TAPE: Temporal Attention-based Probabilistic human pose and shape Estimation

Nikolaos Vasilikopoulos[1,2][0000−0002−2340−8039], Nikos Kolotouros[0000−0003−4885−4876], Aggeliki Tsoli[2][0000−0002−7254−3747], and Antonis Argyros[1,2][0000−0001−8230−3192]

[1] Computer Science Department, University of Crete
[2] Foundation for Research and Technology - Hellas (FORTH)
{nvasilik,aggeliki,argyros}@ics.forth.gr
nikoskolot@gmail.com

Abstract. Reconstructing 3D human pose and shape from monocular videos is a well-studied but challenging problem. Common challenges include occlusions, the inherent ambiguities in the 2D to 3D mapping and the computational complexity of video processing. Existing methods ignore the ambiguities of the reconstruction and provide a single deterministic estimate for the 3D pose. In order to address these issues, we present a **T**emporal **A**ttention based **P**robabilistic human pose and shape **E**stimation method (TAPE) that operates on an RGB video. More specifically, we propose to use a neural network to encode video frames to temporal features using an attention-based neural network. Given these features, we output a per-frame but *temporally-informed* probability distribution for the human pose using Normalizing Flows. We show that TAPE outperforms state-of-the-art methods in standard benchmarks and serves as an effective video-based prior for optimization-based human pose and shape estimation.

Keywords: 3D human pose · human shape · normalizing flows · probabilistic · human body reconstruction

1 Introduction

Human pose and shape estimation from RGB video is a key problem in computer vision with a wide variety of applications, such as AR/VR, surveillance, human-robot interaction and more. This problem is particularly challenging due to the large number of Degrees of Freedom of the human body in terms of pose and shape, the self-occlusions among body parts and the inherent ambiguity in 3D skeletal pose estimation given only 2D video observations. Moreover, processing a video sequence requires increased computational resources compared to single frame processing.

State-of-the-art previous work relies on deep learning and has mostly approached human pose and shape estimation from video as independent estimation of human pose and shape from a single RGB image [16,17,12,2,29]. However, the limited previous efforts that take into account the temporal aspect of a

Fig. 1. Left (Green): 3D body pose and shape estimation with MPS-Net [34]. Right (White): 3D body pose and shape estimation with the proposed method (TAPE). As it can be verified, the 3D human shape and pose estimated by TAPE is in better agreement with the shape and pose of the imaged person.

video [4,15,13] have shown increased accuracy and temporal coherence compared to methods that operate using a single RGB image as input. Most of the datasets used for training have only 2D annotations for the human body joints [36,14] and datasets with 3D annotations are mostly captured in controlled lab settings [9]. Contrary to the vast majority of existing methods that provide deterministic outputs of human pose and shape from a single RGB image [16,17,12,2], recent efforts for probabilistic human pose estimation [19] have shown increased performance due to dealing more effectively with the 2D-3D ambiguity of the observations in existing datasets.

In this paper, we propose the first temporal and probabilistic method for human pose and shape estimation from video. The human body is represented using the widely used SMPL model [23]. We propose a deep learning architecture where static ResNet-50 [8] features are extracted for each frame in the video and are then converted to temporal features using an attention-based temporal encoder and then integrated to one temporal feature as proposed by Wei *et al.* [34]. The output is a probability distribution for the human pose using Normalizing Flows and point estimates for the human shape and camera parameters inspired by Kolotouros et al [19]. Extensive experimental results on well-established datasets show increased accuracy for the task of regressing human pose and shape from visual data using a neural network. In addition, we demonstrate that our work serves as an effective video-based prior for optimization-based human prediction. In summary, our contributions are the following:

– We propose a temporal probabilistic model for human body and shape estimation from video input.

– We extend a state-of-the-art optimization method for human model fitting to 2D observations using our model as a video-based prior.
– We show state-of-the-art 3D pose estimation in standard benchmarks.

2 Related Work

2.1 Human 3D shape and pose from a single RGB image

Regression: Most of previous work on human pose and shape estimation from video input has treated the problem as human prediction from a single image handling each frame independently. These methods typically follow the regression paradigm where the parameters of a parametric model [23,30] are regressed from a deep neural network, given a single image as input [7,18,31,5,6,10]. A representative and baseline method is HMR [12], which regresses SMPL [23] parameters by minimizing the 3D to 2D keypoint reprojection loss while also using a pose discriminator for adversarial training. Similarly to HMR, we use a multi-layer perceptron for estimating human shape and camera parameters.

