17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Image Information Assistance Neural Network for VideoPose3D-based Monocular 3D Pose Estimation

Hao Wang, Dingli Luo, Takeshi Ikenaga

Graduate School of Information, Production and Systems, Waseda University,
Kitakyushu 808-0135, Japan
wanghao@asagi.waseda.jp

Abstract

3D pose estimation based on a monocular camera can be applied to various fields such as human-computer interaction and human action recognition. As a two-stage 3D pose estimator, VideoPose3D achieves state-of-the-art accuracy. However, because of the limitation of two-stage processing, image information is partially lost in the process of mapping 2D poses to 3D space, which results in limited final accuracy. This paper proposes an image-assisting pose estimation model and a back-projection based offset generating module. The image-assisting pose estimation model consists of a 2D pose processing branch and an image processing branch. Image information is processed to generate an offset to refine the intermediate 3D pose produced by the 2D pose processing network. The back-projection based offset generating module projects the intermediate 3D poses to 2D space and calculates the error between the projection and input 2D pose. With the error combining with extracted image feature, the neural network generates an offset to decrease the error. By evaluation, the accuracy on each action of Human3.6M dataset gets an average improvement of 0.9 mm over the VideoPose3D baseline.

1. Introduction

3D human pose estimation with a monocular camera is the task of inferring a 3D pose matching the depicted person. The input is an RGB image or video from a single perspective [1]. This technology has wide application in various tasks such as action recognition, sports analysis, motion capture, etc. [2]

Previous researches proposed different approaches for this target [3, 4, 6]. However, newly proposed methods with deep learning have significantly outperformed conventional methods and have become a mainstream method for 3D pose estimation.

Neural network based methods for 3D pose estimation can be categorized into two main types: (1) end-to-end method (2) two-stage method. End-to-end methods produce a 3D pose without any other tools besides a deep neural network, which can achieve high speed to apply in real-time applications [3]. Two-stage methods use 2D keypoints produced by an off-the-shelf 2D keypoint detector [4]. In most scenarios without constraint on speed, two-stage methods are more commonly adopted because they fully use the advantage of high accuracy of advanced 2D pose estimation and achieve higher average accuracy than end-to-end methods [5]. VideoPose3D [6] by Facebook is one of the great two-stage 3D pose estimators and it achieved state-of-the-art accuracy when published based on temporal convolution and semi-supervised training.

The high accuracy of VideoPose3D proves it is feasible to infer a relatively accurate 3D pose with a 2D pose as input. However, information loss still exists in the mapping from images to 2D poses. And such limitation cannot be compensated by a better 2D pose estimator because there is an upper limit to the amount of information in a 2D pose. This information loss also limits the improvement in the accuracy of VideoPose3D. The reason for the information loss in a two-stage 3D pose estimation approach can be explained as follows. The first stage can be regarded as information compression, extracting the most essential information and abandoning the information that is highly likely to be redundant. But actually, some pixels that do not directly contribute to predicting a 2D pose are helpful for the final stage. For instance, the pixels of the environment can help recognize human action and the luminance of the pixels of limbs can indicate the distance and help determine the projection between 3D and 2D space. Therefore, the information abandoned in the first stage is helpful to the mapping between 2D pose and 3D pose, which is a problem of conventional two-stage 3D pose estimation.

Aiming at improving the accuracy of VideoPose3D with image information, this paper proposes an image-assisting model with more efficient use of image information and a back-projection based offset generating module. Combining the proposals, the network can achieve an improvement on VideoPose3D baseline.

2. Image-assisting neural network

This section shows the proposals and the network architecture for the implementation of them. The overview of the image assisting network is shown in Figure 1. In the processing pipeline, image information is processed by a backbone to extract the feature. On the other side, intermediate 3D poses from the 2D pose processing network are projected to 2D space. Next, the error between projection and input 2D pose is sent to offset generating network along with extracted image feature. In the last stage, the offset is added to the intermediate 3D pose and the final result is generated.

Subsection 2.1 explains the image assisting 3D pose estimation model and the merit of it comparing with directly

