VRMask: 소셜 VR 플랫폼의 유해 컨텐츠 탐지 및 마스킹*

김종섭[1], 김동은[1], 김진우[2†]

[1,2]광운대학교 (학부생, 교수)

VRMask: Detection and Masking of Harmful Content in Social VR Platforms

Jong-Seop Kim[1], Dong-Eun Kim[1], Jin-Woo Kim[2]

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

요 약

메타버스 시장이 커지고 관련 기술이 발전하면서 소셜 VR 플랫폼에 대한 접근성이 점점 증가하고 있다. 그러나 소셜 VR 플랫폼의 가장 큰 특징인 UGC (User Generated Content)가 문제가 되고 있는데, 성적인 컨텐츠를 포함한 유해 컨텐츠가 무분별하게 제작 및 배포되고 있기 때문이다. 본 논문에서는 소셜 VR 플랫폼 상에서 유해 컨텐츠를 실시간으로 탐지하고 이를 마스킹함으로써 사용자를 보호할 수 있는 솔루션인 VRMask에 대해 소개한다. 이를위해 실제 소셜 VR 플랫폼인 VRChat에서 유해 컨텐츠를 수집하고 이를 기반으로 비전 모델인 YOLO v5를 학습시켰다. 또한 모델이 탐지한 영역을 Unity 엔진에서 마스킹하는 시스템을 구현하였다.

I. 서론

소셜 VR 플랫폼의 특징 중 하나는 사용자가 직접 컨텐츠를 생산할 수 있다는 점이다. 개발자는 단순히 사용자에게 기본적인 플랫폼과 배포할 수 있는 환경만 제공하고 실질적인 컨텐츠(예: 개인 아바타, 월드 등)는 사용자가 직접 제작한다. 이러한 점은 사용자가 플랫폼을 즐기는데 있어 적극적인 참여를 유도하고, 본인의 입맛대로 메타버스를 꾸려나갈 수 있게 하며, 다양한 문화 및 환경을 다른 사용자들에게 경험하게 할 수 있다. 이러한 특징을 User Generated Contents (UGC)라고 한다.

그러나, UGC는 장점만 가지는 것은 아니다. 사용자가 직접 컨텐츠를 생산할 수 있는 만큼 유해한 컨텐츠(폭력적인, 성적인, 잘못된 이념에 대한 컨텐츠)들도 무분별하게 제작 및 배포가 되고 있다. 특히 VRChat과 같은 소셜 VR 플랫폼에서 아바타를 직접 제작할 수 있는 점을 악용한 유해 아바타를 많이 볼 수 있다(Fig. 1).

Fig. 1 Examples of (masked) harmful avatars

이러한 아바타들은 자극적인 컨텐츠를 통해 사용자의 정신 건강에 큰 해를 끼치게된다. 예를들어 현재 대표적인 소셜 VR 플랫폼인 VRChat의 소비층 중 적지않은 사용자가 아동 및 청소년이다. 낮은 연령층에게 자극적이고 현실과 괴리가 있는 유해한 컨텐츠는 성인에 비해 더욱 큰 영향을 미칠 가능성이 높다[1]. 그럼에도 불구하고, VRChat은 현재 유해 아바타에 대해 큰 제재를 가하지 않는다.

본 연구에서는 소셜 VR 플랫폼 내의 성적으로 유해한 컨텐츠를 실시간으로 탐지하고 이를 마스킹하는 솔루션인 VRMask를 소개한다. VRMask는 Unity 상에서 Render Texture 모드를 통해서

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

† 교신저자(jinwookim@kw.ac.kr)

Game Scene의 카메라 시점(VR Camera)을 캡처한다. 이때 WatchDog 라이브러리를 통해서 이를 감지 하며, 학습된 YOLO v5 비전 모델을 통해 화면의 유해 컨텐츠를 추론한다. Unity에서는 해당 영역에 Mask Object을 생성하여 유해한 컨텐츠에 대해서 마스킹(masking)을 해주는 역할을 한다. 이를 통해 사용자를 유해 컨텐츠로부터 보호할 수 있다.

II. 배경지식

2.1 User Generated Content (UGC)

소셜 VR 플랫폼의 두드러지는 특징은 사용자가 착용하는 아이템, 아바타, 그리고 자신만의 가상 세계를 직접 구축할 수 있게 한다는 점이다. 이는 큰 장점이만, 다른 사용자들에게 불쾌감을 주거나, 혐오감을 주는 컨텐츠 또한 사용자가 마음대로 생산할 수 있다. 특히 컨텐츠 생산이 고수준의 프로그래밍이 필요한 작업이 아니기에 악의적인 사용자가 유해 컨텐츠를 쉽게 생산할 수 있다는 점이 가장 큰 문제이다.

