LiveCap:
Real-time Human Performance Capture from Monocular Video

MARC HABERMANN, Max Planck Institute for Informatics
WEIPENG XU, Max Planck Institute for Informatics
MICHAEL ZOLLHÖFER, Stanford University
GERARD PONS-MOLL, Max Planck Institute for Informatics
CHRISTIAN THEOBALT, Max Planck Institute for Informatics

Fig. 1. We propose the first real-time human performance capture approach that reconstructs dense, space-time coherent deforming geometry of people in their loose everyday clothing from just a single monocular RGB stream, e.g., captured by a webcam.

We present the first real-time human performance capture approach that reconstructs dense, space-time coherent deforming geometry of entire humans in general everyday clothing from just a single RGB video. We propose a novel two-stage analysis-by-synthesis optimization whose formulation and implementation are designed for high performance. In the first stage, a skinned template model is jointly fitted to background subtracted input video, 2D and 3D skeleton joint positions found using a deep neural network, and a set of sparse facial landmark detections. In the second stage, dense non-rigid 3D deformations of skin and even loose apparel are captured based on a novel real-time capable algorithm for non-rigid tracking using dense photometric and silhouette constraints. Our novel energy formulation leverages automatically identified material regions on the template to model the differing non-rigid deformation behavior of skin and apparel. The two resulting non-linear optimization problems per-frame are solved with specially-tailored data-parallel Gauss-Newton solvers. In order to achieve real-time performance of over 25Hz, we design a pipelined parallel architecture using the CPU and two commodity GPUs. Our method is the first real-time monocular approach for full-body performance capture. Our method yields comparable accuracy with off-line performance capture techniques, while being orders of magnitude faster.

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

This work was funded by the ERC Consolidator Grant 4DRepLy (770784). Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. © 2019 Association for Computing Machinery. 0730-0301/2019/1-ART1 $15.00 https://doi.org/10.1145/nnnnnn.nnnnnn

1 INTRODUCTION

Dynamic models of virtual human actors are key elements of modern visual effects for movies and games, and they are invaluable for believable, immersive virtual and augmented reality, telepresence, as well as 3D and free-viewpoint video. Such virtual human characters ideally feature high-quality, space-time coherent dense models of shape, motion and deformation, as well as appearance of people, irrespective of physique or clothing style. Creating such models at high fidelity often requires many months of work of talented artists. To simplify the process, marker-less performance capture methods were researched to reconstruct at least parts of such models from camera recordings of real humans in motion.

Existing multi-camera methods capture human models at very good quality, but often need dense arrays of video or depth cameras and controlled studios, struggle with complex deformations, and need pre-captured templates. Only few multi-view methods achieve real-time performance, but no real-time method for single RGB performance capture exists. Many applications in interactive VR and AR, gaming, virtual try-on [Hilsmann and Eisert 2009; Pons-Moll et al. 2017; Sekine et al. 2014], pre-visualization for visual effects, 3DTV or telepresence [Orts-Escolano et al. 2016] critically depend on real-time performance capture. The use of complex camera arrays and studios restricted to indoor scenes presents a practical barrier to these applications. In daily use, systems should ideally require only one camera and work outdoors.

ACM Reference format:
Under these requirements, performance capture becomes a much harder and much more underconstrained problem. Some methods have approached this challenge by using multiple [Collet et al. 2015; Dou et al. 2016; Wang et al. 2016] or a single low-cost consumer-grade depth (RGB-D) [Newcombe et al. 2015; Yu et al. 2017] camera for dense non-rigid deformation tracking. While these methods are a significant step forward, RGB-D cameras are not as cheap and ubiquitous as color cameras, often have a limited capture range, do not work well under bright sunlight, and have limited resolution. Real-time human performance capture with a single color camera would therefore greatly enhance and simplify performance capture and further democratize its use, in particular in the aforementioned interactive applications of ever increasing importance. However, dense real-time reconstruction from one color view is even harder, and so today’s best monocular methods only capture very coarse models, such as bone skeletons [Mehta et al. 2017; Sun et al. 2017].

In this paper, we propose the - to our knowledge - first real-time human performance capture method that reconstructs dense, space-time coherent deforming geometry of people in their loose everyday clothing from a single video camera. In a pre-processing step, the method builds a rigid surface and appearance template from a short video of the person in a static pose, on which regions of skin and pieces of apparel are automatically identified using a new multi-view segmentation that leverages deep learning. The template is fitted to the video sequence in a new coarse-to-fine two-stage optimization, whose problem formulation and implementation are rigorously designed for best accuracy at real-time performance. In its first stage, our new real-time skeleton pose optimizer fits the skinned template to (1) 2D and 3D skeleton joint positions found with a CNN, to (2) sparse detected facial landmarks, and (3) to the foreground silhouette.

In a second stage, dense non-rigid 3D deformations of even loose apparel is captured. To this end, we propose a novel real-time capable algorithm for non-rigid analysis-by-synthesis tracking from monocular RGBD data. It minimizes a template-to-image alignment energy jointly considering distance-field based silhouette alignment, dense photometric alignment and spatial and temporal regularizers, all designed for real-time performance. The energy formulation leverages the shape template segmentation labels (obtained in the pre-processing stage) to account for the varying non-rigid deformation behavior of different clothing during reconstruction. The non-linear optimization problems in both stages are solved with specially-tailored GPU accelerated Gauss-Newton solvers. In order to achieve real-time performance of over 25 Hz, we design a pipelined solver architecture that executes the first and the second stage on two GPUs in a rolling manner. Our approach captures high-quality models of humans and their clothing in real-time from a single monocular camera. We demonstrate intriguing examples of live applications in 3D video and virtual try-on. We show qualitatively and quantitatively that our method outperforms related monocular on-line methods and comes close to off-line performance capture approaches in terms of reconstruction density and accuracy.

In summary, our contributions are: (1) We propose the first real-time system for monocular human performance capture. In order to achieve real-time performance, we not only made specific algorithmic design choices, but also contribute several new algorithmic ideas, e.g., the adaptive material based regularization and the displacement warping to guarantee high quality results under a tight real-time constraint. (2) We also show how to efficiently implement these design decisions by combining the compute power of two GPUs and the CPU in a pipelined architecture and how dense and sparse linear systems of equations can be efficiently optimized on the GPU. (3) To evaluate our approach on a wide range of data, we show high quality results on an extensive new dataset of more than 20 minutes of video footage captured in 11 scenarios, which contain different types of loose apparel and challenging motions.

2 RELATED WORK
Performance capture methods typically use multi-view images or depth sensors. We focus here on approaches to capture 3D humans in motion, and leave out the body of work on 2D pose and shape capture. Most monocular-based methods ignore clothing and are restricted to capturing the articulated motion and the undressed shape of the person. Since there are almost no works that do performance capture from monocular video we focus here on multi-view and depth-based methods and approaches that capture pose and undressed shape from single images.

