지 시스템 구현

모델은 YOLOv7를 경량화한 YOLOv7-tiny를 사용하여, VR 기기에서도 무리 없이 작동할 수 있게 하였다. 사전학습된 YOLOv7-tiny의 가중치에서 수집한 데이터셋으로 전이학습을 진행하고 ONNX 모델로 추출하였다.

Unity에서 OpenXR SDK에는 XR 기기를 위한 카메라를 사용할 수 있다. 이 카메라는 사용자의 VR 기기 움직임과 시야각 정보를 바탕으로 Unity 월드에서의 사용자 시야 정보와 동기화하여 VR 기기에 화면을 출력한다. 출력은 Render Texture 형식으로 이루어지며, 이를 통해 사용자가 바라보는 화면 이미지를 가져올 수 있다. 화면 이미지는 텐서 형태로 변환하여

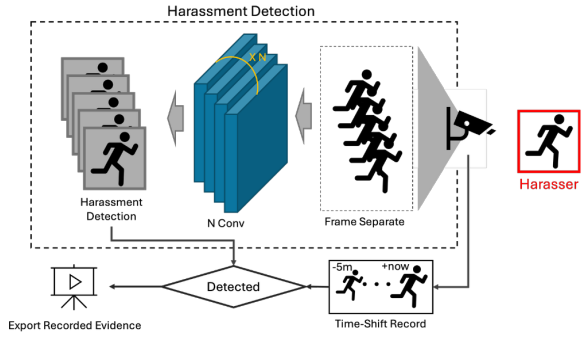

Fig. 2 HarassWatch system design

Sentis를 통해 YOLOv7-tiny 모델로 추론된다. 추론 결과를 토대로 사용자가 처한 상황에 따라 영상을 기록할지 결정한다. Fig. 2는 이러한 HarassWatch 시스템의 설계도이다.

3.5 괴롭힘 녹화 기능 구현

괴롭힘 탐지 후 녹화를 탐지 이후부터 시작할 경우, 이전 상황에 대한 증거가 없어 적절한 신고와 대처가 어려울 수 있다. 이를 해결하기 위해 NVIDIA의 Highlight 기능을 참고하였다[8]. 이 기능은 과거 프레임을 임시 저장하다가, 이벤트 발생 시 과거 프레임부터 현재까지의 영상을 저장하는 기능이다. 본 연구에서는 사용자가 지정한 N분 동안의 사용자의 시야 Render Texture를 큐(queue)에 저장하여 Render Texture를 유지하였다. 이후 괴롭힘이 탐지되었을 때 이벤트를 발생시켜 큐에 저장된 Render Texture를 영상으로 변환하였다. Render Texture를 영상으로 변환하기 위해 FFmpeg 라이브러리를 사용하였다[9].

IV. 평가

4.1 탐지 모델 성능

Fig. 3와 Fig. 5의 결과는 모델이 세 가지 클래스 모두를 높은 정확도로 예측할 수 있음을 보여준다. Fig. 3에서 비록 오탐률(bg FP, 배경을 클래스로 탐지)이 높은 편이나, 괴롭힘 상황을 놓치는 미탐률(bg FN, 클래스를 배경으로 탐지)이 낮아 오히려 감지 가능성을 높여 시스템의 목적에 부합한다. 또한, Fig. 5에 나타난 정밀도(precision)와 재현율(recall)이 각각 0.93과 0.92로, 괴롭힘 상황을 높은 정확도로 탐지

Fig. 3 YOLOv7-tiny confusion matrix

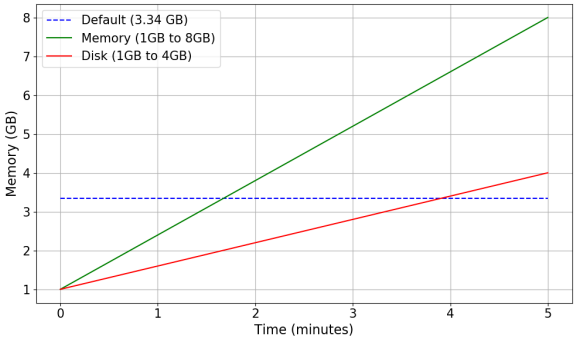

Fig. 4 Memory/disk usage during 5 minutes of video recording

하고 있음을 보여준다. 다만 1인칭 사용자 시점의 특성상 여러 각도에서 객체를 바라볼 수 있으므로, 다양한 각도에서 충분한 데이터를 확보하는 것이 중요하다. 현재 데이터셋은 다양한 각도에서의 데이터가 충분하지 않아, 배경 경계를 정확히 추론하지 못하는 한계가 있다. 그럼에도 높은 정확도 덕에 괴롭힘 탐지에는 문제없다.

