EventCap: Monocular 3D Capture of High-Speed Human Motions using an Event Camera (Supplementary)

Anonymous CVPR submission

Paper ID 1573

	Duration (s)	Intensity frame rate (*fps*)	No. Polarity events	Reference frame rate (*fps*)	No. Reference images
wave	14.20	12	3,187,382	100	1420
ninja	5.84	7	4,267,810	250	1460
javelin	1.78	20	819,647	500	890
boxing	2.60	25	570,345	500	1300
karate	2.00	25	589,437	1000	2000
dancing	1.72	25	684,200	1000	1720
shake	8.10	15	1,720,861	No	No
run_1	3.60	15	1,258,166	No	No
punch	6.40	15	685,477	No	No
throw	5.20	15	875,967	No	No
jump	2.80	15	1,145,612	No	No
run_2	3.70	15	1,375,098	No	No

Table 1: Statistics and basic metrics of the EventCap dataset.

1. Dataset Detail

Our EventCap dataset consists of 12 sequences of 6 actors performing challenging fast non-linear motions. The basic statistics related to both the event camera and the reference camera for each sequence are reported in Table. 1.

For 6 sequences in our dataset, we provide high resolution reference images captured at high frame rate. The reference images of the "wave" sequence are captured using one camera from the multi-view markerless motion capture system [1] at 100 *fps*, which provides accurate 3D motions of the actors for quantitative evaluation. The reference images of the "ninja", "javelin", "boxing", "karate" and "dancing" sequences are captured using a Sony RX0 camera at high frame rates ranging from 250 to 1000 *fps* with various lighting conditions for sufficient evaluation. Furthermore, the "ninja" sequence provides an extremely challenging case, which captures an actor in black ninja suite outdoor at night. Note that due to the inherent limitation of the on-chip memory, the Sony RX0 camera can only record about 4 seconds when the capturing frame rate is set to be 500 or 1000 *fps*. Nevertheless, even in such a short capture duration, our dataset successfully provides various challenging fast motions with reference view for qualitative analysis.

Moreover, our dataset provides 6 additional sequences with longer capture duration and various challenging motions, including "shake", "run_1", "punch", "throw", "jump" and "run_2". For fair evaluation, the frame rates of the intensity image stream for all these 6 sequences are set to be the same (15 *fps*). In such setting, the longer exposure time of the intensity images intensifies the motion blur caused by fast non-linear motions of the actors, making our dataset more challenging.

2. More Results

Qualitative Results. Recall that in the Fig. 5 of the main manuscript, we provided the qualitative results of the 6 sequences with reference. Note that for each sequence, we evenly slice the time duration between two adjacent low frame rate intensity images to enable 1000 *fps* capture. For those sequences with reference views, we further interpolate the 1000 *fps* tracking motions into the reference frame rate, so as to provide qualitative evaluation according to the reference images. The qualitative results of the other sequences without reference are provided in Fig. 1, which demonstrate the effectiveness of our method to accurately capture the high frequency motion details, even though the intensity images from the event camera suffers from severe motion blur.

Quantitative Results. Here we provide more numerical details for the comparison between our EventCap and the baseline methods. Recall that *Mono_all* and *HMR_all* denote applying MonoPerfCap [3] and HMR [2] on all the reconstructed latent images, respectively. *Mono_linear* and *HMR_linear* denote applying the baselines only on the raw intensity images, followed by linearly upsampling operation. *Mono_refer* and *HMR_refer* denote applying the baselines to the high frame rate reference images directly. Note that for fair comparison, we downsample the reference images into the same resolution of the intensity images from event camera.