Multi-view. Many multi-view methods use stereo and shape from silhouette cues to capture the moving actor [Collet et al. 2015; Matsuk et al. 2000; Starck and Hilton 2007; Waschbüsch et al. 2005], or reconstruct via multi-view photometric stereo [Vlasic et al. 2009]. Provided with sufficient images some methods directly non-rigidly deform a subject specific template mesh [Cagniart et al. 2010; Carranza et al. 2003; De Aiguar et al. 2008] or a volumetric shape representation [Allain et al. 2015; Huang et al. 2016]. Such methods are free-form and can potentially capture arbitrary shapes [Mustafa et al. 2016] as they do not incorporate any skeletal constraints. Such flexibility comes at the cost of robustness. To mitigate this, some methods incorporate a skeleton in the template to constrain the motion to be nearly articulated [Gall et al. 2009; Liu et al. 2011; Vlasic et al. 2008]. This also enables off-line performance capture from a stereo pair of cameras [Wu et al. 2013]. Some systems combine reconstruction and segmentation to improve results [Bray et al. 2006; Brox et al. 2010; Liu et al. 2011; Wu et al. 2012]. Such methods typically require a high resolution scan of the person as input. To side step scanning, a parametric body model can be employed. Early models were based on simple geometric primitives [Metaxas and Terzopoulos 1993; Plankers and Fua 2001; Sigal et al. 2004; Sminchisescu and Triggs 2003]. Recent ones are more accurate, detailed and are learned from thousands of scans [Anguelov et al. 2005; Hasler et al. 2010; Kadleccke et al. 2016; Kim et al. 2017; Loper et al. 2015; Park and Hodgins 2008; Pons-Moll et al. 2015]. Capture approaches that use a statistical body model typically ignore clothing or treat it as noise [Balan et al. 2007] or explicitly estimate the shape under the apparel [Balan and Black 2008; Yang et al. 2016; Zhang et al. 2017]. The off-line human performance capture approach of Huang et al. [2017] fits the SMPL body model to 2D joint detections and silhouettes in multi-view data. Some of the recent off-line multi-view approaches jointly track facial expressions [Joo et al. 2018] and hands [Joo et al. 2018; Romero et al. 2017]. Even these approaches do not reconstruct dynamic hair. To capture the geometry of the actor
beyond the body shape an option is to non-rigidiy deform the base model to fit a scan [Zhang et al. 2017] or a set of images [Rhodin et al. 2016]. Recently, the approach of Pons-Moll et al. [2017] can jointly capture body shape and clothing using separate meshes; very realistic results are achieved with this method, but it requires an expensive multi-view active stereo setup. All the aforementioned approaches require multi-view setups and are not practical for consumer use. Furthermore, none of the methods runs at real-time frame rates.

Depth-based. With the availability of affordable depth camera sensors such as the Kinect, a large number of depth-based methods emerged. Recent approaches that are based on a single depth camera, such as KinectFusion, enable the reconstruction of 3D rigid scenes [Izadi et al. 2011; Newcombe et al. 2011] and also appearance models [Zhou and Koltun 2014] by incrementally fusing geometry in a canonical frame. The approach proposed in [Newcombe et al. 2015] generalized KinectFusion to capture dynamic non-rigid scenes. The approach alternates non-rigid registration of the incoming depth frames with updates to the incomplete template, which is constructed incrementally. Such template free methods [Guo et al. 2017; Immann et al. 2016; Newcombe et al. 2011; Slavcheva et al. 2017] are flexible, but are limited to capturing slow and careful motions. One way to make fusion and tracking more robust is by using a combination of a high frame rate/low resolution and a low frame rate/high resolution depth sensor [Guo et al. 2018], improved hardware and software components [Kowdle et al. 2018], multiple Kinects or similar depth sensors [Dou et al. 2017, 2016; Orts-Escolano et al. 2016; Ye et al. 2012; Zhang et al. 2014], or multi-view data [Collet et al. 2015; Leroy et al. 2017; Prada et al. 2017] and registering new frames to a neighboring key frame; such methods achieve impressive reconstructions, but do not register all frames to the same canonical template and require complicated capture setups. Another way to constrain the capture is to pre-scan the object or person to be tracked [De Aguiar et al. 2008; Ye et al. 2012; Zollhöfer et al. 2014], reducing the problem to tracking the non-rigid deformations. Constraining the motion to be articulated is also shown to increase robustness [Yu et al. 2017, 2018]. Some works use simple human shape or statistical body models [Bogo et al. 2015; Helten et al. 2013; Wei et al. 2012; Weiss et al. 2011; Ye and Yang 2014; Zhang et al. 2014], some of which exploit the temporal information to infer shape. Typically, a single shape and multiple poses are optimized to exploit the temporal information. Such approaches are limited to capture naked human shape or at best very tight clothing. Depth sensors are affordable and more practical than multi-view setups. Unfortunately, they have a high power consumption, do not work well under general illumination and most of the media content is still in the format of 2D images and video. Furthermore, depth-based methods do not directly generalize to work with monocular video.

Monocular 3D Pose and Shape Estimation. Most methods to infer 3D human motion from monocular images are based on convolutional neural networks (CNNs) and leverage 2D joint detections and predict 3D joint pose in the form of stick figures, e.g., [Popa et al. 2017; Rogez et al. 2017; Sun et al. 2017; Tome et al. 2017; Zhou et al. 2017]. Tekin et al. [2016] directly predict the 3D body pose from a rectified spatio-temporal volume of input frames. The approach of Tekin et al. [2017] learns to optimally fuse 2D and 3D image cues. These approaches do not capture the dense deforming shape. We also leverage a recent CNN-based 3D pose estimation method [Mehta et al. 2017], but we only employ it to regularize the skeletal motion estimation. Some works fit a (statistical) body surface model to images using substantial manual interaction [Guan et al. 2009; Jain et al. 2010; Rogge et al. 2014; Zhou et al. 2010] typically for the task of image manipulation. Shape and clothing is recovered in [Chen et al. 2013; Guo et al. 2012], but the user needs to click points in the image, select the clothing types from a database and dynamics are not captured. Instead of clicked points, Kraevoy et al. [2009] propose to obtain the shape from contour drawings. With the advance of 2D joint detections, the works of [Bogo et al. 2016; Kanazawa et al. 2018; Lassner et al. 2017] fit a 3D body model [Loper et al. 2015] to them; since only model parameters are optimized, the results are constrained to the shape space. More recent work [Varol et al. 2018] directly regresses a coarse volumetric body shape. Correspondences from pixels of an input image to surface points on the SMPL body model can also be directly regressed [Güler et al. 2018]. Capturing 3D non-rigid deformations from monocular video is very hard. In the domain of non-rigid structure from motion, model-free methods using rigidity and temporal smoothness priors can capture coarse 3D models of simple motions and medium-scale deformations [Garg et al. 2013; Russell et al. 2014]. Some methods [Bartoli et al. 2015; Salzmann and Fua 2011; Yu et al. 2015] can non-rigidly track simple shapes and motions by off-line template fitting; but they were not shown to handle highly articulated fast body motions, including clothing, as we do. Specifically for faces, monocular performance capture methods were presented, for example [Cao et al. 2015; Garrido et al. 2016]. However, monocular full-body capture faces additional challenges due to more frequent (self-)occlusions and much more complex and diverse clothing and appearance. To the best of our knowledge, the only approach that has shown 3D performance capture of the human body including the non-rigid deformation of clothing from monocular video is the approach of Xu et al. [2018]. Its space-time formulation can resolve difficult self-occluded poses at the expense of temporally oversmoothing the actual motion. But at over 1 minute runtime per frame, it is impractical for many applications such as virtual try-on, gaming or virtual teleportation. It is also challenged by starkly non-rigidly moving clothing. Reducing the processing time without compromising accuracy introduces challenges in formulation and implementation of model-based performance capture, which we address in this work. We present, for the first time, a real-time full-body performance capture system that just requires a monocular video as input. We show that it comes close in accuracy to the best off-line monocular and even multi-view methods, while being orders of magnitude faster.

