NeRF 기반 3차원 모델링을 통한 자세 추정

박찬, 김형주, 문남미
호서대학교 컴퓨터공학과
chan.park941003@gmail.com, kimhyungju01@gmail.com, nammee.moon@gmail.com

Pose Estimation through 3D modeling based on NeRF

Chan Park, Hyungju Kim, Nammee Moon
Dept. of Computer Science and Engineering, Hoseo University

요 약

2차원 이미지 또는 영상을 통한 자세 추정의 경우, 영상 내에서 발생할 수 있는 탐지 오류, 피사체 잘림, 폐색(Occlusion) 등으로 인해 자세 추정 정확도가 감소할 수 있다. 본 논문에서는 4장 이상의 다양한 각도로 촬영한 이미지를 NeRF(Neural Radiance Fields)를 통해 이미지 합성(Image synthesis)을 진행하여 3차원 모델을 생성한다. 이후 DeepLabCut을 사용하여 관절 좌표와 골격(Skeleton)을 구축한다. 구축한 골격을 인공지능에 학습시킨 뒤 2차원 영상에서의 관절 좌표 인식, 골격 구축, 자세 추정을 진행한다. 2차원 영상 테스트 데이터를 통해, 3차원 모델을 사전 학습한 인공지능 모델과 기존 2차원 이미지를 사용하여 학습한 인공지능 모델의 자세 추정 정확도를 비교한다.

1. 서론

자세 추정은 이미지 또는 영상을 사용하여 신체의 관절, 뼈대 등 움직이는 지점(Keypoints)을 특정하여 골격을 구축하고, 이를 탐지하여 자세를 인식하는 연구이다. 이를 통해 행동 인식, 객체 추적, 인간-컴퓨터 상호작용, 게임, 감시 시스템 등 다양한 분야에 활용하고 있다[1]. 하지만 2차원 이미지 또는 영상 데이터를 사용하는 기존 자세 추정 연구의 경우, 관절 지점의 모호함, 탐지 오류, 피사체 잘림, 폐색 등으로 인해 자세 추정 정확도가 하락하는 한계점이 있다[2].

2차원 자세 추정의 정확도 하락을 보완하기 위해 3차원 모델을 통한 자세 추정 연구를 제안하고 있다. 깊이(Depth)를 반영하는 양안 입체 영상 카메라를 사용하여 사람을 촬영한 뒤, 3차원 모델을 생성하고 생성한 모델을 통해 별도의 골격 구축 없이 자세 추정을 진행하였다[3]. 하지만 제안된 연구는 특수 카메라를 통해 촬영을 진행해야 하며, 깊이 인식에 오류가 발생하여 사물과 사람이 겹치는 한계점이 있다.

이러한 기존 연구의 한계점을 극복하기 위해 본 논문에서는 2차원 이미지 합성을 통한 3차원 모델 생성을 제안한다. 2차원 이미지 합성을 제안한 연구로는 GAN(Generative Adversarial Networks)을 사용하여 정면에서 촬영한 이미지를 측면 이미지로 구현하는 연구가 제안되었으며[4], 다른 연구에서는 다양한 각도로 물체 또는 배경을 촬영한 여러 장의 이미지를 딥러닝을 사용한 이미지 합성을 진행하여 3D 모델링을 생성하는 연구가 제안되었다[5].

따라서 본 논문에서는 다양한 각도로 촬영한 이미지를 통해 2차원 이미지 합성을 진행하여 3차원 모델을 생성한 뒤, 해당 모델을 기반으로 자세 추정을 진행하고자 한다.

2. 관련 연구

2-1. 3차원 자세 추정

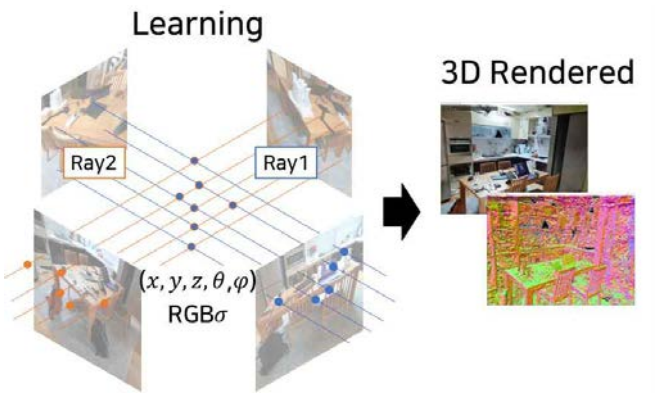
자세 추정은 관절, 뼈대 등 움직이는 지점을 특정하여 골격을 구축하고, 이를 인공지능에 학습시켜 2차원 이미지 또는 영상에서 동물 또는 사람의 신체 부위를 자동으로 찾는 것을 목표로 한다[6].