Optimization: Optimization-based methods estimate iteratively the parameters of a body model, such that it is consistent with a set of 2D cues. Several cues have been employed, e.g., silhouettes [20], POFs [35], dense correspondences [7] or contact [28]. The most widely used cues are 2D keypoints with SMPLify [2] being a representative optimization-based approach which predicts SMPL parameters from 2D keypoints.

Hybrid Optimization-Regression: Optimization-based approaches are slower than regression-based ones, but can be more accurate given a good initialization. Thus, it is a common practice that a regression-based method is used to initialize an optimization-based method. SPIN [16] is a hybrid method which uses regression and optimization in the training loop. EFT [11] combines optimization and regression but also updates the network weights during the fitting procedure. In this work, we demonstrate how our probabilistic model can leverage video-based information to effectively guide keypoint-based optimization.

Non Parametric: METRO [21] and Mesh Graphormer [22] regress the 3D mesh vertices of the human body and do not predict the model parameters, directly. Also, they use the HRNet [37] backbone instead of ResNet [8]. Therefore, despite their SOTA performance, these methods cannot be directly compared with our method and the methods we compare with.

2.2 Human 3D shape and pose from RGB video

Most related to our work are methods for human pose and shape prediction leveraging the temporal aspect of a video. Kanazawa *et al.* [13] proposed a regression-based method to learn human motion kinematics by predicting past and future

frames with a temporal encoder. Kocabas *et al.* [15] proposed VIBE, a temporal encoder with Gated Recurrent Units (GRUs) to capture the motion between the static features and a motion discriminator trained with the AMASS [25] dataset for adversarial training. TCMR [4] reduced drastically the acceleration error of VIBE by using a temporal encoder which consists of 3 GRUs, one for the past, one for the future and one for the current frame and then integrating the features to produce a smooth video result with better pose and acceleration estimation than previous works. Recently, MPS-Net [34] proposed a motion continuity attention (MoCA) module which captures the continuity between frames and a hierarchical attentive feature integration (HAFI) module to effectively combine adjacent past and future feature representations to strengthen temporal correlation and refine the feature representation of the current frame. MPS-Net achieves temporally coherent pose estimates without penalizing the accuracy on pose prediction. Our method adopts the MoCA and the HAFI modules from MPS-Net as well as the motion discriminator in VIBE.

2.3 Multiple hypotheses

Biggs *et al.* [1] extend HMR [12] with N prediction heads. This leads to a discrete set of hypotheses, instead of a full probability of poses as we do. In a concurrent work, Sengupta *et al.* [32] use a Gaussian posterior to model the uncertainty in the parameter prediction. Recently, ProHMR [19] used Normalizing Flows to predict a distribution of 3D poses conditioned on the provided 2D input. The probabilistic modeling of ProHMR is efficient at computing the most likely pose comparable to the SOTA methods and also outperforms previous work on optimization tasks such as 2D keypoint fitting and multi-view refinement. Our method uses the probabilistic model of ProHMR for video input.

3 Method

Given an input video $V = \{I_t\}_{t=1}^{T}$ of length T, we resize each frame at a resolution of 224×224. The output of TAPE when applied to this input is a per-frame probability distribution for the human pose and point estimates for the human shape as well as camera parameters. We represent the human body using the SMPL [23] model. SMPL provides a function $M(\theta, \beta)$ that takes as input the pose parameters $\theta \in \mathbb{R}^{72}$ and the shape parameters $\beta \in \mathbb{R}^{10}$, and returns the body mesh $M \in \mathbb{R}^{N \times 3}$, with $N = 6890$ vertices. The proposed deep learning architecture is illustrated in Figure 2 and detailed in the following sections.

3.1 Temporal encoding

We use a pretrained ResNet-50 [8] backbone from ProHMR that was trained on standard human pose and shape estimation datasets. The backbone extracts static features for each frame of the video. We consider $T = 16$ frames at a time and, following Wei *et al.* [34], the features are sent to the Motion Continuity

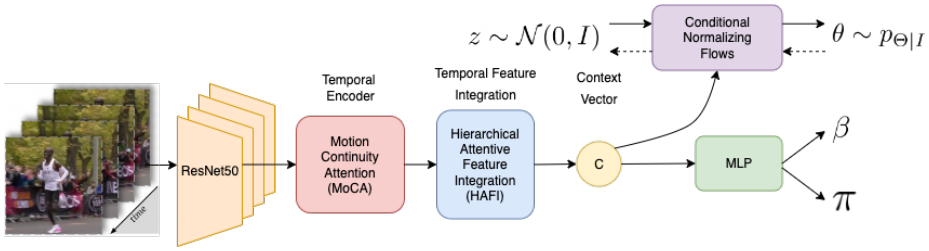

Fig. 2. TAPE architecture: We extract static ResNet-50 features from video which are used as input for the Temporal Encoder (MoCA). MoCA outputs temporal features, and the HAFI module integrates them to a single feature, the hidden vector c, which is used as the conditioning input to the flow model. In parallel, it is also decoded to shape parameters β and camera parameters π. Our flow model learns an invertible mapping which allows for two processing directions; depending on the desired function, we can perform both sampling and fast likelihood computation.

Attention module that outputs T temporal features. The T temporal features are integrated to a single feature context vector c by the Hierarchical Feature Integration module proposed by Wei *et al.* [34].

3.2 Normalizing flows

Following Kolotouros *et al.* [19], we model $p(\theta|I)$ using Conditional Normalizing Flows. We learn a mapping $f : \mathbb{R}^d \times \mathbb{R}^c \rightarrow \mathbb{R}^d$ that is bijective in the latent variable z and the pose parameters θ, and is parametrized by the image features $g = c(I)$. More specifically, $\theta = f(z; c)$ and $z = f^{-1}(\theta; c)$. The choice of this particular model class enables to perform both fast likelihood computation and sampling from the distribution. Another useful property is that, as shown in Kolotouros *et al.* [19], the mode of the output distribution is the transformation on $z = 0$, i.e.,

$$\theta^* = argmax_\theta(\theta|c) = f(0; c), \tag{1}$$

which means that in the absence of additional evidence, the probabilistic model can be used to make predictions by choosing the sample with the maximum probability.

3.3 MLP

Camera parameters π and SMPL shape parameters β are regressed by a small pretrained MLP as in ProHMR. The MLP takes as input the context vector c and outputs matrices of shape and camera parameters per frame.

3.4 Motion discriminator

We train a motion discriminator as in Kocabas *et al.* [15] using the AMASS dataset [25]. The motion discriminator enforces the generator, i.e. our network,

to produce plausible human motions and shapes. The discriminator takes as input the poses Θ that the generator predicted and outputs a value $\in [0, 1]$ representing the probability that $\hat{\Theta}$ belongs to the manifold of plausible human motions. The motion discriminator consists of a GRU with 2 layers and hidden size 1024. The aggregation of hidden states is done by a self-attention mechanism. The objective function for training the Motion Discriminator is:

$$L_{DM} = E_{\Theta\ p_R}[(D_M(\Theta) - 1)^2] + E_{\Theta\ p_G}[D_M(\hat{\Theta})^2], \tag{2}$$

where p_R is a real motion sequence from the AMASS dataset, while p_G is a generated motion sequence. Since D_M is trained on ground-truth poses, it also learns plausible body pose configurations. Therefore, the pose discriminator that was used in ProHMR is not needed in our method.

3.5 Training objective

Due to lack of ground truth SMPL annotations in some of the datasets, we use mixed training following ProHMR. When SMPL data are available, we minimize the negative log-likelihood of the ground truth examples θ_{gt}

$$L_{nll} = -\ln p_{\Theta|V}(\theta_{gt}|c). \tag{3}$$

In every dataset, we minimize the reprojection loss jointly with an adversarial motion prior based on the motion discriminator described in Section 3.4. In notation,

$$L_{exp} = E_{z\ p_z}[L_{2D}(f(z;c), \beta, \pi) + L_{adv}(f(z;c),] \tag{4}$$

where

$$L_{adv} = E_{\Theta\ p_G}[(D_M(\hat{\Theta}) - 1)^2]. \tag{5}$$

For each image I, the mode θ_I^* of the output distribution corresponds to the transformation of $z = 0$. We do this by explicitly supervising θ_I^* with all the available annotations as in a standard regression framework and minimize

$$L_{mode} = L_{3D}(\theta_I^*, \beta_+ L_{2D}) + L_{adv}(\theta_I^*). \tag{6}$$

Following ProHMR, we use 6D representation [37] to model rotations and the L_{orth} loss to force the 6D representations of the samples drawn from the distribution to be close to the orthonormal 6D representation.

