사진-SMPL 추정을 통한 3 차원 생성 아바타의 자세 변경

이재석, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

Text-to-3D Avatar Pose Manipulation with Image-to-SMPL Estimation

Jaeseok Lee, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

AvatarCraft 는 단일 객체에 대한 다중 시점 사진 집합을 통해 3 차원 표면 복원이 가능한 암시 신경 표면(Neural Implicit Surfaces; NeuS)과 텍스트 입력으로 사진 생성이 가능한 생성 모델인 디퓨전 모델(Diffusion Mosdel)의 결합으로 텍스트 기반 3 차원 아바타(text-to-3D avatar) 생성 모델(generative models)을 렌더링할 수 있다. 생성한 아바타는 대(大)자 자세를 취하고 있기 때문에 SMPL(Skinned Multi-Person Linear Model) 매개변수를 입력하여 원하는 자세로 변경할 수 있다. 그러나 SMPL 매개변수의 각각의 관절은 상대적인 회전 값이기 때문에 사용자가 직접 매개변수를 조작하는 것은 어렵다. 따라서 본 논문에서는 원하는 자세를 취한 사진으로부터 3 차원 SMPL 자세를 추출하고, 추정한 SMPL 매개변수로부터 AvatarCraft 모델이 생성한 아바타의 자세로 조작하는 파이프라인(pipeline)을 제안한다. 실험 결과 손이나 발 등의 세밀한 부분은 표현력이 떨어졌으나 전체적인 자세를 잘 재현하였음을 확인하였다.

Ⅰ. 서 론

텍스트-3차원 생성 모델은 다시점에서의 사진으로부터 3 차원에서의 물체의 표면을 복원하는 암시 신경 표면[1]과 텍스트를 입력하여 고해상도 2차원 사진을 생성하는 디퓨전 모델 [2]을 결합하여 입력 문장으로부터 3차원 물체를 생성하는 과업이다. 생성한 물체는 폴리곤 메시(polygon mesh) 등으로 추출이 가능하여 AR(Artificial Reality)이나 VR(Virtual Reality) 등에 활용할 수 있다.

그러나 텍스트-3차원 생성 모델은 사람 모형에 대한 표현 능력이 떨어지는 한계점이 존재한다. 따라서 AvatarCraft[3] 모델은 SMPL[4]을 이용하여 텍스트-3 차원 생성 모델을 확장, 텍스트를 입력하여 아바타를 임의의 자세(pose)로 생성할 수 있도록 텍스트-3차원 아바타 생성 모델을 제안하였다.

모델이 생성한 아바타는 대(大)자 자세를 취한 표준 공간(Canonical space) 상에 있기 때문에 SMPL 자세 매개변수를 이용하여 관측 공간(Observation space)으로 자세를 변형해야 한다. 하지만 SMPL 자세 매개변수는 관절에 대한 상대적인 3차원 회전 값이기 때문에 사용자가 원하는 포즈를 생성하기 위해 SMPL 매개변수를 직접 입력하기 어렵다.

따라서 본 논문에서는 사용자가 인체에 대한 사진을 입력하였을 시 사진에 대한 SMPL 포즈를 추정, 생성된 아바타가 해당 자세와 일치하도록 하는 파이프라인을 제안한다. 그 결과 SMPL 매개변수를 직접 조작하지 않아도 원하는 자세로 아바타를 변형할 수 있었다.

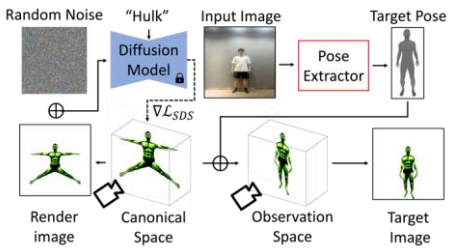


그림 1. 제안 모델 구조

Ⅱ. 본 론

암시 신경 표면은 부호화 거리 함수(Signed Distance Function)을 이용하여 불투명도를 예측한다. 암시 신경 표면은 카메라 매개변수 c 와 암시 신경 표면 모델의 매개변수 Φ를 이용하여 3 차원 공간 상에서 아바타 사진 $g(\Phi, c)$를 렌더링한다. Avatarcraft 모델에서는 대(大)자 모형 아바타로 초기화된 암시 신경 표면이 임의 카메라 시점으로부터 사진을 예측한다.

