MonoPerfCap: Human Performance Capture From Monocular Video

WEIPENG XU, AVISHEK CHATTERJEE, and MICHAEL ZOLLHÖFER, Max Planck Institute for Informatics
HELGE RHODIN, EPFL
DUSHYANT MEHTA, HANS-PETER SEIDEL, and CHRISTIAN THEOBALT,
Max Planck Institute for Informatics

Fig. 1. We present the first marker-less approach for temporally coherent performance capture given just monocular video as input. The reconstructed surface model captures the full articulated motion of the human body as well as medium-scale non-rigid deformations of the surface.

We present the first marker-less approach for temporally coherent 3D performance capture of a human with general clothing from monocular video. Our approach reconstructs articulated human skeleton motion as well as medium-scale non-rigid surface deformations in general scenes. Human performance capture is a challenging problem due to the large range of articulation, potentially fast motion, and considerable non-rigid deformations, even from multi-view data. Reconstruction from monocular video alone is drastically more challenging, since strong occlusions and the inherent depth ambiguity lead to a highly ill-posed reconstruction problem. We tackle these challenges by a novel approach that employs sparse 2D and 3D human pose detections from a convolutional neural network using a batch-based pose estimation strategy. Joint recovery of per-batch motion allows us to resolve the ambiguities of the monocular reconstruction problem based on a low-dimensional trajectory subspace. In addition, we propose refinement of the surface geometry based on fully automatically extracted silhouettes to enable medium-scale non-rigid alignment. We demonstrate state-of-the-art performance capture results that enable exciting applications such as video editing and free viewpoint video, previously infeasible from monocular video. Our qualitative and quantitative evaluation demonstrates that our approach significantly outperforms previous monocular methods in terms of accuracy, robustness, and scene complexity that can be handled.

This work is supported by ERC StartingGrant "CapReal" (335545).
Authors' addresses: W. Xu, A. Chatterjee, M. Zollhöfer, D. Mehta, H.-P. Seidel, and C. Theobalt, Max Planck Institute for Informatics, Saarland Informatics Campus, Campus E1 4, 66123 Saarbrücken, Germany; emails: {wxu, achatter, mzollhoef, dmetha, hpseidel, theobalt}@mpi-inf.mpg.de; H. Rhodin, EPFL IC CVLAB, BC 302, Station 14, CH-1015 Lausanne, Switzerland; email: helge.rhodin@epfl.ch.

https://doi.org/10.1145/3181973

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Motion capture**;

Additional Key Words and Phrases: Monocular performance capture, 3D pose estimation, human body, non-rigid surface deformation

ACM Reference format:
Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. MonoPerfCap: Human Performance Capture From Monocular Video. *ACM Trans. Graph.* 37, 2, Article 27 (May 2018), 15 pages.
https://doi.org/10.1145/3181973

1 INTRODUCTION

Marker-free human performance capture has been a highly relevant and challenging research topic in the computer vision and computer graphics communities for the past decade. Its goal is to track the motion of a moving subject and reconstruct a temporally coherent representation of its dynamically deforming surface from unmodified videos. Capturing the motion of humans is a challenging problem due to the high level of articulation, potentially fast motion, and considerable non-rigid deformations. A robust and highly accurate solution to this problem is a necessary precondition for a broad range of applications in not only computer animation, visual effects, and free-viewpoint video, but also other fields, such as medicine or biomechanics. Especially, with the recent popularity of virtual reality (VR) systems and telepresence, there comes a rising demand of lightweight performance capture solutions.

In the literature of marker-less performance capture, multi-view methods (Bray et al. 2006; Brox et al. 2006, 2010; Cagniart et al. 2010; De Aguiar et al. 2008; Gall et al. 2009; Liu et al. 2011; Mustafa et al. 2015; Vlasic et al. 2008; Wu et al. 2013) have been well studied. These techniques allow us to obtain accurate results,

but require expensive dense camera setups and controlled studios that are only available to a few technical experts. With the recent commoditization of RGB-D sensors such as the Microsoft Kinect, many depth-based approaches (Bogo et al. 2015; Shotton et al. 2011) demonstrate the possibility of low-cost performance capture with commodity hardware. Even real-time tracking (Zollhöfer et al. 2014) and reconstruction (Innmann et al. 2016; Newcombe et al. 2015) of general deforming objects has been demonstrated. While reliable in controlled indoor settings, the active depth sensing modality of such devices hinders their application in direct sunlight. Given their higher energy consumption, they are not as widely distributed as standard RGB cameras, especially on mobile devices. Furthermore, depth-based approaches cannot be applied to existing video footage, e.g., from YouTube.

In this article, we tackle the problem of human performance capture from monocular RGB video sequences (see Figure 1), even outdoors with general background, to overcome the outlined limitations of depth cameras and multi-view setups. Reconstruction from monocular video per se is a highly challenging and ill-posed problem due to strong occlusions and the lack of depth information. Although several recent works target monocular tracking and reconstruction of specific, e.g., human faces (Garrido et al. 2016; Thies et al. 2016), and general (Garg et al. 2013; Russell et al. 2014; Yu et al. 2015) deformable surfaces, they only target objects undergoing relatively small deformations. To the best of our knowledge, our approach is the first to handle the problem of automatic 3D full human body performance capture from monocular video input. Similar to many existing performance capture approaches, ours employs an actor-specific template mesh. The deformation of the template mesh, obtained by image-based reconstruction prior to recording, is parameterized with a kinematic skeleton and a medium-scale deformation field. Given this shape representation, we estimate the deformation of the actor for each frame in the input video, such that the deformed template closely matches the input frame. The resulting algorithm allows us to generate a temporally coherent surface representation of the actor's full-body performance.

To robustly capture the fast and highly articulated motion of the human body, we leverage 2D discriminative joint predictions from a convolutional neural network (CNN) as landmarks for registering the 3D skeleton to the image. However, due to the lack of explicit depth input, 3D pose estimation suffers from a "forward/backward flipping" ambiguity at the revolute joints (Sminchisescu and Triggs 2003b). Therefore, the estimated 3D pose is often incorrect, even though the 2D projections of the skeleton joints accurately match the predictions. We tackle the flipping ambiguity with the help of a second CNN, which is trained to regress 3D joint positions from monocular images. To further resolve the inherent depth ambiguity of the monocular reconstruction problem, we constrain the 3D poses in temporal space with a low-dimensional linear trajectory subspace, which has proven effective in the context of non-rigid structure from motion (Park et al. 2015). In addition, we compute a non-rigid deformation field based on automatically extracted silhouettes to capture non-rigid surface deformation due to loose clothing, and accurately overlay the deformed template mesh onto the input image frames.

In summary, our monocular performance capture approach has the following main contributions:

- The first human 3D performance capture approach that relies only on monocular video input,
- a combination of discriminative 2D and 3D detections and batch-based motion optimization to solve the inherent flipping ambiguities of monocular 3D pose estimation,
- plausible recovery of non-rigid surface deformations with automatically extracted monocular silhouettes,
- a benchmark dataset consisting of around 40k frames, which covers a variety of different scenarios.

2 RELATED WORK

Performance capture has received considerable attention in computer vision and computer graphics. Here, we focus on the works that are most related to our approach.

Multi-view performance capture. Detailed surface geometry can be reconstructed using shape-from-silhouette and stereo constraints from multi-view footage (Matusik et al. 2000; Starck and Hilton 2007; Waschbüsch et al. 2005) and based on photometric stereo in a light stage (Vlasic et al. 2009). These model-free approaches require succeeding surface tracking to obtain temporal correspondence, e.g., using Cagniart et al. (2010). Reconstructions with temporally consistent topology are obtained with model-based solutions that deform an actor-specific shape template to match silhouette and multi-view constraints (Bradley et al. 2008; Carranza et al. 2003; De Aguiar et al. 2008). Incorporation of a kinematic skeleton model further regularizes the solution (Gall et al. 2009; Liu et al. 2011; Vlasic et al. 2008; Wu et al. 2012), and combined reconstruction and segmentation further improves accuracy (Bray et al. 2006; Brox et al. 2006, 2010; Liu et al. 2011; Mustafa et al. 2015; Wu et al. 2013). The required actor model can be computed fully automatically using parametric models (Anguelov et al. 2005; Balan et al. 2007; Hasler et al. 2010; Loper et al. 2014, 2015; Plänkers and Fua 2001; Sminchisescu and Triggs 2003a; Song et al. 2016), also in general environments (Rhodin et al. 2016). These methods obtain high quality under controlled studio conditions, often with a green screen, but they do not work in general outdoor scenarios and the utilized multi-view and stereo constraints do not generalize to performance capture from a single consumer-level camera.

Depth-based performance capture. Modern RGB-D sensors simultaneously capture synchronized color and depth at real-time frame rates. This triggered the development of depth-based reconstruction approaches that fit articulated template models (Bogo et al. 2015; Guo et al. 2015; Helten et al. 2013; Li et al. 2009; Zhang et al. 2014) that overcome many of the ambiguities of monocular RGB techniques. Even real-time template-based non-rigid tracking (Zollhöfer et al. 2014) and template-free reconstruction (Innmann et al. 2016; Newcombe et al. 2015; Xu et al. 2015) of general deforming scenes has been demonstrated. Multi-view depth-based reconstruction obtains even higher accuracy and robustness (Collet et al. 2015; Dou et al. 2013, 2016; Wang et al. 2016; Ye et al. 2012). While very reliable indoors, the active sensing modalities of consumer-level depth sensors hinders their application in

direct sunlight, their high-energy consumption is a drawback for mobile applications, and they are not yet as widely distributed as RGB cameras, which are already integrated in every smartphone. Passive stereo depth estimation helps to overcome some of these limitations (Plänkers and Fua 2001; Wu et al. 2013), but the required camera baseline is impractical for consumer-level applications and the quality of estimated depth is highly dependent on the amount of texture features in the reconstructed scene.

