Towards Accurate Marker-less Human Shape and Pose Estimation over Time

Yinghao Huang[1], Federica Bogo[2], Christoph Lassner[3,*], Angjoo Kanazawa[4]
Peter V. Gehler[5,*], Javier Romero[3], Ijaz Akhter[6], Michael J. Black[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Microsoft, [3]Body Labs Inc., [4]UC Berkeley
[5]University of Würzburg, [6]Australian National University

***Abstract*—Existing markerless motion capture methods often assume known backgrounds, static cameras, and sequence specific motion priors, limiting their application scenarios. Here we present a fully automatic method that, given multi-view videos, estimates 3D human pose and body shape. We take the recently proposed SMPLify method [12] as the base method and extend it in several ways. First we fit a 3D human body model to 2D features detected in multi-view images. Second, we use a CNN method to segment the person in each image and fit the 3D body model to the contours, further improving accuracy. Third we utilize a generic and robust DCT temporal prior to handle the left and right side swapping issue sometimes introduced by the 2D pose estimator. Validation on standard benchmarks shows our results are comparable to the state of the art and also provide a realistic 3D shape avatar. We also demonstrate accurate results on HumanEva and on challenging monocular sequences of dancing from YouTube.**

***Keywords*-3d reconstruction; shape and pose estimation; multi-view; marker-less; body model**

I. Introduction

The markerless capture of human motion (mocap) has been a long term goal of the community. While there have been many proposed approaches and even commercial ventures, existing methods typically operate under restricted environments. Most commonly, such methods exploit background "subtraction," assuming a known and static background, and the most accurate methods employ strong prior assumptions about the motion of the actor. In many cases, the best results on benchmarks like HumanEva [42] are obtained by training on the same motion by the same actor as is evaluated at test time [5]. Here we provide a solution for markerless mocap that is more accurate than the recent state of the art but is also less restrictive.

There are four key components to our approach. First our approach exploits SMPL [28], a realistic, low-dimensional, 3D parametric model of the human body. Second we use a Convolutional Neural Network (CNN) to compute putative 2D joint locations in multiple camera images. We then fit the 3D parametric model to the 2D joints robustly. This extends the SMPLify approach for pose and shape estimation [12] from a single image to multi-camera data.

Third, we go beyond SMPLify [12] to use a deep CNN to also segment people from images [26]. This removes the need for a background image and makes the approach more general. We fit our 3D body model to both the 2D joints and the estimated silhouettes and show that the silhouettes provide significantly improved accuracy and realism to the mocap.

Since 2D joints estimated by CNNs sometimes confuse left and right parts of the body, the image evidence alone is not enough for a reliable 3D solution. Consequently we exploit temporal information to resolve such errors. This leads to the fourth component in which we exploit a generic temporal prior based on the insight that human motions can be captured by a low-dimensional Discrete Cosine Transform (DCT) basis [4]. We implement this DCT temporal term robustly and show that it improves performance yet requires no training data.

We call the method MuVS (Multi-View SMPLify) and evaluate it quantitatively on HumanEva [42] and Human3.6M [23]. We find that MuVS gives an error comparable with any published result and more realistic meshes (see Figure 1), while having fewer restrictions. We evaluate the method with an ablation study on HumanEva to determine which design decisions are most important.

Additionally, our approach also works in the monocular camera setting. We find that the temporal coherence term enables reasonable reconstruction of pose from monocular video even with a moving camera, complex background, and challenging motions. We evaluate this quantitatively on HumanEva [42] and some challenging dancing video sequences from Youtube. The software will be made available for research purposes.

II. Related Work

The majority of previous works only handle one aspect of the two closely related problems: markerless 3D human body shape and pose estimation. Some of these target 3D pose estimation [5], [16], [17], [19], [38], [43], [53]. They formulate it as a discriminative problem, directly inferring

[*] The work was performed at MPI for Intelligent Systems.

