HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar Supplementary Material

Shih-Po Lee[†,*], Niraj Prakash Kini[†], Wen-Hsiao Peng[†], Ching-Wen Ma[†], Jenq-Neng Hwang[‡]
[†]National Yang Ming Chiao Tung University, Taiwan
[‡]University of Washington, USA
{mapl0756051.cs07g, nirajnctu.cs06g, machingwen}@nctu.edu.tw, wpeng@cs.nctu.edu.tw, hwang@uw.edu

Table 1: Dataset Details

Settings	Value
Total Sequences	235
Total Duration	235 minutes
One Sequence Duration	1 minute
Total Triplets	141000
Triples in a Sequence	600
Subjects in Dataset	6
Number of Actions	3
Number of scenes	1
Training Sequences	193
Validation Sequences	21
Test Sequences	21
Subject Occluded	No
Single Subject in Sequence	Yes

1. Dataset Details

Table 1 describes the details of our dataset. We collect a total of 235 sequences. There are 6 human subjects and one of them only appears in test sequences. In any given sequence, there is only one person performing an specific action. Our dataset contains three actions which are standing with fixed postures, standing with waving hand(s), and walking with waving hand(s). Each triplet has 3 component: RGB camera frame, horizontal radar frame, and vertical radar frame. We form a triplet by a process of synchronization over timestamps.

We use an economical and easily available mmWave radar module and a RGB camera to acquire our dataset. Our camera captures the images of resolution 512 $\times$ 512 and we downscale all of them into 256 $\times$ 256 as the input for the image-based HPE network to generate the ground-truths. We use a pre-trained image-based 2D pose network, HRNet [2], to label the training and test sequences. The labels generated by HRNet [2] are almost as accurate as manually generated labels; i.e., the image-based 2D pose network achieves the AP of 99%. Our network shows very promising results on human pose estimation (HPE) task. To the best of our knowledge, Radar-based HPE has not been widely explored. We contribute our dataset to draw an attention of the community and look forward to seeing advancement.

1.1. Synchronization

In this section, we describe how we synchronize the camera with two radar sensors. Both radars can be triggered precisely with a digital sync signal provided by an external sync signal generator. We design a circuit that provides an accurate sync signal to trigger radar frame acquisition at precise timing, i.e., one frame is 100ms with our 10FPS setting to synchronize the FPS of our RGB camera. As both radars are operated at the same frequency band (77GHz-81GHz), we triggered them in an alternate switching mechanism to avoid interference, i.e., the horizontal radar is active only for the first 50ms of the frame, and the vertical radar is active for the next 50ms.

2. Experimental Details

Table 2 shows the accuracy of every keypoint. Our proposed method (VRDAEMap) still outperforms the traditional pre-processing method under a strcter evaluation metric AP^{75}. Our predicted keypoints, especially the fast-moving keypoints like wrists and elbows, achieving lower MPJPE than mmMesh [3]. However, the performance of torso keypoints such as head, neck, and shoulders is quite limited. The may because the radar signal is noisy, resulting in unstable predicted results. The pointcloud-based method, mmMesh, first performs denoising by converting the radar signal into point cloud, obtaining much lower MPJPE of head, neck, and shoulders (30.4, 23.3, and 31.7, respectively).

*This work is supported by National Center for High-performance Computing, Taiwan.

Table 2: Comparison of pre-processing methods. Total denotes the average precision over keypoints.

		AP								AP	AP^{50}	AP^{75}
Pre-processing	Model	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle		Total	
RAEMap	RF-Pose [4]	64.3	67.7	51.0	13.6	6.0	72.7	66.8	60.8	40.6	86.5	31.0
RAEMap	Ours	**80.6**	**84.6**	**75.1**	40.9	17.4	**86.9**	80.7	70.1	61.6	**98.6**	71.0
VRDAEMap	Ours	79.9	82.3	69.7	**45.6**	**23.5**	85.0	**81.9**	**72.5**	**64.3**	98.5	**76.7**

Table 3: Comparison of 3D keypoint performance based on MPJPE in millimeters. Ours + VideoPose3D means that we adopt our proposed method to generate 2D keypoints, which are lifted to 3D by VideoPose3D.

Model	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
mmMesh [4]	**30.4**	**23.3**	**31.7**	112.9	218.2	18.4	**33.6**	**57.4**	71.3
Ours + VideoPose3D [1]	71.4	43.2	44.8	**85.3**	**156.4**	**17.4**	41.6	73.9	**68.2**

References

[1] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[2] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. 2021.

[4] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.