Sparse skeletal pose reconstruction. We make use of current advances in skeleton pose estimation, in particular, from single views to bootstrap our surface reconstruction approach. Motion capture solutions based on a generative image formation model require manual initialization (Wren et al. 1997) and pose correction. Wei and Chai (2010) obtain high-quality 3D poses from challenging sport video sequences using physical constraints, but require manual joint position annotations for each keyframe (every 30 frames). Also, simpler temporal priors have been applied (Sidenbladh et al. 2000; Urtasun et al. 2005, 2006). With recent advances in convolutional neural networks (CNNs), fully automatic, high-accuracy 2D pose estimation (Jain et al. 2014; Newell et al. 2016; Pishchulin et al. 2016; Toshev and Szegedy 2014; Wei et al. 2016) is feasible from a single image. Lifting the 2D detections to the corresponding 3D pose is common (Akhter and Black 2015; Li et al. 2015; Mori and Malik 2006; Simo-Serra et al. 2012; Taylor 2000; Wang et al. 2014; Yasin et al. 2016) but is a hard and underconstrained problem (Sminchisescu and Triggs 2003b). Bogo et al. (2016) employ a pose prior based on a mixture of Gaussians in combination with penetration constraints. The approach of Zhou et al. (2015) reconstructs 3D pose as a sparse linear combination of a set of example poses. Direct regression from a single image to the 3D pose is an alternative (Ionescu et al. 2014a; Li and Chan 2014; Mehta et al. 2016; Pavlakos et al. 2016; Tekin et al. 2016; Zhou et al. 2016a), but it leads to temporally incoherent reconstructions.

Promising are hybrid approaches that combine discriminative 2D- (Elhayek et al. 2015) and 3D-pose estimation techniques (Rosales and Sclaroff 2006; Sminchisescu et al. 2006) with generative image formation models, but these approaches require multiple views of the scene. Recently, a real-time 3D human pose estimation approach has been proposed (Mehta et al. 2017), which also relies on monocular video input. It is a very fast method, but it does not achieve the temporal stability and robustness to difficult poses of our approach. In contrast to this previous work, our method not only estimates the 3D skeleton more robustly, by leveraging the complimentary strength of 2D and 3D discriminative models, and trajectory subspace constraints, but also recovers medium-scale non-rigid surface deformations that cannot be modeled using only skeleton subspace deformation. We extensively compare to the approach of Mehta et al. (2017) in Section 6.

Dense monocular shape reconstruction. Reconstructing strongly deforming non-rigid objects and humans in general apparel given just monocular input is an ill-posed problem. By constraining the solution to a low-dimensional space, coarse human shape can be reconstructed based on a foreground segmentation (Chen et al. 2010; Grest et al. 2005; Guan et al. 2009; Jain et al. 2010; Rogge et al. 2014; Zhou et al. 2010). Still, these approaches rely on manual initialization and correction steps. Fully automatic approaches combine generative body models with discriminative pose and shape estimation, e.g., conditioned on silhouette cues (Sigal et al. 2007) and 2D pose (Bogo et al. 2016) but can also only capture skin-tight clothing without surface details. The recent work of Huang et al. (2017), which fits a parametric human body model to the 2D pose detection and the silhouettes over time, has demonstrated compelling results on both multi-view and monocular data. But again, their method is not able to model loose clothing. Model-free reconstructions are based on rigidity and temporal smoothness assumptions (Garg et al. 2013; Russell et al. 2014) and only apply to medium-scale deformations and simple motions. Template-based approaches enable fast sequential tracking (Bartoli et al. 2015; Salzmann and Fua 2011; Yu et al. 2015), but they are unable to capture the fast and highly articulated motion of the human body. Automatic monocular performance capture of more general human motion is still an unsolved problem, especially if non-rigid surface deformations are taken into account. Our approach tackles this challenging problem.

3 METHOD OVERVIEW

Non-rigid 3D reconstruction from monocular RGB video is a challenging and ill-posed problem, since the subjects are partially visible at each time instance and depth cues are implicit. To tackle the problem of partial visibility, similar to many previous works, we employ a template mesh, pre-acquired by image-based monocular reconstruction of the actor in a static pose. When it comes to the scenario of capturing the full motion of a human body, the problem is even more challenging, due to the high degree of non-rigidity of humans, ranging from complex articulated motion to non-rigid deformations of skin and apparel. We propose the first marker-less performance capture approach for temporally coherent reconstruction of 3D articulated human motion as well as 3D medium-scale surface deformations from just monocular videos recorded outside controlled studios. To this end, we parameterize human motion based on a two-level deformation hierarchy. On the coarser level, the articulated motion is captured in skeleton deformation space. On the finer level, a deformation field parameterized by an embedded deformation graph models medium-scale non-rigid deformations of the surface. Correspondingly, motion capture is performed in a coarse-to-fine manner, based on two subsequent steps, namely batch-based pose estimation (see Section 4) and silhouette-based refinement (see Section 5). As shown in Figure 2, we first estimate the skeleton deformations in the input video, using a novel batch-based 3D pose estimation approach that exploits the discriminative 2D and 3D detections from trained CNNs and a linear trajectory subspace to obtain robust reconstruction. The resulting temporally coherent reconstructions well reproduce articulated motion but lack non-rigid surface deformations of the apparel and skin. Consequently, there exists a noticeable misalignment between the skeleton-deformed model boundary and the image silhouette. To alleviate this problem, we propose a surface refinement approach to better align the deformed model with automatically estimated actor silhouettes that are found by a model-guided foreground segmentation strategy (see Section 5.1).

To obtain the person-specific template mesh, we first record a high-resolution video with a handheld camera orbiting around the

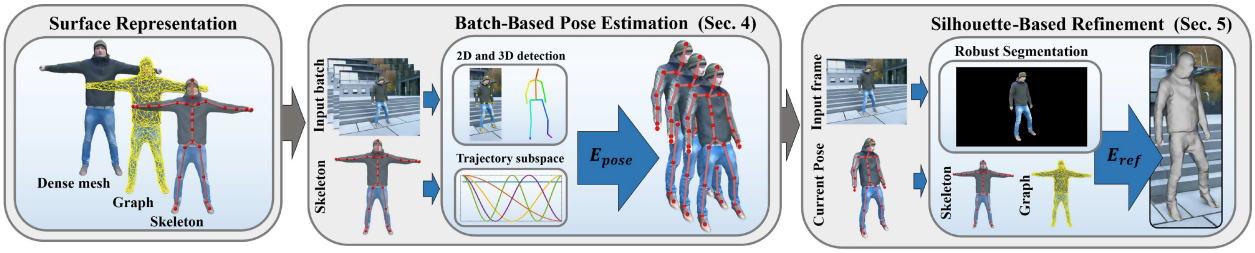

Fig. 2. Given a monocular video and a personalized actor rig, our approach reconstructs the actor motion as well as medium-scale surface deformations. The monocular reconstruction problem is solved by joint recovery of temporally coherent per-batch motion based on a low-dimensional trajectory subspace. Non-rigid alignment based on automatically extracted silhouettes is used to better match the input.

actor standing in a T-pose, and uniformly sample 60 images (see Figure 3). Afterwards, a triangulated surface with the corresponding texture is automatically reconstructed using the image-based reconstruction software Agisoft Photoscan.[1]

4 BATCH-BASED 3D HUMAN POSE ESTIMATION

We parameterize articulated human motion based on a low-dimensional skeleton subspace (Lewis et al. 2000). The skeleton $\mathcal{S} = \{\mathbf{t}, \mathbf{R}, \boldsymbol{\Theta}\}$ with $N_d = 16$ joints J_i is parameterized by the position $\mathbf{t} \in \mathbb{R}^3$ and rotation $\mathbf{R} \in \mathbf{SO}(3)$ of its root joint, and 27 angles stacked in $\boldsymbol{\Theta} \in \mathbb{R}^{27}$. This leads to a 33-dimensional deformation subspace. The high-resolution actor mesh is rigged to the skeleton based on dual quaternion skinning (Kavan et al. 2007).

The skinning weights of our templates are automatically computed using Blender. For the skirt and long coat templates, we manually correct the skinning weights to reduce artifacts.

Given the monocular input video $\mathcal{V} = \{I_f\}_{f=1}^{N}$ with N image frames I_f, the goal of 3D pose estimation is to recover the skeleton parameters $\mathcal{S}_f$ for all input frames. Since the problem of 3D human pose estimation is highly underconstrained given only a single RGB input frame I_f, we propose a novel batch-based approach that jointly recovers the motion for a continuous window in time:

$$\mathcal{B} = \{\mathcal{S}_f \mid f_{start} \leq f \leq f_{end}\}, \tag{1}$$

where f_{start} specifies the index of the first and f_{end} of the last frame included in the current batch. In all our experiments, a constant batch size $|\mathcal{B}| = 50$ is used, and the input video is partitioned into a series of overlapping batches (10 frames overlap). Each batch is processed independently and afterwards the per-batch skeleton reconstruction results are combined in the overlap region based on a linear blending function.

We phrase the problem of estimating the articulated motion of each batch $\mathcal{B}$ as a constrained optimization problem:

$$\begin{aligned} \mathcal{B}^* = \underset{\mathcal{B}}{\text{argmin}} \quad & E_{\text{pose}}(\mathcal{B}), \\ \text{subject to} \quad & \boldsymbol{\Theta}_{min} \leqslant \boldsymbol{\Theta}_f \leqslant \boldsymbol{\Theta}_{max}, \\ & \forall f \in [f_{start}, f_{end}], \end{aligned} \tag{2}$$

where the hard constraints on the per-frame joint angles $\boldsymbol{\Theta}_f$ are physically motivated and ensure the reconstruction of plausible human body poses by forcing the joint angles to stay inside their anatomical lower $\boldsymbol{\Theta}_{min}$ and upper $\boldsymbol{\Theta}_{max}$ bounds (Stoll et al. 2011). The proposed batch-based pose estimation objective function E_{pose} consists of several data fitting and regularization terms:

$$E_{\text{pose}}(\mathcal{B}) = \underbrace{E_{\text{2d}}(\mathcal{B}) + w_{\text{3d}} E_{\text{3d}}(\mathcal{B})}_{\text{data fitting}} + \underbrace{w_d E_{\text{d}}(\mathcal{B})}_{\text{regularization}}. \tag{3}$$

The data-fitting terms ensure that the reconstructed motion closely matches the input: A 2D joint alignment term E_{2d} based on joint detections in image space and a 3D joint alignment term E_{3d} based on regressed 3D joint positions. The discriminative detections are obtained using CNNs that have been trained for 2D and 3D joint localization. The motion of the skeleton is regularized on batch level by E_{d} using a low-dimensional trajectory subspace based on the discrete cosine transform. This enforces the intra-batch motion to be temporally smooth, adds robustness against failed detections and further resolves depth ambiguity. The weights $w_\bullet$ balance the relative importance of the different terms. We provide more details in the remaining part of this section.

