단안 이미지로부터 3D 사람 자세 추정을 위한 순서 깊이 기반 연역적 약지도 학습 기법

이영찬[1], 이규빈[1], 유원상[1*]
[1]선문대학교 정보통신공학과 인공지능 영상처리 연구실(AIIP Lab)
lyc19960618@sunmoon.ac.kr, askailak@sunmoon.ac.kr, wyou@sunmoon.ac.kr

Ordinal Depth Based Deductive Weakly Supervised Learning for Monocular 3D Human Pose Estimation

Youngchan Lee[1], Gyubin Lee[1], Wonsang You[1*]
[1]AIIP Lab, Information and Communications Engineering, Sun Moon University
* corresponding author

요 약

3D 사람 자세 추정 기술은 다양한 응용 분야에서의 높은 활용성으로 인해 대량의 학습 데이터가 수집되어 딥러닝 모델 연구가 진행되어 온 반면, 동물 자세 추정의 경우 3D 동물 데이터의 부족으로 인해 관련 연구는 극히 미진하다. 본 연구는 동물 자세 추정을 위한 예비연구로서, 3D 학습 데이터가 없는 상황에서 단일 이미지로부터 3D 사람 자세를 추정하는 딥러닝 기법을 제안한다. 이를 위하여 사전 훈련된 다중 시점 학습모델을 사용하여 2D 자세 데이터로부터 가상의 다중 시점 데이터를 생성하여 훈련하는 연역적 학습 기반 교사-학생 모델을 구성하였다. 또한, 키포인트 깊이 정보 대신 2D 이미지로부터 레이블링 된 순서 깊이 정보에 기반한 손실함수를 적용하였다. 제안된 모델이 동물 데이터에서 적용 가능한지 평가하기 위해 실험은 사람 데이터를 사용하여 이루어졌다. 실험 결과는 제안된 방법이 기존 단안 이미지 기반 모델보다 3D 자세 추정의 성능을 개선함을 보여준다.

1. 서론

3D 사람 및 동물 자세 추정 모델 훈련의 최종 목표는 이미지나 비디오와 같은 입력 데이터로부터 정확한 골격의 위치 및 깊이를 추정하는 것이다. 3D 사람 자세 추정 기술은 행동 분석(motion analysis), 증강 및 가상 현실(augmented and virtual reality), 스포츠 분석을 포함한 다양한 응용 분야에서 활용될 수 있어 많은 연구가 진행되어 온 반면[1], 데이터 부족 문제로 인해 3D 동물 자세 추정 기술에 관한 연구는 전 세계적으로 극히 미진하다.

기계 학습 및 딥러닝 모델의 정확성은 고품질 훈련 데이터에 크게 의존하기 때문에, 데이터 라벨링(data labeling) 구축은 매우 중요한 작업이다[3]. 그러나 라벨을 생성하는 작업은 많은 시간과 비용이 소모되는 문제가 있다. 이러한 문제를 해결하기 위해 약지도 학습(weakly-supervised learning)은 라벨링 데이터가 완전하지 못한 상태에서 모델을 훈련한다[4]. 3D 사람 자세 추정 분야에서는 다중 시점 기하학, 무작위 투영, 연역적 방법, 순서 정보 등을 사용한 약지도 학습 방식들을 제안하였다. Kocabas M *et al.*은 다중 시점 기하학(multi-view geometry)을 통해 추정된 가상의 3D 자세 라벨을 사용하여 단안 이미지로부터 3D 자세를 추론하는 데 있어서 좋은 성능을 보였다[5]. Iqbal U *et al.*은 다중 시점의 일관성(multi-view consistency)을 사용한 종단간 방식의 약지도 학습 방식을 제안하였다[6]. 또한, Chen X *et al.*은 3D 주석이 없는 약지도 3D 사람 자세 추정 방식의 보정된 카메라 모델의 일반화 문제를 완화하기 위해 연역적 약지도 학습 방식을 채택하였다[7]. Pavlakos G *et al.*은 3D 주석의 제한된 가용성 문제를 해결하기 위해 최초로 순서 깊이 정보 기반의 손실을 사용한 약지도 학습 방식을 제안하였다[8].