3 METHOD
The input to our method is a single color video stream. In addition, our approach requires a textured actor model, which we acquire in a preprocessing step (Sec. 3.1) from a monocular video sequence. From this input alone, our real-time human performance capture approach automatically estimates the articulated actor motion and the non-rigid deformation of skin and clothing coarse-to-fine in two
subsequent stages per input frame. In the first stage, we estimate the articulated 3D pose of the underlying kinematic skeleton. To this end, we propose an efficient way to fit the skeletal pose of the skinned template to 2D and 3D joint positions from a state-of-the-art CNN-based regressor, to sparse detected face landmarks, and to the foreground silhouette (Sec. 3.3). With this skeleton-deformed mesh and the warped non-rigid displacement of the previous frame as initialization, the second stage captures the surface deformation of the actor using a novel real-time template-to-image non-rigid registration approach (Sec. 3.4). We express non-rigid registration as an optimization problem consisting of a silhouette alignment term, a photometric term, and several regularization terms; the formulation and combination of terms in the energy is geared towards high efficiency at high accuracy despite the monocular ambiguities. The different components of our approach are illustrated in Fig. 2. In order to achieve real-time performance, we tackle the underlying optimization problems based on dedicated data-parallel GPU optimizers (Sec. 4). In the following, we explain all components.

### 3.1 Actor Model Acquisition

Similar to many existing template-based performance capture methods, e.g., [Allain et al. 2015; Cagniart et al. 2010; Gall et al. 2009; Vlasic et al. 2008; Xu et al. 2018], we reconstruct an actor model in a preprocessing step. To this end, we take a set of \( M \) images \( I_{\text{rec}} = \{I_{\text{rec}}, \ldots, I_{\text{rec}_M}\} \) of the actor in a static neutral pose from a video captured while walking around the person, which covers the entire body. For all our templates we used around \( M = 70 \) images. With these images, we generate a triangulated template mesh \( \hat{V} \in \mathbb{R}^{N \times 3} \) (\( N \) denotes the number of vertices in the mesh) and the associated texture map of the actor using an image-based 3D reconstruction software\(^1\). We downsample the reconstructed geometry to a resolution of approximately \( N = 5000 \) by using the Quadric Edge Collapse Decimation algorithm implemented in MeshLab\(^2\). The vertex colors of the template mesh \( C \in \mathbb{R}^{N \times 3} \) are transferred from the generated texture map. Then, skeleton joints and facial markers are manually placed on the template mesh resulting in a skeleton model. The template mesh is rigged to this skeleton model via dual quaternion skinning [Kavan et al. 2007], where the skinning weights are automatically computed using Blender\(^3\) (other auto-rigging tools would be feasible). This allows us to deform the template mesh using the estimated skeletal pose parameters (Sec. 3.3). An important feature of our performance capture method is that we model material-dependent differences in deformation behavior, e.g. of skin and apparel during tracking (see Sec. 3.4). To this end, we propose a new multi-view method to segment the template into one of seven non-rigidity classes. We first apply the state-of-the-art human parsing method of Gong et al. [2017] to each image in \( I_{\text{rec}} \) separately to obtain the corresponding semantic label images \( L_{\text{rec}} = \{L_{\text{rec}}, \ldots, L_{\text{rec}_M}\} \). The semantic labels \( L \in \{1, \ldots, 20\}^N \) for all vertices \( V \) are computed based on their back-projection into all label images, and a majority vote per vertex. The materials are binned into 7 non-rigidity classes, each one having a different per-edge non-rigidity weight in the employed regularization term (Sec. 3.4). Those weights were empirically determined by visual observation of the deformation behavior under different weighting factors. The different classes and the corresponding non-rigidity weights are shown in Tab. 1. We use a very high weight for rigid body parts, e.g., the head, medium weights for the less rigid body parts, e.g., skin and tight clothing, and a low weight for loose clothing. We use a high rigidity weight for any kind of hair style, since we do not consider and track hair dynamics. We map the per-vertex smoothness weights to per-edge non-rigidity weights \( s_{i,j} \) by averaging the weights of vertex \( V_i \) and \( V_j \).

---

\(^{1}\)http://www.agisoft.com
\(^{2}\)http://www.meshlab.net/
\(^{3}\)https://www.blender.org/
3.2 Input Stream Processing

After the actor model acquisition step, our real-time performance capture approach works fully automatically and we do not rely on a careful initialization, e.g., it is sufficient to place the T-posed character model in the center of the frame. The input to our algorithm is a single color video stream from a static camera, e.g., a webcam. Thus, we assume camera and world space to be the same. We calibrate the camera intrinsics using the Matlab calibration toolbox. Our skeletal pose estimation and non-rigid registration stages rely on the silhouette segmentation of the input video frames. To this end, we leverage the background subtraction method of Zivkovic and van der Heijden [2006]. We assume the background is static, that its color is sufficiently different from the foreground, and a few frames of the empty scene are recorded before performance capture commences. We efficiently compute distance transform images \( I_{DT} \) from the foreground silhouettes, which are used in the skeletal pose estimation and non-rigid alignment step.

3.3 Skeletal Pose Estimation

We formulate skeletal pose estimation as a non-linear optimization problem in the unknown skeleton parameters \( S \):

\[
S^* = \arg\min_S E_{\text{pose}}(S).
\]  

The set \( S = \{\theta, R, t\} \) contains the joint angles \( \theta \in \mathbb{R}^{27} \) of the \( J \) joints of the skeletal model, and the global pose \( R \in \text{SO}(3) \) and translation \( t \in \mathbb{R}^3 \) of the root. For pose estimation, we optimize an energy of the following general form:

\[
E_{\text{pose}}(S) = E_{\text{2D}}(S) + E_{\text{3D}}(S) + E_{\text{silhouette}}(S) + E_{\text{temporal}}(S) + E_{\text{anatomic}}(S).
\]

Here, \( E_{\text{2D}} \) and \( E_{\text{3D}} \) are alignment constraints based on regressed 2D and 3D joint positions, respectively. In addition, \( E_{\text{silhouette}} \) is a dense alignment term that fits the silhouette of the actor model to the detected silhouette in the input color images. At last, \( E_{\text{temporal}} \) and \( E_{\text{anatomic}} \) are temporal and anatomical regularization constraints that ensure that the speed of the motion and the joint angles stay in physically plausible ranges. To better handle fast motion, we initialize the skeleton parameters before optimization by extrapolating the poses of the last two frames in joint angle space based on an explicit Euler step. In the following, we explain each energy term in more detail.