Discriminative joint alignment terms. For each input image I_f and each of the $N_d = 16$ joints J_i, we estimate the 2D joint position $\mathbf{d}_{f,i}^{\text{2d}}$ in image space and the 3D joint position $\mathbf{d}_{f,i}^{\text{3d}}$. To this end, we use the Resnet (He et al. 2016)-based CNN joint position regression method of Mehta et al. (2016), to which we add detections for the toes. This better constrains the rotation of the feet and leads to higher quality reconstruction results. Our 2D pose network is trained on the MPII Human Pose (Andriluka et al. 2014) and LSP (Johnson and Everingham 2011) datasets, and the

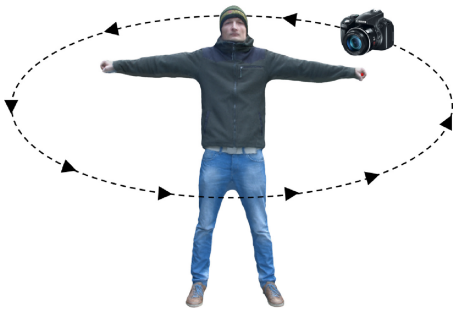

Fig. 3. Acquisition of a textured template mesh from handheld video footage of the actor in a static pose.

[1] http://www.agisoft.com.

3D pose network is fine-tuned from the 2D pose network on the H3.6M (Ionescu et al. 2014a) and 3DHP (Mehta et al. 2016) datasets. Our approach lifts the loose CNN detections to produce a coherent skeleton (parameterized by angles) and enforces constant bone length. In contrast to previous works, e.g., the 2D-to-3D lifting approach of Zhou et al. (2015), we incorporate both 2D and 3D constraints into our generative framework to allow for more robust pose estimation. Our 2D joint alignment term is a re-projection constraint enforcing that the projected joint positions $J_i(\mathcal{S}_f)$ closely match the corresponding 2D detections $\mathbf{d}_{f,i}^{2d}$:

$$E_{2d}(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{\mathcal{S}_f \in \mathcal{B}} \frac{1}{N_d} \sum_{i=1}^{N_d} \left\| \Pi\left(J_i(\mathcal{S}_f)\right) - \mathbf{d}_{f,i}^{2d} \right\|_2^2, \tag{4}$$

where the mapping $\Pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ implements the full perspective camera projection. We apply this constraint independently for every frame of $\mathcal{B}$. In addition, we propose a 3D joint alignment term based on the regressed 3D joint positions $\mathbf{d}_{f,i}^{3d}$:

$$E_{3d}(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{\mathcal{S}_f \in \mathcal{B}} \frac{w_f}{N_d} \sum_{i=1}^{N_d} \left\| J_i(\mathcal{S}_f) - \left(\mathbf{d}_{f,i}^{3d} + \mathbf{t}_f\right) \right\|_2^2. \tag{5}$$

Since the 3D joint detections $\mathbf{d}_{f,i}^{3d}$ are normalized for a skeleton with average bone length and are predicted relative to the root joint, rather than in camera space, they have to be rescaled to match the actor model, and mapped to their corresponding camera space position based on an unknown per-frame global translation $\mathbf{t}_f$. To prune frames with low 3D detection confidence, we measure the per-frame PCK error (Toshev and Szegedy 2014) PCK_f between the 2D joint detections and the projected 3D detections and apply a per-frame binary weight w_f to the 3D data term:

$$w_f = \begin{cases} 1 & \text{if } \text{PCK}_f < thres_{pck}, \\ 0 & \text{else}, \end{cases} \tag{6}$$

where $thres_{pck} = 0.4$ is an empirically determined, but constant, distance threshold. Note, the 2D detections are always included in the optimization, since they have a higher reliability.

Batch-based motion regularization. Up to now, all poses $\mathcal{S}_f$ are temporally independent and sometimes inaccurate, since the monocular reconstruction problem is highly underconstrained. To alleviate this problem, we impose temporal smoothness by forcing the trajectory of each skeleton parameter to lie on a low-dimensional linear subspace. Specifically, we couple all pose estimates $\mathcal{S}_f \in \mathcal{B}$ by minimizing the distance to a $K = 8$-dimensional linear subspace $\mathbf{DCT} \in \mathbb{R}^{K \times |\mathcal{B}|}$ (Park et al. 2015) spanned by the K lowest frequency basis vectors of the discrete cosine transform (DCT):

$$E_d(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \left\| \mathbf{\Lambda} \mathbf{S}_{\mathcal{B}} \, \text{Null}(\mathbf{DCT}) \right\|_F^2. \tag{7}$$

Here, $\text{Null}(\mathbf{DCT})$ denotes the nullspace of the $\mathbf{DCT}$ matrix, and the matrix $\mathbf{S}_{\mathcal{B}}$ stacks all parameters $\mathcal{S}_f$ of the current batch:

$$\mathbf{S}_{\mathcal{B}} = \left[\mathcal{S}_{f_{start}}, \ldots, \mathcal{S}_{f_{end}} \right] \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{B}|}. \tag{8}$$

The diagonal matrix $\mathbf{\Lambda} = \text{diag}([\lambda_\mathbf{t}, \lambda_\mathbf{R}, \lambda_\mathbf{\Theta}])$ balances the motion smoothness of the global translation, rotation and joint angle components. In all our experiments, $\lambda_\mathbf{t} = 1 \cdot \mathbf{1}_3$, $\lambda_\mathbf{R} = 600 \cdot \mathbf{1}_3$ and $\lambda_\mathbf{\Theta} = 600 \cdot \mathbf{1}_{27}$, where $\mathbf{1}_k$ is the k-dimensional row vector of ones. $\| \cdot \|_F$ denotes the Frobenius norm.

Initialization and optimization. The optimization problem proposed in Equation (2) is non-linear due to the involved camera projection and the hierarchical parameterization of articulated motion based on joint angles. We solve this constrained non-linear least squares optimization problem using the Levenberg Marquardt (LM) algorithm provided by Ceres[2].

Since the optimization problem is non-convex, LM requires an initialization close to the global optimum for convergence. To this end, we resort to a per-frame initialization strategy by finding the $\mathcal{S}_f$ that minimize a joint alignment energy function $E_{2d} + E_{3d}$.

5 SILHOUETTE-BASED REFINEMENT

As mentioned before, our batch-based pose optimization does not capture non-rigid surface deformation due to apparel and skin, and thus it leads to misalignments between the skeleton-deformed template mesh and the input images, particularly at the boundaries. To alleviate this problem, we propose a pose and surface refinement method based on automatically extracted silhouettes.

5.1 Automatic Silhouette Extraction

Given an input frame $\mathbf{I}_f$, we estimate the silhouette of the actor through a foreground segmentation method based on GrabCut (Rother et al. 2004). GrabCut requires a user-specified initialization $\mathbf{T} = \{\mathbf{T}_b, \mathbf{T}_{ub}, \mathbf{T}_{uf}, \mathbf{T}_f\}$, where $\mathbf{T}_f$ and $\mathbf{T}_b$ denote the known foreground and background masks and the segmentation is computed over the remaining uncertain foreground $\mathbf{T}_{uf}$ and background $\mathbf{T}_{ub}$ regions. The original GrabCut interactively initializes the masks based on a user-specified bounding box $\mathbf{B}$ and sets $\mathbf{T}_{uf} = \mathbf{T}_{ub} = \mathbf{B}$, $\mathbf{T}_b = \overline{\mathbf{B}}$, and $\mathbf{T}_f = \emptyset$. In contrast, we propose a fully automatic initialization strategy for $\mathbf{T}$ based on the skeleton parameters $\mathcal{S}_f$ obtained by our batch-based pose estimation. To this end, we first rasterize the skeleton and the deformed dense actor template $V(\mathcal{S}_f)$ to obtain two masks $\mathbf{R}$ and $\mathbf{M}$, respectively. Then, we set the masks $\mathbf{T}$ as follows:

$$\begin{aligned} \mathbf{T}_f &= R \cup erosion(\mathbf{M}), \\ \mathbf{T}_b &= \overline{dilation(\mathbf{M})}, \\ \mathbf{T}_{uf} &= \mathbf{M} - \mathbf{T}_f, \\ \mathbf{T}_{ub} &= dilation(\mathbf{M}) - M, \end{aligned} \tag{9}$$

where $erosion(\cdot)$ and $dilation(\cdot)$ denote the image erosion and dilation operator. In Figure 4 (Initialization Mask), the masks $\{\mathbf{T}_b, \mathbf{T}_{ub}, \mathbf{T}_{uf}, \mathbf{T}_f\}$ are illustrated in red, blue, yellow, and green. We show our robust model-based segmentation result on the complete sequence in the accompanying video.

To improve the robustness of our segmentation method, we extend the original GrabCut objective function, by incorporating motion cues. Specifically, we extract the temporal per-pixel color gradients between adjacent frames and encourage neighboring pixels with small temporal gradients to belong to the same region. As shown in Figure 4, our model-based strategy and the extension with motion cues lead to fully automatic and significantly improved segmentation.

[2] http://ceres-solver.org.