본 논문에서는 기존 연역적 약지도 학습 기반 모델을 소개하고, 순서 정보를 사용하여 3D 자세 추정 모델의 훈련 성능을 향상시키는 교사-학생 모델을 제안한다. 연역적 약지도 학습 기반 모델을 백본 네트워크로 하는 교사-학생 구조로서, 두 단계로 구성되어 있다. 첫 번째 단계에서는 사전 학습 모델을 사용하여 새로운 시점의 자세 데이터를 생성하고,

두 번째 단계에서는 기존 시점 데이터와 생성된 새로운 시점 데이터를 사용하여 3D 자세를 추정한다. 또한, 약지도 학습 방식의 성능적인 한계를 보완하기 위해 추가적인 순서 깊이 정보를 사용하였다.

본 연구는 사람과 비교하여 현저히 부족한 동물 데이터에 대해 다중 시점 데이터를 생성하고, 3D 동물 자세 추정 연구에 대해 본 논문에서 제안하는 접근 방식의 적용 가능성에 대하여 평가한다. 또한, 순서 깊이 정보 활용을 통한 기존 모델의 성능 향상을 확인하는 것에 목적이 있다.

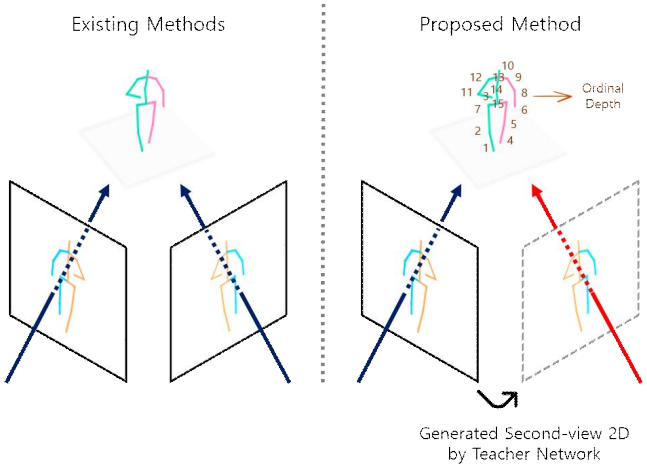

그림 1. 기존 및 제안된 3D 자세 추정 방식 요약

2. 방법

2.1. 연역적 약지도 학습 기반 모델

본 연구에 기반이 되는 연역적 약지도 학습 기반 모델(Deductive Weakly-Supervised Learning, DWSL)은 3D 약지도 학습 과정에서 보정된 카메라 매개 변수의 필요성과 자연 이미지에 대해 제한된 모델의 일반화 문제를 해결하기 위해 제안되었다. DWSL은 그림 1과 같이 3D 사람 자세를 추정하는 3D Pose Estimator와 시점 변환에 사용되는 변환 행렬을 예측하는 Transformation Predictor로 구성된다[7]. 3D Pose Estimator는 J. Martinez *et al.* 이 제안한 2D-3D 리프팅(lifting) 모델을 사용하여 2D 단안 사람 자세 데이터로부터 3D 사람 자세를 추정한다[9]. 2D-3D 리프팅 모델은 2D 사람 자세 좌표 $p_{\text{input}} = (\mathrm{x}, \mathrm{y})$를 입력으로 $\hat{P}_{\text{input}} = (X, Y, Z)$를 예측하는 모델로서 다층 신경망(multilayer neural network), 배치 정규화(batch normalization), 드롭아웃(dropout) 및 정류 선형 유닛(rectified linear unit)과 잔차 연결(residual connections)로 구성된 간단한 3D 자세 추정 모델이다.