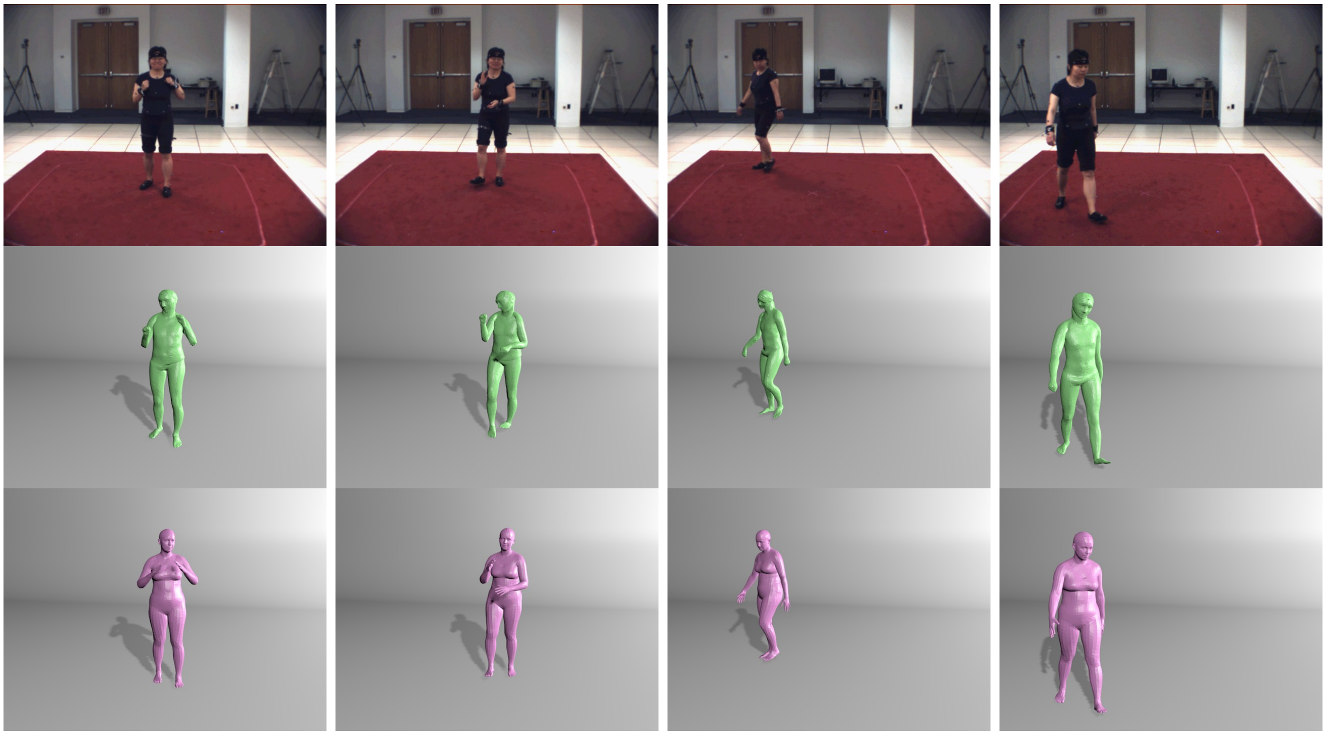


Figure 1: Given multi-view videos, our method can not only yield more accurate 3D pose estimation results, but also more realistic and natural meshes than the state of the art. The entire process is fully automatic. From up to bottom: example input frames; meshes returned by [39]; meshes generated by our method.

3D pose from 2D image features, assuming no explicit human body model. Amin et al. [5] extend single-view based pictorial structure to multi-view cases, jointly infer 2D joint location of all views, then use linear-triangulation to obtain the 3D joints. Yao et al. [53] propose a stochastic gradient-based method for a Gaussian Process Latent Variable Model (GPLVM), which shows good optimization properties. Uncertainty over estimated 2D image features has also been considered. Zhou et al. [55] introduce sparsity prior over human pose, and jointly handle the pose and 2D location uncertainty, while Kazemi et al. [25] address the body part correspondence problem by optimizing latent variables. Similar ideas are proposed by Simo-Serra et al. [44], in which they also estimate 2D and 3D pose at the same time. Twin Gaussian processes [11] have also been used on this problem. Most recently deep learning methods achieve the most accurate pose estimation results [17], [31], [37], [47], [48]. To address their need for huge amounts of training data, Yasin et al. [54] propose a dual-source approach. Pavlakos et al. [33] directly regress 3D pose from RGB image via CNNs in a coarse-to-fine manner.

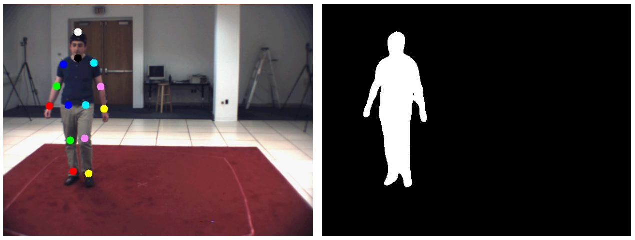
The second major set of approaches use an explicit intermediate human body representation, which effectively assists pose estimation but often lacks realism [10], [15], [43], [46]. Common human body representations include the Articulated Human Body Model [16], 3D Pictorial Structures [10], [43], the sum-of-Gaussians model [46], and the Triangulated Mesh Model [41]. These models are usually utilized to represent the structure of the human body, thus facilitating the inference of pose parameters. Sometimes the body mesh is also considered, but in an abstract or coarse way, without consideration of the shape details.

Estimating both the pose and surface mesh, usually requires complex global optimization [19], [20]. Often the silhouette of the body is assumed to be known [7] and manual initialization or a pre-scanned surface mesh is required [3], [9], [14], [21], [22], [24], [36], [45], [51], [49]. Balan et al. [8] address this problem by fitting a SCAPE body model [6] to multi-view silhouettes. Their initialization method is complex and they do not integrate information over time. Another very recent work, concurrent with ours, is the one proposed in [34]. They also use CNNs to detect 2D joints, then fit a 3D pictorial structures model to the detections. Their method only returns 3D joints as output, while ours estimates body shape and pose together. The method proposed in [30] simultaneously regresses 2D and 3D joints from monocular video via one CNN, then fits a skeleton model to the 3D joint estimations, achieving real-time performance.

The most similar recent work addresses fully automatic estimation of 3D pose and shape from monocular images [12] and multi-views videos [39]. The SMPLify algorithm proposed by Bogo et al. [12] makes it possible to simultaneously obtain 3D pose and convincing body shape from a single image, without requiring any user intervention, and without assuming background extraction or complex optimization techniques. Based on the state-of-the-art 3D human body model, SMPL [28], they infer human shape and pose parameters by fitting the projection of 3D SMPL joints to 2D joints estimated via a 2D joint detector like DeepCut or CPM [35], [50]. Ambiguity issues are handled by applying priors learned from the large-scale public CMU dataset [2], which is vital for their method to yield valid results. Rhodin et al. [39] propose a method that works on multi-view videos. Built upon a sum-of-Gaussian shape model [40], [46], their algorithm first initializes the pose of each Gaussian blob, then refines the pose and shape of each blob via the body contour approximation with image gradients. As in Bogo et al. [12], they use deep learning to estimate 2D joints to get rough joint locations in each view. Also they enforce temporal coherence by penalizing acceleration between frames.

