음악에 어울리는 춤 자동 생성 및 실시간 춤 모션 판정

박소현 [1], 정유진 [1], 박근영 [1], 강지우 [2]
[1] 숙명여자대학교 IT 공학전공 학부생
[2] 숙명여자대학교 인공지능공학부 교수
jwkang@sookmyung.ac.kr

Music-Driven Choreography Generation and Real-time Motion Assessment

So-Hyun Park[1], Yu-Jin Jeong[1], Kuen-Young Park[1], Ji-Woo Kang [2]
[1]Department of IT Engineering, Sookmyung Women's University
[2]Division of Artificial Intelligence Engineering, Sookmyung Women's University

요 약

최근 화제인 가상 아이돌의 춤 제작에 많은 자원 및 비용이 발생한다. 만일 춤을 자동으로 생성해 3D 모델에 피팅하면 이러한 비용을 줄일 수 있으며, 다양하고 복잡한 춤의 구현도 가능할 것이다. 또한, 댄스 게임을 통해 춤을 배우고 즐기는 사람들이 많지만, 경험할 수 있는 춤이 한정적이며, 모션 인식 정확도가 낮다는 단점이 있다. 따라서 본 논문에서는 트랜스포머 구조의 인공지능 모델을 통해 음악에 어울리는 3D 춤 모션을 자동으로 생성하고, 3D 자세 추정 모델을 사용해 사용자의 모션을 추정한 후, 두 모션의 유사도를 랜드마크 3D 좌표로 계산하여 판정하고자 한다. 이는 1 인 댄스 룸 또는 댄스 게임에 활용되어 발전 가능하다.

1. 서론

최근 화제인 가상 아이돌 (virtual idol)의 춤은 모션 (motion) 캡쳐 데이터를 이용하여 제작되기 때문에 많은 자원과 비용이 발생한다. 만일 음악에 어울리는 춤을 자동으로 생성해 3D 모델에 피팅하면 이러한 비용을 크게 줄일 수 있으며, 물리적인 제약과 유연함의 한계를 극복한 다양하고 복잡한 춤의 구현도 가능할 것이다. 음악에 어울리는 춤 생성 연구[1][2]는 트랜스포머 (transformer) 구조를 변형하고 추가하는 것으로 진행되고 있다. TV 프로그램 '스트릿 우먼/맨 파이터'의 인기로 춤에 대한 관심이 증가하면서, 춤을 배우고 즐기는 사람들이 많아졌다. 이를 위해 동영상과 게임을 활용하는 것이 선호되며, 특히 유비소프트의 '저스트 댄스'는 수백만 명이 이용하고 있다. 하지만 정해진 춤만 경험할 수 있으며, 콘솔 게임기를 통한 모션 인식으로 정확도가 낮다는 단점이 있다. 따라서 본 논문에서는 (그림 1)과 같이, 음악에 어울리는 춤을 자동으로 생성하고, 생성된 춤과 사용자 춤의 유사도를 계산하여 실시간으로 판정하고자 한다.

2. SMPL

SMPL (Skinned Multi-Person Linear Model)[3]은 사람의 체형 (shape)을 모델링하고, 자세 (pose)의 변화에 따른 체형의 변화를 예측하여 실제 사람과 같도록 하는 3 차원 인체 선형 모델이다. 파라미터 β 로 체형을, 파라미터 θ 로 자세를 제어할 수 있으며, 6,890 개의 정점 (vertex)과 24 개의 관절 (joint)로 이루어져 있다.

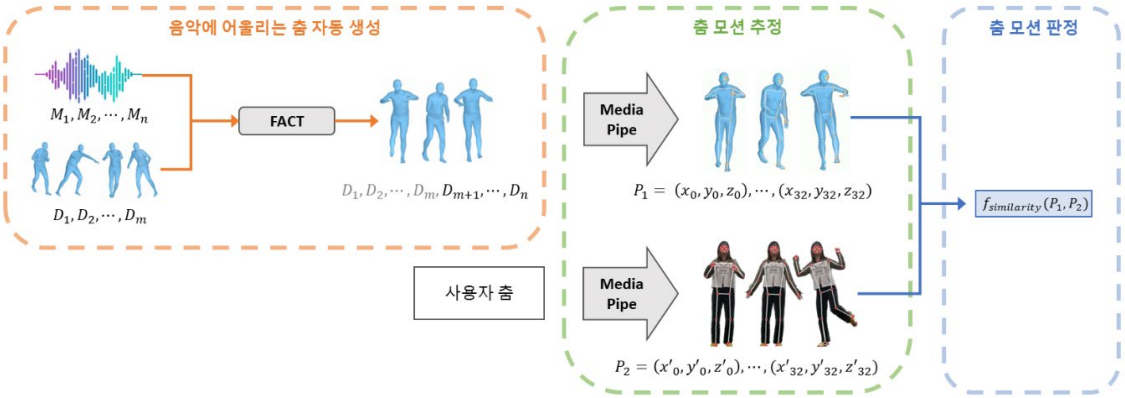
SMPL 모델에서 각 관절의 회전은 축 회전 (axis angle) 표현으로 나타내며, 다음과 같은 *Rodrigues* 공식을 통해 3×3 회전행렬 R로 변환한다.

$$R = I + K \sin\theta + K^2 \cos\theta$$

여기서 I는 3×3 항등 행렬이고, θ는 회전 각도, K는 3×3 반대칭 행렬로, 회전축의 방향 벡터이다.