디퓨전 모델은 임의의 잡음으로부터 사진을 예측하는 생성 모델로, Avatarcraft 모델은 점수 증류 샘플링(Score Distillation Sampling; SDS)[5] 손실 함수를 통하여 암시 신경 표면 모델로 역전파를 진행한다.

$$\nabla SDS = \mathbb{E}\left[w(t)(\hat{\epsilon}(x_t, t, y) - \epsilon)\frac{\partial g(\Phi, c)}{\partial x_t}\right] \quad (1)$$

이 때 $\hat{\epsilon}$는 디퓨전 모델이 예측한 잡음, y 는 입력 텍스트, t는 디퓨전 모델의 시점(time step), ϵ는 암시 신경 표면이 렌더링한 사진에 삽입한 실제 잡음, x_t는 잡음이 추가된 사진, $w(t)$는 디퓨전 모델의 타임스텝에 따른 가중치 함수를 의미한다.

SMPL은 24개의 관절을 이용하여 3차원에서의 사람 모형을 생성하는 모델이다. SMPL은 72개로 구성된 자세 매개변수 θ 와 10개로 구성된 형상 매개변수 β 를 조합하여 6,890개의 정점(vertex)를 가진 3차원 폴리곤 메시 형태로 표현한다. 변형하고자 하는 자세의 SMPL 매개변수를 $(\theta_{obj}, \beta_{obj})$ 라 한다면, 선형 혼합 스키닝(Linear Blend Skinning)[4] 함수와 변형 알고리즘

표 1. 각 자세에 대한 사용자 선호도 조사 결과

	1	2	3	4	5	6	7	8
점수	4.4	4.3	4.8	3.2	3.9	4.3	4.1	3.5
종합	4.1							

(Deformation Algorithm)[4] 함수 T_P를 통해 3차원 공간상에서 원하는 자세의 사람 모형을 M으로 나타낼 수 있다.

$$T_P(\theta_{obj},\beta_{obj}) = \bar{T} + B_S(\beta_{obj}) + B_p(\theta_{obj}) \quad (2)$$

$$M(\theta_{obj},\beta_{obj}) = LBS(T_P(\theta_{obj},\beta_{obj}),J(\beta_{obj}),\theta_{obj},\mathcal{W}) \quad (3)$$

이 때 $\bar{T}$는 표준 공간에 있는 기본 대(大)자 자세, B_S는 형태 혼합 함수(blend shape function), B_p는 자세 혼합 함수, LBS는 선형 혼합 스키닝 함수, J는 관절의 위치, $\mathcal{W}$는 가중치를 의미한다.

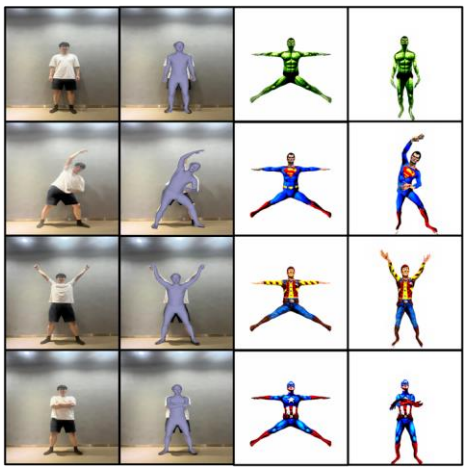
본 논문에서는 θ_{obj},β_{obj}의 직접적인 입력 없이 신경망 기반 자세 추출 모델[6]을 이용하여 사진을 입력하였을 때 사진에 있는 사람의 SMPL 매개변수를 추출하고 아바타의 자세를 변환하는 파이프라인을 제안한다. 파이프라인은 [그림 1]과 같다.

Ⅲ. 실험 및 논의

정성적 지표는 사진에 나타난 사람의 자세가 아바타의 자세를 잘 따르는지에 대한 사용자 선호도 조사를 실시하였다. 8가지의 자세에 대하여 점수를 1점부터 5점까지 측정하여 평균을 기록하였다. 사용자 선호도 조사 결과는 [표 1]과 같다.