The general performance capture method in [52] is also closely related to ours, and works on monocular videos. A specific mesh and skeleton for each actor is required in advance and sometimes manual labour is needed. In contrast, our method runs fully automatically and determines the skeleton configuration and surface mesh together with the pose in the process.

We go beyond SMPLify by extending it to multi-view and monocular videos in a principled way. We show that there are important additional cues other than 2D joints to utilize, like silhouettes and temporal coherence. Though conceptually similar to the method in [39], our framework differs in important respects. First, we use explicit segmentation to obtain the body contour. Second, we use a DCT basis as the temporal prior model. As a general temporal smoothness model, DCT can be applied in any video sequence, without the need of learning from a training dataset. Third, in contrast to the sum-of-Gaussian model [40], we use the SMPL [28] body model, which naturally encodes the statistical shape and pose dependency between different body parts in a holistic way. This enables our method to, not only estimate accurate 3D joint locations, but also a realistic body mesh. This facilitates future modification and animation. In comparison, a volumetric skinning approach is utilized in [39] to estimate the actor body surface from the Gaussian representation. Their surface is coarser and does not allow for detailed deformations. Finally, we demonstrate that our method can be applied on monocular videos, unlike the method in [39].



(a) Estimated 2D joints (b) Segmented silhouette

Figure 2: Automatically estimated 2D joint locations using DeepCut [35] and the silhouette estimated via [26]; here shown on the HumanEva dataset [42].

III. 2D Joints and Contour Segmentation

Our method takes as input a set of 2D body joints and segmentation of the body from the background. For a direct quantitative comparison with SMPLify on standard test datasets, we use the same CNN-based joint estimator, DeepCut [35]. For more complex videos from the Internet, we use the real-time version of the CPM method [13] since we find it is more reliable than DeepCut. We also use a CNN trained to estimate human segmentations [26]. Both of these are fully automatic and computed by CNNs [35], [50] trained on generic databases, which do not overlap with any of our test data. Illustrative joint estimation and human body segmentation results are shown in Figure 2.

IV. Multi-view SMPLify

Here we first extend SMPLify to multiple camera views, then further extend it over time. Given the 2D joints and silhouettes for all the input frames for each camera view, we estimate the 3D pose for each time instant. We then combine information from all the views to estimate a consistent 3D human shape over time. Consequently, our algorithm is composed of two consecutive stages described in detail below.

In the first stage, a separate SMPL model is fit to all views independently at each time instant. The extension of the public SMPLify code to multiple views is straightforward: we estimate the shape and pose using information from all camera views. This gives a fully automatic approach to multi-camera marker-less motion capture. In the case of the original single-view SMPLify, the 3D pose and shape may be quite ambiguous given 2D joints and the method relied heavily on priors to prevent interpenetration. In contrast, with as few as 2 views, many of these ambiguities go away. After that, the silhouette is used to refine the estimated shape, which is then more faithful to the observed body.

In the second stage, we first estimate a consistent 3D shape across the entire sequence by taking the median of all the shape parameters obtained in the first stage. The pose parameters for each frame are initialized with their values from the first stage. We then consider a set of consecutive frames together and regularize the motion in time. We do this by minimizing the projected joint error while encouraging the trajectory of each 3D joint to be well represented by a set of low-d DCT basis vectors [4]. This temporal smoothing helps remove errors caused by inaccurate 2D joint estimates, which may be noisy and contain errors. In particular CNNs sometimes detect spurious points or suffer from left/right ambiguity.

A. Stage One: Per-frame Fitting

As in SMPLify, we use SMPL as our underlying shape representation. SMPL is a state-of-the-art statistical human body model [28], which is controlled by two sets of parameters, one for body shape, the other for pose. More formally, SMPL is defined as $M(\boldsymbol{\beta},\boldsymbol{\theta};\Phi)$, where $\boldsymbol{\beta}$ is a vector of shape parameters that are responsible for the 3D body shape due to identity, and $\boldsymbol{\theta}$ is a vector of pose parameters representing body part rotations in a kinematic tree. The fixed parameters Φ are learned from a large number of 3D body meshes. For the detailed meaning of all these parameters, we refer the reader to [28].

We first estimate the shape and pose parameters of the SMPL model for each time instant. Given the corresponding 2D joint estimates $\{J_{est}^1, J_{est}^2, \ldots, J_{est}^{|V|}\}$ for all the different views V, we formulate the energy function as the following:

$$E_M(\boldsymbol{\beta},\boldsymbol{\theta}) = E_P(\boldsymbol{\beta},\boldsymbol{\theta}) + \sum_{v=1}^{V} E_J(\boldsymbol{\beta},\boldsymbol{\theta};K_v,J_{\text{est}}^v)\,, \quad (1)$$

where E_p is the prior term, K_v are the camera parameters for view v, and E_J is the joint fitting term (i.e. the data term). Note that here we remove the other priors used in SMPLify, because in multi-view cases the solution is better constrained. E_P is composed of two terms: a shape prior E_β and a pose prior E_θ. The pose prior is learned from the CMU dataset [2], while the shape prior is leaned from the SMPL body shape training data.

$$E_P(\boldsymbol{\beta},\boldsymbol{\theta}) = \lambda_\theta E_\theta(\boldsymbol{\theta}) + \lambda_\beta E_\beta(\boldsymbol{\beta}). \quad (2)$$

The joint fitting term is formulated as follows:

$$E_J(\boldsymbol{\beta},\boldsymbol{\theta};K_v,J_{\text{est}}^v) = \sum_{\text{joint } i} w_i \rho_{\sigma_1}(\mathit{\Pi}_{K_v}(R_\theta(J_i(\boldsymbol{\beta}))) - J_{\text{est},i}^v)\,, \quad (3)$$

where $J(\cdot)$ is the joint estimation function, which returns joint locations, R is the rotation function, $\mathit{\Pi}$ the projection function, w_i the confidence yielded by the 2D joint detection CNN. Considering the inevitable detection noise and errors in the entire process, instead of the standard squared error we use a robust Geman-McClure error function, which is defined by:

$$\rho_\sigma(e) = \frac{e^2}{\sigma^2 + e^2}\,, \quad (4)$$

here e is the residual error, and σ is the robustness constant carefully chosen.

