VIBE: Video Inference for Human Body Pose and Shape Estimation

Laleh Samadfam
s6lasama@uni-bonn.de

Institute for Informatics,
University of Bonn

3D Human body shape and pose estimation is a classic computer vision problem and it is fundamental to many applications like robotics and behavior understanding. This problem has proceeded in various representations and methods. Some works use generic 3d object representations such as voxel-based representations, while more recent works use SMPL [3] which represents body as a function of shape and pose parameters. From another point of view, some works use a single image to infer a 3D representation for human body, while others use a video input to also exploit motion cues for the same purpose.
For instance, Bodynet [6] applies volumetric shape estimation networks on a single RGB image to estimate a human body in a voxel-based representation. HMR[2] uses regression network on a single image to estimate SMPL parameters and power up their regression using a discriminator trained with a pool of real 3D human meshes. Sim2real[1] extends HMR to video input. It uses LSTM units for regression, and a synthetic 3d annotated dataset to discriminate real and fake human body.

In this paper, we use a similar approach to HMR and Sim2real. We train an Encoder together with a discriminator to estimate sequences of 3D body shapes and poses in SMPL model format, from in-the-wild videos. Our encoder is also supervised by 2D keypoint annotations. Figure 1 shows an abstract model of our method. Superiority of our work to state of the art methods comes from 2 key contributions. 1- We use a self attention mechanism in discriminator to weight the frames by their importance in the final representation and show how it enhances the final result. 2- We train our discriminator using AMASS dataset[4], a large scale Motion capture dataset represented by SMPL model to get more accurate and real-like results at the end.

Our **Temporal Encoder** consists of a spatial CNN followed by a bidirectional GRU layer and, topped with regression networks. The CNN takes a video of a fixed length T as input and extracts it's per-frame features $f_1, f_2, ..., f_T$. Then, the GRU layer estimates latent variables $g_1, g_2, ..., g_T$ containing information incorporated from past and future frames. These variables are then sent to T regressors with iterative feedback which are in charge of estimating the camera, shape and pose parameters Θ. Regressors are initialized with mean shape $\hat{\Theta}$.
The loss function of Temporal Encoder is calculated by equation 1 and includes regression loss from 2D, 3D, and SMPL shape and pose parameters β and θ when they are available, and an adversarial loss from the discriminator.

$$L_E = L_{2D} + L_{3D} + L_{SMPL} + L_{adv} \tag{1}$$

Each of the regression loss functions are calculated as follows. Having 3D parameters Θ from the regressed model, 3D joint locations $\hat{X}$ are estimated from body vertices using a pre-trained regressor. 2D predictions for joint locations $\hat{x}$ are computed by projecting the joint locations of the 3D model using a weak-perspective camera with scale and transition s, t. R is showing the global rotation, and Π is representing orthographic projection.

$$L_{3D} = \sum_{t=1}^{T} ||X - \hat{X}||_2 \tag{2}$$

$$L_{2D} = \sum_{t=1}^{T} ||x - \hat{x}||_2, \;\; \hat{x} = s\Pi(R\hat{X}(\Theta)) + t \tag{3}$$

$$L_{SMPL} = ||\beta - \hat{\beta}||_2 + \sum_{t=1}^{T} ||\theta - \hat{\theta}||_2 \tag{4}$$

The **Motion Discriminator** is composed of a multi-layer GRU network which estimates a latent variable h_i for each frame. Then. a self-attention Network composed of linear MLP layers ϕ learns a linear combination of the latent variables as the final representation r. Finally, a linear layer produces a value $\in [0,1]$ deciding the probability of r belonging to a plausible human motion.
The last hidden layer in a recurrent netowrk is already a summary of all the hidden layers from last time steps, but this summary is weighted by time steps, and the last layers have more dominant effect on it. Self-attention mechanism ensures the final representation r is weighted by learnt weights that account for the importance of frames in the final representation. The weights are later normalized by soft-max to form a probability distribution.

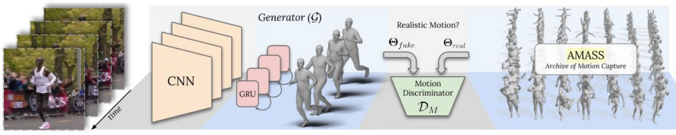

Figure 1: VIBE estimates SMPL parameters from a frame sequence using a temporal encoder trained alongside with a discriminator. The discriminator has access to a large corpus of human motions in SMPL format.

$$r = \sum_{i=1}^{N} a_i h_i, \;\; a_i = \frac{e^{\phi(h_i)}}{\sum_{t=1}^{N} e^{\phi(h_t)}} \tag{5}$$

The motion discriminator is trained to minimizes the objective function 6. Inputs parameters sampled from AMASS dataset P_R are labeled as real, and data generated by the temporal encoder is labeld as fake input.

$$L_{D_M} = E_{\Theta \sim P_R}[(D_M(\Theta) - 1)^2] + E_{\Theta \sim P_G}[(D_M(\hat{\Theta}))^2] \tag{6}$$

The adversarial loss function for the Encoder is calculated by equation 7.

$$L_{adv} = E_{\Theta \sim P_G}[(D_M(\hat{\Theta}) - 1)^2] \tag{7}$$

This loss ensures that parameters estimated by the Temporal encoder belong to the manifold of plausible human motion. We compare our results with the state-of-the-art methods and outperform all of them in considered datasets. Comparison of the results with some of SOTA methods is shown in Table 1.

	3DPW	H36M
Kanazawa *et al.[2]*	76.7	56.8
Doersch *et al.[1]*	74.7	-
Sun *et al.[5]*	69.5	42.4
VIBE	**51.9**	**41.4**

Table 1: Comparison of VIBE and SOTA methods on 3DPW and H36M datasets reported by Procrustes-aligned mean per joint position error.

In this paper we discuss VIBE in better details and give implementation technicalities. We also learn a new motion prior from AMASS and show it also helps training but is less powerful than the discriminator. In addition, we perform ablation experiments to examine the enhancement achieved by each part of our proposed model.

[1] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In *NeuRIPS*, 2019.

[2] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of CVPR*, 2018.

[3] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6), 2015.

[4] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of ICCV*, October 2019.

[5] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of IEEE ICCV*, 2019.

[6] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of ECCV*, 2018.