Anatomy and Geometry Constrained One-Stage Framework for 3D Human Pose Estimation

Xin Cao and Xu Zhao

Department of Automation, Shanghai Jiao Tong University, China
Institute of Medical Robotics, Shanghai Jiao Tong University, China
{xinc1024,zhaoxu}@sjtu.edu.cn

1 Introduction

This supplementary material presents: (1) the detail description the network architecture and data augmentation strategies; (2) additional qualitative experimental results from different camera views.

2 Implementation details

In this section, we provide more details regarding the data augmentation strategies used in the training process and the network architecture of our *backbone* and *feature extractor*.

Network Architecture. We use ResNet-50 followed by three deconvolution layers as our backbone [1]. After that an average pooling layer and three FC layers are applied to acquire the final root joint position, translation parameters and rotation parameters. The exact architecture is summarized in Table 1.

Table 1. Detailed network architectures. ConvT corresponds to a layer performing transposed Convolution, N denotes the number of channels, K denotes the kernel size, S denotes the stride size, and P denotes the padding size. BN denotes the Batch Normalization operation.

	Layer	Module
Backbone	1	ResNet50
	2	ConvT-(N256,K4,S2,P1), BN, ReLU
	3	ConvT-(N256,K4,S2,P1), BN, ReLU
	4	ConvT-(N256,K4,S2,P1), BN, ReLU
Feature Extracter	1	Avgpool(K64)
	2	Reshape(Batch, 64)
Output	1	FC(64, 17), Sigmoid
	2	FC(64, 51), Sigmoid
	3	FC(64, 3), Sigmoid

Data Augmentation. We apply data augmentation strategies to strengthen the robustness of our model. Training data is augmented by random rotations

of $\pm 30^{\circ}$ and scaled by a factor between 0.75 and 1.25. Additionally, we utilize synthetic occlusions [2] to make the network robust to occluded joints.

3 Qualitative Results

In this section, we provide more qualitative results from different views.

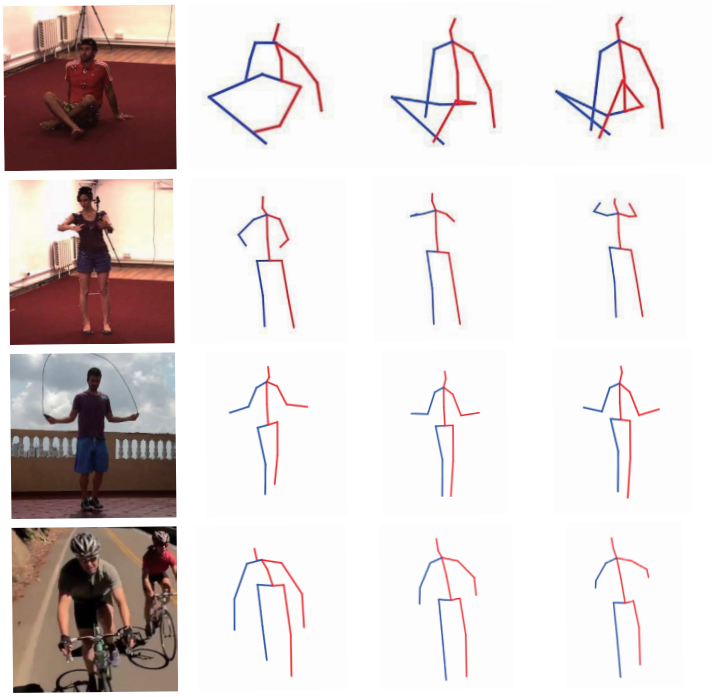

Fig. 1. Qualitative results on Human3.6m datasets and MPII datasets, results are represented from multiple views.

References

1. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (ECCV). (2018)
2. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: How robust is 3d human pose estimation to occlusion? arXiv preprint arXiv:1808.09316 (2018)