After obtaining the initial pose and shape estimation via fitting SMPL to 2D joints, we further refine it by adding silhouette information. The fitting error between the contour rendered from the SMPL model and the CNN-segmented one is defined as:

$$E_S(\boldsymbol{\beta},\boldsymbol{\theta};K_v,U_v) = \sum_{x\in\hat{S}(\boldsymbol{\beta},\boldsymbol{\theta})} l(x,U_v)^2 + \sum_{x\in U_v} l(x,\hat{S}(\boldsymbol{\theta},\boldsymbol{\beta}))\,, \quad (5)$$

where $l(x,S)$ denotes the absolute distance from a point x to a silhouette S; the distance is zero when the point is inside S. The first term computes the distance from points on the projected model $\hat{S}(\boldsymbol{\beta},\boldsymbol{\theta})$ to the estimated silhouette U_v for the v-th view, while the second term computes the distance from points in the estimated silhouette U_v to the model $\hat{S}(\boldsymbol{\beta},\boldsymbol{\theta})$. As in [26], an L_1 distance metric is used in the second term to make it more robust to noise, while the first term uses the common L_2 distance. Combined with the 2D joint fitting term, the final energy function is:

$$E_1(\boldsymbol{\beta},\boldsymbol{\theta}) = E_M(\boldsymbol{\beta},\boldsymbol{\theta}) + \sum_{v\in V} E_S(\boldsymbol{\beta},\boldsymbol{\theta};K_v,U_v)\,, \quad (6)$$

We found faster convergence to better solutions was obtained using a hierarchical optimization strategy: firstly fitting SMPL to 2D joints can yield a coarse estimation of pose and shape parameters efficiently, then adding the silhouette fitting term can further improve accuracy.

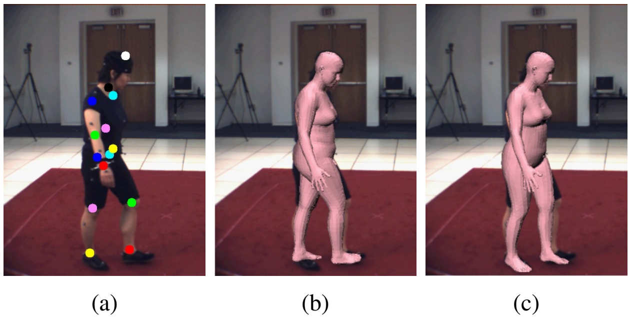

(a) (b) (c)

Figure 3: DCT based temporal prior helps alleviate the leg swap problem. a): Pose detection with leg swap; b): MuVS without DCT prior; c): MuVS with DCT prior.

B. Stage Two: Temporal Fitting

One obvious shortcoming of the algorithm used in the first stage is that it does not take into account the temporal relationship between consecutive frames, while in real life human motions usually present consistency. What is more, due to the lack of texture, occlusion, similarity to the background and other noise, the joint estimator can be erroneous in ambiguous cases. One of these errors is leg swap, which is demonstrated in Figure 3. Sometimes these errors can be difficult to automatically correct in single frame settings. By processing several consecutive frames simultaneously, we can greatly alleviate these types of errors.

To make our algorithm more efficient, in this stage, we do not consider the silhouette, and only use 2D joints. The silhouette's value is in estimating the body shape in the first stage. We study the effect of this choice in our ablation study. Using the obtained median shape $\hat{\boldsymbol{\beta}}$ and pose parameters $\boldsymbol{\Theta}$ from the first stage, we optimize the following objective, which is composed of the 2D joint fitting term and low-dimensional DCT reconstruction term $\boldsymbol{B}$ with corresponding coefficients $\boldsymbol{C}$:

$$\begin{aligned} E_2(\boldsymbol{\Theta}, \boldsymbol{C}; \hat{\boldsymbol{\beta}}, N) = \sum_{n=1}^{N} E_M(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}_n) + \\ \sum_{\text{joint } e} \sum_{d \in \{X,Y,Z\}} \lambda_T E_T(\boldsymbol{C}_{e,d}, \boldsymbol{D}_{e,d}; \hat{\boldsymbol{\beta}}, \boldsymbol{B}, N), \end{aligned} \quad (7)$$

here $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N\}$ is the set of pose parameters for the N frames, $\boldsymbol{C}$ are the corresponding DCT coefficients, $\boldsymbol{D}$ is the collection of all 3D SMPL joints across these frames, while $\boldsymbol{D}_{e,d}$ represents the vector constructed from d-coordinate of the e-th SMPL joints, which is defined as:

$$\boldsymbol{D}_{e,d} = [R_{\theta_1}(J_d(\hat{\boldsymbol{\beta}}))_e, R_{\theta_2}(J_d(\hat{\boldsymbol{\beta}}))_e, \ldots, R_{\theta_N}(J_d(\hat{\boldsymbol{\beta}}))_e]$$