The final training objective becomes:

$$\begin{aligned} L = &\lambda_{nll} L_{nll} \\ &+ \lambda_{exp,2D} L_{exp,2D} + \lambda_{exp,adv} L_{exp,adv} \\ &+ \lambda_{mode,2D} L_{mode,2D} + \lambda_{mode,adv} L_{mode,adv} \\ &+ \lambda_{mode,\theta} L_{mode,\theta} + \lambda_{mode,\beta} L_{mode,\beta} \\ &+ \lambda_{mode,3D} L_{mode,3D} + \lambda_{orth} L_{orth}. \end{aligned} \tag{7}$$

The loss weights are defined as: $\lambda_{nll} = 0.001$, $\lambda_{exp,2D} = 0.001$, $\lambda_{exp,adv} = \lambda_{mode,adv} = 0.01$, $\lambda_{mode,2D} = 0.01$, $\lambda_{mode,3D} = 0.05$, $\lambda_{mode,\theta} = 0.001$, $\lambda_{mode,\beta} = 0.0005$ and $\lambda_{orth} = 0.1$.

3.6 Optimization-based model fitting

Extending the model fitting procedure in ProHMR to the temporal domain, we use a video-based pose prior that models the likelihood of the pose at a specific frame conditioned on the video evidence:

$$E_{\theta|V} = -\ln p_{\Theta|V}(\theta|c). \tag{8}$$

As initialization for the fitting, we use the mode θ_I^* of the conditional distribution calculated from the regression step.

Following SMPLify [2], E_J penalizes the weighted 2D distance between the projected model joints and the detected joints and E_β is a quadratic penalty on the shape coefficients.

The final objective function for the optimization is:

$$E = \lambda_J E_J - \lambda_V E_{\theta|V} + \lambda_\beta E_\beta. \tag{9}$$

4 Experiments

4.1 Training procedure

The sequence length during training is $T = 16$ frames with a mini batch size of 32. Each frame gets augmented as described in Section 3.1 and then static features are computed using ResNet-50. We initialize our network with pretrained versions of the ResNet-50 network as well as the Normalizing Flows based on the ProHMR checkpoint that is publicly available. We keep the ResNet-50 features fixed, but continue training the Normalizing Flows. During training, we draw 2 samples from the distribution and not only the mode. The motion discriminator takes as input the predicted parameters Θ with the ground-truth data from AMASS and is trained to predict a single fake/real probability for each sample. We train our network as well as the motion discriminator using Adam optimizers with learning rate equal to $5e-5$. Training requires at least 7 epochs and takes about 1 hour on a single NVidia GTX1080Ti GPU.

4.2 Datasets

We train and evaluate our network for human prediction using the following datasets:

1. MPI-INF-3DHP [27]: This dataset contains videos of human motion, mostly indoors, at 30fps. MPI-INF-3DHP is augmented with SMPL parameters every 10 frames, derived from SPIN.
2. Human3.6M [9,3]: Human3.6M is a large-scale dataset captured indoors with an optical motion capture system and the training set consists of 8 different subjects performing various actions. Human3.6M was captured at 50fps and we subsample it to 25fps to match MPI-INF-3DHP. We optionally use SMPL annotations acquired with Mosh [24].

Models	3DPW				MPI-INF-3DHP			Human3.6M		
	PA-MPJPE	MPJPE	MPVE	ACCEL-ERR	PA-MPJPE	MPJPE	ACCEL-ERR	PA-MPJPE	MPJPE	ACCEL-ERR
SPIN	59.2	96.9	116.4	29.8	67.5	105.0	-	41.1	-	18.3
ProHMR	59.8	-	-	-	65.0	-	-	41.2	-	-
Biggs	59.9	-	-	-	-	-	-	41.6	-	-
VIBE	56.5	93.5	113.4	27.1	63.4	97.7	29.0	41.5	65.9	18.3
TCMR	55.8	95.0	111.5	**6.7**	62.8	96.5	**9.5**	41.1	62.3	**5.3**
MPS-NET	**54.0**	91.6	**109.6**	7.5	-	-	-	-	-	-
TAPE (Ours)	56.6	**89.3**	112.5	10.7	**56.7**	**94.0**	12.4	**39.5**	**60.0**	6.5

Table 1. Evaluation of state-of-the-art single image-based and video-based methods on the 3DPW, Human3.6M, and MPI-INF-3DHP datasets. Training has been performed on the MPI-INF-3DHP, Human3.6M datasets (not 3DPW).