2.2 관련연구

Guo 등[2]은 Vision Language Model (VLM)을 이용하여 홍보 이미지로 사용되는 유해한 UGC를 탐지하는 시스템을 제안하였다. 특히 프롬프트 엔지니어링을 통해 유해한 컨텐츠 여부를 판별하였다. 본 논문에서는 UGC를 이용하는 사용자 입장에서 게임에 몰입하며 시각으로 직접 렌더링되는 화면에 대해서 탐지하여 마스킹하는 것이 목적이다.

Wang 등[3]은 소셜 VR 플랫폼에서의 괴롭힘에 대하여 분석하였다. 괴롭힘을 감지하기 위해서 음성, 제스처를 인식하고 분석하였다. 사용자의 안전을 보장하려는 목표는 본 논문과 유사하지만, 해당 연구에서는 주로 괴롭힘 행동에 대해서 실시간으로 감지하는데 초점을 두었다면, 본 논문에서는 성적으로 유해한 컨텐츠에 대한 탐지와 차단에 집중하였다.

III. VRMask

유해 컨텐츠로부터 사용자를 보호하고 안전한 메타버스 환경을 구축하기 위한 솔루션인 VRMask는 아래와 같은 컴포넌트로 구성된다 (Fig. 2).

Renderer: 사용자 측에서 VR 기기를 사용하며 얻는 렌더링된 화면으로 Detector의 이벤트

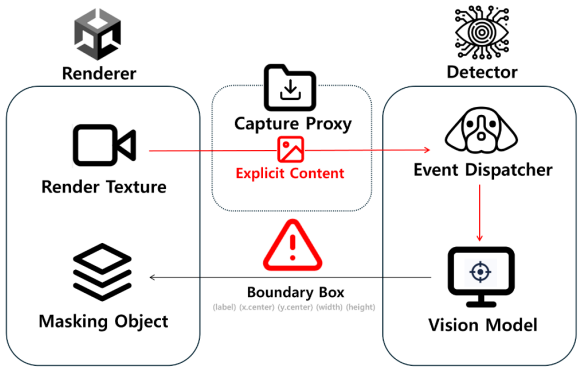

Fig. 2 VRMask system overview

트리거가 발생할 수 있도록 Render Texture를 통해 Capture Proxy에 기록한다. Detector가 제공하는 경계 박스 정보를 받으면 해당 영역에 대한 마스크 랜더링을 진행하여 사용자를 보호한다. 추가적으로 마스크가 유해 컨텐츠 탐지를 방해하는 현상을 방지하기 위해서, 별도의 Renderer를 통해 마스킹을 하지않은 유해 contents를 Detector에게 지정된 프레임 단위로 끊이지 않게 전달하며, 추론을 지연시키지 않게 구성하였다.

Capture Proxy: Renderer에서 화면에 대한 프레임을 생성하여 저장하는 컴포넌트이다. 해당 화면에는 마스크가 렌더링되지 않아 사용자 화면에 마스크가 유해 컨텐츠가 가려지더라도 해당 프레임에는 유해 컨텐츠가 제대로 랜더링되어 탐지에 문제가 없게 설정했다.

Detector: 실질적인 유해 컨텐츠를 탐지하는 부분으로 사전에 모델을 로드하고 대기상태에 빠진다. 대기상태에는 Event Dispatcher가 Capture Proxy를 감시하고 있고, 만약 해당 디렉토리에 프레임 생성을 탐지한다면, 생성된 프레임에 대해 추론을 진행하고 다시 대기상태로 빠진다. 추론 결과는 Renderer에 전달하여 성적 컨텐츠의 경계 상자를 전달하여 Fig. 3과 같이 해당 영역에 대해 마스킹 오브젝트를 생성한다.

IV. 평가

평가를 위해 Unity 기반 테스트 환경을 Intel i9-12900K CPU와 NVIDIA GeForce RTX 3090 GPU가 장착된 머신에서 구성하였다.

Detector의 비전 모델로 YOLO v5를 이용하였으며, 학습 하기위한 데이터는 직접 VRChat의 유해한 월드에서 수집하여 수작업으로 800여개의 데이터를 라벨링하였다. 클래스는 총 4가지로, Behave, Naked, Underwear, Background

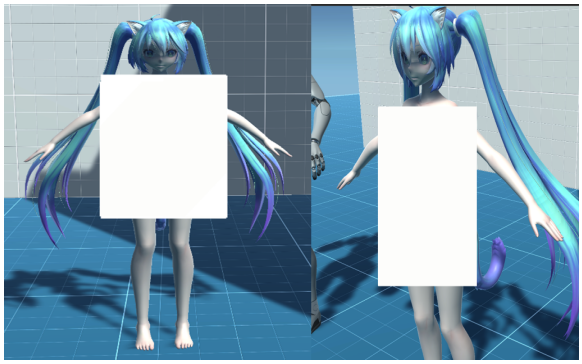

Fig. 3 The harmful avatars masked by VRMask

로, 각각 성적인 행위, 아바타의 중요 부위 노출, 선정적인 코스튬, 마지막으로 선정적인 배경에 대한 클래스이다. 테스트 환경의 아바타는 학습에 쓰이지 않은 유해 아바타로 실제 VRChat에서 사용할 수 있는 아바타로 제작하였다.