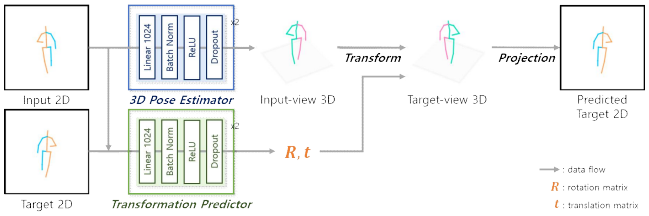

그림 2. 연역적 약지도 학습 기반 모델

Transformation Predictor는 3D Pose Estimator와 똑같이 구성되어있는 반면, 두 시점의 2D 사람 자세 좌표를 입력으로 하고 출력층을 변형하여 카메라 외부 파라미터인 회전 행렬 $\boldsymbol{R}$과 이동 행렬 $\boldsymbol{t}$을 예측한다. 또한, 식 (1)과 같이 P_{input}을 다른 시점의 3D 좌표 $P_{\text{target}} = (\mathrm{X}', \mathrm{Y}', \mathrm{Z}')$로 변환 및 투영하여 다른 시점 2D 데이터 $\hat{p}_{target} = (x', y')$을 생성하고, 생성된 2D 및 3D를 내부적으로 감독하여 훈련한다.

$$P_{target} = R \cdot P_{\text{input}} + \mathrm{t} \tag{1}$$

2.2. 연역적 약지도 교사-학생 모델

연역적 약지도 교사-학생 모델은 그림 2와 같이 두 개의 Stage로 구성되며 각 Stage의 백본 네트워크로서 DWSL 모델이 사용되었다. Stage 1의 사전 학습된 교사 네트워크는 Stage 2에서 학습될 학생 네트워크에 예측된 다른 시점의 2D 자세를 제공할 뿐만 아니라(빨간색 화살표) 입력 시점의 3D 자세 추정을 통해 지도한다.

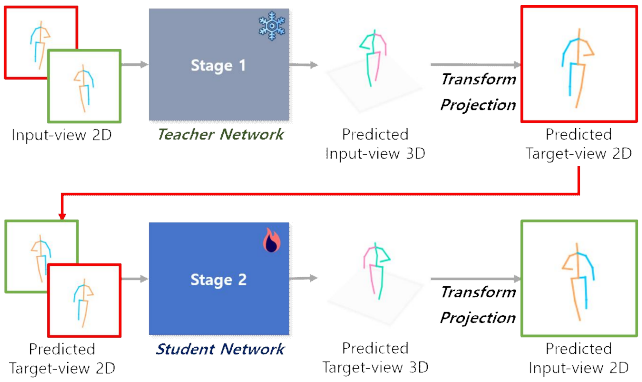

그림 3. 연역적 약지도 교사-학생 모델

학생 모델의 훈련을 위해, Human3.6M 데이터 세트[10]를 사용하여 표준 프로토콜에 따라 다섯

가지 주제 (S1, S5, S6, S7, S8)에서 추출한 2D 자세 좌표를 사용하였고 두 가지 주제 (S9, S11)에 대해 평가하였다. 배치 크기는 1024, 최대 반복 수는 150, 초기 학습률은 0.001로 설정하였고 최적화 기법은 Adam을 사용하였다.

2.3. 순서 깊이 정보

약지도 학습의 성능 향상을 위해 순서 깊이 정보의 차이를 손실함수(L_{ord})로 정의하였다. 순서 깊이 행렬 $O_{i,j}$은 정확한 3D 자세 좌표가 아닌 각 카메라 관점에서 각 관절 J_i의 인접 관절 J_j과의 깊이 관계를 나타낸다. 이러한 깊이의 순서는 실제 3D 깊이 정보와 달리, 사람이 시각적으로 쉽게 획득할 수 있다는 장점이 있다. 예를 들어, J_i가 J_j보다 가깝다면 -1, 멀다면 1, 동일 선상에 존재한다면 0으로 표시하며 식 (3)과 같다.

$$O_{i,j} = \begin{cases} -1, & sgn(J_i - J_j) < 0 \\ 0, & sgn(J_i - J_j) = 0 \\ 1, & sgn(J_i - J_j) > 0 \end{cases} \quad (2)$$

