Motion Diffusion Model 을 활용한 텍스트 기반 언리얼 엔진 런타임 애니메이션 생성 플러그인[1]

박수호 [1*], 이재훈 [2*], 조용현 [3*], 제해찬 [4*], 차승종 [5*], 구형준 [6**]
성균관대학교(기계공학 [1], 건설환경공학 [2], 물리학 [3], 시스템경영공학 [4], 신소재공학 [5]) 학부생
성균관대학교 소프트웨어학 교수 [6]

suho5721@daum.net[1], hoon7617@naver.com[2], gocks0802@g.skku.edu[3], 4ipodapp@gmail.com[4], daniel1426551@gmail.com[5], kevin.koo@skku.edu[6]

An Unreal Engine Plugin for Text-based Runtime Animation Generation with a Motion Diffusion Model

Suho Park[1], Jaehoon Lee[2], YongHyeon Jo[3], Haechan Je[4], Daniel Cha[5], Hyungjoon Koo[6]
Department of Mechanical Engineering[1], Civil Architectural Engineering[2], Physics[3], System Management Engineering[4], Advanced Material Science[5], Computer Science and Engineering[6]
Sungkyunkwan University

요 약

언리얼 엔진 기반의 메타버스나 실시간 게임 환경에서 캐릭터의 맞춤형 동작이 필요한 경우가 있다. 본 논문은 모션 디퓨전 모델을 활용하여 특정 동작을 자동 생성하는 기능을 제공하는 언리얼 엔진 플러그인을 제시한다. 특히 사용자가 텍스트로 신체 동작이나 감정 표현을 기술해 입력값으로 제공하면 서버에서 모션 디퓨전 모델로 애니메이션을 실시간으로 생성한 후, 언리얼엔진 클라이언트에서 후처리하여 사용자의 캐릭터에 실시간으로 적용하는 방식으로 구현했다.

1. 서론

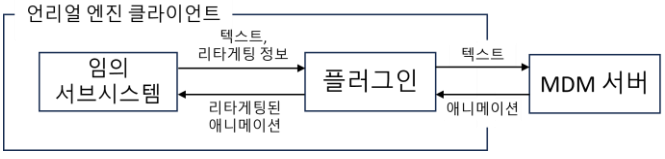
최근 메타버스와 비디오 게임 산업이 크게 발전하고 경쟁이 가속화되면서 사용자들에게 더욱 개인화된 경험을 제공하는 기능이 중요하다[1]. 차별화된 경험의 중요한 구현방식 중 하나는 감정 표현을 포함한 캐릭터 애니메이션인데, 주로 모션 캡쳐나 키프레이밍 방식으로 제작된다. 그러나 이러한 제작 방식은 특수 장비 및 전문 인력을 필요로 하여 애니메이션 제작 비용이 커서 개발 접근성에 한계가 있다. 또한 기술의 특성상 사용자가 시스템 런타임에서 원하는 애니메이션을 직접 제작할 수 없기에 공급자로부터 한정된 애니메이션을 일방적으로 제공받게 된다.

본 연구에서는 이러한 문제점을 극복하고 사용자의 개인화된 경험을 극대화할 수 있도록 MDM (Motion Diffusion Model) [2]을 활용하여 언리얼 엔진에서 실시간으로 애니메이션 생성 기능을 추가해 주는 플러그인을 제안한다.

2. 관련 연구: 모션 디퓨전 모델

MDM 은[2] 디퓨전과 트랜스포머를 기반으로 하여 text-to-motion, action-to-motion 등의 다양한 인간형 캐릭터 애니메이션의 생성 및 수정 기능을 제공하는 생성 모델로서 각 디퓨젼 단계에서 노이즈가 아닌 샘플을 예측하는데, 이를 통해 모션 위치 및 속도에 대한 geometric loss 를 쉽게 사용할 수 있다. 또한 가벼운 리소스로 학습할 수 있음에도 텍스트-모션에 대한 주요 벤치마크에서 우수한 결과를 산출해내는 이점이 있다. HumanML3D 는[3] MDM 훈련에 사용된 3D 인간형 모델을 다양하게 표현하는 데이터셋으로, 인간 포즈 예측, 3D 재구성 등 관련 분야에서 널리 사용되고 있다. 이 데이터셋은 관절의 위치, 연결성, 메시 토폴로지 등 다양한 정보를 포함하며 데이터 생성과정, 하드웨어 설정 등의 자세한 설명을 제공하고 있어 정확도와 일관성을 확보할 수 있다.

3. 언리얼 엔진 런타임 애니메이션 생성 플러그인



<그림 1> 애니메이션 생성 플러그인 동작 개요

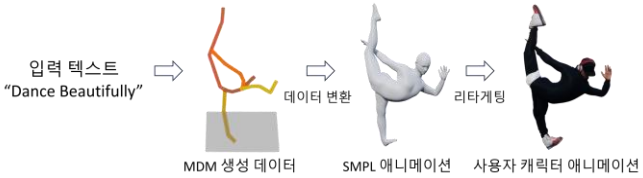
[1] 본 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 (No. 2022-0-01199, 융합보안대학원(성균관대학교))과 LINC 3.0 사업단 지원으로 수행한 연구임 *: 제1저자, **: 교신저자

애니메이션 생성은 언리얼 엔진 클라이언트와 MDM 을 탑재한 서버간 API 를 활용하여 사용자의 텍스트와 이를 기반으로 생성된 애니메이션을 교환하는 방식으로 동작한다. 서버-클라이언트 모델을 사용함으로써 사용자로부터 애니메이션 생성에 필요한 GPU 자원 소모를 최소화할 수 있다. 구현 내용을 <그림 1>과 같이 언리얼 엔진 플러그인 형태로 패키징하고 배포하면 편의성과 이식성을 획득할 수 있다. 개발자는 프로젝트에서 본 플러그인을 활성화하는 것으로 손쉽게 런타임 애니메이션 생성 기능을 기존 또는 신규 프로젝트에 추가하고, 플러그인 API 를 활용하여 원하는 방식으로 사용자에게 기능을 전달할 수 있다.