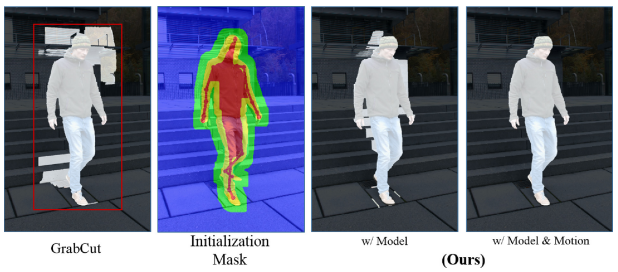

Fig. 4. Our fully automatic model-based initialization (middle/left) significantly improves the segmentation (middle/right) compared to manual initialization based on a bounding box (left). In addition, we use motion cues to further improve the results (right).

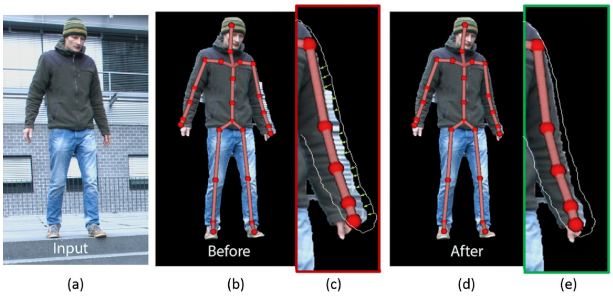

Fig. 5. Our silhouette-based pose refinement improves both pose estimation and the silhouette segmentation. The error in pose estimation causes inaccurate background subtraction near the left arm (b, c). Our silhouette pose refinement pulls the mesh (in white) to the silhouette (c) and therefore moves the arm skeleton leftwards to the correct position (d). The silhouette segmentation is significantly refined after the second silhouette extraction based on the refined pose (e).

5.2 Silhouette-Based Pose Refinement

Our silhouette-based pose refinement is performed in an Iterative Closest Point (ICP) manner. In an ICP iteration, for each boundary point of the projected surface model, we search for its closest point on the image silhouette that shares a similar normal direction. Then, we refine the pose by solving the following non-linear least squares optimization problem:

$$E_{\text{ref}}(\mathcal{S}_f) = \underbrace{E_{\text{con}}(\mathcal{S}_f)}_{\text{data fitting}} + \underbrace{w_{\text{stab}} E_{\text{stab}}(\mathcal{S}_f)}_{\text{regularization}}, \tag{10}$$

where E_{con} aligns the mesh boundary with the input silhouette, E_{stab} constrains the solution to stay close to the batch-based results and w_{stab} balances the importance of the two terms. We initialize the iterative pose refinement with the batch-based pose estimates, and typically perform three iterations.

Silhouette alignment constraint. The closeness of corresponding points is enforced as follows:

$$E_{\text{con}}(\mathcal{S}_f) = \frac{1}{|\mathrm{S}|} \sum_{k \in \mathrm{S}} \left[\mathbf{n}_k^T \cdot \left(\Pi\left(\mathbf{v}_k(\mathcal{S}_f)\right) - \mathbf{s}_k \right) \right]^2, \tag{11}$$

where S is the boundary of the actor model, $\mathbf{v}_k$ the position of vertex k and $\mathbf{s}_k \in \mathbb{R}^2$ the corresponding silhouette point in the image with 2D normal $\mathbf{n}_k$.

Pose stabilization constraint. We enforce the refined skeleton pose to be close to its initialization based on the following soft-constraint:

$$E_{\text{stab}}(\mathcal{S}_f) = \frac{1}{N_d} \sum_{i=1}^{N_d} \left\| J_i(\mathcal{S}_f) - J_i(\hat{\mathcal{S}}_f) \right\|_2^2, \tag{12}$$

where $\hat{\mathcal{S}}_f$ are the joint angles after batch-based pose estimation and $J_i(\cdot)$ computes the 3D position of joint J_i.

After the iterative pose refinement, we perform the silhouette extraction of Section 5.1 for a second time, to further improve the segmentation. As shown in Figure 5, our iterative pose refinement not only improves the pose estimates but also significantly increases the accuracy of the silhouette segmentation, which allows for the more accurate non-rigid surface alignment of Section 5.3.

5.3 Silhouette-Based Non-Rigid Surface Refinement

Given the silhouette segmentation improved by iterative pose refinement, we perform a surface refinement step based on a medium-scale deformation field to closely align the model to the extracted image silhouettes. This captures the non-rigid surface deformations of apparel and skin that are visible in the silhouette outline. Refinement of the interior is hard due to a potential lack of strong photometric cues.

We parameterize the medium-scale warp field using an embedded deformation graph (Sumner et al. 2007). The deformation graph $\mathcal{D}$, consisting of $M \approx 1000$ nodes, is generated from the template mesh using a uniform mesh decimation/simplification strategy. We assign a radius of influence for each deformation node by computing the maximum geodesic distance to its connected graph nodes. Each node defines a local warp field W_i that rotates $\mathbf{R}_i \in \mathbf{SO}(3)$ and translates $\mathbf{t}_i \in \mathbb{R}^3$ points $\mathbf{x} \in \mathbb{R}^3$ in the surrounding space:

$$W_i(\mathbf{x}) = \mathbf{R}_i(\mathbf{x} - \hat{\mathbf{g}}_i) + \hat{\mathbf{g}}_i + \mathbf{t}_i, \tag{13}$$

where $\hat{\mathbf{g}}_i \in \mathbb{R}^3$ is the canonical position of node i, computed with the result of the pose refinement. We refer to the graph and its associated degrees of freedom as

$$\mathcal{D} = \{(\mathbf{R}_i, \mathbf{t}_i) | i \in [0, M)\}. \tag{14}$$

We apply the medium-scale deformation field to the dense actor model by linear blending of the per-node warp fields:

$$\mathbf{v}_i = W(\hat{\mathbf{v}}_i) = \sum_{k \in \mathbf{F}_i} b_{i,k}(\mathbf{x}) \cdot W_k(\hat{\mathbf{v}}_i). \tag{15}$$

Here, $\mathbf{v}_i \in \mathbb{R}^3$ is the deformed vertex position, $\hat{\mathbf{v}}_i \in \mathbb{R}^3$ is the canonical position of vertex i and $\mathbf{F}_i$ is the set of deformation nodes that influence vertex i. We compute the blending weights $b_{i,k}$ based on an exponential distance falloff and make them a partition of unity.

Given the embedded deformation graph, our silhouette-based surface refinement is expressed as the following optimization

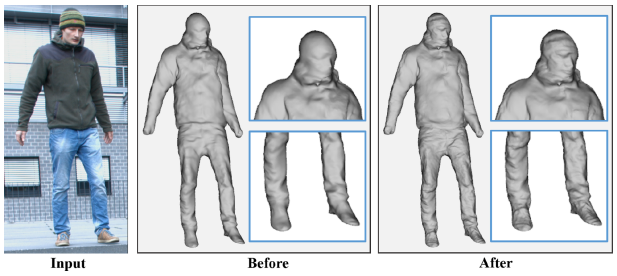

Fig. 6. Our reconstruction results can optionally be refined to add fine-scale surface detail.

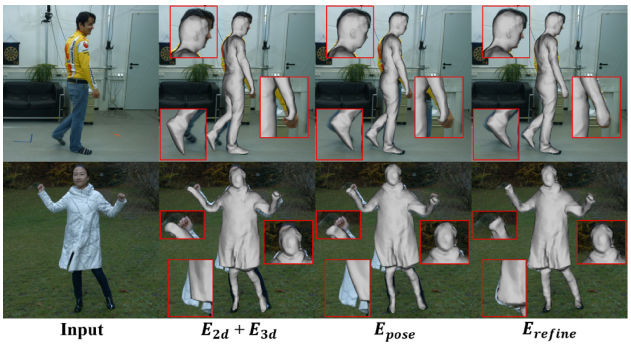

Fig. 7. Qualitative evaluation of components: The batch-based pose optimization E_{pose} significantly improves alignment over the discriminative energy $E_{2\text{d}} + E_{3\text{d}}$. Note the improved rotation of the feet. Residual non-rigid deformations are compensated via surface-based silhouette refinement E_{refine}.

problem:

$$E_{\text{surf}}(\mathcal{D}) = \underbrace{E_{\text{con}}(\mathcal{D})}_{\text{data fitting}} + \underbrace{w_{\text{arap}} E_{\text{arap}}(\mathcal{D})}_{\text{regularization}}. \tag{16}$$

Here, E_{con} is the silhouette alignment term, E_{arap} an as-rigid-as-possible regularization term (Sorkine and Alexa 2007), and w_{arap} balances the two terms.

Our silhouette alignment term E_{con} encourages the actor model to tightly align with the input silhouette:

$$E_{\text{con}}(\mathcal{D}) = \frac{1}{|\mathbf{S}|} \sum_{k \in \mathbf{S}} \left[\mathbf{n}_k^T \cdot \left(\Pi\big(\mathbf{v}_k(\mathcal{D})\big) - \mathbf{s}_k \right) \right]^2, \tag{17}$$

where $\mathbf{S}$ is the model silhouette, $\mathbf{v}_k$ the position of vertex k and $\mathbf{s}_k \in \mathbb{R}^2$ its corresponding silhouette point with normal $\mathbf{n}_k \in \mathbb{R}^2$.

The as-rigid-as-possible term regularizes the non-rigid surface deformation of the graph nodes:

$$E_{\text{arap}}(\mathcal{D}) = \frac{1}{M} \sum_{i=1}^{M} \sum_{j \in \mathcal{N}_i} \left\| (\mathbf{g}_i - \mathbf{g}_j) - \mathbf{R}_i(\hat{\mathbf{g}}_i - \hat{\mathbf{g}}_j) \right\|_2^2. \tag{18}$$

Here, $\mathbf{g}_i = W_i(\hat{\mathbf{g}}_i) = \hat{\mathbf{g}}_i + \mathbf{t}_i$ is the deformed position of node $\hat{\mathbf{g}}_i$ and $\mathcal{N}_i$ is its 1-ring neighbourhood.