2.4. 손실함수

딥러닝 모델에 적용된 손실함수(loss function)로서 식 (3)과 같이 여러 손실함수의 가중합으로 정의하였다. 예측된 $\hat{p}_{target}$과 p_{target} 사이의 차이가 최소화되도록 평균 제곱 오차(mean squared error)를 통해 손실함수(L_{rec})를 정의한다. 또한, 다른 시점 2D 데이터의 잘못된 예측으로 인해 3D 자세 추정이 크게 잘못될 수 있으므로 신체 구조적인 규칙을 이용한 손실함수를 사용한다. $\hat{P}_{\text{input}}$을 양의 깊이로 제한하는 손실함수(L_{pos}), 팔과 다리에 대한 대칭 길이 손실함수(L_{sym}), 유효하지 않은 각도를 제거할 수 있는 손실함수(L_{ang})를 정의하였다. DWSL과 달리 제안된 모델은 교사-학생 네트워크를 통한 손실함수(L_{bridge})와 순서 깊이 정보 기반 손실함수(L_{ord})를 추가적으로 사용하였다. 이때, 손실함수 가중치 λ_{rec}, λ_{pos}, λ_{sym}, λ_{ang}는 기존 DWSL 모델과 같이 1.0, 1.0, 10.0, 0.001로 설정하였고, λ_{bridge}, λ_{ord}는 1.0, 1.0로 설정하였다.

$$L_{dwsl} = \lambda_{rec}L_{rec} + \lambda_{pos}L_{pos} + \lambda_{sym}L_{sym} + \lambda_{ang}L_{ang} + \lambda_{bridge}L_{bridge} + \lambda_{ord}L_{ord} \quad (3)$$

3. 결과

본 논문에서는 모델 성능의 정량적 평가지표로서 MPJPE(Mean Per Joint Position Error)[10] 및 PMPJPE(Procrustes Analysis MPJPE)를 사용한 자세 유사성을 평가하고, PCK(Percentage of Correct Keypoints), AUC(Area Under Curve)를 통해 정확도를 측정하였다. 표 1은 DWSL[7], 교사-학생 모델 및 순서 정보 손실함수의 유무에 따른 3D 사람 자세 추정 결과를 정량적으로 보여준다.

표 1. 3D 자세 추정 정량적 평가 결과
(OD: Ordinal Depth, TS: Teacher-Student)

감독	모델 구분	PMPJPE	MPJPE	PCK	AUC
full	Lifting [9]	52.1	62.9	-	-
weak	DWSL [7]	64.1	67.6	0.93	0.58
	DWSL + OD[7]	62.3	65.8	0.94	0.58
	DWSL-TS	72.1	76.4	0.91	0.53
	DWSL-TS + OD (Ours)	69.9	74.5	0.92	0.54

표 1과 같이 전체적인 결과는 지도 학습 기반 리프팅 모델이 약지도 학습 모델보다 성능이 높은 것을 볼 수 있다. 그러나, 이러한 결과는 학습을 감독하는 방식 차이에서 발생하는 한계로 볼 수 있다. 교사-학생 모델은 기존 DWSL 모델과 비교하여 성능이 향상되지 않았다. 하지만 정량적 및 시각적 평가 결과를 보았을 때 많은 차이가 나지 않는 것을 볼 수 있으며 (PMPJPE=69.9, PCK=0.92), 이러한 결과는 새로운 시점 데이터 생성 및 추론이 충분히 가능할 것으로 보인다(그림 5). 또한, 각 모델에 순서 깊이 정보 기반 손실함수를 적용한 결과는 정량적 및 정성적으로 모두 성능 향상을 보였다(그림 4).