Fig. 4는 모델의 탐지 결과이다. Underwear와 Naked 클래스가 각각 0.842, 0.695로 준수한 성능을 보인다. Behave는 0.608로 낮은 성능을 보이고 있는데 이는 모션을 탐지하기 어려움이 있었기 때문이다. 클래스 간 혼동은 0.2~0.5로 낮은 수치를 보인다.

4.2 탐지 지연 시간

유해 컨테츠를 모델이 탐지 후 Unity 엔진이 마스킹하는 시점 까지의 지연시간을 측정하였다. 높은 지연 시간은 사용자가 유해 컨텐츠에 오래 노출될 확률을 높여주기 때문이다. 측정 시간은 1) 전처리 시간, 2) 추론 시간, 3) NMS (Non-Maximum Suppression) 시간으로 각각 분류하였다. NMS는 객체 탐지 과정에 객체가 존재하는 위치 주변에 높은 스코어를 가진 여러개의 Boundary Box 중 가장 정확도가 높은 Box를 선택하는 알고리즘이다.

측정된 지연 시간 결과는 Fig. 5와 같다. 추론 시간으로 수집된 197개의 데이터는 CDF 분포에 따라 80%의 확률로 8ms 이하의 지연시간을 가진다. 또한 Unity에서 객체의 유해 컨텐츠 영역에 대한 마스크 생성 시간은 Unity 상에서는 탐지하지 못할 정도로 미미했다. 총 지연시간은 8ms로, 이는 VR상에서 몰입감을 해치지 않고 유해 컨텐츠를 차단하기에 충분하다.

또한 VRMask를 사용하여 마스킹한 결과는 아바타의 상반신과 하반신을 모두 마스킹해주

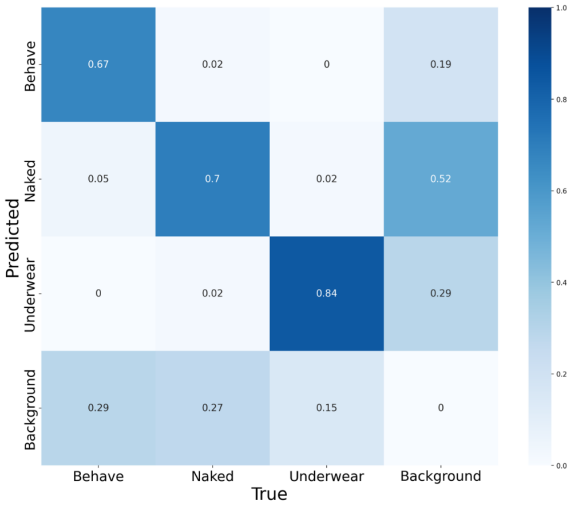

Fig. 4 Confusion matrix of YOLO v5 vision model

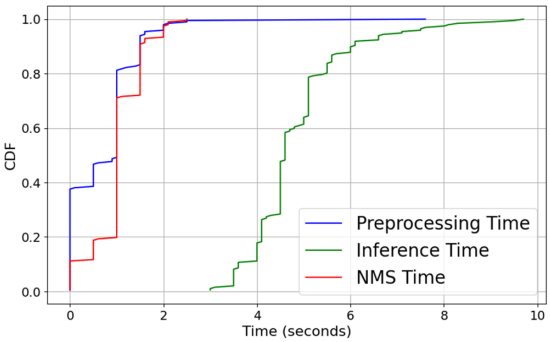

Fig. 5 Latency of YOLO v5 Vision Model

며, 유해 컨텐츠를 충분히 차단하는 모습을 보인다(FIg. 3).

Ⅴ. 결론

본 논문에서는 직접 수집한 유해 콘텐츠 데이터셋을 기반으로 YOLO v5 모델을 학습시키고, 이를 통해 실시간 유해 컨텐츠 탐지 및 이에 대한 마스킹 처리를 통해 사용자를 보호하는 시스템인 VRMask를 제안했다.

[참고문헌]

[1] "Metaverse app allows kids into virtual strip clubs", https://www.bbc.com/news/technology-60415317

[2] Guo, Keyan, et al. "Moderating Illicit Online Image Promotion for Unsafe User-Generated Content Games Using Large Vision-Language Models." USENIX Security Symposium (2024).

[3] Wang, Na, et al. "HardenVR: Harassment Detection in Social Virtual Reality." IEEE Virtual Realtiy and 3D user Interfaces (2024)