Silhouette-based Synthetic Data Generation for 3D Human Pose Estimation with a Single Wrist-mounted 360° Camera

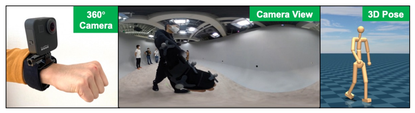
Ryosuke Hori*, Ryo Hachiuma* , Hideo Saito* , Mariko Isogawa†, Dan Mikami†

* Keio University, † NTT

Paper ID : 2554

Summary

Our Goal:
3D human pose estimation (HPE) with a single wrist-mounted camera.



Difficulties:

- High data preparation cost
  There is no existing dataset for a new camera setting. Existing works with body-mounted cameras solved it by generating synthetic dataset. The cost, however, to bridge domain gaps between real data and synthetic data remains high.
- Body parts are occluded
  Wearable camera-based 3D HPE is quite challenging as some human body parts are occluded from the camera's line of sight.

Proposed Method:
Main Idea: Dimensionality Reduction via Binary Silhouette image

Key Point 1: Low-cost synthetic training data generation

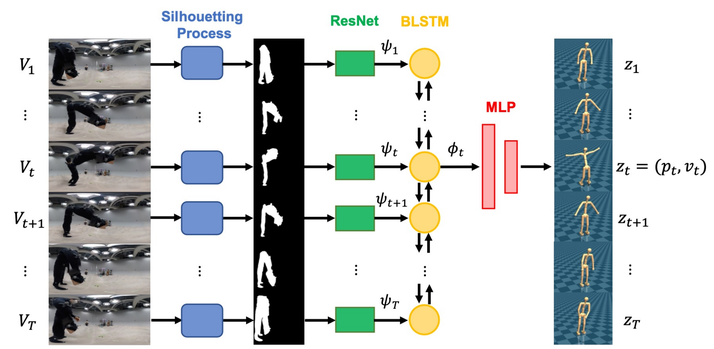
A silhouette-based 3D HPE reduces the data preparation cost.

The network is trained only with synthetic silhouette data generated at a lower cost than conventional methods.

Key Point 2 : 3D HPE from a 360° camera image sequence

Estimate 3D human poses from images captured by a single wrist-mounted 360° camera using a convolutional neural network-based framework following [1].

To bridge the domain gap, silhouetting process is applied to images captured in real-world for inference.



Contributions:

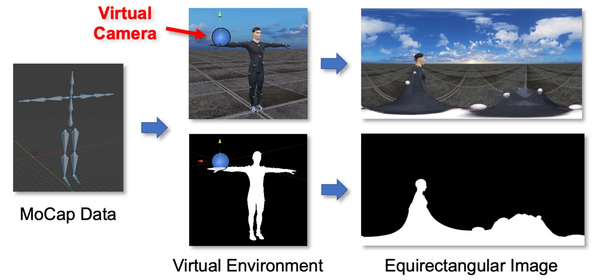
- A 3D human pose estimation framework given only a single wrist-mounted 360° camera.
- A silhouette-based synthetic data generation method, which enables us to bridge the domain gap and reduces the data preparation cost

Synthetic Training Data Generation

We generate silhouette equirectangular image sequences given only existing motion capture (MoCap) data to train the network.

How to generate the synthetic silhouette images:

1. Fix a virtual 360° camera at the avatar's wrist position in a virtual environment, such as Unity.
2. Set the avatar's body to white and the background to black.
3. Capture silhouette equirectangular images by making the virtual avatar move with the MoCap data.

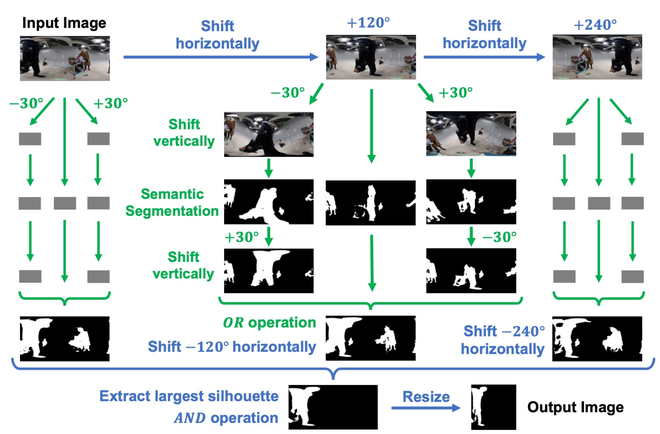


Silhouetting Process for Inference

We apply a silhouetting process to the equirectangular images captured by the wrist-mounted 360° camera for inference.

How to generate the silhouette image:

1. Shift the input image horizontally and vertically to extract the human silhouette accurately.
2. Apply semantic segmentation to each shifted image and extract the region labeled as human as a silhouette.
3. Merge the images obtained in the previous step and output a single binary image of a human silhouette.

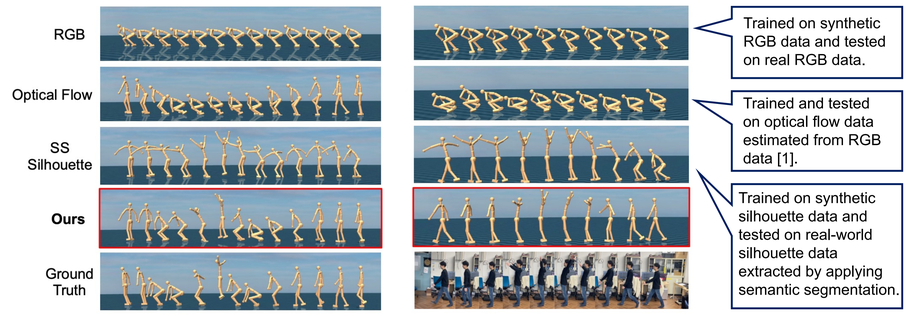


Experiments and Results

Experiments were conducted on the following datasets. The results show that our method outperforms other baseline methods.

MoCap Test Data: 360° camera images with MoCap data.

In-the-Wild Data: 360° camera images with 2D joint position data obtained from side view images to verify the effectiveness of our method in real-world environments.



MPJPE: Euclidean distance between the estimated and the ground-truth 3D poses.

$\mathbf{E}_{\text{key}}$: Euclidean distance between the estimated and the ground-truth 2D poses.

Method	MoCap Test Data (MPJPE)					In-the-Wild Data
	Walk	Jump	Crouch	Raise hand	All Frames	$\mathbf{E}_{\text{key}}$
RGB	0.346	0.311	0.284	0.407	0.339 ± 0.068	0.330 ± 0.074
Optical Flow	0.118	0.192	0.145	0.128	0.132 ± 0.070	0.352 ± 0.091
SS Silhouette	0.227	0.256	0.229	0.173	0.227 ± 0.057	0.275 ± 0.073
Ours	**0.106**	**0.147**	**0.138**	**0.106**	$\mathbf{0.115 \pm 0.053}$	$\mathbf{0.198 \pm 0.083}$

Conclusion

- Our pose estimation network is trained only on synthetic silhouette image data
- Silhouette-based approach reduces the data generation cost and bridges the domain gap between synthetic and real-world data.
- We achieved higher estimation accuracy quantitatively and qualitatively compared with other baseline methods.

Reference

[1] Y. Yuan and K. Kitani, "Ego-Pose Estimation and Fore-casting as Real-Time PD Control," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10081–10091.

Contact Information

- Ryosuke Hori : hori-rysk@keio.jp
- Hideo Saito : hs@keio.jp
- Mariko Isogawa : mariko.isogawa@ieee.org

Dataset available