4.2 녹화 기능 오버헤드

녹화 기능의 경우, N분 동안 사용자의 시야 Render Texture를 저장하므로 메모리 사용량은 N × 60 × FPS × Render Texture로 계산된다. 실험에서는 Render Texture의 해상도를 1920 × 1080(FHD)으로 설정하고 녹화시간은 5분으로 설정하였다. 직접 구현한 소셜 VR 월드는 3.34GB의 메모리를 점유한다. Render Texture를 관리하는 큐를 모두 메모리에서 관리할 경

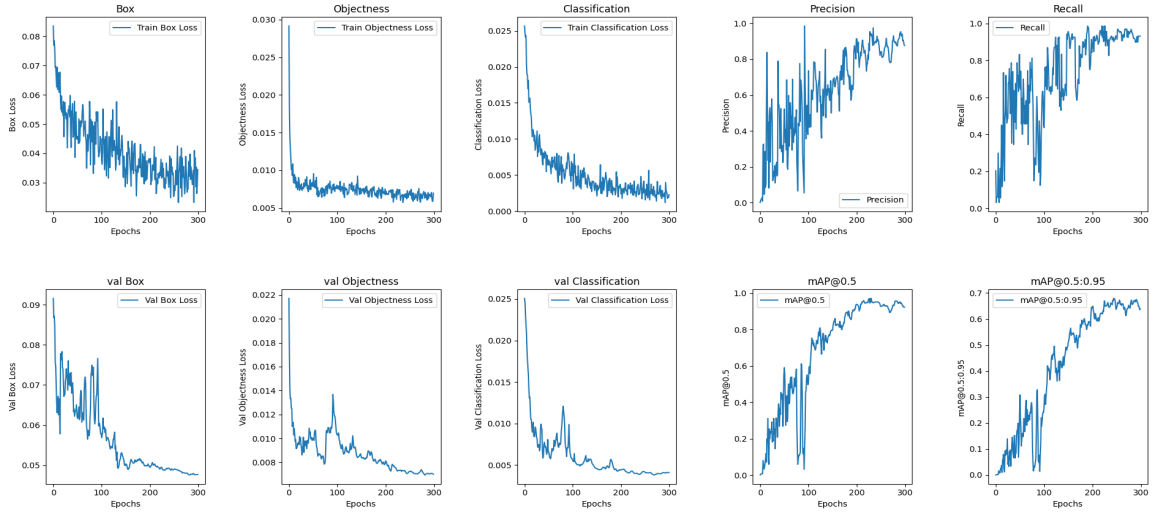

Fig. 5 Training and validation metrics over epochs.

우, Fig. 4와 같이 5분 동안 4.66GB(8GB - 3.34GB)의 메모리를 사용하게 된다. Meta Quest Pro의 메모리는 12GB로, 녹화 기능으로 인해 전체 메모리의 38%를 사용하게 된다. 디스크에 임시로 저장하여 Render Texture를 관리할 경우, 기기의 디스크를 자주 사용하게 되지만 메모리 사용량은 0.7GB(4GB - 3.34GB)로 크게 줄어들어 훨씬 효율적이었다.

V. 결론

최근에는 VR 기술뿐만 아니라 AR, MR, XR 등 다양한 기술이 빠른 속도로 발전하고 있다. 이와 함께 Meta Horizon World와 같은 소셜 메타버스 플랫폼이 증가함에 따라, 플랫폼 내의 괴롭힘 문제도 대두되고 있다. 그러나 수동적으로 신고 및 차단하는 방법만 존재하며[10], 자동으로 괴롭힘을 탐지하고 증거를 생성하는 기능은 부족하다. 본 연구에서 제안한 괴롭힘 탐지 시스템은 괴롭힘을 자동으로 탐지하고 증거를 생성하여 문제 해결에 기여할 수 있다.

향후에는 VR과 AR에서 동시에 작동하는 탐지 시스템을 개발하고, 기기 및 컨트롤러의 좌표 데이터를 수집하여 XR 행동 탐지 모델을 구현하고자 한다. 이를 통해 XR 환경에서 사용 가능한 괴롭힘 탐지 시스템을 구축하여 안전한 메타버스 환경 조성에 기여하고자 한다.

[참고문헌]

[1] “VRChat - Steam Charts”, https://steamcharts.com/app/438100

[2] “Police investigate virtual sex assault on girl's avatar”, https://www.bbc.com/news/technology-67865327

[3] “Interpol working out how to police the metaverse”, https://www.bbc.com/news/technology-64501726

[4] Deldari, Elmira, et al. "An investigation of teenager experiences in social virtual reality from teenagers', parents', and bystanders' perspectives." Symposium on Usable Privacy and Security (SOUPS). 2023.

[5] Abhinaya, S. B., Aafaq Sabir, and Anupam Das. "Enabling Developers, Protecting Users: Investigating Harassment and Safety in VR." USENIX Security Symposium, 2024.

[6] Wang, Na, et al. "HardenVR: Harassment detection in social virtual reality." 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). IEEE, 2024.

[7] “Unity Sentis: Empowering Real-time AI Inference in 3D Applications.”, https://unity.com/blog/games/create-next-gen-ai-models-with-unity-sentis, 2023.

[8] “Nvidia Highlights”, https://developer.nvidia.com/highlights

[9] “FFmpeg Github Repository”, https://github.com/FFmpeg/FFmpeg, 2024.

[10] “Use Personal Boundary in Meta Horizon Worlds”, https://www.meta.com/ko-kr/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/personal-boundary-horizon-worlds