Sparse 2D and 3D Alignment Constraint. For each input frame \( I \), we estimate the 2D and 3D joint positions \( P_{2D,i} \in \mathbb{R}^2 \) and \( P_{3D,i} \in \mathbb{R}^3 \) of the \( J \) joints using the efficient deep skeleton joint regression network of the VNeck algorithm [Mehta et al. 2017] trained with the original data of [Mehta et al. 2017]. However, with these skeleton-only joint detections, it is not possible to determine the orientation of the head. Therefore, we further augment the 2D joint predictions of [Mehta et al. 2017] with a subset of the facial landmark detections of [Saragih et al. 2009], which includes the eyes, nose and chin. We incorporate the 2D detections \( P_{2D,i} \in \mathbb{R}^2 \) based on the following:

\[
E_{\text{silhouette}}(S) = \lambda_{\text{silhouette}} \sum_{i \in B} b_i \left[ |I_{DT}(\pi(V_i(\theta, R, t)))| \right]^2.
\]

Here, \( V_i \) is the \( i \)-th boundary vertex of the skinned actor model, and the image \( I_{DT} \) stores the Euclidean distance transform with respect to the detected silhouette in the input image. The \( b_i \in \{-1, +1\} \) are directional weights that guide the optimization to follow the right direction in the distance field. In the minimization of the term in Eq. 5, silhouette model points are pushed in the negative direction of the distance transform image gradient \( \nabla z \in \mathbb{R}^2 \). By

\[
E_{\text{2D}}(S) = \lambda_{\text{2D}} \sum_{i=1}^{J+1} \lambda_i \| \pi(P_{3D,i}(\theta, R, t)) - P_{2D,i} \|^2.
\]

Here, \( P_{3D,i} \) is the 3D position of the \( i \)-th joint/facade marker of the used kinematic skeleton and \( \pi : \mathbb{R}^3 \to \mathbb{R}^2 \) is a full perspective projection that maps 3D space to the 2D image plane. Thus, this term enforces that all projected joint positions are close to their corresponding detections. \( \lambda_i \) are detection-based weights. We use \( \lambda_{\text{head}} = 0.326 \) for the facial landmarks and \( \lambda_{\text{all}} = 1.0 \) for all other detections to avoid that the head error dominates all other body parts. To resolve the inherent depth ambiguities of the re-projection constraint, we also employ the following 3D-to-3D alignment term between model joints \( P_{3D,i}(\theta, R, t) \) and 3D detections \( P_{3D,i} \):

\[
E_{\text{3D}}(S) = \lambda_{\text{3D}} \sum_{i=1}^{J} \| P_{3D,i}(\theta, R, t) - (P_{3D,i} + t') \|^2.
\]

Here, \( t' \in \mathbb{R}^3 \) is an auxiliary variable that transforms the regressed 3D joint positions \( P_{3D,i} \) from the root centered local coordinate system to the global coordinate system. Note that the regressed 3D joint positions \( P_{3D,i} \) are in a normalized space. Therefore, we rescale the rescaled skeleton according to the bone lengths of our parameterized skeleton model.

Dense Silhouette Alignment Constraint. We enforce a dense alignment between the boundary of the skinned actor model and the detected silhouette in the input image. In contrast to the approach of Xu et al. [2018] that requires closest point computations we employ a distance transform based constraint for efficiency reasons. Once per frame, we extract a set of contour vertices \( B \) from the current deformed version of the actor model. Afterwards, we enforce that all contour vertices line well to the interface between the detected foreground and background:

\[
E_{\text{silhouette}}(S) = \lambda_{\text{silhouette}} \sum_{i \in B} b_i \left| I_{DT}(\pi(V_i(\theta, R, t))) \right|^2.
\]
definition, $z$ points in the direction of the nearest image silhouette (IS) contour. If model points fall outside of the IS they will be dragged towards the nearest IS contour as desired. When model points fall inside the IS however there are two possibilities: 1) the model point normal $n$ follows roughly the same direction as $z$ or 2) it does not. In case 1) the normal at the nearest IS point matches the direction of the model point normal. This indicates that $z$ is a good direction to follow. In case 2) however, the normal at the nearest IS point follows the opposite direction, indicating that $z$ is pointing towards the wrong IS contour, see Fig. 3. Therefore, in case 2) we follow the opposite direction $p = -z$ by setting $b_i = -1$. This is preferable over just following $n$, since $n$ is not necessarily pointing away from the wrong IS contour. Mathematically, we consider that we are in case 2) when $n^Tz < 0$. For all the other cases, we follow the direction of $z$ by setting $b_i = +1$.

**Temporal Stabilization.** To mitigate temporal noise, we use a temporal stabilization constraint, which penalizes the change in joint position between the current and previous frame:

$$E_{\text{temporal}}(S) = \lambda_{\text{temporal}} \sum_{i=1}^{J} \| p_{\text{3D},i}(\theta, R, t) - p_{\text{3D},i}^{t-1}(\theta, R, t) \|^2 .$$

Here the $\lambda_i$ are joint-based temporal smoothness weights. We use $\lambda_i = 2.5$ for joints on the torso and the head, $\lambda_i = 2.0$ for shoulders, $\lambda_i = 1.5$ for knees and elbows, and $\lambda_i = 1.0$ for the hands and feet.

**Joint Angle Limits.** The joints of the human skeleton have physical limits. We integrate this prior knowledge into our pose estimation objective based on a soft-constraint on $\theta \in \mathbb{R}^{27}$. To this end, we enforce that all degrees of freedom stay within their anatomical limits $\theta_{\text{min}} \in \mathbb{R}^{27}$ and $\theta_{\text{max}} \in \mathbb{R}^{27}$:

$$E_{\text{anatomic}}(S) = \lambda_{\text{anatomic}} \sum_{i=1}^{27} \Psi(\theta_i) .$$

Here, $\Psi(x)$ is a quadratic barrier function that penalizes if a degree of freedom exceeds its limits:

$$\Psi(x) = \begin{cases} (x - \theta_{\text{max},i})^2, & \text{if } x > \theta_{\text{max},i} \\ (\theta_{\text{min},i} - x)^2, & \text{if } x < \theta_{\text{min},i} \\ 0, & \text{otherwise} \end{cases} .$$

This term prevents un-plausible human pose estimates.

### 3.4 Non-rigid Surface Registration

The pose estimation step cannot capture realistic non-rigid deformations of skin and clothing that are not explained through skinning. The model therefore does not yet align with the image well everywhere, in particular in cloth and some skin regions. Hence, starting from the pose estimation result, we solve the following non-rigid surface tracking energy:

$$E_{\text{non-rigid}}(V) = E_{\text{data}}(V) + E_{\text{reg}}(V) .$$

The energy consists of several data terms $E_{\text{data}}$ and regularizations $E_{\text{reg}}$, which we explain in the following. Our data terms are a combination of a dense photometric alignment term $E_{\text{photo}}$ and a dense silhouette alignment term $E_{\text{silhouette}}$:

$$E_{\text{data}}(V) = E_{\text{photo}}(V) + E_{\text{silhouette}}(V) .$$

**Dense Photometric Alignment.** The photometric alignment term measures the re-projection error densely:

$$E_{\text{photo}}(V) = \sum_{i \in V} \sigma_{\text{Gauss}}(I_{\text{Gauss}}(\pi(V_i)) - C_i)^2 ,$$

where $C_i$ is the color of vertex $V_i$ in the template model and $\sigma_{\text{Gauss}}(\cdot)$ is a robust kernel that prunes wrong correspondences according to color similarity by setting residuals that are above a certain threshold to zero. More specifically, we project every visible vertex $V_i \in V$ to screen space based on the full perspective camera model $\pi$. The visibility is obtained based on the skinned mesh after the pose estimation step using depth buffering. In order to speed up convergence, we compute the photometric term based on a 3-level pyramid of the input image. We perform one Gauss-Newton iteration on each level. We use the projected positions to sample a Gaussian blurred version $I_{\text{Gauss}}$ of the input image $I$ at the current time step for more stable and longer range gradients. The Gaussian kernel sizes for the 3 levels are 15, 9 and 3 respectively.