Similar to pose refinement (see Section 5.2), we perform surface refinement in an ICP-like manner. To this end, we iterate the correspondence search and the model alignment step two times. We initialize the optimization problem based on the pose refinement result and minimize E_{surf} using the LM algorithm. The final results are obtained by temporally smoothing the per-frame results based on a centered window of five frames.

6 RESULTS

In all experiments, we use the following empirically determined parameters to instantiate our energy functions: $w_{3d} = 0.1$, $w_p = 0.1$, $w_d = 50$, $w_{stab} = 0.06$, and w_{arap} is set to 0.6 and 0.2 for the two ICP iterations, respectively. Our approach proved robust to the specific choice of parameters, and thus we use this fixed set in all experiments. We performed all experiments on a desktop computer with a 3.6GHz Intel Xeon E5-1650 processor. Our unoptimized CPU code requires approximately 1.2 minutes to process one input frame. This divides into 10 seconds for batch-based pose estimation and 1 minute for surface refinement. We believe that the runtime of our approach can be greatly improved based on recent progress in data-parallel optimization (Zollhöfer et al. 2014). In the remaining part of this section, we first discribe the proposed benchmark dataset, then present our qualitative results, then evaluate all components of our approach, and finally compare to state-of-the-art monocular and multi-view approaches quantitatively.

6.1 Benchmark Dataset

To evaluate our monocular performance capture approach for a variety of scenarios, we propose a benchmark dataset consisting of 13 sequences (around 40k frames in total), which divides into the following subsets. First, we captured eight video sequences at 30Hz, which cover a variety of different scenarios including indoor and outdoor settings, handheld and static cameras, natural and man-made environments, male and female subjects, as well as tight and loose garments. Second, to further increase the diversity of human motions of our benchmark dataset, we captured 40 additional actions, including daily actions such as walking, jumping as well as highly challenging ones such as rolling, kicking and falling. Each action is repeated multiple times by three subjects. In total, this leads to 120 video clips in three long video sequences, 7 minutes each.

In addition, we included two sequences from prior works, Robertini et al. (2016) and Wu et al. (2013), in our benchmark dataset. These two sequences provide accurate surface reconstruction from multiview images, which can be used as ground truth for quantitative evaluation.

To evaluate our results in a different view from the one used as input, we also captured a side view using a second camera and calibrated the extrinsic parameters for both cameras. We also provide manually labeled silhouettes for a subset of the benchmark, which are used to compute silhouette overlap as a metric for quantitative evaluation of the performance capture results. The benchmark dataset and our performance capture results will be made publicly available.

6.2 Qualitative Results

Our qualitative results are shown in Figure 8. We refer to the accompanying video for the complete results on our entire benchmark dataset.

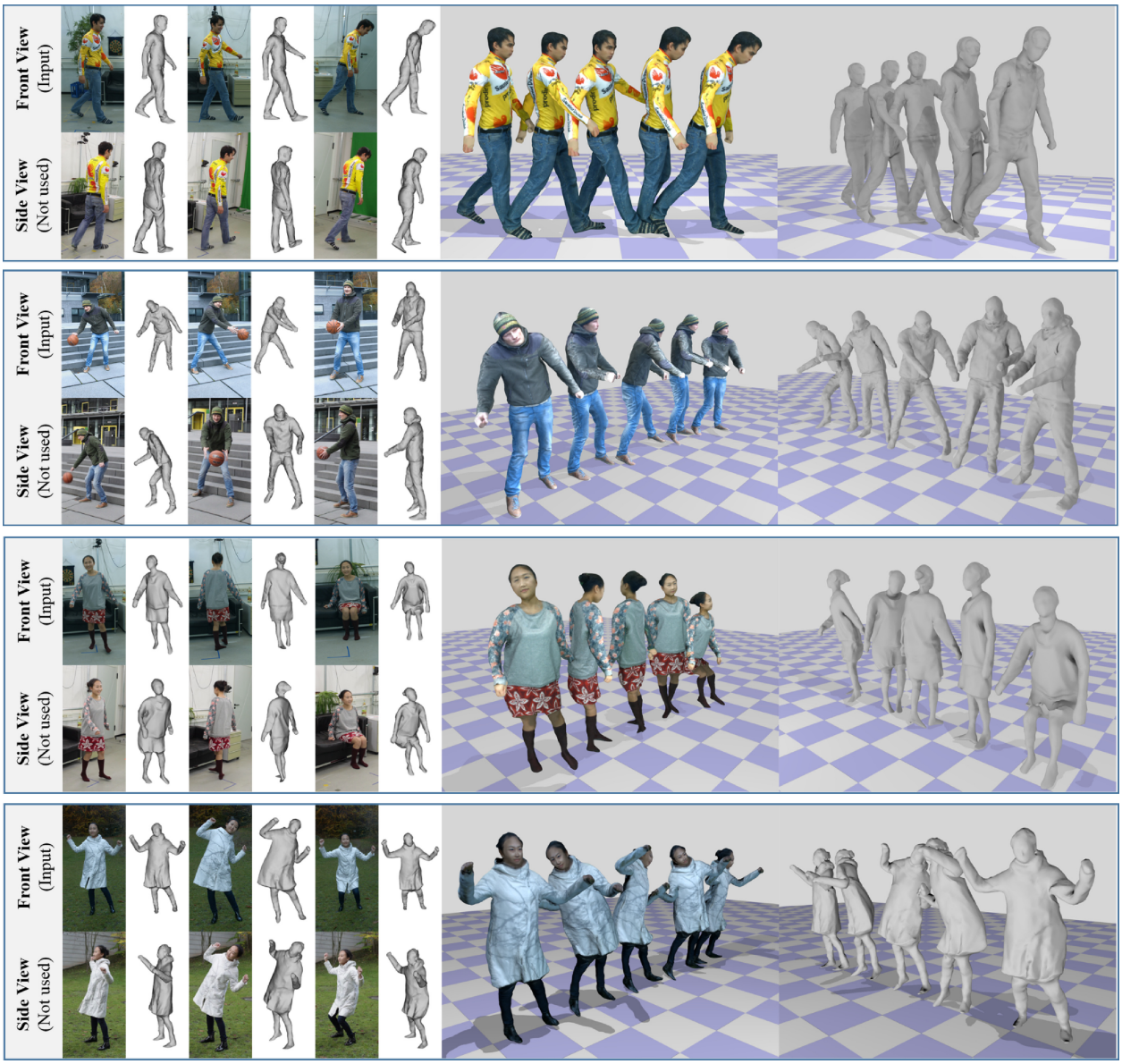

Fig. 8. Qualitative Results: We demonstrate compelling monocular performance capture results on a large variety of challenging scenes (left) that span indoor and outdoor settings, natural and man-made environments, male and female subjects, as well as tight and loose garments. Our reconstructions match the real world even when viewed from the side. Note, the side views are just used for reference and are not used as input to our approach. The reconstructions obtained by our approach are the basis for free-viewpoint video (right). For more results on our benchmark dataset, we refer to the accompanying video.

Our approach accurately captures the performance of the actors and obtains temporally coherent results on all test sequences, even for challenging motions with 360 degrees of rotation and sitting down on a sofa. Even continuous interactions with objects, e.g., a basketball, are allowed. To the best of our knowledge, no previous monocular approach could handle such scenarios. As shown in Figure 9, from the reference view, in spite of a small offset between the projected meshes and the actor due to some ambiguities in monocular depth estimation, our approach is able to accurately recover the full 3D deforming pose and the shape of the actor. The textured spatio-temporal reconstructions are the basis for free-viewpoint videos and can be rendered from arbitrary viewpoints; see Figure 8 (right). Our reconstructed models also allow us to employ shading-based refinement using estimated lighting (Wu et al. 2013) to recover fine-scale surface detail, see Figure 6, but this is not the focus of our work. We also tested our approach on the three long video sequences, 7 minutes each. Our approach is able to continuously track each sequence in one go without restarting. This demonstrates its robustness and generality. Our complete results on all three sequences, which contain very

Fig. 9. Despite a small depth offset between the reconstruction and the actor in the reference view (not used for tracking) due to the remaining monocular depth ambiguity, our approach is able to accurately recover the deforming pose and shape.

challenging motions that bring our approach to its limits, are provided in the second supplementary video. Furthermore, one of the sequences in our benchmark dataset is captured with a handheld camera, demonstrating the effectiveness of our approach even for non-static cameras.

6.3 Evaluation of Algorithmic Components

The three main steps of our approach are: Frame-to-frame 3D skeleton pose initialization based on the 2D/3D predictions (E_{2d} + E_{3d}), batch-based pose estimation (E_{pose}) and silhouette-based pose and surface refinement (E_{refine}). We demonstrate the importance of all steps by comparing the results qualitatively in Figure 7. While the joint-detection-based initialization ($E_{2d} + E_{3d}$) yields plausible results, the following batch-based pose estimation step (E_{pose}), which exploits temporal smoothness, improves the overlay and also removes the temporal jitter of the temporally incoherent 2D/3D CNN joint detections (see the accompanying video). Note, in contrast to our coherent skeletal pose reconstruction, the CNN joint detections do not enforce a temporally constant bone length. Furthermore, our silhouette-based refinement step, which produces our final results, significantly improves the overlay. In addition, we also quantitatively evaluate the contribution of each component. To this end, we made use of the multi-view performance capture method described in Robertini et al. (2016), which has demonstrated convincing results in capturing both pose and surface deformations in outdoor scenes. We select a single view of one of their multi-view sequences (*Pablo* sequence) as a test set, and use their results as ground truth for quantitative evaluation. As shown in Figure 10 (average errors are given in the legend), our three main steps gradually improve the per-frame surface-to-surface mean error, while the complete approach (E_{refine}) has the lowest error over almost all frames. Note that, for this evaluation, we aligned the reconstruction to the ground truth with a translation to eliminate the global depth offsets shown in Figure 9.