모델 생성. HumanML3D 에는 감정을 표현하는 단어의 수가 불균형하여 long-tail 문제가 발견되었다. 이 문제를 해결하기 위해 자체적으로 데이터 증강을 진행하였다. 출현 빈도가 높은 감정 표현 단어를 연관된 의미를 지닌 출현 빈도가 적은 단어로 교체하는 방식으로 데이터의 개수를 증가시켰다. 이러한 방식으로 표현력이 좋은 감정 분포를 획득하여 long-tail 문제를 해결하면서 애니메이션의 감정 표현 능력을 향상시켰다. 이를 제외한 다른 모델 훈련 조건은 기존 MDM 과 동일하게 설정하여 기대 성능을 보장하였다. 훈련에 RTX Titan 으로 72 시간이 소요되었다.

데이터 변환. MDM 을 포함한 기계학습에 사용되는 애니메이션 데이터 형식은 게임 및 메타버스 등의 산업에서 쓰이는 형식과 큰 차이가 있어 MDM 이 출력한 데이터를 언리얼엔진에서 사용하려면 여러 단계의 변환이 필요하다. MDM 이 생성한 데이터를 범용성이 있는 SMPL 3D 인체 모델 형식과 FBX 파일 형식으로 변환하면 언리얼엔진에서 임포트할 수 있다[4][5].

런타임 리타게팅. 생성한 애니메이션은 SMPL 의 골격구조를 사용하고 있고 이는 임의의 게임 및 메타버스 사용자 캐릭터의 골격구조와 다르기에 생성된 애니메이션은 곧바로 사용될 수 없다. 이를 해결하기 위해 <그림 2>와 같이 두 골격구조간 대응되는 각 뼈 마디를 매칭하고, 기존 애니메이션 데이터를 새 골격구조에 맞춰 값을 변환하는 리타게팅 과정을 도입했다.



<그림 2> 애니메이션 데이터 변환 및 리타게팅 예시

4. 구현

MDM 서버는 애니메이션 생성 후 데이터에 Smplify 모델을 적용하여 SMPL 인체 형식 및 FBX 파일 형식으로 변환 후 바이너리 형태로 클라이언트에 반환한다[6]. 플러그인은 ASSIMP 를 사용하여 FBX 데이터로부터 엔진 내부 형식의 애니메이션을 생성한다[7]. 그 후 언리얼 엔진의 IKRig 모듈의 API 를 활용하여 런타임 리타게팅을 구현함으로써 언리얼 엔진 에디터에서 사용되는 리타겟팅 관련 에셋들을 런타임에서 그대로 활용할 수 있도록 하였다. 또한 엔진의 프리징 방지를 위해, 처리에 긴 시간이 소모되는 클라이언트-서버 통신과 FBX 파일 읽기 과정은 각각 비동기 및 멀티스레딩으로 구현하였다.

5. 한계점

캐릭터 애니메이션의 높은 복잡도로 인해 플러그인에서 텍스트 입력 후 애니메이션을 반환받기까지 상용 GPU(RTX Titan)로 현재구현 기준 약 4 분의 시간이 소요되어 런타임에서 사용자와 실시간으로 상호작용하며 사용하기는 쉽지 않다. 같은 이유로 HumanML3D 은 손가락, 발가락 등의 말단 관절을 생략하였기에 해당 관절의 움직임은 표현하지 못한다.

6. 결론

본 연구를 통해 언리얼엔진 기반의 메타버스와 게임에서 사용자의 개인화된 경험을 증대시킬 목적으로 텍스트 기반 런타임 애니메이션 생성 플러그인을 개발하였다. 서버에서 MDM 을 실행하여 사용자 측면에서 GPU 부담 없이 애니메이션을 생성할 수 있으며, 데이터 증강을 통해 보다 풍부한 감정을 표현했다. 하드웨어 및 모델 성능이 발전하여 애니메이션 생성 시간이 단축되고, 더 자세한 표현이 가능해진다면 본 플러그인의 활용성이 더욱 증대될 것으로 기대된다.

참고문헌

[1] Beyond Games. The Evolution of Emotes in Gaming and the Metaverse. Retrieved from https://www.beyondgames.biz/35864/the-evolution-of-emotes-in-gaming-and-the-metaverse/

[2] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., & Bermano, A. H. (2023). Human Motion Diffusion Model. The Eleventh International Conference on Learning Representations.

[3] Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., & Cheng, L. (2022). Generating Diverse and Natural 3D Human Motions from Text.

[4] Loper, Matthew and Mahmood, Naureen and Romero, Javier and Pons-Moll, Gerard and Black, Michael J. (2015). SMPL: A Skinned Multi-Person Linear Model.

[5] https://github.com/softcat477/SMPL-to-FBX

[6] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., & Black, M. J. (2016). Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image.

[7] https://github.com/assimp/assimp

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium.