Illumination-invariant Robust Multiview 3D Human Motion Capture

Nadia Robertini[1,2], Florian Bernard[1], Weipeng Xu[1], Christian Theobalt[1]
[1] MPI for Informatics, [2] Intel Visual Computing Institute, Saarbrücken, Germany
{nroberti,fbernard,wxu,theobalt}@mpi-inf.mpg.de

Abstract

In this work we address the problem of capturing human body motion under changing lighting conditions in a multi-view setup. In order to account for changing lighting conditions we propose to use an intermediate image representation that is invariant to the scene lighting. In our approach this is achieved by solving time-varying segmentation problems that use frame- and view-dependent appearance costs that are able to adjust to the present conditions. Moreover, we use an adaptive combination of our lighting-invariant segmentation with CNN-based joint detectors in order to increase the robustness to segmentation errors. In our experimental validation we demonstrate that our method is able to handle difficult conditions better than existing works.

1. Introduction

Image-based marker-less human motion capture is an important and long-standing problem in computer vision. In the past decades, there has been a constantly growing demand on robust human motion capture algorithms from a wide range of application fields, such as computer animation, video effects (VFX) and biomechanical analysis.

Most of the existing methods for human motion capture rely on a controlled environment, which typically means a uniformly illuminated studio [8, 17, 9, 16, 35, 38]. Some of them even require a studio covered with green screen. In recent years, several methods have proven their effectiveness in general outdoor scenarios [12, 31, 13]. However, since these methods are not robust to complicated lighting conditions, drastic changes of illumination and harsh shadows easily lead to tracking failures. Furthermore, most of the generative methods require consistent appearance of the captured images for tracking, which does not allow for reconstructing the model and tracking the motion in noticeably varying lighting conditions. While recent data-driven methods have been shown to be more robust to difficult lighting conditions due to the strong generalization ability of deep neural networks [26], these methods alone cannot achieve as high accuracy as generative methods [46].

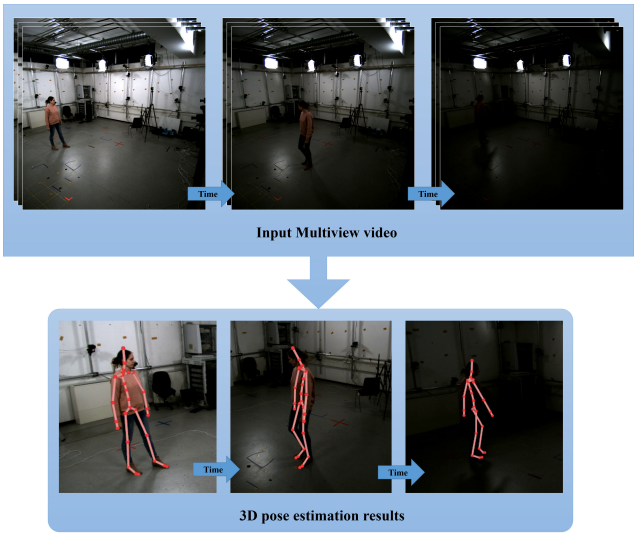

Figure 1. Our method estimates 3D human pose from multiview image sequences. Even under extreme lighting conditions, our method is still able to track the motion successfully.

In this paper, we address the problem of consecutively tracking the articulated motion of a human body through a given image sequence. In order to do so, we introduce a novel human motion capture approach that fits a parameterized human body model to a combination of an abstract intermediate image representation that is robust to lighting changes and joint detections based on a convolutional neural network (CNN). Similar to [12, 31, 13], our method relies on a human body model and multiview image sequences captured with calibrated cameras. The main insight of our work is to factor out the lighting influence from the images by extracting the reflectance (or albedo) component of the image, and then perform motion tracking using the albedo channel. To this end, for each frame of the multiview sequence, the proposed method estimates the albedo channel by solving a lighting-invariant segmentation problem. Our segmentation problem is phrased as a multi-labeling problem based on a pairwise Markov Random Field (MRF) [7], where each "material" of the human body is assigned a unique label that corresponds to its albedo. Lighting-invariance is achieved by dynamically updating the MRF

data term to reflect the changing lighting conditions. This is implemented by combining the image appearance and a pose prediction prior in a suitable way. In order to increase the overall robustness, we additionally employ CNN-based joint position detections, which have been shown to have remarkable generalization ability [11, 49, 23, 39, 25, 28, 26]. Nevertheless, we have found that the CNN detector alone may be noisy and struggles under some conditions. In an attempt to combine the advantages of both worlds, we use an adaptive strategy for setting the relative weighting between the segmentation and the detectors, which are then embedded into a model-based tracking method that is based on a Gaussian blob tracker [38].