Now, we study the effectiveness and robustness of our silhouette extraction method. As shown in Figure 12, benefiting from the model-based initialization and the motion cue, our model-based silhouette extraction method yields significantly more accurate foreground segmentation results than the default GrabCut method. More segmentation comparison is provided in the supplementary video. However, segmentation failure occurs inevitably when the actor is occluded by objects (see Figure 12 row 2) or the

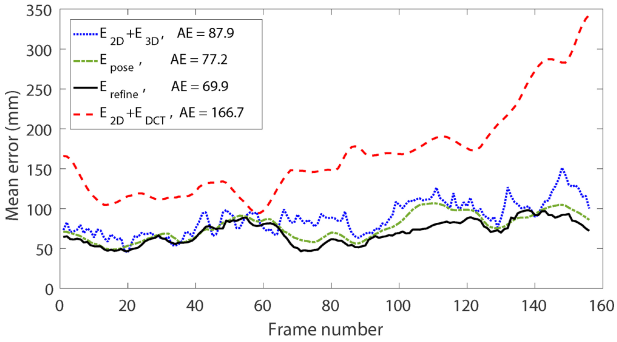

Fig. 10. Quantitative evaluation of components: All steps of our approach improve the surface reconstruction error (in millimeters). The average error (AE) over all frames is given in the legend.

Fig. 11. Qualitative shape comparison on the *Pablo* sequence. Our approach obtains comparable quality as the multi-view approach of Robertini et al. (2016) (eight cameras) and drastically outperforms the template-based monocular tracking approach of Yu et al. (2015). In comparison to rigging the template model with respect to the 3D pose estimation results of Mehta et al. (2017) and Zhou et al. (2016b), our approach yields more accurate results with less artifacts that better overlay with the input.

Fig. 12. Robustness of our model-based silhouette extraction method. Our model-based silhouette extraction method yields significantly more accurate foreground segmentation results than the default GrabCut method. Our correspondence prune strategy improves the robustness of our silhouette-based surface refinement to segmentation errors (row 2). The failure case, in which the pose estimate is far off, cannot be corrected by our silhouette-based refinement (row 3).

background color is too similar to the foreground. In these cases, our correspondence prune strategy improves the robustness of our silhouette-based surface refinement by checking whether the corresponding points are close enough and whether their normal directions are similar. If this condition is not met, then those correspondences are ignored in the ICP-like iterative refinement, which ensures if segmentation is not good at any part, then the mesh is not modified at those parts. In very rare cases, the estimated pose can be wrong due to occlusion or complicated poses (see row 3 in Figure 12). In this case the artifacts cannot be corrected by our silhouette-based refinement, since the wrong pose leads to wrong segmentation. However, our approach instantly recovers once the occluded parts become visible again (see Figure 19 for an example). This shows that our method does not accumulate errors over time.

6.4 Comparisons

Comparison to monocular non-rigid reconstruction. In Figure 13, we provide a qualitative comparison between our approach and the template-based dense monocular non-rigid reconstruction method of Yu et al. (2015) in terms of the full reconstructed surface. Their method fails to track the actor motion within a few frames,

Fig. 13. Qualitative shape comparison between our approach and the template-based monocular non-rigid tracking approach of Yu et al. (2015). Our approach is able to reconstruct the motion of the complete sequence, while (Yu et al. 2015) fails after a few frames.

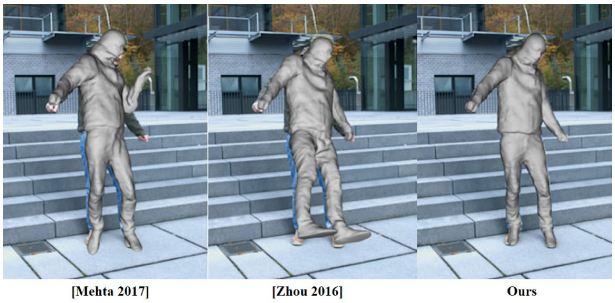

Fig. 14. Qualitative shape comparison to rigging the template model with respect to the 3D pose estimation results of Mehta et al. (2017) and Zhou et al. (2016b); our results are temporally more stable, of higher quality, and better overlay the input. See the accompanying video for more comparisons.

and is not able to recover afterwards, while our method constantly yields accurate tracking results throughout the entire sequence. Note that the approach of Yu et al. (2015) does not rely on a skeleton, and therefore can be applied to general shapes. However, this comparison confirms the benefits of our shape representation in the specific task of human performance capture. In addition, we perform a qualitative and quantitative comparison on the *Pablo* sequence. As shown in Figure 11, the tracking performance of our monocular approach is very close to the multi-view approach of Robertini et al. (2016) that uses eight cameras and drastically outperforms the template-based monocular approach of Yu et al. (2015). In addition, our approach consistently outperforms theirs in terms of mean vertex error compared to the multi-view reconstructions of Robertini et al. (2016) (see Figure 15).

Comparison to monocular 3D joint estimation. We also quantitatively compare our batch-based pose estimation method (E_{pose}) to the state-of-the-art real-time 3D skeleton tracker of Mehta et al. (2017) and the 2D-to-3D lifting approach of Zhou et al. (2016b).

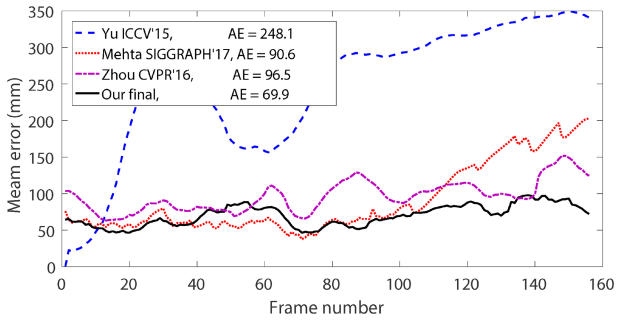

Fig. 15. Quantitative shape comparison on the *Pablo* sequence. Our approach outperforms the template-based tracking approach of Yu et al. (2015) and rigging the template model with respect to the 3D pose estimation results of Mehta et al. (2017) and Zhou et al. (2016b). The average error (AE) in millimeters is given in the legend.

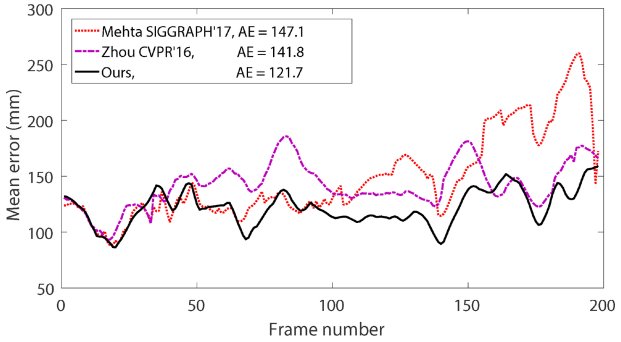

Fig. 16. Quantitative pose evaluation: We compare our average per-joint 3D position error to the approach of Mehta et al. (2017) and the 2D-to-3D lifting approach of Zhou et al. (2016b). Our approach achieves a consistently lower error leading to higher quality results. The average error (AE) in millimeters is given in the legend.

We first compare on the basis of joint positions, as the other methods do not reconstruct a deformable surface model. The ground-truth joint locations for this evaluation are provided by the professional multi-view marker-less motion capture software CapturyStudio.[3] We evaluate the average per-joint 3D error (in millimeters) after similarity transformation for each frame of the *Pablo* sequence. As shown in Figure 16 (average errors are given in the legend), our batch-based approach that uses 2D and 3D pose detections and fits joint angles of a coherent skeleton model obtains consistently lower errors than theirs. This lower error in 3D joint positions translates into higher quality and temporally more stable reconstructions. We also provide a comparison on the publicly available Human3.6M (Ionescu et al. 2014b) dataset. To this end, we applied our approach to four sequences of Human3.6M, and we compare the mean joint error of the 3D skeleton pose results obtained by our batch-optimization to that of Mehta et al. (2016) and Zhou et al. (2016a). To factor out the global pose and bone length scaling, we apply the Procrustes analysis between all pose predictions and the ground truth before computing the errors. As show in Table 1, our method outperforms these state-of-the-art approaches by a large margin (>3cm improvement).

[3] http://www.thecaptury.com/.

Table 1. Quantitative Pose Evaluation on Human3.6M

	Mehta16	Zhou16	**Ours**
S9C2Walking	124.93	133.37	**90.50**
S9C2Posing	109.36	129.28	**71.28**
S11C2Greeting	114.50	116.99	**75.40**
S11C2Sitting	112.90	110.08	**88.70**

Note: We compare the mean joint error (in millimeters) of the 3D skeleton pose results obtained by our batch-optimization to the approach of Mehta et al. (2016) and Zhou et al. (2016a). Our method outperforms theirs by a large margin.

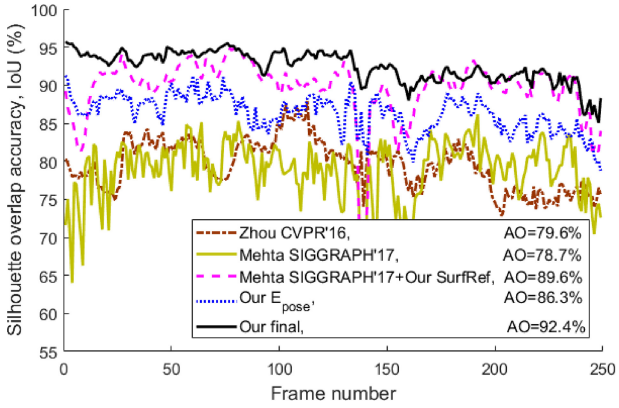

Fig. 17. Quantitative evaluation. We compare our silhouette overlap accuracy to the 2D-to-3D lifting approach of Zhou et al. (2016b), the real-time approach of Mehta et al. (2017), and our surface refinement method applied on the results of Mehta et al. (2017). Our pose estimation consistently outperforms the existing methods, while our surface refinement further significantly improves the overlap accuracy. The average silhouette overlap accuracy (AO) in percentage is given in the legend.