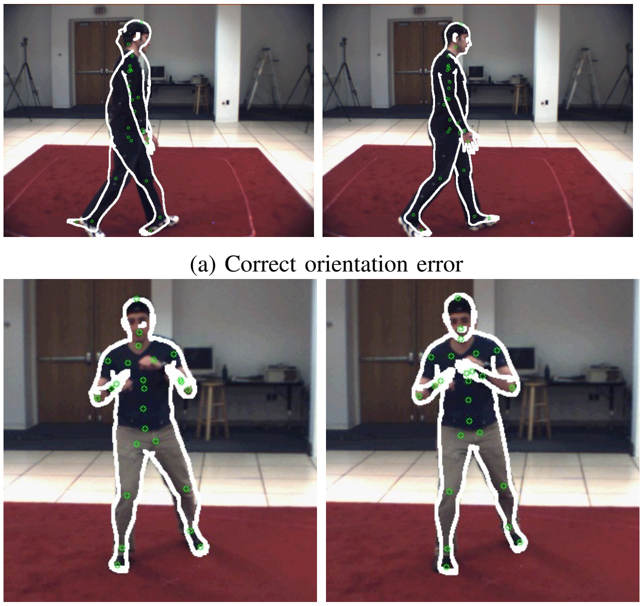
where $e \in \{1, 2, \ldots, N\}$ and $d \in \{X, Y, Z\}$. We encourage the trajectory $\boldsymbol{D}_{e,d}$ across N frames to be well approximated by some low-dimensional DCT basis $\boldsymbol{B}$:

$$E_T(\boldsymbol{c}, \boldsymbol{d}; \boldsymbol{\beta}, \boldsymbol{B}, N) = \sum_{j=1}^{N} \rho_{\sigma_2}(\boldsymbol{d}_j - (\boldsymbol{B}\boldsymbol{c})_j), \quad (8)$$

where ρ is the same function introduced in Eq. 4. Note that the temporal smoothness prior is formulated on the 3D SMPL joint locations.

C. Implementation Details

We implement our entire algorithm in Python. The two involved optimization problems are conducted using Powell's dogleg method [32], OpenDR [29] and Chumpy [1]. In the first stage, all the parameters to optimize are initialized in the same way as [12]. For the second stage, we choose 30 consecutive frames as a unit, and use the first 10 DCT components to act as the bases $\boldsymbol{B}$. For 4 views with 500x500 images, on a normal PC with 12GB RAM and 4 cores, the first stage of our method usually takes around 70 seconds for each frame, while each temporal unit in the second stage takes around 12 minutes. All the weights are empirically chosen by running our method on the training dataset of HumanEva.



(a) Correct orientation error

(b) Better pose

Figure 4: MuVS works better than single-view SMPLify. Left column: results of SMPLify, right column: results of MuVS. The white contour represents the projected mesh.

	Walking			Boxing			Avg
Method	S1	S2	S3	S1	S2	S3	
MuVS^{2}	18.9	19.4	20.7	19.2	13.6	20.0	18.6
MuVS$^{2,\text{ S}}$	14.6	9.6	16.6	15.1	8.5	15.8	13.4
MuVS$^{2,\text{ S, T}}$	14.1	9.3	15.9	15.0	7.4	15.1	12.8
MuVS$^{2,\text{ S, T, H}}$	12.6	5.7	6.9	12.0	**5.4**	**7.8**	8.4
MuVS^{3}	17.6	18.6	20.6	17.2	12.3	19.4	17.6
MuVS$^{3,\text{ S}}$	13.5	9.1	16.0	14.3	7.9	15.8	12.8
MuVS$^{3,\text{ S, T}}$	13.1	8.6	15.3	14.1	5.9	14.9	12.0
MuVS$^{3,\text{ S, T, H}}$	**12.0**	**5.5**	**6.4**	**11.3**	5.8	**7.8**	**8.1**

Table II: Shape estimation error on HumanEva. Error in *mm*.

V. Evaluation

To evaluate the effectiveness of each stage of our method, we perform experiments on two commonly used datasets, HumanEva [42] and Human3.6M [23], and compared with state-of-the-art methods [5], [10], [18], [34], [39], [43]. Both datasets are collected in a controlled lab environment. HumanEva is composed of 4 different subjects and 6 different motions, while Human3.6M collects sequences from 11 subjects, each performing 15 different motions. To keep compatibility with SMPLify, we also use the first 10 shape parameters in all the experiments, and fine tune all the parameters on the training dataset of HumanEva.

A. Ablation study

To analyze the effect of different parts of our algorithm, firstly we performed various ablation experiments on HumanEva. The estimated 3D locations of joints are

	Walking			Boxing			Mean	Median
Method	S1	S2	S3	S1	S2	S3		
MuVS^{2}	59.22	66.81	88.60	79.51	78.68	88.34	76.86	79.10
MuVS$^{2,\text{ S}}$	54.35	56.06	80.95	70.27	72.01	79.01	68.78	71.14
MuVS$^{2,\text{ S, T}}$	50.14	56.11	79.55	68.96	71.73	78.45	67.49	70.35
MuVS$^{2,\text{ S, T, H}}$	39.28	45.81	64.63	55.12	56.49	57.09	53.07	55.81
MuVS^{3}	52.50	62.76	82.51	72.86	73.10	80.42	70.69	72.98
MuVS$^{3,\text{ S}}$	47.21	52.72	75.04	64.88	68.39	71.98	63.37	66.64
MuVS$^{3,\text{ S, T}}$	43.11	53.37	73.56	64.00	67.94	71.44	62.23	65.97
MuVS$^{3,\text{ S, T, H}}$	**35.51**	**44.22**	**61.30**	**49.67**	**53.89**	**51.37**	**49.33**	**50.52**

Table I: Ablation results on HumanEva. 3D joint errors in *mm*. Here labels 2/3 mean using the first 2/3 camera views; S means silhouette fitting term; T means temporal fitting term; and H means adding silhouette fitting term at the second stage. The same notation is used in the rest of the paper.

compared with that of the ground-truth; unlike other work, no similarity transformation is used unless stated. Error is reported in *mm* and the results are shown in Table I. Note that adding the silhouette term in the second stage yields better 3D pose estimation by a large margin, but at the cost of consuming much more running time. To make our algorithm comparable with other methods in running-time, we do not use the silhouette term in the temporal fitting stage.