Contributions: In summary, in this paper we introduce a novel approach for human motion capture under complicated lighting conditions. Our main technical contributions to achieve this goal are:

1. the formulation of suitable *time-varying segmentation problems* that use *frame- and view-dependent appearance costs* in order to obtain lighting-invariant representations of the individual images,
2. the *adaptive combination of an abstract intermediate image representation with CNN-based joint detectors*, and
3. the *integration* of this combined information *into a robust model-based tracker*.

Moreover, we demonstrate the effectiveness of our approach on several challenging sequences, including drastic lighting changes as well as harsh shadows. Our quantitative and qualitative results evidence that our approach accurately tracks the human pose and outperforms the existing methods in such challenging scenarios. We make our multiview sequences publicly available[1].

2. Related Works

3D human motion capture has received a lot of attention in recent decades. Many vision-based marker-less methods have been proposed to address this problem. A large portion of the literature focused on generative model-based multiview motion capture, where the goal is to optimize the overlap of the projected 3D body model with multiview images. In this context, many methods rely on the silhouette input obtained by background subtraction [38, 4, 21, 15, 24, 3]. As the silhouette cue provides a strong constraint and drastically reduces the difficulty of the problem, those methods yield accurate motion capture results. Multiple character tracking, even with complex interaction, has also been enabled [24]. To this end, they first segment the provided foreground image region into different characters based on the color models and the pose prior. Then the different characters are tracked independently. In our method, the pose term for material segmentation is inspired by this work. However, our method segments the images with respect to different materials instead of different characters, and does not rely on the silhouette input. This allows for automatic tracking as well as better estimation of the correspondences between the body model and the images. Although plausible results have been achieved by multiple character tracking methods, their application scenario is restricted to static backgrounds, since the background subtraction does not work on dynamic background.

There also exist several approaches that do not require explicit silhouette input [8, 17, 9, 16, 35, 38]. Even real time 3D tracking is achieved using a Gaussian representation of the images and the body model [38]. However, only the results in a well-controlled studio have been demonstrated. Cluttered background in a general outdoor scenario typically leads to tracking failures.

Recently, two variants of [38] have shown their effectiveness on outdoor human motion capture [13, 32]. Specifically, the occlusion problem is better handled with a novel translucent medium shape representation as used in [32]. However, the varying illumination problem in outdoor motion capture is not addressed. In contrast, Elhayek et al. proposed an improved model-to-image consistency energy in weighted HSV color space, which is more resistant to intensity changes [13]. Although the tracking failure due to illumination changes is alleviated to some extent, we have found in our experiments that using the color consistency energy in HSV space alone is not enough to handle unconstrained complex lighting conditions for outdoor tracking. To handle varying illumination, Wu et al. proposed to simultaneously estimate the illumination and track the motion in a joint optimization framework [44]. However, since the illumination estimation and motion tracking is performed in an alternating manner, and each step relies on the previous step being correct, their method is not able to recover from errors. Besides, it is also worth mentioning that the computational complexity of their method is rather high due the illumination estimation.

In contrast to generative model-based methods, data-driven methods address the 3D pose estimation problem from the perspective of image feature extraction, regression or classification [1, 19, 34, 27, 36]. With the enormous success of deep learning methods and the thus resulting growing popularity, many CNN-based approaches have recently been proposed to predict 3D human pose from monocular images. A common approach is to lift the 2D joint prediction to 3D using temporal constraints and/or pose priors [50, 47, 48], while many other methods directly estimate the 3D pose from single images [11, 49, 23, 39, 25, 28, 26]. Even real-time 3D pose estimation has been achieved [26]. The CNN-based methods are typically more robust to illumination changes in the outdoor scenario than genera-

[1] http://gvv.mpi-inf.mpg.de/projects/IntrinsicMoCap/index.html

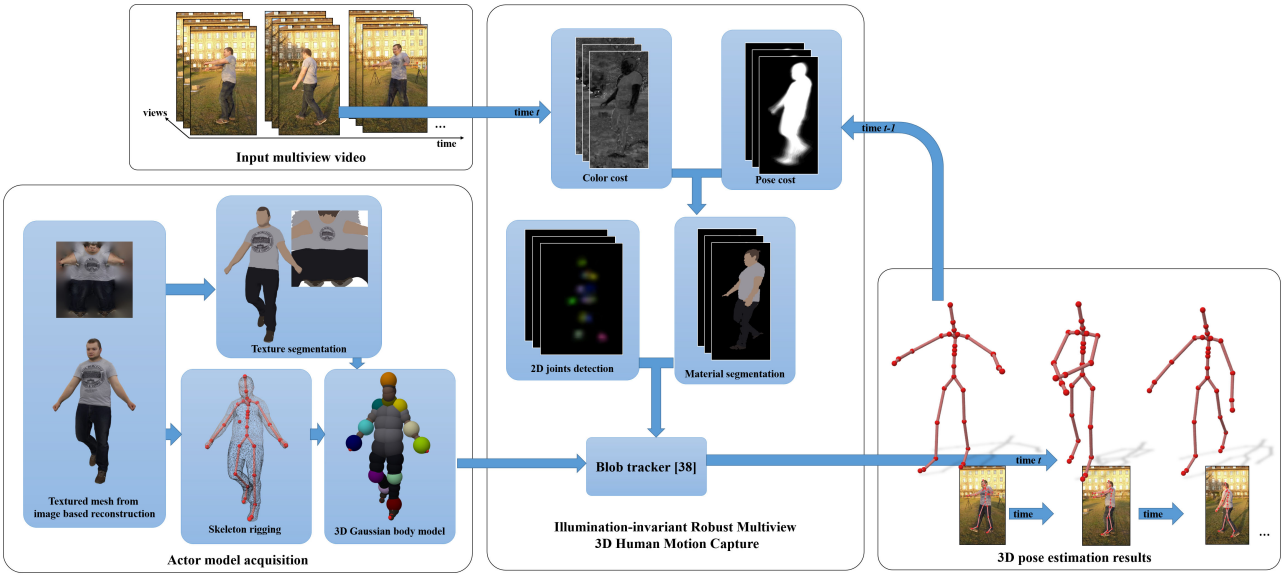

Figure 2. Pipeline of our method.

tive methods due to the strong generalization ability of the deep neural network, but those monocular-based methods are usually less accurate since they suffer from the inherent depth ambiguity. To address such problems, multiview-based discriminative methods [37, 10, 2, 40] have been proposed. However, they typically have lower temporal stability than generative methods and, more importantly, they use a simplified body model with only few degrees of freedom.

Several recent methods combine the generative model and the discriminative approach to benefit from the beauty of both sides [31, 14, 30, 5]. In particular, the methods in [31] and [14] share a similar outdoor multiview motion capture setting with our method. In contrast to their methods, our approach alternatively tracks the skeletal motion and estimates the intrinsic segmentation to factor out the illumination changes. Our experimental results demonstrate that our approach significantly outperforms the existing methods under varying illumination.

3. Methods

We now describe our model-based motion capture approach, which is summarised in Fig. 2. Given a sequence of multiview images capturing the action of a single actor, our goal is to consecutively track the human body pose in each frame, resulting in the temporally coherent skeletal motion across the entire sequence. Similar to existing methods (e.g. [31]), our approach leverages a parameterized human body model. To handle complex lighting condition, for each frame, we estimate a lighting-invariant segmentation of the images, and then incorporate the segmentation into the skeleton tracking. Since the focus of this paper is on the tracking part, we assume that the model is aligned to the image in the first frame of a sequence, similar to [38, 33].

In the rest of this section, we first discuss our method for parameterized appearance model acquisition, then describe our segmentation approach, and finally present the skeleton tracking method.