To further compare our method against the baseline monocular 3D joint estimation methods on surface level, we rigged our template to the pose estimation results of Mehta et al. (2017) and Zhou et al. (2016b). Naive rigging exhibits surface artifacts, as shown in Figures 14 and 11, whereas our approach yields smooth results and improved surface reconstruction quality. For quantitative comparison, we compute the silhouette overlap accuracy (Intersection over Union, IoU) on the sequence shown in Figure 14, based on manually labeled ground truth silhouettes. As shown in Figure 17, benefiting from our batch-based optimization, our pose estimation consistently outperforms previous state-of-the-art methods, while our surface refinement further significantly improves the overlap accuracy. To further evaluate our silhouette-based surface refinement method, we apply it on the meshes rigged with the pose estimation results of Mehta et al. (2017). As shown in Figure 17, the overlap accuracy is improved by 11% comparing to the skinning results. Furthermore, Figure 15 quantitatively shows that our approach outperforms the baseline methods in terms of per-vertex surface reconstruction error. For more comparisons, we refer to the accompanying video.

Comparison to stereo performance capture. We further compare to the stereo-based performance capture approach of Wu et al. (2013) on one of their stereo sequences. This approach leverages

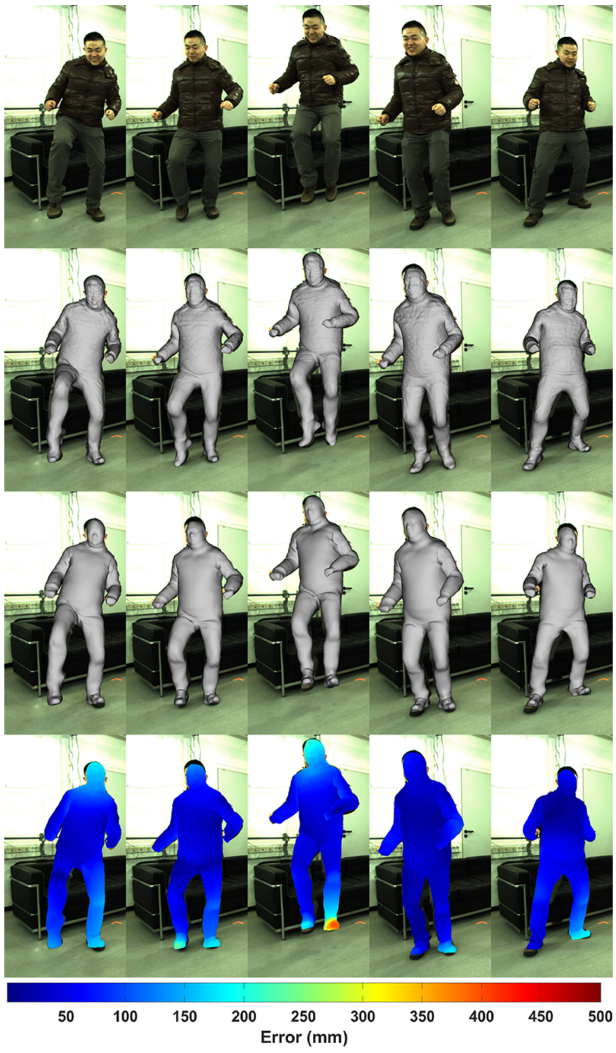

Fig. 18. Qualitative comparison to the binocular stereo performance capture approach of Wu et al. (2013). Our monocular approach obtains comparable quality results without requiring explicit depth cues. The per-vertex differences are color coded in the last row. Note, we do not employ full BRDF-based shading-based refinement to obtain smoother results.

explicit depth cues based on binocular stereo and yields high-quality performance capture results. For comparison, we selected a single camera view (the left camera) to obtain the monocular input video for our approach. As shown in Figure 18, our approach achieves similar accuracy given only monocular input despite the lack of explicit depth cues. Note, we did not employ full BRDF-based shading-based surface refinement as done in their approach, and thus obtain slightly smoother reconstructions.

7 LIMITATIONS

We have demonstrated compelling performance capture results given just monocular video input. Nevertheless, our approach is subject to the following limitations, which can be addressed in future work. First, currently, our approach, similar to previous

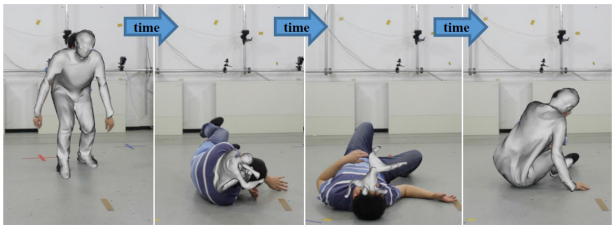

Fig. 19. Strong occlusion and fast motion can lead to tracking failure, but our approach is able to instantly recover as soon as the occluded parts become visible again.

performance capture approaches, requires a person-specific actor rig built in a pre-processing step. However, note that our template can be automatically generated from a video following a circular path around the static actor, which can be recorded within only half a minute. After this pre-process, our approach is fully automatic. The automatic extraction of such a rig from a monocular video sequence containing general motion is currently an unsolved problem, but first progress given only a sparse set of views (Bogo et al. 2016; Rhodin et al. 2016) has been made. Second, strong occlusion in combination with fast motion can still lead to tracking failure in our extremely challenging monocular setting. Nevertheless, our approach is able to instantly recover due to the discriminative joint detections as soon as the occluded parts become visible again; see Figure 19. Third, the capturing of feet is less robust in our results. Existing datasets for 2D human pose estimation only contain annotation for the major human joints, and unfortunately not for the feet. Since our model is pre-trained on these 2D datasets it is less robust on feet detection. Further, since the foot usually has a depth (from toe to the heel) from the camera perspective, wrong correspondences can be generated while assigning correspondence between the silhouette and the mesh boundary, e.g., a point on the toe being matched to a point on the heel, unless the initial foot position is sufficiently accurate. However, this issue can be alleviate in future works by incorporating better foot annotations in the training data. Fourth, topological changes of complex garments, e.g., opening a jacket, will lead to tracking failure, since only the outer layer of the surface is reconstructed during the template reconstruction. This can be resolved by reconstructing a multi-layer template. We would like to remind the reader of the profound difficulty of the monocular reconstruction setting. Despite several remaining limitations, we believe to have taken an important step in monocular performance capture that will inspire follow-up works.

8 CONCLUSION

We have presented the first approach for automatic temporally coherent marker-less human performance capture from a monocular video. The ambiguities of the underconstrained monocular reconstruction problem are tackled by leveraging sparse 2D and 3D joint detections and a low-dimensional motion prior in a joint optimization problem over a batch of frames. The tracked surface geometry is refined based on fully automatically extracted silhouettes to enable medium-scale non-rigid alignment. We demonstrated compelling monocular reconstruction results that enable exciting

applications such as video editing and free viewpoint video previously impossible using single RGB video.

We believe our approach is a significant step toward making marker-less monocular performance capture viable. In the future, a further improved and real-time solution to this challenging problem would have big implications for a broad range of applications in not only computer animation, visual effects, and free-viewpoint video, but also other fields such as medicine or biomechanics.

REFERENCES

Ijaz Akhter and Michael J. Black. 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1446–1455.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape completion and animation of people. *ACM Trans. Graph.* 24, 3, 408–416.

Alexandru O. Balan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. 2007. Detailed human shape and pose from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*. 1–8.

A. Bartoli, Y. Gérard, F. Chadebecq, T. Collins, and D. Pizarro. 2015. Shape-from-template. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 10, 2099–2118. DOI: http://dx.doi.org/10.1109/TPAMI.2015.2392759

Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. 2015. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proceedings of the International Conference on Computer Vision (ICCV'15)*. 2300–2308.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*.

Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. 2008. Markerless garment capture. *ACM Trans. Graph.* 27, 99.

Matthieu Bray, Pushmeet Kohli, and Philip H. S. Torr. 2006. Posecut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In *Proceedings of the European Conference on Computer Vision (ECCV'06)*. Springer, 642–655.

Thomas Brox, Bodo Rosenhahn, Daniel Cremers, and Hans-Peter Seidel. 2006. High-accuracy optical flow serves 3D pose tracking: Exploiting contour and flow-based constraints. In *Proceedings of the European Conference on Computer Vision (ECCV'06)*. Springer, 98–111.

Thomas Brox, Bodo Rosenhahn, Juergen Gall, and Daniel Cremers. 2010. Combined region and motion-based 3D tracking of rigid and articulated objects. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 3, 402–415.

Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. 2010. Free-form mesh tracking: A patch-based approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. IEEE, Los Alamitos, CA, 1339–1346.

Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. 2003. Free-viewpoint video of human actors. *ACM Trans. Graph.* 22, 3, 569–577.

Yu Chen, Tae-Kyun Kim, and Roberto Cipolla. 2010. Inferring 3D shapes and deformations from single views. In *Proceedings of the European Conference on Computer Vision (ECCV'10)*. 300–313.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Trans. Graph.* 34, 4, 69.

Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. In *ACM Trans. Graph.* 27, 98.

Mingsong Dou, Henry Fuchs, and Jan-Michael Frahm. 2013. Scanning and tracking dynamic objects with commodity depth cameras. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR'13)*. IEEE, Los Alamitos, CA, 99–106.

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, and others. 2016. Fusion4D: Real-time performance capture of challenging scenes. *ACM Trans. Graph.* 35, 4, 114.

Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. 2015. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 3810–3818.

Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, Los Alamitos, CA, 1746–1753.

R. Garg, A. Roussos, and L. Agapito. 2013. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. 1272–1279. DOI: http://dx.doi.org/10.1109/CVPR.2013.168

Pablo Garrido, Michael Zollhoefer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Perez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Trans. Graph.* 35, 3 28:1–28:15.

Daniel Grest, Dennis Herzog, and Reinhard Koch. 2005. Human model fitting from monocular posture images. In *Proceedings of the Conference on Vision, Modeling and Visualization (VMV'05)*.