시각화 결과는 [그림 2]와 같다. 자세는 직립 자세, 스트레칭 자세, 만세 자세, 팔짱 낀 자세의 4가지로 구성하였다. 제일 왼쪽부터 차례대로 자세를 추출하고자 하는 사진, 추출된 3차원 자세, 아바타의 기본 대(大)자 자세, 변형된 아바타의 자세를 가리킨다. 아바타는 Hulk, Superman, Woody, Captain America를 사용하였다.

선호도 실험 결과 전체적인 포즈를 잘 재현해 내는 것을 확인하였으나, 시각화 결과에서 손에 대한 포즈가 정확하지 못한 것을 [그림 3]에서 확인할 수 있다. 확인할 수 있었다. 이는 SMPL 포즈 추정 모델이 손에 대한 관절이 부족하기 때문에 손에 대한 표현력이



(1) (2) (3) (4)

그림 2. 시각화 결과.
(1) 원본 사진, (2) 추출한 자세 사진,
(3) 아바타 기본 자세, (4) 최종 결과

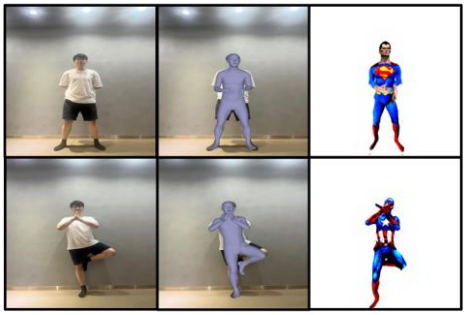


그림 3. 실패 사례.

떨어지고, 뒷짐을 지는 자세 등 보이지 않는 부분에 대해서도 추론하려고 하기 때문이라고 추론할 수 있다.

Ⅳ. 결 론

본 논문에서는 텍스트-3차원 아바타 모델의 SMPL 자세를 변형하기 위해 SMPL 매개변수를 직접 입력하는 대신 사진에서 자세를 예측한 뒤, SMPL 파라미터를 추정하여 아바타 자세를 변형하는 모델을 제안한다.

실험 결과, 전체적인 자세를 잘 재현하였으나 손이나 발 등의 세밀한 포즈를 잘 재현해 내지 못하는 것을 확인할 수 있었다. 추후 연구에는 손 등의 관절까지 표현하는 인간 자세 모델을 사용하는 아바타를 이용하여 이를 개선하고자 한다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.RS-2023-00212484, 복잡한 실제 주행환경에서 설명 가능한 움직임 예측).

참 고 문 헌

[1] WANG, Peng, et al. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106. 10689, 2021.

[2] ROMBACH, Robin, et al. High-resolution image synthesis with latent diffusion models. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 10684-10695.

[3] JIANG, Ruixiang, et al. AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control. arXiv preprint arXiv:2023, 17606, 2023.

[4] LOPER, Matthew, et al. SMPL: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Bolundaries, Volume 2. 2023. p. 851-866.

[5] POOLE, Ben, et al. "Dreamfusion: Text-to-3d using 2d diffusion." arXiv preprint arXiv:2209, 14988, 2022.

[6] CHOUTAS, Vasileios, et al. Monocular expressive body regression through body-driven attention. In: Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X 16, Springer International Publishing, 2020. p. 20-40.

메모 포함[이인1]: 그림 1 글씨 크기는 캡션 크기와 동일하게 키워주세요.

메모 포함[이2R1]: 글씨 크기 맞춰주었습니다

메모 포함[이인3]: 그림 2, 3 순서가 순차적으로 배치시켜주세요.

메모 포함[이4R3]: 순차 배치 완료하였습니다. 또한 댓글이 사라졌는데, 선호도 실험 결과를 종합 평균으로 내는 게 더 나을까요?? 각각의 포즈에 대해서 서술하는 것이 쉽지 않을 것 같습니다.

메모 포함[이인5R3]: 편한걸로 표기하죠.

메모 포함[이6R3]: 둘다 표기하는 방식을 사용하였습니다!