Exploring the Impact of Rendering Method and Motion Quality on Model Performance when Using Multi-view Synthetic Data for Action Recognition (Supplementary Material)

Stanislav Panev[*1], Emily Kim[*1], Sai Abhishek Si Namburu[1], Desislava Nikolova[2], Celso de Melo[3], Fernando De la Torre[1], Jessica Hodgins[1]

[1]Carnegie Mellon University, [2]Technical University of Sofia, [3]Army Research Laboratory

{spanev, ekim2}@andrew.cmu.edu, snamburu@alumni.cmu.edu, dnikolova@tu-sofia.bg, celso.m.demelo.civ@army.mil, {ftorre, jkh}@cs.cmu.edu

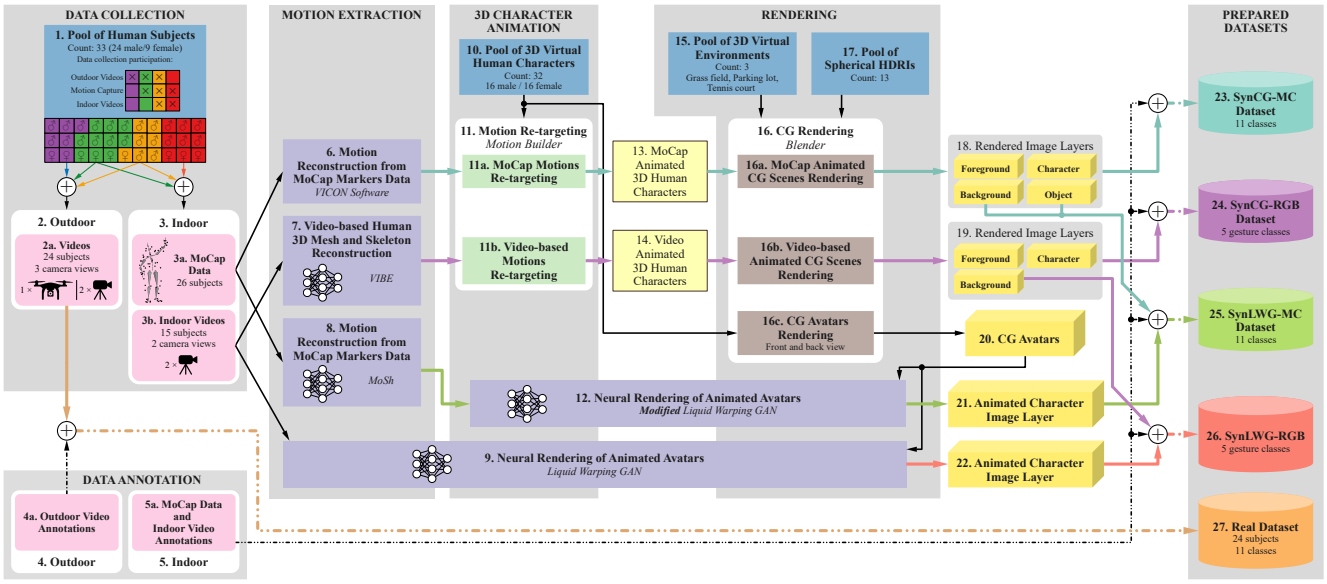

Figure 1. The dataset generation workflow diagram.

1. Datasets generation workflow

Fig. 1 depicts our data generation workflow. It comprises six major stages: *Data collection*, *Data annotation*, *Motion extraction*, *3D character animation*, *Rendering*, and the outcome—the *Prepared datasets*. We explain each of these stages in this section.

Data collection *Block 1* from Fig. 1 represents the pool of human subjects involved in the data collection. In total, 33 subjects (24 male and 9 female) participated in three different data collection procedures—outdoor video recording, indoor motion capture, and simultaneous video recording. Each subject is depicted by a color-coded square box with their gender sign. Four groups of subjects were formed based on which procedures they participated in: the purple boxes represent the participants who were only involved in the outdoor video data collection, the green boxes represent the subjects who participated in outdoor and mocap data collection, the orange boxes represent those who participated in all data collections, and the red boxes represent the participants who took part in the mocap and indoor video data collection.