Effect of multi-view. Clearly multiple views can provide more information about the underlying human body. To verify this, we run our algorithm on HumanEva using different numbers of views. As indicated by the result, adding more views consistently improves 3D pose estimation. As shown in Figure 4 using multiple views helps eliminate incorrectly estimated orientation and improves pose estimation accuracy.

Effect of silhouette fitting. Then we conducted experiments to validate the effectiveness of the silhouette term in our method. Adding silhouettes consistently improves both 3D pose and shape estimation accuracy.

Effect of DCT based temporal prior. Adding the DCT temporal smoothness term also boosts overall performance. As expected its effect diminishes when more views are added, since in this case quite good results can be obtained in the first stage.

Effect on shape estimation. To verify the effect of the aforementioned factors on body shape estimation, we run our method on the validation motion sequences of HumanEva, and compare the estimated meshes with those obtained by MoSh [27]. Prior work by Loper et al. [27] shows the generated reference meshes are quite accurate. As evidenced in Table II, adding silhouette information and the DCT temporal prior consistently improves body shape estimation. With 3 views, the average vertext-to-vertex distance is as low as 12 *mm* without the silhouette term and around 8 *mm* with it.

B. Quantitative comparison

HumanEva: We follow the standard practice of evaluating on the "Walking" and "Boxing" sequences of subjects 1, 2 and 3. As in SMPLify [12], the gender of the subject is assumed known and a gender-specific shape model is used for each motion sequence. The result is shown in Table III. Here *General* means the method is trained on the training dataset of HumanEva, instead of separately training the model for each specific subject, which is referred to *Specific*. For the *General* case, we use the joint regressor distributed with SMPL to obtain 3D joints, and directly compare these with the ground truth joint locations. For the *Specific* case, we use the joint regressor trained on HumanEva with MoSh, which is provided in SMPLify [12]. Then as in [39], we compute the displacement between the estimated joint location and ground-truth in the first frame, then compensate for this difference in the remaining frames.

In the *General* case, with only 2 views, our method is more accurate than all the other methods using all 3 views. With 3 views we obtain a significant improvement relative to the second best method (55.52 vs 63.25). Our method also achieves the lowest error in the *Specific* case. Another advantage of our method over the state-of-the-art is that we return a highly realistic body mesh together with skeleton joints. Though the method proposed by Rhodin et al. [39] also yields a blob-based 3D mesh, we argue that the underlying SMPL model we use is more realistic. A qualitative comparison between our results and those of [39] are shown in Figure 1. For more results please refer to our supplementary materials.

Human3.6M: To further validate the generality and usefulness of MuVS, we also evaluate it on Human3.6M [23]. Human3.6M is the largest public dataset for pose estimation, composed of a wide range of motion types, some of them being very challenging. We use the same parameters trained on HumanEva, then apply MuVS on all the 4 views of subjects S9 and S11. We compare it with SMPLify [12] and other state-of-the-art multi-view pose estimation methods [34]. The result is shown in Table IV. The multi-view version is significantly more accurate than SMPLIfy and

		Walking			Boxing			Mean	Mean (all)
Method	Trained on	S1	S2	S3	S1	S2	S3		
Rhodin et al. [39]	General	74.9				**59.7**		67.3	
Sigal et al. [43]		66.0						66.0	
Belagiannis et al. [10]		68.3				62.7		65.5	
Elhayek et al. [18]		66.5				60.0		63.25	
MuVS$^{2, S, T}$		**50.14**	56.11	79.55	68.96	71.73	78.45	**60.94**	67.49
MuVS$^{3, S, T}$		**43.11**	53.37	73.56	64.00	67.94	71.44	**55.52**	62.23
Amin et al. [5]	Specific	54.5				47.7		51.10	
Rhodin et al. [39]		54.6				**35.1**		44.85	
MuVS$^{3, S, T}$		**33.72**	36.78	60.11	46.85	49.92	46.99	**41.82**	45.73

Table III: Quantitative comparison on HumanEva. 3D joint errors in *mm*.

	Walking			Boxing			Avg
Method	S1	S2	S3	S1	S2	S3	
SMPLify[12]	73.3	59.0	99.4	82.1	79.2	87.2	79.9
MuVS$^{1, S, Sim}$	51.3	48.2	**80.9**	68.4	80.7	88.7	69.7
MuVS$^{1, S, T, Sim}$	**51.2**	**48.1**	81.6	**61.5**	**78.3**	**82.6**	**67.2**

Table V: Comparison with SMPLify on monocular videos from HumanEva in mm. Here Sim means using Procrustes analysis per frame, as with SMPLify.

our 3D joint estimation accuracy is quite close to that of [34], which is concurrent with our work. While they only focus on 3D joint estimation, we address 3D pose and shape estimation at the same time. Our method not only returns 3D joint estimates, but also a realistic body shape model that is faithful to the subjects and which is ready for later modification and animation.

VI. Pose and Shape from Monocular Video

Though we focus on multi-view pose and shape estimation, our method can be applied to monocular video sequences without large modifications, while still being fully automatic. Note manually initialized pose is required for the method in [39] to work on monocular data.