**Dense Silhouette Alignment.** In addition to dense photometric alignment, we also enforce alignment of the projected 3D model boundary with the detected silhouette in the input image:

$$E_{\text{silhouette}}(V) = \sum_{i \in B} b_i \left| I_{\text{DT}}(\pi(V_i)) \right|^2 .$$

After Stage I, we first update the model boundary $B$ and consider all vertices $V_i \in B$. These boundary vertices are encouraged to match the zero iso-line of the distance transform image $I_{\text{DT}}$, and thus be aligned with the detected input silhouette. The $b_i$ are computed similar to the pose optimization step (see Sec. 3.3). Due to the non-rigid deformation that cannot be recovered by our pose estimation stage, in some cases the projection of the mesh from Stage I has a gap between body parts such as arms and torso, while in the input image the gaps do not exist. To prevent image silhouettes being wrongly explained by multiple model boundaries we project the posed model $V^5$ into the current frame and compute a body part mask — derived from the skinning weights. We increase the extent of each body part by a dilation (maximum of 10 pixels, the torso has preference over the other parts) to obtain a conservative region boundary that closes the above mentioned gaps. If a vertex $V_i$ moves onto a region with a differing semantic label, we disable its silhouette term by setting $b_i = 0$. This drastically improves the reconstruction quality (see Fig. 4).

Our high-dimensional monocular non-rigid registration problem with only the data terms is ill-posed. Therefore, we use regularization constraints:

$$E_{\text{reg}}(V) = E_{\text{smooth}}(V) + E_{\text{edge}}(V) + E_{\text{velocity}}(V) + E_{\text{acceleration}}(V) .$$

Here, $E_{\text{smooth}}$ and $E_{\text{edge}}$ are spatial smoothness priors on the mesh geometry, and $E_{\text{velocity}}$ and $E_{\text{acceleration}}$ are temporal priors. In the following, we give more details.
encourages small acceleration between adjacent frames.

\textbf{Displacement Warping.} The non-rigid displacements \( \mathbf{d}^{-1} = \mathbf{V}^{t-1} - \mathbf{V}^{S}_{i} \in \mathbb{R}^{3} \) that are added to each vertex \( i \) after skinning are usually similar from frame \( t-1 \) to frame \( t \). We warp \( \mathbf{d}^{-1} \) back to the rest pose by applying Dual Quaternion skinning with the inverse rotation quaternions given by the pose at time \( t-1 \). We refer to them as \( \mathbf{d}^{-1} \). For the next frame \( t \), we transform \( \mathbf{d}^{-1} \) according to the pose at time \( t \) resulting in a skinned displacement \( \mathbf{d}^{S} \), and initialize the non-rigid stage with \( \mathbf{V}^{t} = \mathbf{V}^{S} + \mathbf{d}^{S} \). This jump-starts the non-rigid alignment step and leads to improved tracking quality. Similarly, we add \( \mathbf{d}^{S} \) to the skinned actor model for more accurate dense silhouette alignment during the skeletal pose estimation stage.

\textbf{Vertex Snapping.} After the non-rigid stage, the boundary vertices are already very close to the image silhouette. Therefore, we can robustly snap them to the closest silhouette point by walking on the distance transform along the negative gradient direction until the zero crossing is reached. Vertex snapping allows us to reduce the number of iteration steps, since if the solution is already close to the optimum, the updates of the solver become smaller, as is true for most optimization problems. Therefore, if the mesh is already close to the silhouette, we ‘snap’ it to the silhouette in a single step, instead of requiring multiple iterations of Gauss-Newton. To obtain continuous results, non-boundary vertices are smoothly adjusted based on a Laplacian warp in a local neighborhood around the mesh contour.

4 DATA PARALLEL GPU OPTIMIZATION

The described pose estimation and non-rigid registration problems are non-linear optimizations based on an objective \( E \) with respect to unknowns \( \mathbf{X} \), i.e., the parameters of the kinematic model \( \mathbf{S} \) for pose estimation and the vertex positions \( \mathbf{V} \) for non-rigid surface deformation. The optimal parameters \( \mathbf{X}^* \) are found via energy minimization:

\[ \mathbf{X}^* = \arg \min_{\mathbf{X}} E(\mathbf{X}) \ . \]  

In both capture stages, i.e. pose estimation (see Sec. 3.3) and non-rigid surface tracking (see Sec. 3.4), the objective \( E \) can be expressed as a sum of squares:

\[ E(\mathbf{X}) = \sum_{i} \left[ F_{i}(\mathbf{X}) \right]^{2} = \left\| \mathbf{F}(\mathbf{X}) \right\|_{2}^{2} . \]  

Here, \( \mathbf{F} \) is the error vector resulting from stacking all residual terms. We tackle this optimization at real-time rates using a data-parallel iterative Gauss-Newton solver that minimizes the total error by linearizing \( \mathbf{F} \) and taking local steps \( \mathbf{X}_{k} = \mathbf{X}_{k-1} + \delta_{k} \) obtained by the solution of a sequence of linear sub-problems (normal equations):

\[ J^{T}(\mathbf{X}_{k-1}) J(\mathbf{X}_{k-1}) \cdot \delta_{k} = -J^{T}(\mathbf{X}_{k-1}) \mathbf{F}(\mathbf{X}_{k-1}) \ . \]  

Here, \( J \) is the Jacobian of \( \mathbf{F} \). Depending on the pose estimation or non-rigid registration, the linear systems have a quite different structure in terms of dimensionality and sparsity. Thus, we use tailored parallelization strategies for each of the problems. Since we use Gauss-Newton instead of Levenberg-Marquardt, the residual has not to be computed during the iterations, thus leading
to faster runtimes and in consequence more iterations are possible within the tight real-time constraint.