In total, 24 subjects participated in the outdoor video data collection. They were recorded with three cameras—one orbiting UAV and two ground static cameras. In total, 26 subjects participated in the motion capture data collec-

[*]Equal contribution to this work.

tion. Fifteen were also recorded with two static video cameras, constituting the indoor video data collection.

Data annotation We manually annotated the temporal segments of all outdoor videos and the motion capture sequences corresponding to the eleven activity classes of interest. Because the indoor videos were recorded during the motion capture data collection, they share the same annotations.

Motion extraction We use two different sources of motion information—motion capture data and motions extracted from videos. The raw motion capture 3D marker locations are processed in two different ways depending on the type of output synthetic dataset. For the *SynCG-MC* dataset, the raw marker data were processed by the *VICON motion capture system* software (*Block 6*), which calculates the motions of each subject and applies them to a skeleton. For the *SynLWG-MC* dataset, the raw marker data were fed into *MoSh* [5]—a model that fits an SMPL parametric model [4] to the locations of the markers' point cloud (*Block 8*). *SynCG-RGB* motions were extracted from the indoor videos using VIBE [1], whereas for the *SynLWG-RGB* dataset the motions were extracted by SPIN [2], which is integrated into the original *Liquid Warping GAN* (LWG) [3] pipeline.

3D character animation Animating the 3D human characters for *SynCG-MC* and *SynLWG-RGB* is performed by Autodesk's MotionBuilder (*Block 11*). This process is done within the models for *SynLWG-MC* and *SynLWG-RGB* (Blocks 9 and 12).

Rendering *Blender* is used to render all image layers (the character, the object held, the foreground, and the background) of *SynCG-MC* and *SynCG-RGB* datasets. They are composited at the post-processing stage to form the final video frames. Fig. 2 illustrates the image layers compositing process for different class groups and camera views. Since the five gesture classes do not include object handling, no object layer is rendered. Moreover, no foreground layer is rendered for the aerial camera view due to the camera's position. Utilizing the layered rendering approach sped up the computationally-heavy rendering process by reusing the background, foreground, and object layers for the other synthetic data variants, as seen in the diagram (Fig. 1). For *SynLWG-MC* and *SynLWG-RGB* datasets, we used the animated character layer frames generated by *LWG* and composited them with the background and object layers from *SynCG-MC* and *SynCG-RGB*, respectively. To render the input character avatars for LWG, we used *Blender*.

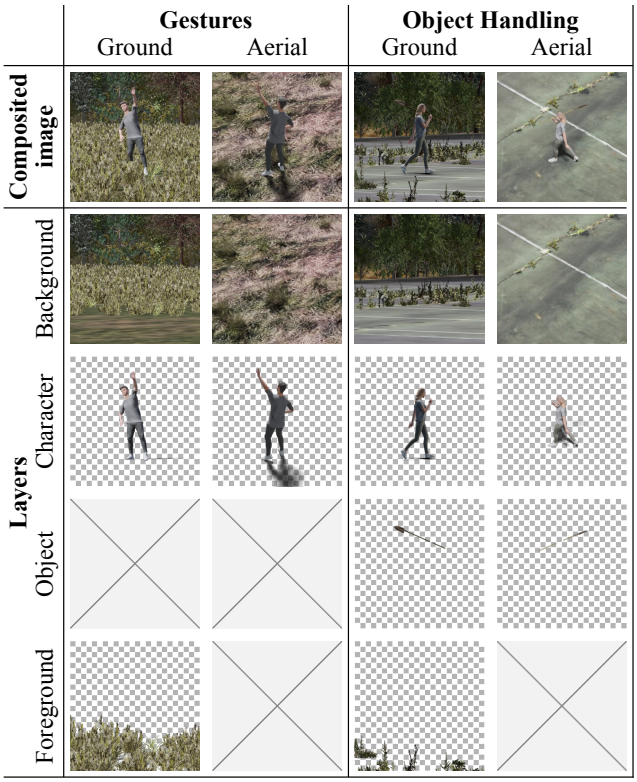

Figure 2. Synthetic CG data image layers compositing.