3. 음악에 어울리는 춤 자동 생성

10 가지 장르의 춤 모션 데이터셋 (AIST++ dataset)으로 학습된 FACT (Full-Attention Cross-modal Transformer network) 모델[1]을 사용하여 음악에 어울리는 3D 춤 모션을 자동으로 생성할 수 있다. FACT 모델은 오디오 트랜스포머와 모션 트랜스포머, 교차모달 (cross-modal) 트랜스포머가 사용되며, 3 개의 트랜스포머는 독립적으로 학습된다. 테스트 시 오디오와 2 초 길이의 시드 (seed) 모션이 주어지면, 자동 회귀 (auto-regressive) 방식으로 연속적인 미래의 모션을 생성한다. 결과적으로 각 24 개 관절의 *Rodrigues* 회전 행렬, SMPL 모델의 전역 (global) 위치 좌표와 6 개의 패딩 (padding) 값으로 이루어진, 총 225 차원 SMPL 모션 벡터가 생성된다.



(그림 1) 오디오 $(M_1, \cdots, M_n)$와 시드 (seed) 모션 $(D_1, \cdots, D_m)$이 주어지면 $(n > m)$ FACT (Full-Attention Cross-modal Transformer network) 모델을 통해 미래의 모션 $(D_{m+1}, \cdots, D_n)$이 생성된다. MediaPipe 자세 추정 모델로 생성된 춤과 사용자 춤의 3D 자세를 추정하고, 그 결과로 얻은 33 개 랜드마크의 3D 좌표로 유사도를 계산하여 모션을 판정한다.

4. 춤 모션 추정

실시간으로 모션을 추정하기 위해 프레임 단위로 자세를 추정한다. 3D 자세 추정 방법으로는 (1) 2D 자세 추정 후 RGBD 카메라로 측정한 각 관절점의 깊이 값을 추가하는 것과 (2) 2D 이미지에서 3D 자세를 추정하는 것이 있고, 그 중 두 번째 방법을 사용한다. MediaPipe[4]의 자세 추정 모델을 사용하여 2D 이미지에서 3D 자세 추정을 한다. MediaPipe 는 학습 과정에서만 히트맵 (heatmap)과 오프셋 손실 (offset loss)을 사용하고, 추론 과정에서는 추적 알고리즘 기반으로 추정하며, 화면에 보이지 않는 부분도 추정이 가능하다. 최종적으로, Blaze 모델의 33 개 랜드마크 (landmark) 3D 좌표와 가시성을 알 수 있다. 랜드마크 (x, y)의 이미지 좌표에 해당하는 깊이 값을 RGBD 카메라로 측정하고, 랜드마크 z 와 동일하게 정규화 하여 차이를 비교한 결과, 차이의 최소값은 0.07, 최대값은 0.36, 평균값은 0.17 이 나왔다.

5. 춤 모션 판정

생성된 춤과 사용자 춤의 유사도를 계산하기 위한 방법으로는 (1) 생성된 춤과 사용자 춤의 3D 자세를 추정하고, 각 관절점을 비교하여 유사도를 구하는 것과 (2) 사용자의 춤을 SMPL 모션 벡터로 변환하여 생성된 춤의 SMPL 모션 벡터와 비교하는 것이 있고, 그 중 첫 번째 방법을 사용한다.

생성된 춤과 사용자의 춤을 FBX 파일로 변환하여 언리얼 엔진 (Unreal Engine)에서 읽고, 두 모션의 3D 자세를 추정한다. 생성된 춤과 사용자의 춤의 크기를 동일하게 하고, 랜드마크 (x, y) 좌표를 벡터로 변환하여 정규화 한 후 코사인 유사도를 계산한다. 추가로 랜드마크 z 좌표 사이의 맨하탄 (manhattan) 거리를 비교하여 두 모션 사이의 유사도를 계산하여 판정할 수 있다.

6. 결론

본 논문에서는 (1) 모션 추정과 판정에서 두 가지 방법 중 실시간성을 만족하는 방법을 선택하고, (2) 자세 추정 결과 랜드마크의 z 값과 Realsense 의 깊이 값을 비교하였으며, (3) 생성된 춤과 사용자 춤의 랜드마크 3D 좌표를 이용해 유사도를 계산하였다.

추후 연구에서 진행할 내용은 먼저, 자세 추정 과정에서 Realsense 의 깊이 값을 이용하여 정확도를 높일 수 있다. 그리고 사용자가 춤을 보며 정확한 시점에 따라하는 것은 반응 속도 때문에 힘들다. 이를 고려하여 유사도를 계산할 때, 두 모션의 프레임을 완전히 동일하게 맞추지 않고, 적절한 보정이 필요하다. 또, 모션 판정 결과를 언리얼 엔진 상에서 시각화 하고, 모든 과정이 실시간으로 이루어지도록 한다. 3D 춤 자동 생성과 실시간 춤 모션 판정은 코인노래방과 같은 1 인 댄스 룸 또는 댄스 게임에 활용되어 발전될 수 있어 활동적인 취미 생활을 유도할 수 있고, 이에 따라 건강하고 활발한 사회의 형성을 기대할 수 있다.

ACKKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1F1A1068704)

참고문헌

[1] Li, Ruilong, et al. "AI Choreographer: Music Conditioned 3D Dance Generation with AIST++." IEEE/CVF International Conference on Computer Vision. 2021. p. 13401-13412.

[2] Kim, Jinwoo, et al. "A Brand New Dance Partner: Music-Conditioned Pluralistic Dancing Controlled by Multiple Dance Genres." IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 3490-3500.

[3] Loper, Matthew, et al. "SMPL: A Skinned Multi-person Linear Model." ACM Transactions on Graphics. 2015. p. 34.6: 1-16.

[4] Bazarevsky, Valentin, et al. Blazepose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204, 2020.