We compare our method with SMPLify on the first camera view of HumanEva, and the result is shown in Table V. Of course given only a single video, it is hard to apply the DCT constraint in depth, since we do not have any trustable evidence in that dimension. Empirically we find our method can still return quite promising results when the performer does not move much in depth. For the videos where no camera information is provided, we manually set the focal length and other imaging parameters to some common value as done in [12]. We qualitatively evaluate our method on some videos downloaded from YouTube, and show the results for specific frames in Figure 5. Figure 6 shows the reconstructed mesh sequence of one of the videos. For the full video, please refer to our supplementary materials.

VII. Conclusion and Future Work

In this paper we present a new marker-less motion capture system, MuVS, that extends SMPLify in a principled and straightforward way. Our method computes relatively accurate 3D pose and also returns a realistic and faithful human body mesh. Unlike previous work that assumes known silhouettes, needs user intervention, or limits the user motion, our algorithm works for general activities seen in daily life. Evaluation on public benchmarks validates the effectiveness and generality of our method. Additionally we apply the approach to monocular video sequences, and achieve promising results.

Future work will address more complex scenarios, like cluttered backgrounds, multiple people, and extreme poses. A key direction to make the method practical is to reduce the computational costs. Finally, other body parts, like faces, hands and feet could be easily combined into our model.

Acknowledgements. We thank the anonymous reviewers for their insightful comments, Naureen Mahmood for help with the figures, Ahmed Osman for voice recording, and Tianye Li and Timo Bolkart for proofreading.

References

[1] Chumpy. http://chumpy.org.

[2] Cmu mocap dataset. http://mocap.cs.cmu.edu.

[3] N. Ahmed, E. De Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Automatic generation of personalized human avatars from multi-view video. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 257–260. ACM, 2005.

[4] I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear spatiotemporal basis models. *ACM Transactions on Graphics (TOG)*, 31(2):17, Apr. 2012.

[5] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3d human pose estimation. In *BMVC*. Citeseer, 2013.

[6] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 408–416. ACM, 2005.

[7] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29. Springer, 2008.

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases	Sit
SMPLify [12]	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3	100.3
MuVS$^{4, S, T, Sim}$	**35.05**	**39.22**	**38.59**	**37.35**	**59.16**	**46.07**	**40.52**	**38.47**	**60.07**
Tekin et al. [47]	102.41	147.72	88.83	125.28	118.02	182.73	112.38	129.17	138.89
MuVS$^{4, S, T}$	44.32	**46.99**	51.75	44.99	67.68	54.56	49.25	**48.90**	**72.82**
Pavlakos et al.[34]	**41.18**	49.19	**42.79**	**43.44**	**55.62**	**46.91**	**40.33**	63.68	97.56
	SitDown	Smoking	Waiting	WalkDog	Walk	WalkTogether	Mean	Median	
SMPLify [12]	137.3	83.4	77.3	79.7	86.8	81.7	82.3	69.3	
MuVS$^{4, S, T, Sim}$	**69.70**	**56.24**	**67.91**	**46.91**	**38.00**	**33.15**	**47.09**	**40.52**	
Tekin et al. [47]	224.9	118.42	138.75	126.29	55.07	65.76	124.97	125.28	
MuVS$^{4, S, T}$	**76.51**	63.70	116.24	55.44	42.94	**37.24**	58.22	51.75	
Pavlakos et al.[34]	119.90	**52.12**	**42.68**	**51.93**	**41.79**	39.37	**56.89**	**46.91**	

Table IV: Quantitative comparison with SMPLify, the methods of Tekin et al. [47] and Pavlakos et al. [34] on H3.6M dataset in *mm*. The accuracy of our method is comparable with that of the recent method proposed in [34].

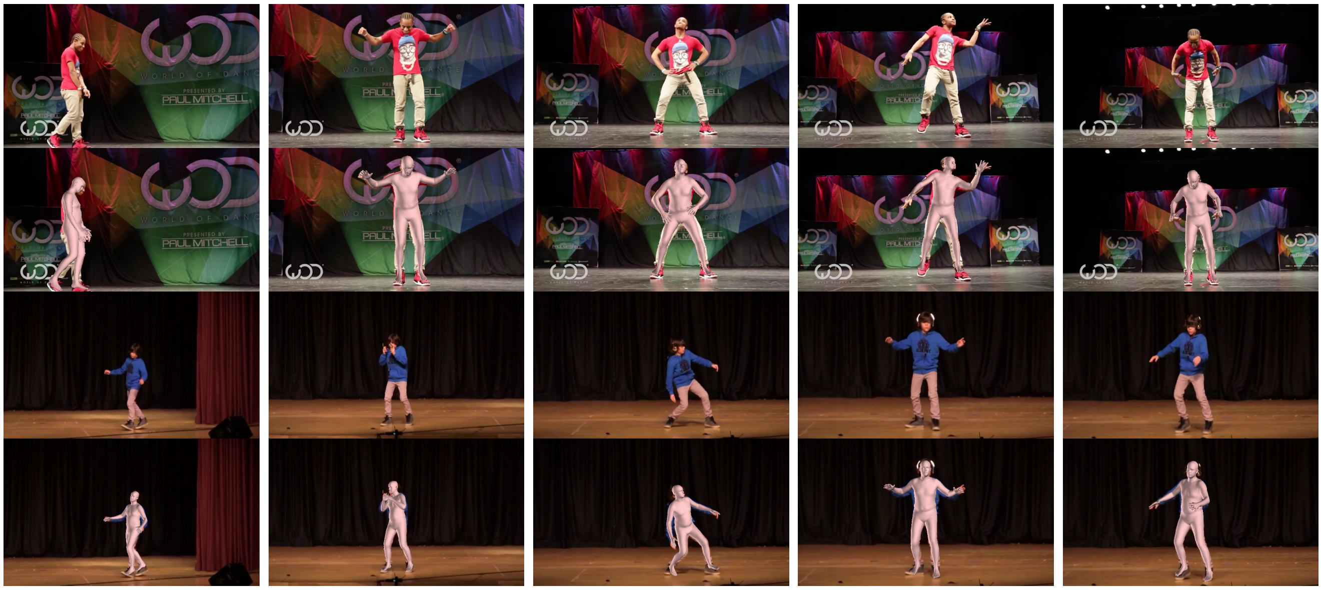


Figure 5: Monocular pose estimation results on videos downloaded from YouTube.

