Exploiting temporal context for 3D human pose estimation in the wild

Anurag Arnab*, Carl Doersch*, Andrew Zisserman

1. Introduction

- Monocular 3D human pose estimation is an inherently ill-posed problem.
- Furthermore, metric ground-truth for "in-the-wild" data very difficult to obtain.
- Temporal consistency from video provides crucial information (Fig. 1) and is discarded by most methods.

Our contributions:

- Propose a form of *bundle adjustment* to encourage temporal consistency throughout whole video.
  - This achieves state-of-the-art performance on Human 3.6M
- Apply our method to about 107 000 YouTube videos in the Kinetics dataset, and automatically create a new "in-the-wild" dataset, which we publicly release.
- Substantially improve a per-frame model using our new dataset.

2. Bundle adjustment

- Allows us to take temporal information, and multi-view geometry, from the whole video into account.
- Assume an orthographic projection, Π, and camera parameters $\mathbf{\Omega}^t = \{s^t, u^t\}$.
- Use the SMPL human body model
  - $\beta \in \mathbb{R}^{10}$ shape parameters, $\theta^t \in \mathbb{R}^{23\times 3}$ pose parameters
  - Shape parameters are constant for all frames in the video.
  - 3D joints, $\mathbf{X}^t = \mathrm{SMPL}(\beta, \theta^t)$. 2D joints, $\mathbf{x}^t = s^t \Pi(R\mathbf{X}^t) + u^t$.
- Objective function has reprojection, temporal and prior terms:

$$E(\beta, \theta, \mathbf{\Omega}) = E_R(\beta, \theta, \mathbf{\Omega}) + E_T(\beta, \theta, \mathbf{\Omega}) + E_P(\theta, \beta)$$

- Use per-frame HMR model [1] to initialise, and 2D keypoints from [2].
- Solve with L-BFGS.

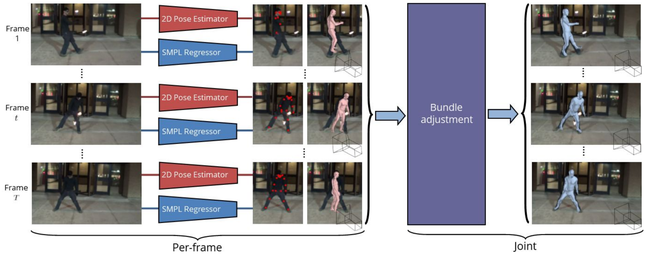

Figure 2. Using initial per-frame estimates of 2D keypoints, SMPL- and camera parameters, we jointly optimise over the whole video to encourage temporal consistency.

Figure 1. Our proposed bundle adjustment method exploits temporal context to prevent major failures (columns 2 and 3) and to resolve ambiguities (column 5) in real-world, YouTube data.

3. Objective function

- Reprojection error encourages 3D keypoints to reproject onto predicted 2D keypoints

$$E_R(\beta, \theta, \mathbf{\Omega}) = \lambda_R \sum_t^T \sum_i^J w_i \rho(\mathbf{x}_i^t - \mathbf{x}_{det,i}^t).$$

- Temporal error encourages smooth motions of 3D joints, $\mathbf{X}$, 2D joints, $\mathbf{x}$, and camera parameters, $\mathbf{\Omega}$, that are typical of videos.

$$E_T(\beta, \theta, \mathbf{\Omega}) = \sum_{t=2}^T \sum_{i=1}^J \lambda_1 \rho(\mathbf{X}_i^t - \mathbf{X}_i^{t-1}) + \lambda_2 \rho(\mathbf{x}_i^t - \mathbf{x}_i^{t-1}) + \lambda_3 \rho(\mathbf{\Omega}^t - \mathbf{\Omega}^{t-1})$$

- Finally, we include a prior term as there are many 3D poses (including some that are not humanly possible) that project correctly onto 2D keypoints and vary slowly through time. The first term of the prior encourages our result to stay close to the HMR initialisation, and the second is the commonly used GMM joint angle prior of [3].