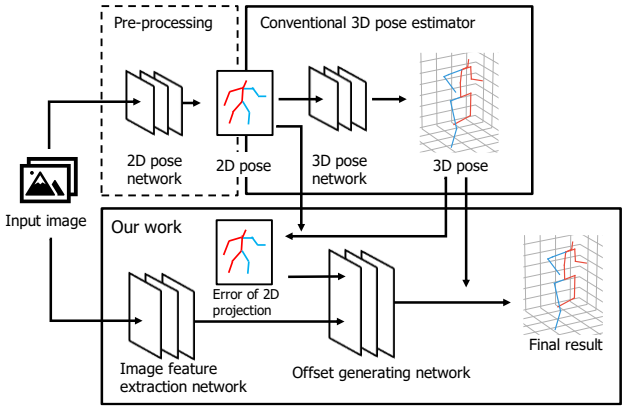
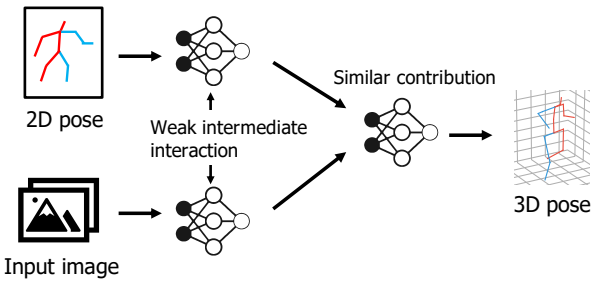


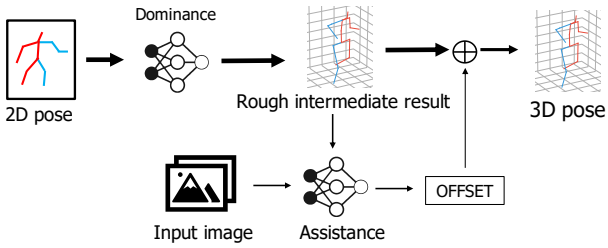
Figure 1. Overview of image-assisting network

combining the features. Subsection 2.2 introduces the back-projection based offset generating module, including its components and its advantages. And subsection 2.3 shows the detail of the network that implements the proposals.

2.1. Image-assisting 3D pose estimation model



(a) Independent processing model



(b) Image-assisting model

Figure 2. Conceptual difference between independent processing model and proposed model

To effectively use image information, it is necessary to find an approach to add image information to a two-stage 3D pose estimation network.

The post-processing method is independent of the pose estimation network and should be trained solely so it can be well applied as a model-agnostic refiner but is not suitable to improve a specific model.

To integrate image processing and VideoPose3D, we can divide approaches into two types as follows. The conceptual difference is demonstrated in Figure 2.

Independent processing model: To process image information and 2D pose equally, we assign two different networks to either input and make them extract the feature respectively. Afterward, we concatenate the feature and do final-stage processing.

In the independent processing method, it is possible to combine image feature and 2D pose feature. However, because of the final concatenating operation, the two inputs have the same status in the whole process. It means in order to optimize a well-trained 2D pose network, the image branch should have a similar capacity to infer a 3D pose, which indicates that it should have the same level of network complexity as an end-to-end network. In conclusion, the balanced processing method has high difficulty for design and high requirement for hardware.

Image-assisting model: Considering that 2D pose can indicate a 3D pose with considerable accuracy and it indirectly contains some information from the original image, we think it is promising to process images as assistance. Based on this, we propose a network structure to learn the mapping between image pixels and offset on a pose for refinement instead of abstract feature for concatenating operation.

Image-assisting model fully uses the advantage of conventional 2D pose network and has an acceptable requirement for hardware. Moreover, this model has strong interaction between the two branches, which enables the information to be well utilized. So, our experiment and evaluation are all based on this method.

2.2. Back-projection based offset generating module

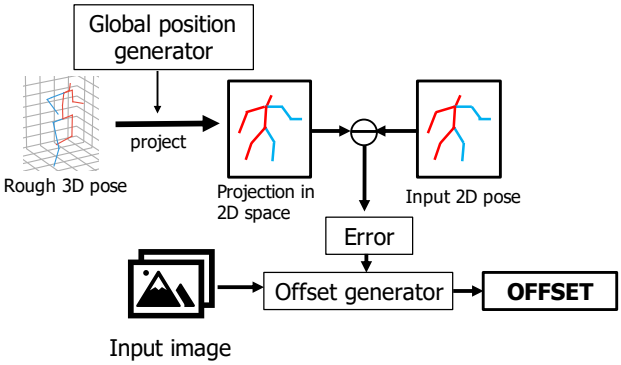


Figure 3. Demonstration of proposed offset generating method based on back-projection

As introduced, we want to use image information as assistance to produce an offset. Some researchers also refine a pose by adding an offset with post-processing network [7] or based on the stable distribution of different kinds of error in 2D pose estimation [8]. However, in 3D pose estimation, there is no discovered law of the occurrence of errors so we cannot do refinement with the same method for 2D pose and post processing method is not suitable as explained in 2.1. To solve this problem, we design a back-projection based offset prediction as shown in Figure 3.

In general, firstly we get an intermediate 3D pose with the 2D processing network and project the pose to 2D space

with the help of a predicted global position and camera parameters. Then we get the error between projection and input 2D pose and process the image feature and the error together to generate an offset. Finally, we get an optimized 3D pose by adding the offset to the intermediate rough pose.