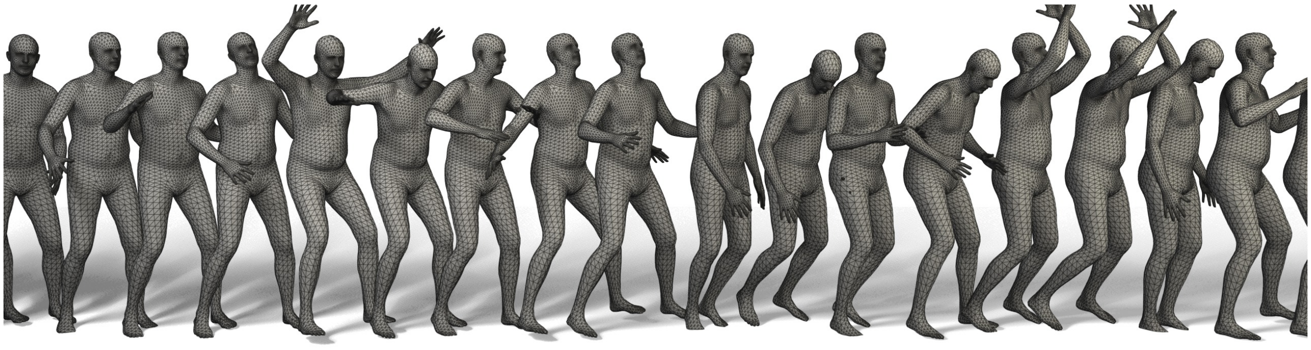


Figure 6: Demonstration of generated meshes for a monocular motion sequence from YouTube.

[8] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[9] L. Ballan and G. M. Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. *3DPVT, Atlanta, GA, USA*, 37, 2008.

[10] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, 2014.

[11] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1):28–52, 2010.

[12] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[13] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017.

[14] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics (TOG)*, volume 27, page 98. ACM, 2008.

[15] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 126–133. IEEE, 2000.

[16] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61(2):185–205, 2005.

[17] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. Marker-less 3d human motion capture with monocular image sequence and height-maps. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[18] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3810–3818. IEEE, 2015.

[19] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *International journal of computer vision*, 87(1-2):75–92, 2010.

[20] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1746–1753. IEEE, 2009.

[21] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1823–1830. IEEE, 2010.

[22] S. Ilic and P. Fua. Implicit meshes for surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):328–333, 2006.

[23] 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 transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2014.

[24] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. Moviereshape: Tracking and reshaping of humans in videos. In *ACM Transactions on Graphics (TOG)*, volume 29, page 148, 2010.

[25] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view body part recognition with random forests. In *2013 24th British Machine Vision Conference, BMVC 2013; Bristol; United Kingdom; 9 September 2013 through 13 September 2013*. British Machine Vision Association, 2013.

[26] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017.

[27] M. Loper, N. Mahmood, and M. J. Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014.

[28] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015.

[29] M. M. Loper and M. J. Black. Opendr: An approximate differentiable renderer. In *European Conference on Computer Vision*, pages 154–169. Springer, 2014.

[30] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, July 2017.

[31] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. *arXiv preprint arXiv:1611.09010*, 2016.

[32] J. Nocedal and S. Wright. Numerical optimization: Springer science & business media. *New York*, 2006.

[33] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[34] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[35] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR'16*.

[36] R. Plankers and P. Fua. Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(CVLAB-ARTICLE-2003-003):63–83, 2003.

[37] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. *arXiv preprint arXiv:1701.08985*, 2017.

[38] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. *Computer Vision–ECCV 2012*, pages 573–586, 2012.

[39] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European Conference on Computer Vision*, pages 509–526. Springer, 2016.

[40] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 765–773, 2015.

[41] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in neural information processing systems*, pages 1337–1344, 2007.

[42] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4–27, 2010.

[43] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International journal of computer vision*, 98(1):15–48, 2012.

[44] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A joint model for 2d and 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3634–3641, 2013.

[45] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 915–922. IEEE, 2003.

[46] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *2011 International Conference on Computer Vision*, pages 951–958. IEEE, 2011.

[47] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3d body poses from motion compensated sequences. *arXiv preprint arXiv:1511.06692*, 2015.

[48] M. Trumble, A. Gilbert, A. Hilton, and J. Collomosse. Deep convolutional networks for marker-less human pose estimation from multiple views. In *Proceedings of CVMP 2016. The 13th European Conference on Visual Media Production*, 2016.

[49] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)*, volume 27, page 97. ACM, 2008.

[50] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[51] C. Wu, K. Varanasi, and C. Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *European Conference on Computer Vision*, pages 757–770. Springer, 2012.

[52] W. Xu, C. Avishek, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *arXiv preprint arXiv:1708.02136*, 2017.

[53] A. Yao, J. Gall, L. V. Gool, and R. Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *Advances in Neural Information Processing Systems*, pages 1359–1367, 2011.

[54] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4956, 2016.

[55] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4966–4975, 2016.