4. Scaling up to Kinetics

- The "in-the-wild" videos from Kinetics cause frequent failures in our initialisation (Fig 1,2,3).
- To handle multiple people, and be more robust to outliers, we modify the reprojection error to:

$$E_R(\beta, \theta^t, \mathbf{\Omega}^t) = \min\left(\min_{p \in P^t} \textstyle\sum_i^J w_i h(\mathbf{x}_i^t - \mathbf{x}_{det,i}^{t,p}), \tau_R\right)$$

  - The "inner min" means that the loss is with respect to the best matching 2D pose.
  - The "outer min" means that if our estimate is too far from the predicted 2D pose, we consider it an outlier and pay a constant penalty.
- We modify the prior term in a similar manner to ignore outliers in the HMR initialisation.

5. Experimental results

- We first apply bundle adjustment to HMR initialisation on Human 3.6M, where we achieve state-of-the-art results among methods using SMPL (Tables 1 and 3).
- Applying our method to 106 589 YouTube videos from Kinetics, we obtain 16 720 videos after thresholding the normalised loss to ignore failures and trivial examples.
  - Of these 4.1M frames, 3.4M are considered inliers with respect to 2D keypoints from [2].
  - These 3.4M frames form our new, public dataset.
- Training with our new dataset improves performance on 3DPW ("in-the-wild" videos) and HumanEVA (mocap) in Table 2.

Table 1. Ablation study on Human 3.6M. All variants of our method consider the whole video.

Method	MPJPE (mm)	PA-MPJPE (mm)
HMR initialisation (per-frame)	85.8	57.5
E_R	154.3	99.7
$E_R + E_P$	79.6	55.3
$E_R + E_P + E_T$	77.8	54.3
Ground truth keypoints		
E_R	89.2	64.5
$E_R + E_P$	66.5	45.7
$E_R + E_P + E_T$	63.3	41.6

Table 2. Improvement on 3DPW and HumanEVA when training HMR (per-frame model) with our automatically generated dataset.

Dataset	Original data	Original + Kinetics 300K	Original + Kinetics 3M
3DPW	77.2	73.8	**72.2**
HumanEVA	85.7	83.5	**82.1**

Table 3. Comparison to other approaches fitting the SMPL model on Human 3.6M. No additional Kinetics data is used. Only our method considers the whole video.

Method	MPJPE (mm)	PA-MPJPE (mm)
SMPLify	--	82.3
Pavlakos CVPR '18	--	75.9
NBF	--	59.9
HMR	88.0	56.8
Ours	**77.8**	**54.3**

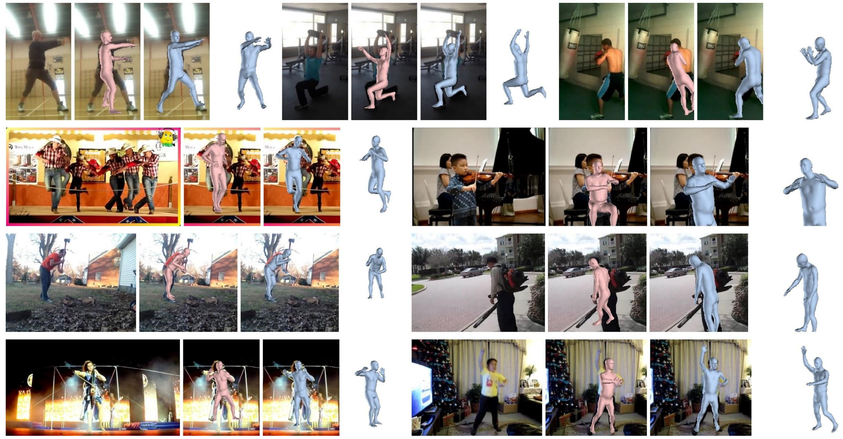

Figure 3. The dataset we automatically generated from Kinetics has diversity not found in mocap. The per-frame HMR method often fails in cases where our bundle adjustment succeeds.

[1] A Kanazawa *et al.* End-to-end Recovery of Human Shape and Pose. In *CVPR* 2018. [2] G Papanderou *et al.* Towards Accurate Multi-Person Pose in the Wild. In *CVPR* 2017 [3] F Bogo *et al.* Keep it SMPL. Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *ECCV* 2016

Dataset and models: https://github.com/deepmind/Temporal-3D-Pose-Kinetics