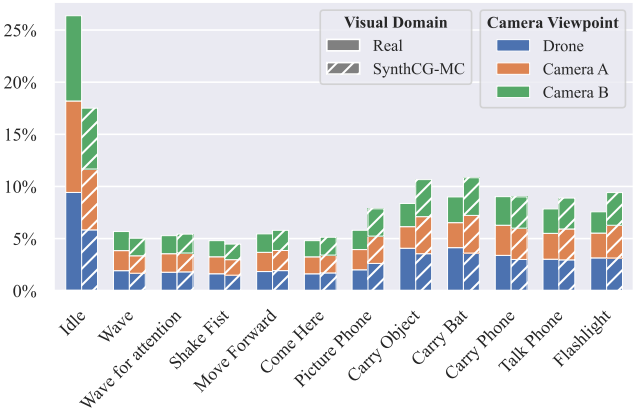

Figure 3. Activity-wise data samples distributions.

Prepared datasets After all image layers were rendered and composited, we paired them with their corresponding annotations. Thus, the five datasets (one real and four synthetic) were created.

2. Datasets samples distribution

Fig. 3 illustrated the class-wise distribution of the samples of the *Real* and *SynCG-MC* datasets. Overall, the samples are relatively uniformly distributed across all classes of

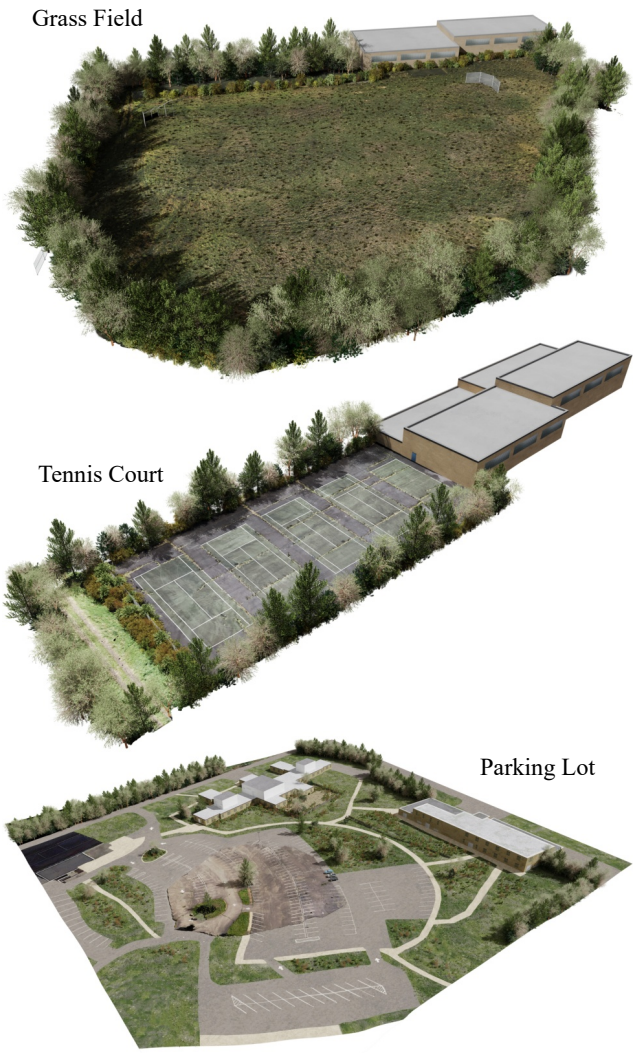

Figure 4. The three synthetic environments created for CG data generation.

interest except for the *Idle* class, which is expected because it contains all the frames that do not belong to any other category. Furthermore, the envelope of the distributions of the real and synthetic datasets have the same character, which is expected as we tried to create the synthetic dataset as close as possible to the real one.

We only present *SynCG-MC* dataset samples distribution because *SynLWG-MC* contains the same type of samples since both are based on motion capture data. The other two video-motion-based synthetic datasets are based on videos, recorded during the motion capture dataset collection and thus their samples are distributed the same way.

3. Synthetic data virtual environments

With the help of 3D artists, we designed three virtual environments (Fig. 4) that were incorporated into our synthetic generation pipeline. They closely resemble the three real-world environments where the outdoor videos were recorded. The ground planes, for example, were recreated by recording aerial videos of the area and incorporating photogrammetry to export the geometry and the textures. Subsequently, various 3D assets were scattered across the scenes, such as vegetation and buildings.

References

[1] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pages 5252–5262, Piscataway, NJ, June 2020. IEEE. 2

[2] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[3] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2

[5] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, Nov. 2014. 2