Models	3DPW				MPI-INF-3DHP			Human3.6M		
	PA-MPJPE	MPJPE	MPVE	ACCEL-ERR	PA-MPJPE	MPJPE	ACCEL-ERR	PA-MPJPE	MPJPE	ACCEL-ERR
VIBE	57.7	91.9	-	27.1	68.9	103.9	27.3	53.3	78.0	27.3
TCMR	52.4	86.5	103.2	**6.8**	63.5	97.6	**8.5**	52.0	76.0	15.3
MPS-NET	52.1	84.3	99.7	7.4	62.8	96.7	9.6	47.4	69.4	**3.9**
TAPE (Ours)	**51.5**	**79.9**	**98.1**	8.9	**59.1**	**94.2**	11.6	**42.1**	**62.6**	6.5

Table 2. Evaluation of state-of-the-art video-based methods on 3DPW, MPI-INF-3DHP, and Human3.6M datasets. Following Choi *et al.* [4], all methods are trained on the training set including 3DPW, but do not use the Human3.6M SMPL parameters obtained with Mosh [24]. The number of input frames follows the original protocol of each method.

3. 3DPW [26]: 3DPW was generated using IMU sensors combined with a 2D pose detector to compute ground truth SMPL parameters. We keep the initial resolution of 3DPW at 30fps. 3DPW is the only dataset we use that consists only of outdoor videos.

4.3 Metrics

We report Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) to evaluate the accuracy of the obtained 3D poses. Mean Per Joint Position Error (MPJPE) takes also into consideration the global translation/orientation of the human and the predicted camera parameters. Mean Per Vertex Error (MPVE) is only available in 3DPW and it helps to measure the accuracy in shape prediction. We compare TAPE with state-of-the-art single-image and temporal methods. Acceleration error (ACCEL-ERR measured in mm/s^2), calculated as the difference in acceleration between the ground-truth and predicted 3D joints can be important in video methods since it evaluates how temporal coherent the prediction between consecutive frames is. Given though that existing datasets contain limited variation in acceleration, we rely more heavily in our evaluation on the rest of the error metrics.

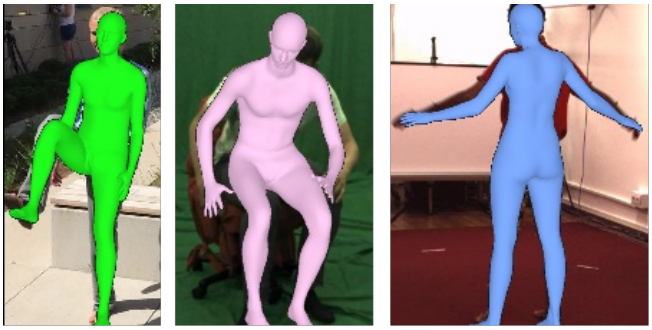

Fig. 3. Human body estimation with TAPE in the *first* scenario. Results are shown for 3DPW (left-green), MPI-INF-3DHP (center-pink), Human3.6M (right-blue).

	n=5		n=10		n=25	
	3DPW	H36M	3DPW	H36M	3DPW	H36M
Biggs	57.1	42.0	56.6	42.2	55.6	42.2
ProHMR	56.5	39.4	54.6	38.3	52.6	36.8
TAPE	**53.9**	**38.1**	**52.0**	**37.0**	**49.5**	**35.4**

Table 3. Multiple hypotheses evaluation. Numbers are PA-MPJPE in mm. We report the minimum error over n samples drawn from the distribution.

4.4 Comparison with state-of-the-art methods

We consider two scenarios for assessing the performance of the proposed method. In the *first* scenario, we train our network using MPI-INF-3DHP and Human3.6M augmented with SMPL annotations. We evaluate our method on all datasets (MPI-INF-3DHP, Human3.6M, 3DPW). Evaluating on 3DPW shows how our method generalizes to unknown data and how it estimates human shape and pose in videos with outdoor activities. In the *second* scenario, we, additionally, use the 3DPW [26] dataset with outdoor scenes during training, but ignore the SMPL annotations in Human3.6M. We evaluate our method again on all datasets. AMASS [25] is used for adversarial training to obtain real samples of 3D human motion in both scenarios.