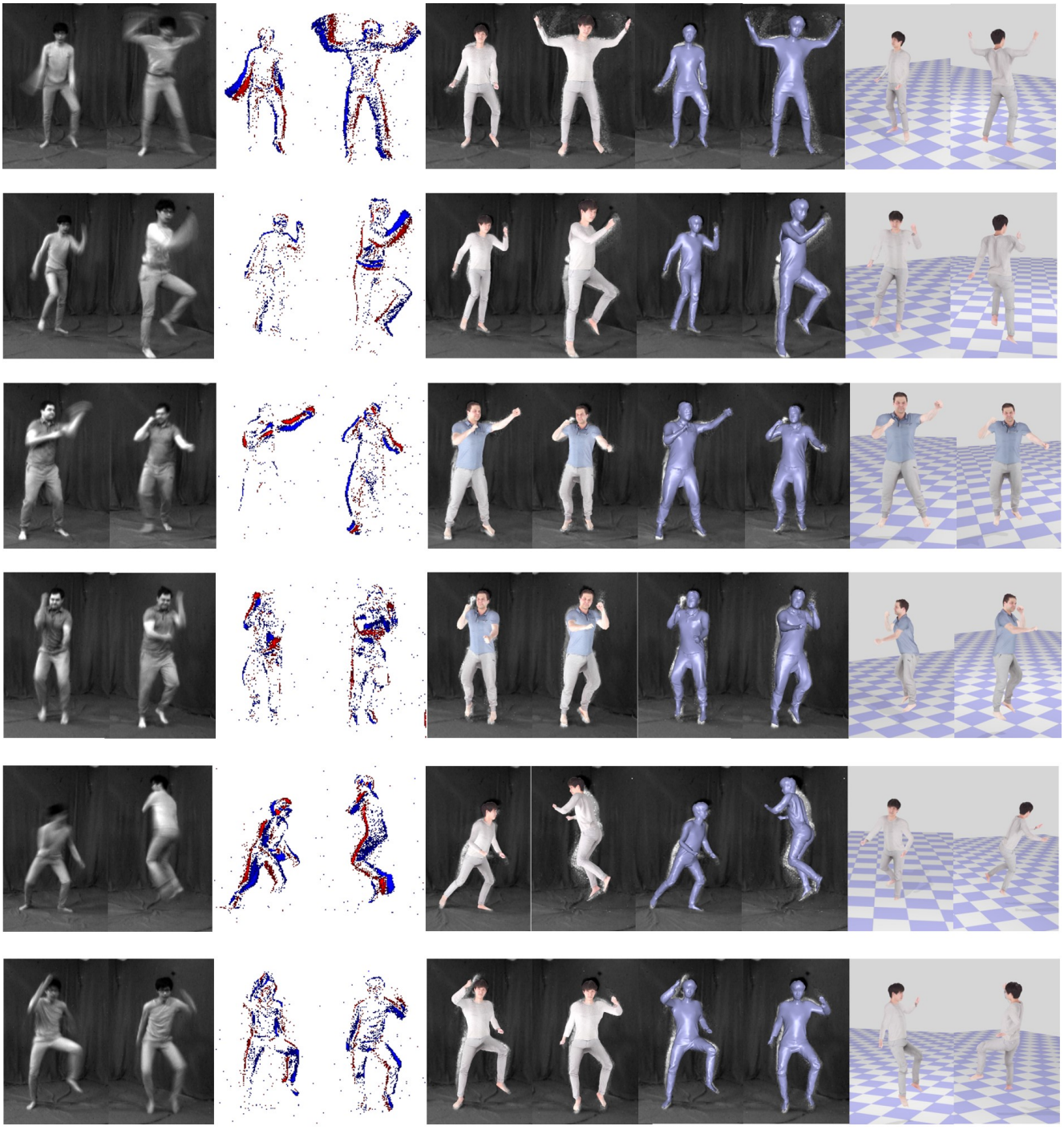

Figure 1: More qualitative results of EventCap on some sequences from our benchmark dataset. From top to down, the results correspond to the following sequences:"shake", "run_1","punch", "throw", "jump" and "run_2". (a,b) The intensity images; (c,d) Polarity events accumulated between the time duration from the previous to the current tracking frames; (e,f) Textured motion capture results overlaid on the reconstructed latent images; (g,h) Geometric motion capture results overlaid on the reconstructed latent images; (i,j) Results rendered in 3D views.

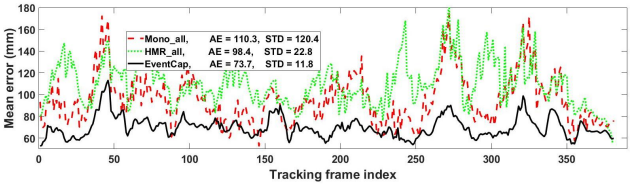

Figure 2: Comparison to *Mono_all* and *HMR_all* in terms of the average per-joint 3D error. Our method consistently achieves the lowest error.

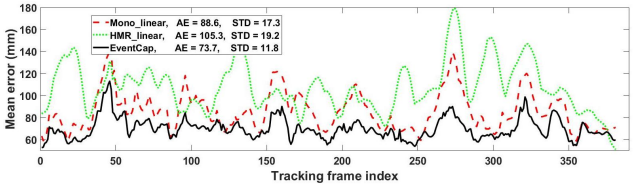

Figure 3: Comparison to *Mono_linear* and *HMR_linear* in terms of the average per-joint 3D error. Our method achieves the lowest error.

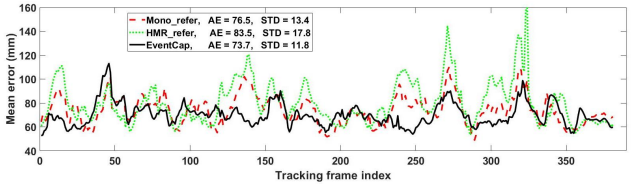

Figure 4: Comparison to *Mono_refer* and *HMR_refer* in terms of the average per-joint 3D error. Our method achieves the lowest error.

All the numerical curves in terms of average per-joint error (AE) compared to the baselines above are reports in Fig. 2, 3 and 4, respectively. When sharing the same input from the event camera, our method outperforms the other baselines and accurately captures the high frequency temporal motion details. In addition, our method achieves similar tracking accuracy compared to *Mono_refer* and consistently outperforms *HMR_refer*. Recall that our method relies upon only 3.4% of the data bandwidth of the reference image-based methods, and even achieves better tracking accuracy.

For further evaluation, we apply MonoPerfCap [3] and HMR [2] to the raw reference images (both high frame rate and high resolution), denoted as *Mono_large* and *HMR_large*, respectively. Not surprisingly, the AE of *Mono_large* and *HMR_large* reach 62.3 and 75.1, respectively. Even under such unfair comparison, our method achieves similar tracking accuracy compared to *HMR_large*, with only 0.45% data bandwidth of *Mono_large* and *HMR_large*.

References

[1] The Captury. http://www.thecaptury.com/. 1

[2] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 3

[3] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37(2):27:1–27:15, 2018. 1, 3