EventHPE: Event-based 3D Human Pose and Shape Estimation Supplementary Material

1. Unsupervised Learning of Optical Flow

The FlowNet can be trained by unsupervised learning via warping loss of two sequential gray-scale images $(I_{t_{i-1}}, I_{t_i})$. The loss functions used to train the model, similar to [2], includes a photo-metric and a smoothness loss.

Given the predicted optical flow O_{t_i}, the warped image $\hat{I}_{t_i}$ can be obtained by warping the second image I_{t_i} to the first image $I_{t_{i-1}}$ via bilinear sampling. The photo-metric loss describes the difference between $I_{t_{i-1}}$ and $\hat{I}_{t_i}$,

$$\mathcal{L}_{\text{photo}}(u, v; I_{t_{i-1}}, I_{t_i}, O_{t_i}) = \sum_{x,y} \rho(I_{t_{i-1}}(x, y) - I_{t_i}(x + u(x, y), y + v(x, y))), \quad (1)$$

where ρ is the Charbonnier loss function defined as $\rho(x) = \sqrt{x^2 + \epsilon^2}$ and (u, v) is the 2D direction of the predicted flow O_{t_i}. The Charbonnier loss is more robust than the absolute difference. The smoothness loss constraints the output flow by minimizing the difference of each in-pixel flow and its neighboring flows,

$$\mathcal{L}_{\text{smooth}}(u, v; O_{t_i}) = \sum_{x,y} \sum_{i,j \in \mathcal{N}(x,y)} \rho(u(x, y) - u(i, j)) + \rho(v(x, y) - v(i, j)), \quad (2)$$

where $\mathcal{N}(x, y)$ is the neighbors of pixel (x, y).

To summarize, FlowNet is trained by minimizing the loss

$$\mathcal{L}_{\text{optical-flow}} = \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{photo}}. \quad (3)$$

2. MMHPSD Dataset Details

Data Acquisition. In data acquisition stage, our multi-camera acquisition system has 12 cameras out of 4 different types of imaging modality: one event camera, one polarization camera, five-view RGB-D cameras. *All the frame-based cameras are soft-synchronized with the gray-scale images of the event camera* and the events between two sequential gray-scale images are collected synchronously. 15 subjects are recruited for the data acquisition, where 11 are male and 4 are female. Each subject is required to perform 3 groups of actions (21 different actions in total, as is shown in Fig. 1) for 4 times, where each group includes actions of fast/medium/slow speed respectively.

Finally, we collect 12 short videos for each subject and each video has around 1,300 frames with 15 FPS, that is 180 videos in total with each video lasting about 1.5 minutes. We conduct the annotations for each video and check manually whether the annotated shape aligns well with multi-view images. We abandon the unsatisfactory annotated shapes. Details on the number of frames per subject and number of annotated frames per subject are presented in Tab. 2. The event camera used is CeleX-V [1] with resolution 1280x800 and the sensor frequency is 20-70 MHz. The MIPI interface supports up to 2.4Gbps transfer rate while the parallel interface supports the maximum readout of 140M pixels/second. The average number of events for the dataset is around 1 million per second. Fig. 1 presents the layout of our multi-camera system and three annotated shapes as examples. Overall, our dataset consists of 240k frames with each frame including a gray-scale image and inter-frame events, a polarization image, five-view color and depth images.

References

[1] Shoushun Chen and Menghan Guo. Live demonstration: Celex-v: A 1m pixel multi-mode event-based sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2019. 1

[2] Alex Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. 1

group	speed	actions
1	medium	jumping, jogging, waving hands, kicking legs, walk
2	fast	boxing, javelin, fast running, shooting basketball, kicking football, playing tennis, playing badminton
3	slow	warming up elbow/wrist ankle/pectoral, lifting down-bell, squating down, drinking water

Table 1. Types of actions in each group. Subjects are required to do each group of actions for 4 times. The order of actions each time is random.

subject	gender	# of original frames	# of annotated frames	# of discarded frames
1	male	15911	15911	0 (0.0%)
2	male	15803	15803	0 (0.0%)
3	male	16071	16071	0 (0.0%)
4	male	16168	16152	16 (0.01%)
5	male	16278	16262	16 (0.01%)
6	male	16715	16384	331 (2.0%)
7	female	16091	16091	0 (0.0%)
8	male	16257	15642	715 (4.4%)
9	male	15467	15461	6 (0.03%)
10	male	16655	16655	0 (0.0%)
11	male	16464	16443	21 (0.13%)
12	male	16186	16186	0 (0.0%)
13	female	16064	14562	1502 (9.4%)
14	female	15726	15166	560 (3.6%)
15	female	14193	14075	118 (0.8%)
total	-	240049	236764	3285 (1.4%)

Table 2. Detail number of frames for each subject and the number of frames that have annotated SMPL pose and shape.

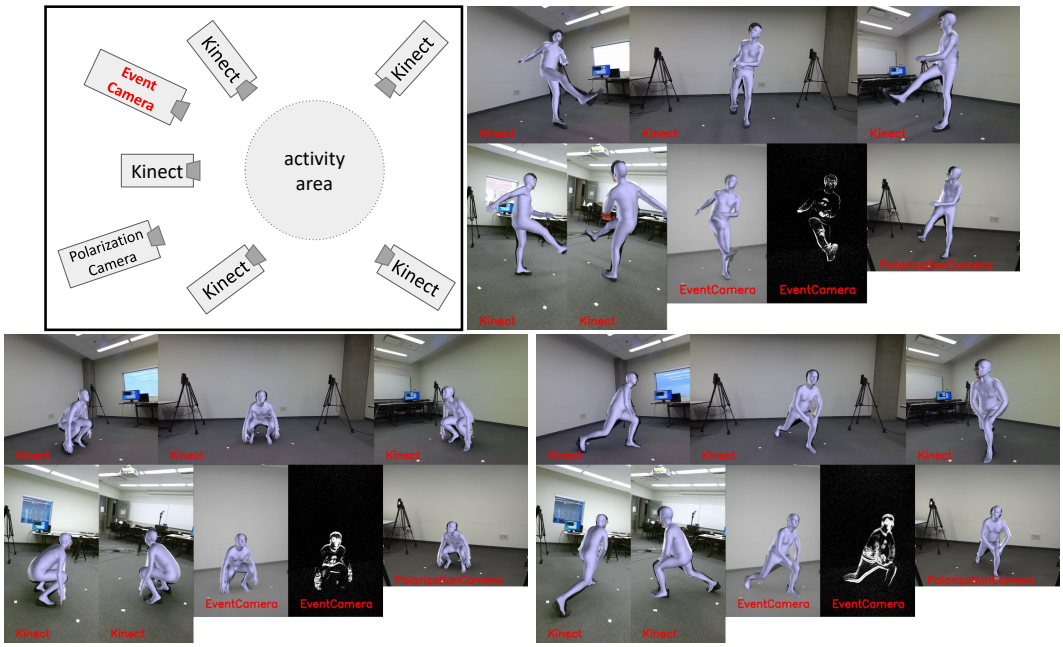

Figure 1. Layout of multi-camera acquisition system and three examples of annotated shapes rendered on multi-view images. The top left figure is the layout, and the other three figure present three examples of pose and shape annotation.