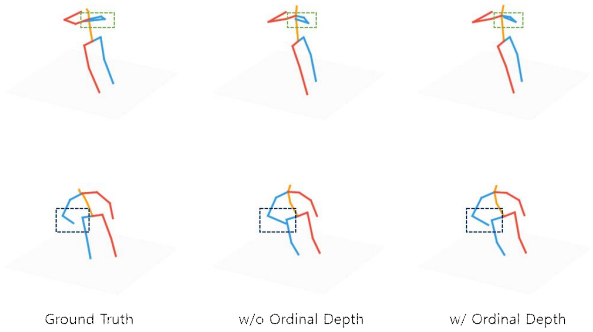

그림 4. 순서 깊이 정보 사용 유무에 따른 3D 사람 자세 추정 결과 비교

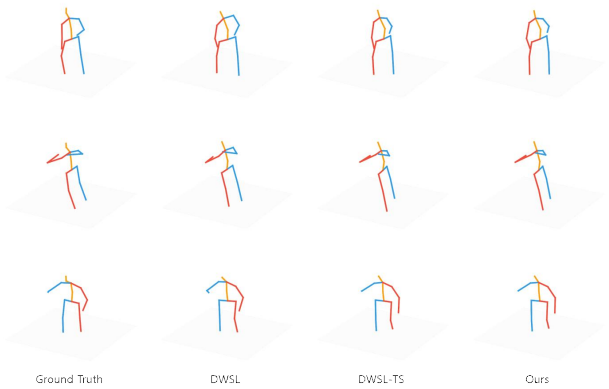

그림 5. 3D 사람 자세 추정 결과 비교

4. 결론 및 향후 연구

본 논문에서는 사람과 비교하여 3D 데이터가 현저히 부족한 동물에 대해 새로운 시점 데이터를 생성하고, 약지도 학습을 통해 3D 동물 자세 추정 모델에 대한 훈련 가능성을 평가하기 위해 DWSL을 백본 네트워크로 하는 연역적 약지도 교사-학생 모델을 사용하였다.

실험 결과에 따르면, 교사-학생 모델의 성능은 지도 학습 기반 리프팅 모델이나 기존 DWSL 모델과 비교하여 낮지만, 비슷한 결과를 보였다. 제안된 방식으로 인한 성능 개선은 미비하지만, 이러한 결과는 동물 데이터에 대하여 새로운 시점 동물 데이터 생성과 다중 시점 동물 데이터를 사용한 모델의 학습 및 추론이 가능할 것으로 보인다.

성능적인 한계를 극복하기 위해 향후 연구에서는 NeRF 및 Diffusion과 같은 모델 구조를 활용하여 다중 시점 데이터 생성 성능을 개선할 계획이다. 또한, 동물 데이터에 대해 순서 깊이 라벨링 데이터를 구축하고, 실제 깊이 데이터 분포와의 연관성을 높이는 방안을 마련하여 독창적인 3D 동물 자세 추정 기술을 개발할 예정이다.

ACKNOWLEDGMENTS

본 연구는 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 기본연구(2022R1 F1A1075204), 4단계 두뇌한국21 사업(4단계 BK21 사업) 및 지자체-대학 협력기반 지역혁신사업(2022RIS-004), 중소기업벤처부의 재원으로 수행된 2021년도 창업성장기술개발사업(S3228660)의 연구결과로 수행되었음.

참고문헌

[1] Zheng C, Wu W, Chen C et al., “Deep Learning-based Human Pose Estimation: A Survey” ACM Comput. Surv. 56, 1, Article 11, 37 pages.

[2] Zhang S, Wang C, Dong W et al., “A Survey on Depth Ambiguity of 3D Human Pose Estimation” Applied Sciences. 2022; 12(20):10591.

[3] Jain A, Patel H, Nagalapatti L et al., “Overview and Importance of Data Quality for Machine Learning Tasks.” In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 2020, 3561 - 3562.

[4] Khoreva A, Benenson R, Hosang J et al., "Simple Does It: Weakly Supervised Instance and Semantic Segmentation," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1665-1674.

[5] M. Kocabas, S. Karagoz and E. Akbas, "Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 1077-1086.

[6] U. Iqbal, P. Molchanov and J. Kautz, "Weakly-Supervised 3D Human Pose Learning via Multi-View Images in the Wild," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 5242-5251.

[7] Chen X, Wei, P, Lin, L, “Deductive Learning for Weakly-Supervised 3D Human Pose Estimation via Uncalibrated Cameras,” Proceedings of the AAAI Conference on Artificial Intelligence, 35(2), 1089-1096.

[8] G. Pavlakos, X. Zhou and K. Daniilidis, "Ordinal Depth Supervision for 3D Human Pose Estimation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7307-7316.

[9] J. Martinez, R. Hossain, J. Romero and J. J. Little, "A Simple Yet Effective Baseline for 3d Human Pose Estimation," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2659-2668.

[10] C. Ionescu, D. Papava, V. Olaru and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1325-1339, July 2014.