Data Generation Of Learning from Synthetic Humans

Cheng Guan

June 12, 2018

1 Data generation

This section presents our SURREAL (Synthetic humans for real tasks) dataset and describes key steps for its generation . We also describe how we obtain ground truth data for real MoCap sequences .

1.1 Synthetic Humans

Our pipeline for generating synthetic data is illustrated in Figure 1. A human body with a random 3D pose, random shape and random texture is rendered from a random viewpoint for some random lighting and a random background image. Below we define what random means in all these cases. Since the data is synthetic, we also generate ground truth depth maps, optical flow, surface normals, human part segmentations and joint locations (both 2D and 3D). As a result, we obtain 6.5 million frames grouped into 67582 continuous image sequences. See Table 1.1 for more statistics.

Table 1: SURREAL dataset in numbers. Each MoCap sequence is rendered 3 times (with 3 different overlap ratios). Clips are mostly 100 frames long. We obtain a total of 6,5 million frames

Topic	Subjects	Sequences	Clips	Frames
Train	115	1964	55001	5342090
Test	30	703	12528	1194662
Total	145	2607	67582	**6536752**

Body model. Synthetic bodies are created using the SMPL body model [3]. SMPL is a realistic articulated model of the body created from thousands of high-quality 3D scans, which decomposes body deformations into pose (kinematic deformations due to skeletal posture) and shape (body deformations intrinsic to a particular person that make them different from others). SMPL is compatible with most animation packages like Blender.

Body shape. In order to render varied, but realistic, body shapes we make use of the CAESAR dataset , which was used to train SMPL. To create a body shape, we select one of the CAESAR subjects at random and approximate their shape with the first 10 SMPL shape principal componets.Ten shape components explain more than 98% of the shape variance in CAESAR (at the resolution of our mesh) and produce quite realistic body shapes.

Body pose. To generate images of people in realistic poses, we take motion capture data from the CMU MoCap database . CMU MoCap contains more than 2000 sequences of 23 high-level action categories, resulting in more than 10 hours of recorded 3D locations of body markers.

Human texture. We use two types of real scans for the texture of body models. First, we extract SMPL texture maps from CAESAR scans, which come with a color texture per 3D point. These maps vary in skin color and person identities, however, their quality is often low due to the low resolution, uniform tight-fitting clothing, and visible markers placed on the face and the body. Anthropometric markers are automatically removed from the texture images and inpainted. To provide more variety, we extract a second set of textures obtained from 3D scans of subjects with normal clothing.

Light. The body is illuminated using Spherical Harmonics with 9 coefficients . The coefficients are randomly sampled from a uniform distribution between -0.7 and 0.7, apart from the ambient illumination coefficient (which has a minimum value of 0.5) and the vertical illumination component, which is biased to encourage the illumination from above. Since Blender does not provide Spherical Harmonics illumination, a spherical harmonic shader for the body material was implemented in Open Shading Language.

1.2 Generating ground truth for real human data

Human3.6M dataset [1, 2] provides ground truth for 2D and 3D human poses. We complement this ground truth and generate predicted body-part segmentation and depth maps for people in Human3.6M. Here again we use MoSh to fit the SMPL body shape and pose to the raw MoCap marker data. This provides a good fit of the model to the shape and the pose of real bodies. Given the provided camera calibration, we project models to images. We then render the ground truth segmentation, depth, and 2D/3D joints as above, while ensuring correspondence with real pixel values in the dataset. As MoSh provides almost perfect fits of the model, we consider this data to be ground truth. We use this ground truth for the baseline where we train only on real data, and also for fine-tuning our models pre-trained on synthetic data. In the rest of the paper, all frames from the synthetic training set are used for synthetic pre-training.

References

[1] C. Ionescu, F. Li, and C. Sminchisescu. Latent structured models for human pose estimation. In *ICCV*, 2011. 1

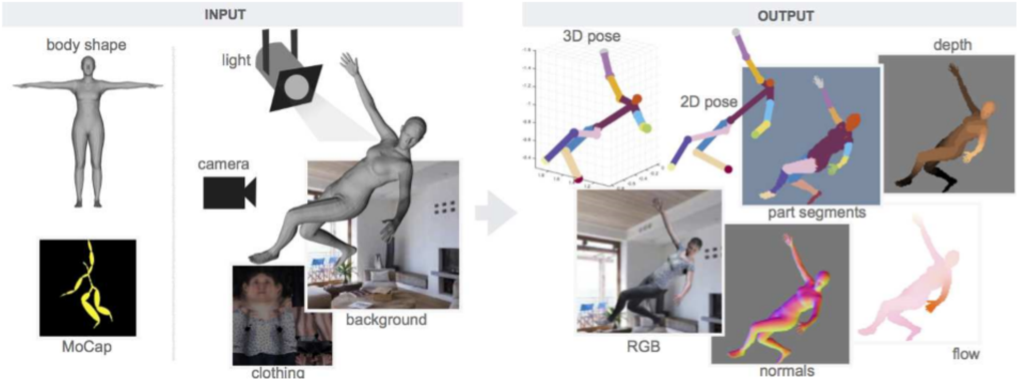

Figure 1: Our pipeline for generating synthetic data. A 3D human body model is posed using motion capture data and a frame is rendered using a background image, a texture map on the body, lighting and a camera position. These ingredients are randomly sampled to increase the diversity of the data. We generate RGB images together with 2D/3D poses, surface normals, optical flow, depth images, and body-part segmentation maps for rendered people.

[2] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 2014. 1

[3] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 2015. 1