The projection generating module functions to output a global position, namely the position of root joint of predicted 3D pose. The reason we need a global position of a 3D pose is that a 3D pose corresponds to numerous 2D poses because different points with different depths are likely to have the same position in 2D space. So, we need to determine the 2D space first. One way to manage it is to obtain the camera parameter and the root joint position in 2D space. The parameter of the camera is available in both our training set and daily application. Also, VideoPose3D proves it is feasible to infer a global joint with a 2D pose as input. This is the working principle of the projection generating module.

To extract image features, we use a simple convolutional neural network as a backbone. Then we combine the extracted feature and error of projection so we obtain an output containing both image information and defect of the 2D processing network. In the next stage, this intermediate feature is further processed and the whole module will output an offset to be added to rough 3D pose as optimization.

In our work, we use the error between projected 2D pose and input 2D pose to represent the defect of the rough 3D output. But actually, the usual way to represent the error is to calculate the difference between predicted 3D pose and ground truth. We made this adjustment based on the following reasons:

- Input 2D pose estimator can be approximately regarded as ground truth so theoretically a perfect output of 3D pose should have a projection same as the input 2D pose. Therefore, the projection error can also represent the deviation of the output and to make the projection consistent with the input can also optimize the 3D pose output
- 2D key points contain much less information than 3D key points so processing 2D error requires much lower network complexity than processing 3D error. We want to propose a hardware-friendly method that has the capacity to optimize VideoPose3D and the potential for further adjustment.

2.3. Implementation on network

The network architecture is shown in Figure 4. We divide the whole network into 3 parts including the original network of VideoPose3D. They are as follows:

2D pose processing network: The input 2D pose is processed by the VideoPose3D network with a reception field of 243 frames. It consists of 9 blocks of 1D convolutional layers, BN layers, ReLU activation layers and dropout layers.

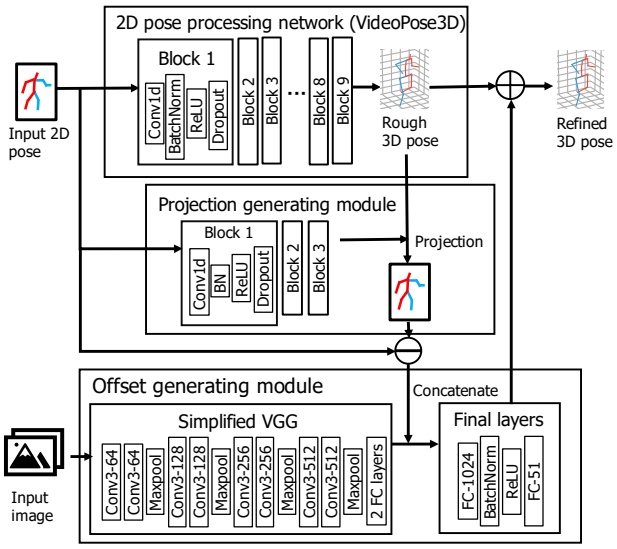


Figure 4. Network architecture

Projection generating module: Inspired by the semi-supervised training method of VideoPose3D, we use 3 blocks of 1D convolutional layers to generate a global position with a 2D pose as input. With the global position and camera parameters, the output of VideoPose3D can be projected to 2D space for the next step of processing.

Offset generating module: We use a simplified VGG network as a backbone to extract image features. Original VGG net consists of 5 blocks of convolutional layers and max-pooling layers along with 3 fully connected layers. We use a simplified version with only 4 blocks and 2 fully connected layers for faster training and less consumption of hardware resources. And we calculate the difference between projection and input 2D pose and then concatenate the image feature and the difference as an intermediate feature, which combines both image information and predicted error of the 2D pose processing network. In the final stage, the intermediate feature is processed by two fully connected layers and the output is added to a rough 3D pose to generate the final output, a refined 3D pose.

3. Experiment

3.1. Experiment method

Loss function of our network is mean per joint position

Table 1. Reconstruction error (MPJPE)

	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Sun et al.[10]	52.8	54.8	54.2	54.3	61.8	67.2	53.1	53.6	71.7	86.7	61.5	53.4	61.6	47.1	53.4	59.1
Yang et al.[11]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	**43.6**	60.1	47.7	58.6
Lee et al.[12]	**40.2**	49.2	47.8	52.6	50.1	75.0	50.2	**43.0**	**55.9**	73.9	54.1	55.6	58.2	43.3	43.3	52.8
VideoPose3D[7]	45.9	47.5	44.3	46.4	49.0	56.9	45.6	44.6	58.8	66.8	47.9	44.7	49.6	33.1	34.0	47.6
Ours	44.3	**46.9**	**43.2**	**45.0**	**47.1**	**56.5**	**44.3**	44.5	57.9	**66.3**	**47.0**	**43.4**	48.1	**32.6**	**33.3**	**46.7**