3.1. Actor Model

Our approach relies on a person-specific human body model, which consists of a triangulated mesh model with associated texture segmentation, see Fig. 2. To obtain the actor model, we capture images of the actor from different view points. Then the textured mesh model is reconstructed using the image-based 3D reconstruction software *Agisoft PhotoScan*[2]. Afterwards, the texture is semi-automatically segmented into different regions according to the albedos of different materials (such as skin, shirt, etc.). To this end, we first apply the image smoothing method of [45] on the texture image to remove the high frequency shading components while preserving features in the image. Then we manually annotate the smoothed texture image, resulting in the material texture segmentation that assigns a unique label to each material. Note that while we assume that each material is homogeneous, non-homogeneous parts (e.g. a shirt with a logo) can be modelled by introducing sub-materials.

The articulated motion of the actor is represented based on a kinematic skeleton (see Fig. 2), as done in [22]. The skeleton has 24 joints and is parameterized with a 48-dimensional vector $\mathcal{S}$ containing translation and rotation of the root joint, and 42 joint angles (each joint has between 1 and 3 degrees of freedom). The used actor mesh is rigged to the skeleton based on dual quaternion skinning [20] where the skinning weights are computed automatically.

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

Figure 3. Examples of our segmentation.

3.2. Lighting-invariant Segmentation

We now describe how to obtain a lighting-invariant representation of a multiview frame $(I_{1,t}, \ldots, I_{V,t})$ at time $t{>}0$.

Problem statement: The objective of the lighting-invariant segmentation (Fig. 3) is to assign to each pixel i (with position $x_i \in \Omega$) of image I a label $\ell_i \in \mathcal{L}$ that indicates which material is seen in that pixel (for the sake of simplicity we consider the background to be a material). By $L := |\mathcal{L}|$ we denote the total number of labels. A labelling $\boldsymbol{\ell} \in \mathcal{L}^{|I|}$ for image I is obtained by minimizing an energy of the form

$$E(\boldsymbol{\ell}) = \sum_{i=1}^{|I|} E_i(\ell_i) + \sum_{i \sim j} E_{ij}(\ell_i, \ell_j)\,, \tag{1}$$

where $E_i(\ell)$ is the data term that measures the cost for assigning label ℓ to pixel i, and E_{ij} is the smoothness term that penalizes neighboring pixels i and j that are assigned different labels ($i \sim j$ indicates that i and j are neighbors).

3.2.1 Data term

The data term $E_i(\ell)$ measures the cost of assigning the material $\ell \in \mathcal{L}$ to pixel i. It is defined by weighting the appearance cost $E_i^{\mathrm{a}}(\ell)$ with the pose cost $E_i^{\mathrm{p}}(\ell)$, i.e.

$$E_i(\ell) = E_i^{\mathrm{a}}(\ell) \cdot E_i^{\mathrm{p}}(\ell)\,, \tag{2}$$

which are to be introduced below.

Pose costs: In order to define the pose costs, we make use of pose preditions. To this end, for each material $\ell \in \mathcal{L}$ we estimate a pose probability image $H_\ell : \Omega \to [0,1]$, where $H_\ell(x_i)$ denotes the probability that pixel i belongs to material ℓ (note that $\sum_{\ell \in \mathcal{L}} H_\ell(x) = 1$ for each x). In order to estimate the pose probability image for the current frame, we sample 50 random pose parameters from a Gaussian distribution around the current pose parameter prediction $\mathcal{S}^t$ at time t, which is obtained based on the acceleration computed from the pose parameters of the previous two frames, i.e. $\mathcal{S}^{t-1}$ and $\mathcal{S}^{t-2}$. For each of the 50 mesh samples, we project the mesh onto the image plane. Combining the projection and the material texture of the mesh, we compute the 2D pose probability image from these 50 projections, which are normalized so that they sum up to one for each pixel. By thresholding H_ℓ, we extract a pose prediction mask $J_\ell : \Omega \to \{0,1\}$, where $J_\ell(x_i) = 1$ means that the pose prediction at time t based on the pose at times $t{-}1$ and $t{-}2$ would allow that pixel i belongs to material ℓ.