Peng Guan, Alexander Weiss, Alexandru O Bălan, and Michael J Black. 2009. Estimating human shape and pose from a single image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'09)*. 1381–1388.

Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. 2015. Robust non-rigid motion tracking and surface reconstruction using L0 regularization. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*. 3083–3091.

Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. 2010. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. IEEE, Los Alamitos, CA, 1823–1830.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the EEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

Thomas Helten, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. 2013. Real-time body tracking with one depth camera and inertial sensors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'13)*.

Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. 2017. Towards accurate marker-less human shape and pose estimation over time. In *Proceedings of the International Conference on 3D Vision (3DV'17)*.

Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. 2016. VolumeDeform: Real-time volumetric non-rigid reconstruction. In *Computer Vision—ECCV 2016*. Springer, 17.

Catalin Ionescu, Joao Carreira, and Cristian Sminchisescu. 2014a. Iterated second-order label sensitive pooling for 3D human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. 1661–1668.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014b. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 7, 1325–1339.

Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. 2010. MovieReshape: Tracking and reshaping of humans in videos. *ACM Trans. Graph.* 29, 5, Article 148. DOI: http://dx.doi.org/10.1145/1866158.1866174

Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. 2014. Modeep: A deep learning framework using motion features for human pose estimation. In *Proceedings of the Asian Conference on Computer Vision (ACCV'14)*. 302–315.

Sam Johnson and Mark Everingham. 2011. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. 2007. Skinning with dual quaternions. In *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games (I3D'07)*.

J. P. Lewis, Matt Cordner, and Nickson Fong. 2000. Pose Space Deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'00)*. 165–172.

Hao Li, Bart Adams, Leonidas J. Guibas, and Mark Pauly. 2009. Robust single-view geometry and motion reconstruction. *ACM Trans. Graph.* 28, 5, Article 175. DOI: http://dx.doi.org/10.1145/1618452.1618521

Sijin Li and Antoni B Chan. 2014. 3D human pose estimation from monocular images with deep convolutional neural network. In *Proceedings of the Asian Conference on Computer Vision (ACCV'14)*. 332–347.

Sijin Li, Weichen Zhang, and Antoni B Chan. 2015. Maximum-margin structured learning with deep networks for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*. 2848–2856.

Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. 2011. Markerless motion capture of interacting characters using multi-view image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, Los Alamitos, CA, 1249–1256.

Matthew Loper, Naureen Mahmood, and Michael J. Black. 2014. MoSh: Motion and shape capture from sparse markers. *ACM Trans. Graph.* 33, 6, 220.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* 34, 6, Article 248.

Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. 2000. Image-based visual hulls. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. 369–374.

Dushyant Mehta, Helge Rhodin, Dan Casas, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2016. Monocular 3D human pose estimation using transfer learning and improved CNN supervision. arXiv:1611.09813.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Trans. Graph.* 36, 4, 14.

Greg Mori and Jitendra Malik. 2006. Recovering 3D human body configurations using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 7, 1052–1062.

Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. 2015. General dynamic scene reconstruction from multiple view video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*.

Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. 2015. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*.

Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. arXiv:1603.06937.

Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh. 2015. 3D trajectory reconstruction under perspective projection. *Int. J. Comput. Vision* 115, 2, 115–135. DOI: http://dx.doi.org/10.1007/s11263-015-0804-2

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. 2016. Coarse-to-fine volumetric prediction for single-image 3D human pose. arXiv:1611.07828.

Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. 2016. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

Ralf Plänkers and Pascal Fua. 2001. Tracking and modeling people in video sequences. *Comput. Vision Image Understand.* 81, 3, 285–302.

Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. 2016. General automatic human shape and motion capture using volumetric contour cues. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*. 509–526.

Nadia Robertini, Dan Casas, Helge Rhodin, Hans-Peter Seidel, and Christian Theobalt. 2016. Model-based outdoor performance capture. In *Proceedings of the International Conference on Computer Vision (3DV'16)*.

Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann, and Marcus Magnor. 2014. Garment replacement in monocular video sequences. *ACM Trans. Graph.* 34, 1, 6.

Rómer Rosales and Stan Sclaroff. 2006. Combining generative and discriminative models in a framework for articulated pose estimation. *Int. J. Comput. Vis.* 67, 3, 251–276.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* 23, 309–314.

Chris Russell, Rui Yu, and Lourdes Agapito. 2014. *Video Pop-up: Monocular 3D Reconstruction of Dynamic Scenes*. Springer International Publishing, Cham, 583–598. DOI: http://dx.doi.org/10.1007/978-3-319-10584-0_38

Mathieu Salzmann and Pascal Fua. 2011. Linear local models for monocular reconstruction of deformable surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 5, 931–944. DOI: http://dx.doi.org/10.1109/TPAMI.2010.158

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. 1297–1304.

Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. 2000. Stochastic tracking of 3D human figures using 2D image motion. In *Proceedings of the European Conference on Computer Vision (ECCV'00)*. 702–718.

Leonid Sigal, Alexandru Balan, and Michael J. Black. 2007. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, 1337–1344.

Edgar Simo-Serra, Arnau Ramisa, Guillem Alenyà, Carme Torras, and Francesc Moreno-Noguer. 2012. Single image 3D human pose estimation from noisy observations. In *Proceedings of the EEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. IEEE, Los Alamitos, CA, 2673–2680.

Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. 2006. Learning joint top-down and bottom-up processes for 3D visual inference. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, Los Alamitos, CA, 1743–1752.

Cristian Sminchisescu and Bill Triggs. 2003a. Estimating articulated human motion with covariance scaled sampling. *Int. J. Robot. Res.* 22, 6, 371–391.

Cristian Sminchisescu and Bill Triggs. 2003b. Kinematic jump processes for monocular 3D human tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Vol. 1. IEEE, Los Alamitos, CA, I–69.

Dan Song, Ruofeng Tong, Jian Chang, Xiaosong Yang, Min Tang, and Jian Jun Zhang. 2016. 3D body shapes estimation from dressed-human silhouettes. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 147–156.

Olga Sorkine and Marc Alexa. 2007. As-rigid-as-possible surface modeling. In *Proceedings of the 5th Eurographics Symposium on Geometry Processing (SGP'07)*.

Jonathan Starck and Adrian Hilton. 2007. Surface capture for performance-based animation. *IEEE Comput. Graph. Appl.* 27, 3, 21–31.

Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. 2011. Fast articulated motion tracking using a sums of Gaussians body model. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'11)*. 951–958.

Robert W. Sumner, Johannes Schmid, and Mark Pauly. 2007. Embedded deformation for shape manipulation. *ACM Trans. Graph.* 26, 3, 80.

Camillo J. Taylor. 2000. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)*, Vol. 1. 677–684.

Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. 2016. Structured prediction of 3D human pose with deep neural networks. In *Proceedings of the British Machine Vision Conference (BMVC'16)*.

J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. IEEE, Los Alamitos, CA.

Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'14)*. 1653–1660.

Raquel Urtasun, David J. Fleet, and Pascal Fua. 2005. Monocular 3D tracking of the golf swing. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 932–938.

Raquel Urtasun, David J. Fleet, and Pascal Fua. 2006. Temporal motion models for monocular and multiview 3D human body tracking. *Comput. Vision Image Understand.* 104, 2, 157–177.

Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. 2008. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.* 27, 97.

Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. 2009. Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph.* 28, 5, 174.

Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L. Yuille, and Wen Gao. 2014. Robust estimation of 3D human poses from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. 2361–2368.

Ruizhe Wang, Lingyu Wei, Etienne Vouga, Qixing Huang, Duygu Ceylan, Gerard Medioni, and Hao Li. 2016. Capturing dynamic textured surfaces of moving targets. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*.

Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus Gross. 2005. Scalable 3D video of dynamic scenes. *Visual Comput.* 21, 8–10, 629–638.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

Xiaolin Wei and Jinxiang Chai. 2010. Videomocap: Modeling physically realistic human motion from monocular video sequences. *ACM Trans. Graph.* 29, 42.

Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. 1997. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 7, 780–785.

Chenglei Wu, Carsten Stoll, Levi Valgaerts, and Christian Theobalt. 2013. On-set performance capture of multiple actors with a stereo camera. *ACM Trans. Graph.* 32, 161:1–161:11. DOI: http://dx.doi.org/10.1145/2508363.2508418

Chenglei Wu, Kiran Varanasi, and Christian Theobalt. 2012. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *Proceedings of the European Conference on Computer Vision (ECCV'12)*. 757–770.

Weipeng Xu, Mathieu Salzmann, Yongtian Wang, and Yue Liu. 2015. Deformable 3D fusion: From partial dynamic 3D observations to complete 4D models. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*. 2183–2191.

Hashim Yasin, Umar Iqbal, Björn Krüger, Andreas Weber, and Juergen Gall. 2016. A dual-source approach for 3D pose estimation from a single image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

Genzhi Ye, Yebin Liu, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt. 2012. Performance capture of interacting characters with handheld kinects. In *Proceedings of the European Conference on Computer Vision (ECCV'12)*, Vol. 7573 LNCS. 828–841. DOI: http://dx.doi.org/10.1007/978-3-642-33709-3_59

Rui Yu, Chris Russell, Neill D. F. Campbell, and Lourdes Agapito. 2015. Direct, dense, and deformable: Template-based non-rigid 3D reconstruction from RGB video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*.

Qing Zhang, Bo Fu, Mao Ye, and Ruigang Yang. 2014. Quality dynamic human body modeling using a single low-cost depth camera. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR'14). IEEE, 676–683.

Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. 2010. Parametric reshaping of human bodies in images. *ACM Trans. Graph. (TOG)* 29, 4 (2010), 126.

Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas Daniilidis. 2015. 3D shape estimation from 2D landmarks: A convex relaxation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 4447–4455.

Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. 2016a. Deep kinematic pose regression. *arXiv Preprint arXiv:1609.05317* (2016).

Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. 2016b. Sparseness meets deepness: 3D human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4966–4975.

Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rhemann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, and Marc Stamminger. 2014. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Trans. Graph.* 33, 4, Article 156.

Received September 2017; revised February 2018; accepted February 2018