**Pose Estimation.** The normal equations of the pose optimization problem are small, but dense, i.e., the corresponding system matrix is small, rectangular and dense. Handling each non-linear Gauss-Newton step efficiently requires a specifically tailored parallelization and optimization strategy. First, in the beginning of each Gauss-Newton step, we compute the system matrix \( J^T J \) and right hand side \(-J^T F\) in global memory on the GPU. Afterwards, we ship the small system of size \(36 \times 36\) (\(36 = 3^2 + 3 \times 3\), 3 DoFs for \(R\), 3 for \(t\), 27 for \(\theta\), and 3 for \(\psi\)) to the CPU and solve it based on QR decomposition. The strategy of splitting the computation to CPU and GPU is in spirit similar to [Tagliasacchi et al. 2015]. To compute \( J^T J \) on the GPU, we first compute \( J \) fully in parallel and store it in device memory based on a kernel that launches one thread per matrix entry. We perform a similar operation for \( F \). \( J^T F \) is then computed based on a data-parallel version of a matrix-matrix multiplication that exploits shared memory for high performance. The same kernel also directly computes \( J^T F \). We launch several thread blocks per element of the output matrix/vector, which cooperate in computing the required dot products, e.g., between the \(i\)-th and \(j\)-th column of \(J\) or the \(i\)-th column of \(J\) and \(F\). To this end, each thread block computes a small subpart of the dot product based on a shared memory reduction. The per-block results are summed up based on global memory atoms. In total, we perform 6 Gauss-Newton steps, which turned out to be a good trade-off between accuracy and speed.

**Non-rigid Surface Registration.** The non-rigid optimization problem that results from the energy \( E_{\text{non-rigid}} \) has a substantially different structure. It leads to a large sparse system of normal equations, i.e., the corresponding system matrix is sparse and has a low number of non-zeros per row. Similar to [Innmann et al. 2016; Zollhöfer et al. 2014], during GPU-based data-parallel Preconditioned Conjugate Gradient (PCG) we parallelize over the rows (unknowns) of the system matrix \( J^T J \) using one thread per block row \((x-, y-, z\)-entry of a vertex). Each thread collects and handles all non-zeros in the corresponding row. We use the diagonal of \( J^T J \) as a preconditioner. We perform 3 Gauss-Newton steps and solve the linear system based on 4 PCG iterations, which turned out to be a good trade-off between accuracy and speed.

**Pipelined Implementation.** To achieve real-time performance, we use a data-parallel implementation of our entire performance capture algorithm in combination with a pipeline strategy tailored for our problem. To this end, we run our approach in three threads on a PC with two GPUs. Thread 1 uses only the GPU, which is responsible for data preprocessing. Thread 2 computes the CNN-based human pose detection on the first graphics card, thread 3 solves the pose optimization problem and estimates the non-rigid deformation on the second graphics card. Our distributed computation strategy induces a 2 frame delay, but for most applications it is barely noticeable.

## 5 RESULTS

For all our tests, we employ an Intel Core i7 with two Geforce GTX 1080Ti graphics cards. Our algorithm runs at around 25 FPS, which fulfills the performance requirement of many real-time applications.

In all our experiments, we use the same set of parameters that are empirically determined: \( \lambda_{2D} = 460, \lambda_{3D} = 28, \lambda_{\text{silhouette}} = 200, \lambda_{\text{temporal}} = 1.5, \lambda_{\text{anatomic}} = 10^6, w_{\text{photo}} = 10000, w_{\text{silhouette}} = 600, w_{\text{smooth}} = 10.0, w_{\text{edge}} = 30.0, w_{\text{velocity}} = 0.25 \) and \( w_{\text{acceleration}} = 0.1 \). In the following, we first introduce our new dataset, evaluate our approach on several challenging sequences qualitatively and quantitatively, and compare to related methods. Then, we perform an ablation evaluation to study the importance of the different components of our approach. Finally, we demonstrate several live applications. More results are shown in our two supplementary videos, which in total show over 20 minutes of performance capture results. We applied smoothing with a filter of window size 3 \( (\text{stencil: } [0.15, 0.7, 0.15]) \) to the trajectories of the vertex coordinates as a post process for all video results except in the live setup.

### 5.1 Dataset

In order to qualitatively evaluate our method on a wide range of settings we recorded several challenging motion sequences. These contain large variations in non-rigid clothing deformations, e.g., skirts and hooded sweaters, and fast motions like dancing and jumping jacks. In total, we captured over 20 minutes of video footage split in 11 sequences with different sets of apparel each worn by one of seven subjects. All sequences were recorded with a Blackmagic video camera \((30\text{fps}, 540 \times 960\text{ resolution})\). We provide semantically segmented, rigid and textured templates, calibrated camera parameters and an empty background image for all sequences. In addition, we provide the silhouettes from background subtraction, our motion estimates and the non-rigidly deformed meshes. For eight of the sequences we captured the subject from a reference view, which we will also make available, to evaluate the tracking quality. Fig. 5 shows some of the templates and example frames of the captured sequences. All templates are shown in the supplementary video. We will make the full dataset publicly available.

### 5.2 Qualitative and Quantitative Results

In total, we evaluated our approach on our new dataset and five existing video sequences of people in different sets of apparel. In addition, we test our method with 4 subjects in a live setup (see Fig. 1) with a low cost webcam. Our method takes frames at \(540 \times 960\) resolution as input. To better evaluate our non-rigid surface registration method, we used challenging loose clothing in these sequences, including skirts, dresses, hooded sweatshirts and baggy pants. The sequences show a wide range of difficult motions (slow to fast, self-occlusions) for monocular capture. Additionally, we compare our approach to the state-of-the-art monocular performance capture method of Xu et al. [2018] on two of their sequences and on one of our new captured sequences.

**Qualitative Evaluation.** In Fig. 5, we show several frames from live performance capture results. We can see that our results precisely overlay the person in the input images. Note that body pose, head-orientation, and non-rigid deformation of loose clothing, are accurately captured. Both the side-by-side comparison to RGB input and the accurate overlay with the reconstructed mesh show the high quality of the reconstruction. Also note that our reconstruction results match the images captured from a laterally displaced
Fig. 5. Qualitative results. We show several live monocular performance capture results of entire humans in their loose everyday clothing. (a) shows the template models. (b) shows input images to our method, while (c) shows that the corresponding results precisely overlay the person in the input images. Our results can be used to render realistic images (d) or free viewpoint video (e).
Fig. 6. Our real-time approach even tracks challenging and fast motions, such as jumping and a fast 360° rotation with high accuracy. The reconstructions overlay the input image well. For the complete sequence we refer to the supplemental video.

reference view which is not used for tracking (see supplemental video). This further evidences the fidelity of our 3D performance capture results, also in depth, which shows that our formulation effectively meets the non-trivial underconstrained monocular reconstruction challenge. To evaluate the robustness of our method, we included many fast and challenging motions in our test set. As shown in Fig. 6, even the fast 360° rotation (see the first row) and the jumping motion (see the second row) are successfully tracked. This illustrates the robustness of our algorithm and its efficient and effective combined consideration of sparse and dense image cues, as well as learning-based and model-based capture, which in this combination were not used in prior work, let alone in real-time.

Comparison to Related Monocular Methods. In Fig. 7, we provide a comparison to 3 related state-of-the-art methods: The fundamentally off-line, monocular dense (surface-based) performance capture method of Xu et al. [2018], called MonoPerfCap, and two current monocular methods for 3D skeleton-only reconstruction, the 2D-to-3D lifting method of Zhou et al. [2016] and the real-time VNect algorithm [Mehta et al. 2017]. For the latter two, we show the skinned rendering of our template using their skeleton pose. The test sequence is provided by Xu et al. [2018] with manually labeled ground truth silhouettes. Our method’s results overlay much better with the input than the skeleton-only results of Zhou et al. [2016] and Mehta et al. [2017], confirming our much better reconstructions. Also a quantitative comparison on this sequence in terms of the silhouette overlap accuracy (Intersection over Union, IoU), Fig. 8, shows that our method achieves clearly better results and outperforms [Zhou et al. 2016] and [Mehta et al. 2017] by 8.5% and 9.4%, respectively. Using the same metric, our IoU is only 4.3% smaller than Xu et al. [2018], which is mainly caused by the fact that their foreground segmentation is more accurate than ours due to their more advanced but offline foreground segmentation strategy (see Fig. 10). But please note that our method is overall orders of magnitude faster than their algorithm which takes over 1 minute per frame and our reconstructions are still robust to the noisy foreground segmentation. To compare against MonoPerfCap more thoroughly, we also compare against them on one of our sequences (see Sec. 5.1), which shows more challenging non-rigid dress deformations in combination with fast motions (see bottom rows of Fig. 9). On this sequence, the accuracy of the foreground estimation is roughly the same leading to the fact that our approach achieves an IoU of 86.86% (averaged over 500 frames) which is almost identical to the one of [Xu et al. 2018] (86.89%). As shown in Fig. 9, we achieve comparable reconstruction quality and overlay while being orders of magnitude faster.
LiveCap: Real-time Human Performance Capture from Monocular Video

Fig. 9. Qualitative comparison to MonoPerfCap [Xu et al. 2018]. We achieve comparable reconstruction quality and overlay while being orders of magnitude faster.

MonoPerfCap’s window-based optimizer achieves slightly better boundary alignment and more stable tracking though some difficult, convolved, self-occluded poses, but is much slower. Our reconstruction of head and feet is consistently better than [Xu et al. 2018] due to the additional facial landmark alignment term and the better pose detector that we adopted. We provide a qualitative comparison showing highly challenging motions in the supplementary video.

Surface Reconstruction Accuracy. To evaluate our surface reconstruction error, also relative to multi-view methods, we use the Pablo sequence from the state-of-the-art multi-view template-based performance capture method of Robertini et al. [2016] (they also provide the template). As shown in Fig. 12, our real-time monocular method comes very close in quality to the results of the fundamentally off-line multi-view approach of Robertini et al. [2016] and the monocular off-line method of Xu et al. [2018]. In addition, it clearly outperforms the monocular non-rigid capture method of Yu et al. [2015] and the rigged skeleton-only results of the 3D pose estimation methods of Zhou et al. [2016] and Mehta et al. [2017].

Fig. 10. Comparison of the foreground segmentation of Xu et al. [2018] and our method. Note that our silhouette estimates are less accurate than the ones of [Xu et al. 2018]. Nevertheless, our reconstruction results are robust to the noisy foreground estimates and look plausible.

Fig. 11. Quantitative comparison of the surface reconstruction accuracy on the Pablo sequence. Our real-time monocular approach comes very close in quality to the results of the monocular off-line method of Xu et al. [2018]. It clearly outperforms the monocular non-rigid capture method of Yu et al. [2015] and the rigged skeleton-only results of the 3D pose estimation methods of Zhou et al. [2016] and Mehta et al. [2017].
Fig. 12. Qualitative comparisons of the surface reconstruction accuracy on the Pablo sequence. Our real-time monocular approach comes very close in quality to the results of the fundamentally off-line multi-view approach of Robertini et al. [2016] and the monocular off-line method of Xu et al. [2018]. It clearly outperforms the monocular non-rigid capture method of Yu et al. [2015] and the rigged skeleton-only results of the 3D pose estimation methods of Zhou et al. [2016] and Mehta et al. [2017].

[2015] and the rigged skeleton-only results of the 3D pose estimation methods of Zhou et al. [2016] and Mehta et al. [2017] (latter two as described in the previous paragraph). This is further evidenced by our quantitative evaluation on per-vertex position errors (see Fig. 11). We use the reconstruction results of Robertini et al. [2016] as reference and show the per-vertex Euclidean surface error. Similar to [Xu et al. 2018], we aligned the reconstruction of all methods to the reference meshes with a translation to eliminate the global depth offset. The method of Xu et al. [2018] achieves slightly better results in terms of surface reconstruction accuracy. Similar to our previous experiment (see Fig. 10), we observed that our foreground estimates are slightly worse than the ones of [Xu et al. 2018] which caused the lower accuracy.

Skeletal Pose Estimation Accuracy. We also compare our approach in terms of joint position accuracy on the Pablo sequence against VNect [Mehta et al. 2017], [Zhou et al. 2016] and MonoPerfCap [Xu et al. 2018]. As reference, we use the joint positions from the multi-view method of Robertini et al. [2016]. We report the average per-joint 3D error (in millimeters) after aligning the per-frame poses with a similarity transform. As shown in Fig. 13, our method outperforms the three other methods, most notably the skeleton-only methods [Mehta et al. 2017; Zhou et al. 2016]. This shows that our combined surface and skeleton reconstruction also benefits 3D pose estimation quality in itself.

Ablation Study. We first qualitatively evaluate the importance of all algorithmic components in an ablation study on a real video sequence. To this end, we compare the results of: 1) our pose estimation without facial landmark alignment term and the silhouette term, which we refer to as $E_{2Dw} + E_{3D}$, 2) our pose estimation without the silhouette term ($E_{2D} + E_{3D}$), 3) our complete pose estimation ($E_{pose}$) and 4) our full pipeline ($E_{pose} + E_{non-rigid}$). As shown in Fig. 14, 1) the facial landmark alignment term significantly improves the head orientation estimation (red circles), 2) the misalignment of $E_{2D} + E_{3D}$ is corrected by our silhouette term in $E_{pose}$ (yellow circles), 3) the non-rigid deformation on the surface, which cannot be modeled by skinning, is accurately captured by our non-rigid registration method $E_{non-rigid}$ (blue circles). Second, we also quantitatively evaluated the importance of the terms on a sequence where high-quality reconstructions based on the multi-view performance capture results of De Aguiar et al. [2008] are used as ground truth. The mean vertex position error shown in Fig. 15 clearly demonstrates the consistent improvement by each of the algorithmic components of our approach. The non-rigid alignment stage obtains on average better results than the pose-only
Fig. 14. Ablation study. 1) the facial landmark alignment term significantly improves the head orientation estimation (red circles), 2) the misalignment of $E_{2D} + E_{3D}$ is corrected by our silhouette term in $E_{pose}$ (yellow circles), 3) the non-rigid deformation on the surface, which cannot be modeled by skinning, is accurately captured by our non-rigid registration method $E_{non-rigid}$ (blue circles).

Fig. 15. Ablation study. The mean vertex position error clearly demonstrates the consistent improvement by each of the algorithmic components of our approach. Our full approach consistently obtains the lowest error.

Fig. 16. Improvement of the non-rigid stage ($E_{pose} + E_{non-rigid}$) over pose-only deformation ($E_{pose}$). Top row: Four monocular input images. On the bottom row, for each image, we show the per-vertex error of the pose only results minus the per-vertex error of our method. Consequently, negative means pose only is better and is colored in red. Positive means our method is better and is colored in blue. As expected our method achieves most improvement on the non-rigid skirt part—which is around 20mm for the shown frames.