Video-based methods: Results in Table 1 show a comparison of our method with the state-of-the-art methods on 3DPW, MPI-INF-3DHP and Human 3.6M for the *first* scenario. Our method outperforms in PA-MPJPE and MPJPE every other method in Human3.6M and MPI-INF-3DHP. Testing on 3DPW without the use of the dataset during training shows how accurately the presented methods can generalize to outdoor videos. In 3DPW, our method produces comparable PA-MPJPE with the state-of-the-art and outperforms them in MPJPE.

In Table 2, we see the results for the *second* training scenario. Ignoring the SMPL annotations in Human3.6M during training drops the performance

	SPIN	SPIN+ SMPLify	ProHMR	ProHMR+ fitting	TAPE	TAPE+ SMPLify	TAPE+ fitting
PA-MPJPE	41.8	43.8	41.2	34.8	39.5	38.3	**32.8**
ACCEL-ERR	-	-	-	-	6.5	6.2	**5.4**

Table 4. Evaluation of different model fitting methods. The fitting algorithms are initialized by the corresponding regression results. All numbers are PA-MPJPE in mm.

Fig. 4. Qualitative results on 3DPW test set [26]. From left to right: original input, MPS-Net [34] (light-green), TAPE (white), TAPE + fitting on OpenPose [33] detections (purple).

slightly for all methods. However, our method outperforms all other methods in the PA-MPJPE, MPJPE and MPVPE metrics for all datasets, which is also facilitated by the fact that 3DPW has been added in the training set. These metrics show that our method predicts better poses, body orientation and human shape.

In both scenarios (Tables 1, 2), the acceleration error of our method is comparable to MPS-NET.

Single frame methods: Our approach outperforms single frame methods (SPIN, Biggs, ProHMR) and this shows that we effectively capture temporal information from the video input.

Multiple hypotheses: In Table 3, we compare the representational power of TAPE with Biggs [1] and ProHMR [19] that provide non-deterministic outputs for different number of random samples drawn from the distribution. We consider 5, 10 and 25 samples and report the minimum PA-MPJPE out of all selected

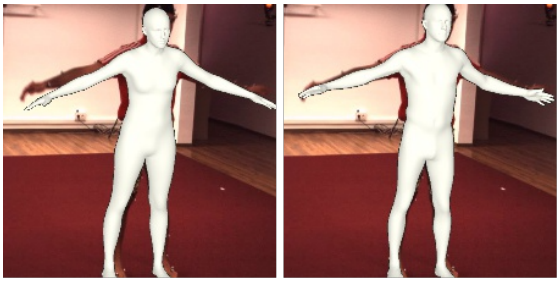

Fig. 5. (Left) Human body estimation using TAPE on Human3.6M. (Right) Refinement of TAPE prediction using the proposed optimization framework.

samples. Our method outperforms both previous work methods for every number of samples.

Optimization task: In Table 4, we compare SPIN[17], ProHMR[19] and TAPE in conjunction with optimization-based model fitting frameworks considering 2D keypoints. Evaluation is performed on the Human3.6M dataset. In all cases, optimization is initialized using the regression output of the neural network in each method. The output of SPIN is fitted on 2D keypoints using SMPLify [2]. The output of ProHMR is used in the optimization proposed in the corresponding paper. The output of TAPE is used in the proposed optimization framework. We observe that our proposed optimization framework drops noticeably the already reduced PA-MPJPE of TAPE and is more effective than the widely used SMPLify framework. The acceleration error is improved as well.

5 Qualitative Evaluation

We provide qualitative results on the performance of our method against previous work and on various datasets. In Figure 1, we show a comparison with the SOTA method MPS-Net on the challenging dataset 3DPW without using the dataset in training (*first* scenario). We observe that our method produces more accurate pose predictions than the current SOTA. Figure 5 shows the visual impact of the proposed optimization framework at refining the output of TAPE. It is clear that the fitting optimization improves the pose and shape accuracy. Optimization on both ground-truth and open-pose keypoints improves the prediction as shown in Figure 4 on 3DPW dataset and in Figure 6 on MPI-INF-3DHP. Finally, Figure 3 shows examples of the performance of TAPE on all datasets based on the same scenario.