2차원 이미지 또는 영상을 사용하여 3차원 자세 추정을 진행하는 다양한 연구가 제안되었다. 여러 대의 카메라로 동시 촬영을 진행하여 2차원 골격을 3차원 골격으로 생성하는 연구가 제안되었고[7], 여러 대의 카메라로 동물의 움직임을 촬영한 영상 데

이터에 DeepLabCut을 사용하여 3차원 골격을 구성하여 자세 추정을 진행한 연구가 제안되었다[8]. 본 논문에서는 3차원 동물 자세 추정 연구에 사용된 DeepLabCut을 사용하여 이미지 합성을 통해 생성된 3차원 모델의 골격을 구축하고 이를 사용한 자세 추정을 제안하고자 한다.

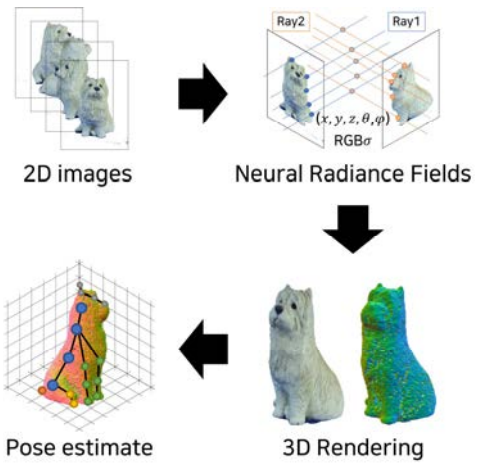
2-2. NeRF 기반 이미지 합성

이미지 합성을 통한 3차원 모델링은 다수의 이미지 데이터를 GAN, 딥러닝을 통한 이미지 합성 과정을 거쳐 3차원 모델로 구성하는 방법이다[9]. 제안된 이미지 합성 방법 중, NeRF는 공간적 위치(x, y, z)에서 보는 방향(θ, φ), 즉 5차원 좌표(x, y, z, θ, φ)를 입력 값으로 하여 RGBσ를 도출하는 딥러닝 학습을 진행한다. 이후 도출된 RGBσ값을 통해 3D 렌더링(Rendering)을 거쳐 3차원 모델을 생성하는 기법이다[5].



(그림 1) NeRF를 통한 이미지 합성 과정

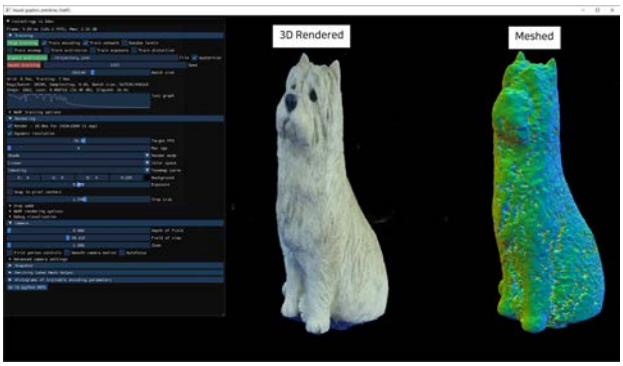
3. NeRF 기반 이미지 합성을 통한 자세 추정



(그림 2) NeRF 기반 이미지 합성을 통한 자세 추정 프로세스

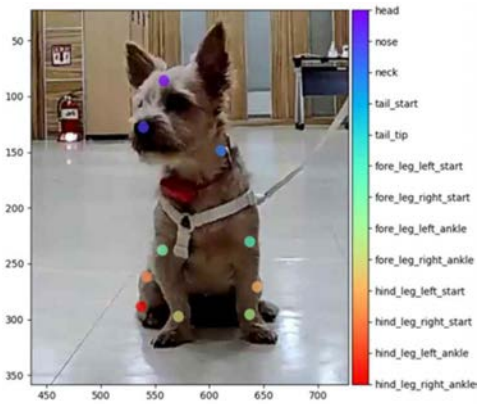
본 논문에서는 (그림 2)와 같이 NeRF 기반 이미지 합성을 통한 자세 추정 연구를 진행한다. 먼저, 강아지를 중심으로 360도 촬영을 진행한 이미지 데이터 셋인 THU-MVS Datasets를 사용한다.