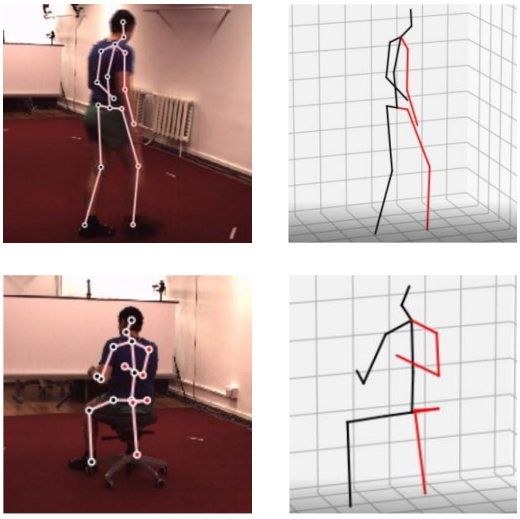


Figure 5. Visualized results

error (MPJPE), which is the mean of Euclidean distance between ground truth and prediction for all the joints.

Our training dataset is Human3.6M dataset [9]. It contains 3.6 million frames of videos performed by 11 subjects and each frame is annotated with 3D positions of 17 joints. The input 2D pose is generated by a CPN model fined tuned on Human3.6M, which is provided by [6]. Our global position generator is trained independently and functions as a ready-made module. The training of image network and 2D pose network takes about 5 days on a GeForce GTX 1080Ti.

3.2. Experiment result

The accuracy on each action and comparison with conventional work is shown in Table 1. And some visualized results are shown in Figure 5. The demonstrated accuracy of VideoPose3D is evaluated without test time augmentation. Our global position generator gets an average accuracy of 5 mm. Our network achieves an average MPJPE of 46.7 mm on Human3.6M dataset, which is a 0.9 mm improvement on the original VideoPose3D network without image input. The highest average accuracy increase is 1.9 mm on phoning action and the lowest average accuracy increase is 0.1 mm on purchasing action.

4. Conclusion and future work

This paper proposes an image-assisting neural network to improve VideoPose3D. The proposed image-assisting pose estimation model uses image information as the assistance of a 2D pose processing network. The proposed back-projection based offset generating module represents the error of intermediate rough pose in 2D space. By projecting 3D poses to 2D space and calculating the bias from input 2D pose, the error can indicate the accuracy but has lower computational complexity than 3D error. A network architecture is proposed to implement the conceptual proposals. By experiments, this work gets an average accuracy of 46.7 mm on Human3.6M, which is an improvement of 0.9 mm compared with VideoPose3D baseline.

For future research, the structure of the proposed network has no strong dependency on the VideoPose3D network. Therefore, this work has the potential to be applied to other two-stage 3D pose estimation and it is planned to do further experiments based on it.

Acknowledgement

This work was supported by KAKENHI (21K11816) and Waseda University.

References

[1] S. C. Babu, "A 2019 guide to 3D Human Pose Estimation", nanonets.com/blog/ human-pose-estimation-3d-guide/

[2] H. Liu, D. Luo, S. Du and T. Ikenaga, "Resolution Irrelevant Encoding and Difficulty Balanced Loss Based Network Independent Supervision for Multi-Person Pose Estimation," *2020 13th International Conference on Human System Interaction*, Tokyo, Japan, 2020

[3] S. Li, A. Chan. "3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network"，*Asian Conference on Computer Vision*, 2014

[4] D. Luo, S. Du and T. Ikenaga, "End-to-End Feature Pyramid Network for Real-Time Multi-Person Pose Estimation," *2019 16th International Conference on Machine Vision Applications*, Tokyo, Japan, 2019

[5] C. Chen and D. Ramanan, "3D Human Pose Estimation = 2D Pose Estimation + Matching," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5759-5767，2017

[6] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. "3D human pose estimation in video with temporal convolutions and semi-supervised training," *Conference on Computer Vision and Pattern Recognition* 2019.

[7] M. Fieraru, A. Khoreva, L. Pishchulin and B. Schiele, "Learning to Refine Human Pose Estimation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 318-31809，2018

[8] G. Moon, J. Y. Chang and K. M. Lee, "PoseFix: Model-Agnostic General Human Pose Refinement Network," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019

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

[10] X. Sun, J. Shang, S. Liang, and Y. Wei. "Compositional human pose regression, " International Conference on Computer Vision, pages 2621–2630, 2017.

[11] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. "3d human pose estimation in the wild by adversarial learning," *Conference on Computer Vision and Pattern Recognition*, volume 1, 2018

[12] K. Lee, I. Lee, and S. Lee. "Propagating LSTM: 3d pose estimation based on joint interdependency," *European Conference on Computer Vision*, pages 119–135, 2018