6 Conclusions

We propose TAPE, the first probabilistic temporal model for human pose, shape estimation and camera prediction from video input. We combine an attention-based temporal encoder with a probabilistic model based on normalizing flows

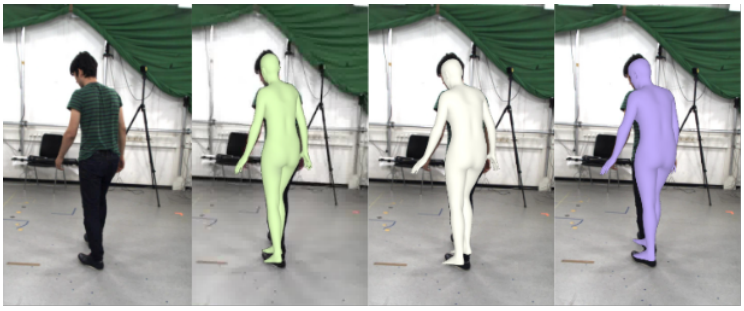

Fig. 6. Qualitative on MPI-INF-3DHP test set [27]. From left to right: original input, MPS-Net [34] (light-green), TAPE (white), TAPE + fitting on OpenPose [33] detections (purple).

and show increased accuracy compared to state-of-the-art and real-time performance. We, additionally, show that optimizing for human pose and shape estimation using TAPE as a video-based prior for human pose outperforms the widely used SMPLify method for 2D keypoint fitting that is image-agnostic by a large margin. Future work includes extending our method to temporal human pose estimation from multiple views and experimenting with Transformer-based temporal encoders to further increase the 3D shape and pose estimation accuracy.

Acknowledgements

This work is partially supported by the Greek Secretariat for Research and Innovation and the EU, Project SignGuide: Automated Museum Guidance using Sign Language T2EDK-00982 within the framework of "Competitiveness, Entrepreneurship and Innovation" (EPAnEK) Operational Programme 2014-2020. It was also partially supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "1st Call for HFRI Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment", project I.C.Humans, number 91.

References

1. Biggs, B., Ehrhart, S., Joo, H., Graham, B., Vedaldi, A., Novotny, D.: 3D multibodies: Fitting sets of plausible 3D models to ambiguous image data. In: NeurIPS (2020)
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Springer International Publishing (Oct 2016)
3. Catalin Ionescu, Fuxin Li, C.S.: Latent structured models for human pose estimation. In: International Conference on Computer Vision (2011)

4. Choi, H., Moon, G., Chang, J.Y., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
5. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: European Conference on Computer Vision. pp. 20–40. Springer (2020)
6. Georgakis, G., Li, R., Karanam, S., Chen, T., Košecká, J., Wu, Z.: Hierarchical kinematic human mesh recovery. In: European Conference on Computer Vision. pp. 768–784. Springer (2020)
7. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10884–10894 (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
9. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (jul 2014)
10. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5579–5588 (2020)
11. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In: 3DV (2020)
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)
13. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Computer Vision and Pattern Recognition (CVPR) (2019)
14. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Computer Vision and Pattern Recognition (CVPR) (2019)
15. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
16. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
17. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019)
18. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4501–4510 (2019)
19. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV (2021)
20. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6050–6059 (2017)
21. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR (2021)
22. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: ICCV (2021)

23. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
24. Loper, M.M., Mahmood, N., Black, M.J.: MoSh: Motion and shape capture from sparse markers. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **33**(6), 220:1–220:13 (Nov 2014). https://doi.org/10.1145/2661229.2661273, `http://doi.acm.org/10.1145/2661229.2661273`
25. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
26. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018)
27. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3D Vision (3DV), 2017 Fifth International Conference on. IEEE (2017). https://doi.org/10.1109/3dv.2017.00064, `http://gvv.mpi-inf.mpg.de/3dhp_dataset`
28. Muller, L., Osman, A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9990–9999 (2021)
29. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). pp. 484–494. IEEE (2018)
30. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
31. Pavlakos, G., Kolotouros, N., Daniilidis, K.: Texturepose: Supervising human mesh estimation with texture consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 803–812 (2019)
32. Sengupta, A., Budvytis, I., Cipolla, R.: Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from Images in the Wild. In: International Conference on Computer Vision (October 2021)
33. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
34. Wei, W.L., Lin, J.C., Liu, T.L., Liao, H.Y.M.: Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
35. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10974 (2019)
36. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: 2013 IEEE International Conference on Computer Vision. pp. 2248–2255 (2013). https://doi.org/10.1109/ICCV.2013.280
37. Zhou, Y., Barnes, C., Jingwan, L., Jimei, Y., Hao, L.: On the continuity of rotation representations in neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)