Fig. 17. Importance of our material-based non-rigid deformation adaptation strategy. With a low global regularization weight the deformation of the skirt is well reconstructed, but the head is distorted (left). A high deformation weight preserves the shape of the head, but prevents tracking of the skirt motion (middle). Our new semantic weight adaptation strategy enables the reconstruction of both regions with high accuracy and leads to the best results (right).

alignment. Since non-rigid deformations are most of the time concentrated in certain areas, e.g., a skirt, and at certain frames when articulated motion takes place, we also measure the per-frame and per-vertex improvement of the proposed non-rigid stage. To this end, we measure the improvement of $(E_{pose} + E_{non-rigid})$ over $(E_{pose})$ by computing the per-vertex error of the pose only results minus the per-vertex error of our method. Consequently, positive means our method is better than the pose-only deformation. As demonstrated in Fig. 16, the non-rigid stage significantly improves the reconstruction of the skirt and the arm. The improvement is especially noticeable for frames where the deformation of the skirt significantly differs from the static template model, since such motion cannot be handled by the pose-only step. On the same dataset we also evaluated the influence of 1) the warping of the non-rigid displacement of the previous frame, 2) the proposed body part masks used in the dense silhouette alignment, and 3) the proposed vertex snapping. Those algorithmic changes respectively lead to 2.4%, 1.7% and 1.7% improvement in average 3D vertex error which sums up to a total improvement of 5.8%. The importance of our material-based non-rigid deformation adaptation strategy is shown in Fig. 17. With constantly low non-rigidity weights ($s_{i,j} = 2.0$) in all regions, the deformation of the skirt is well reconstructed, but the head is severely distorted (left). In contrast, with high global non-rigidity weights ($s_{i,j} = 50.0$), the head shape is preserved, but the skirt cannot be
We have demonstrated compelling real-time full-body human performance capture. This illustrates the potential of our method in several of the aforementioned live application domains. In Fig. 19, we demonstrate a real-time virtual try-on application based on our performance capture method. We replace the texture corresponding to the trousers on the template and visualize the tracked result in real time. With such a system, the users can see themselves in clothing variants in real time with live feedback, which could be potentially used in VR or even AR online shopping.

5.3 Applications

Our monocular real-time human performance capture method can facilitate many applications that depend on real-time capture: interactive VR and AR, human-computer interaction, pre-visualization for visual effects, 3D video or telepresence. We exemplify this through two application demonstrators. In Fig. 18, we show that our method allows live free-viewpoint video rendering and computer animation of the performance captured result from just single color input. This illustrates the potential of our method in several of the aforementioned live application domains. In Fig. 19, we demonstrate a real-time virtual try-on application based on our performance capture method. We replace the texture corresponding to the trousers on the template and visualize the tracked result in real time. With such a system, the users can see themselves in clothing variants in real time with live feedback, which could be potentially used in VR or even AR online shopping.

6 DISCUSSION AND LIMITATIONS

We have demonstrated compelling real-time full-body human performance capture results using a single consumer-grade color camera. Our formulation combines constraints used, individually, in different image-based reconstruction methods before. But the specific combination we employ embedded in a hierarchical real-time approach is new and enables, for the first time, real-time monocular performance capture. Further, our formulation geared rigorously for real-time use differs from the related, but off-line MonoPerfCap [Xu et al. 2018] method in several ways: In Stage I, the facial landmarks as well as the displacement warping which is also added during pose tracking improve the pose accuracy of our real-time method. Further, we track the pose per frame instead of a batch-based formulation which reduces the computation time and allows faster motions. Further improvement in terms of efficiency are achieved by our GPU-based pose solver. In Stage II, our dense photometric term that adds constraints for non-boundary vertices and our adaptive material based regularization improve reconstruction quality. Our non-rigid fitting stage is faster due to the more efficient combination of spatial regularizers that requires a much smaller number of variables than the as-rigid-as-possible regularizer. We directly solve for the vertex displacements instead of estimating the embedded graph rotations/translations. We found that this formulation is better suited for a parallel implementation on the GPU and it also gives a more flexible representation. Due to our real-time constraint, we make use of an efficient distance transform-based representation, instead of the ICP-based approach that requires expensive search of correspondences between the model boundary and the image silhouettes. Our experiments show that our method achieves a similar reconstruction quality compared to the off-line performance capture approach of Xu et al. [2018] while being orders of magnitude faster.

Nonetheless, our approach is subject to some limitations, see also Fig. 20. Due to the ambiguities that come along with monocular performance capture, we rely on an accurate template acquisition since reconstruction errors and mislabeled part segmentations in the template itself cannot be recovered during tracking. Further, we cannot handle topological changes that are too far from the template, e.g., removing of some clothes and deformations along the camera viewing axis can only be partially recovered by our photometric term. The latter point could be addressed by an additional term that involves shading and illumination estimation. As is common for learning methods, the underlying 3D joint regression deep network fails for extreme poses not seen in training. Our model fitting can often, but not always correct such wrong estimates which produces glitches in the tracking results. However, our performance capture approach robustly recovers from such situations, see Fig. 20 (top). Since our method uses foreground/background segmentation, strong shadows and shading effects, objects with similar color to the performer, and changing illumination situations can cause suboptimal segmentation; thus leading to noisy data association in the silhouette alignment term, which manifests itself as high-frequency jitter. Our approach is robust to some degree of miss-classifications, but can get confused by big segmentation outliers. This could be alleviated in the future by incorporating more sophisticated background segmentation strategies, e.g., based on deep neural networks. Strong changes in shading or shadows, specular materials or non-diffuse lighting can also negatively impact the color alignment term. A joint optimization for scene illumination and material properties could alleviate this problem. Even though we carefully orchestrated the components of our method to achieve high accuracy and temporal stability in this challenging monocular setting, even under non-trivial occlusions, extensive (self-)occlusion are still fundamentally difficult. Our estimates for occluded parts will be less accurate than with multi-view methods due to the lack of image evidence. While pose and silhouette plausibly constrain the back-side of the body, fully-occluded limbs may have incorrect
poses. Additional learned motion priors could further resolve such ambiguous situations. Fortunately, our approach recovers as soon as the difficult occlusions are gone, see Fig. 20 (bottom).

7 CONCLUSION

We have presented the first monocular real-time human performance capture approach that reconstructs dense, space-time coherent deforming geometry of entire humans in their loose everyday clothing. Our novel energy formulation leverages automatically identified material regions on the template to differentiate between different non-rigid deformation behaviors of skin and various types of apparel. We tackle the underlying non-linear optimization problems at real-time based on a pipelined implementation that runs two specially-tailored data-parallel Gauss-Newton solvers, one for pose estimation and one for non-rigid tracking, at the same time. We robustly recover from such situations. Bottom row: Our estimates for occluded parts will be less accurate than with multi-view methods due to the lack of image evidence. While pose and silhouette plausibly constrain the back-side of the body, fully-occluded limbs may have incorrect poses.

REFERENCES


Lee, Yinghui; Jia, Xin; Liu, Song; Wang, Peng; and Zhang, Shihao. 2017. 3D Tracking by Detection. In Proc. CVPR.


LiveCap: Real-time Human Performance Capture from Monocular Video


Received September 2018; accepted January 2019; final version January 2019.