Instant-NGP는 기존 NeRF의 5차원 좌표(x, y, z, θ, φ)를 입력 값으로 변환하는 과정을 해쉬 인코딩과 선형 보간을 사용한다[10]. 이를 통해 5차원 좌표를 빠르게 도출하여, 3차원 모델링까지의 학습 시간이 단축된 딥러닝 모델이다. 따라서 Instant-NGP를 사용하여 이미지 합성을 진행하고, 3차원 모델을 생성한 뒤, 3차원 모델의 메쉬(Mesh)를 추출한다.



(그림 3) Instant-NGP를 통한 이미지 합성 과정

이후 DeepLabCut을 사용하여 3차원 모델의 각 관절 좌표를 눈, 코, 머리, 척추, 슬개골 등의 10개 이상의 좌표로 라벨링하여 골격을 구축한다. 구축한 골격을 기반으로 인공지능 학습을 진행한 뒤, 2차원 영상에서의 자세 추정 정확도를 산출한다.



(그림 4) DeepLapCut을 사용한 관절 좌표 라벨링

정확도 산출 방법으로는 관절의 추정 좌표와 GT(Ground truth)좌표의 거리 평균을 산출하는 MPJPE(Mean Per Joint Position Error) 등의 지표를 사용한다. 이를 통해 3차원 골격을 사전 학습한

인공지능 모델과 기존 2차원 이미지를 사용하여 학습한 인공지능 모델의 자세 추정 정확도를 비교한다.

4. 결론

NeRF 기반 2차원 이미지 합성 기술을 통한 3차원 모델링 생성 및 이를 통한 자세 추정을 진행하는 방법을 제안하였다. THU-MVS Datasets의 경우, 실제 동물이 아닌 모형을 360도로 촬영한 데이터 셋으로써 실제 동물의 자세와 부정확하다는 한계점이 있다. 향후 실제 동물의 다양한 자세, 이미지 데이터 셋을 확보함으로써 제안한 자세 추정 방법을 통해 자세별 분류를 진행할 수 있을 것이다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부와 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (2019-0-01834)

참고문헌

[1] Munea, T. L., Jembre, Y. Z., Weldegebriel, H. T., Chen, L., Huang, C., Yang, C., "The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation", IEEE Access, Vol. 8, pp. 133330-133348, 2020.

[2] Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C., "Pose Flow: Efficient online pose tracking", arXiv preprint, arXiv:1802.00977, pp. 1-12, 2018.

[3] Hassan, M., Choutas, V., Tzionas, D., Black, M. J., "Resolving 3D human pose ambiguities with 3D scene constraints", In Proceedings of the IEEE/CVF international conference on computer vision, Korea, pp. 2282-2292, 2019.

[4] Chan, E. R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G., "pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p. 5799-5809, 2021.

[5] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., "Nerf: Representing scenes as neural radiance fields for view synthesis", Communications of the ACM, Vol. 65, No. 1, p. 99-106, 2021.

[6] Dang, Q., Yin, J., Wang, B., Zheng, W., "Deep learning based 2d human pose estimation: A survey", Tsinghua Science and Technology, Vol. 24, No. 6, p. 663-676, 2019.

[7] Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., Zhou, X., "Fast and robust multi-person 3d pose estimation and tracking from multiple views", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, No. 10, pp. 6981-6992, 2021.

[8] Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., Mathis, M. W., "Using DeepLabCut for 3D markerless pose estimation across species and behaviors", Nature protocols, Vol. 14, No. 7, pp. 2152-2176, 2019.

[9] Schwarz, K., Sauer, A., Niemeyer, M., Liao, Y., Geiger, A., "VoxGRAF: Fast 3D-Aware Image Synthesis with Sparse Voxel Grids", arXiv preprint, arXiv:2206.07695, pp. 1-22, 2022.

[10] Müller, T., Evans, A., Schied, C., Keller, A., "Instant neural graphics primitives with a multiresolution hash encoding", arXiv preprint, arXiv:2201.05989, pp. 1-15, 2022.