With that, we define the predicted pose cost as

$$E_i^{\mathrm{p}}(\ell) := 1 - H_\ell(x_i)\,. \tag{3}$$

Image features: In order to (partially) factor out global illumination, instead of using the RGB color space we employ the *hue* and *saturation* components in HSV color space and ignore the *value* component. Moreover, since the *hue* component is represented as angle, by using its sine and cosine we deal with the periodicity in order to ease further processing. With that, we obtain the feature image $\Phi : \Omega \to [0,1]^3$, where the first and second components are the sine and cosine of the *hue*, respectively, and the third component is the *saturation*. Since the same material may have a different appearance when viewed from different cameras we treat each view v independently, and thus assume in the remainder of this section that the view v and the time t are fixed. Hence, for notational convenience we write Φ or I in place of $\Phi_{v,t}$ or $I_{v,t}$.

Frame-dependent appearance costs: In order to improve upon the (limited) lighting-invariance that we gain when considering the feature image Φ in place of the RGB image I, we consider a (frame-dependent) robustified version of the *Mahalanobis distance* to measure the discrepancy between the observed feature vector $\Phi(x_i)$ of a given pixel i and the expected feature vector μ_ℓ for material ℓ. By dynamically updating μ_ℓ in each frame we take the (possibly changing) lighting conditions implicitly into account.

Using the mask J_ℓ, we extract all feature vectors for material ℓ, which we denote as $X_\ell := \{\Phi(x) : J_\ell(x) = 1\}$. We use the *geometric median* $\mu_\ell \in [0,1]^3$ as robust representation of the "typical" feature vector in X_ℓ. The geometric median is given by

$$\mu_\ell := \mu(X_\ell) = \arg\min_y \sum_{x \in X_\ell} \|x - y\|_2\,, \tag{4}$$

which admits an efficient solution [43, 18]. Note that when the ℓ_2-norm in (4) is replaced by the squared ℓ_2-norm, one obtains the mean, whereas using the ℓ_1-norm results in the coordinate-wise median. In addition to μ_ℓ, we estimate a robust "covariance matrix" of X_ℓ based on the geometric median μ_ℓ, which is given by

$$C_\ell := \frac{1}{N-1} \sum_{x \in X_\ell} (x - \mu_\ell)(x - \mu_\ell)^T. \tag{5}$$

Using the dynamically updated geometric median μ_ℓ and the "covariance" C_ℓ, for all foreground materials $\ell_1, \ldots, \ell_{L-1}$ we define the appearance cost E^{a}, in the spirit

of the *Mahalanobis distance*, as

$$E_i^{\mathrm{a}}(\ell) := (\Phi(x_i) - \mu_\ell)^T C_\ell^{-1} (\Phi(x_i) - \mu_\ell)\,. \qquad (6)$$

Since the background (having label ℓ_L) is in general inhomogeneous, a single-modal model in the Φ-feature space as assumed in (6) is inappropriate. Instead of using a multimodal model, for the background we consider a lifted feature vector Φ_{bg} by augmenting Φ with the already predicted foreground costs, i.e.

$$\Phi_{\mathrm{bg}}(x_i) = [\Phi(x)^T, E_i^{\mathrm{a}}(\ell_1), \ldots, E_i^{\mathrm{a}}(\ell_{L-1})]^T\,. \qquad (7)$$

By embedding the computed foreground costs into the $(3{+}(L{-}1))$-dimensional background feature vector Φ_{bg}, we have found that a single-modal model in this higher-dimensional feature space is able to provide sufficient discriminability for the background. The background appearance costs are then computed as in (6) with Φ_{bg} in place of Φ, and μ_L and C_L being computed from the predicted background mask using $X_L = \{\Phi_{\mathrm{bg}}(x) \;:\; J_L(x) = 1\}$.

3.2.2 Smoothness term

In order to achieve a piecewise constant labelling $\boldsymbol{\ell}$ of image I, we use a smoothness term that penalizes neighboring pixels that are assigned different labels. The Potts model [29] is a robust discontinuity-preserving interaction potential that is given by

$$E(\ell, \ell') := \min(1, |\ell - \ell'|)\,. \qquad (8)$$

The pairwise term used in (1) is now given by the generalized Potts model [6]

$$E_{ij}(\ell, \ell) := \lambda_{\mathrm{s}} \omega_{ij} E(\ell, \ell') \quad \forall\; i \sim j\,. \qquad (9)$$

The scalar $\lambda_{\mathrm{s}} > 0$ is a fixed weight, $\omega_{ij} \geq 0$ is a weight that depends on neighboring pixels i, j and is defined as

$$\omega_{ij} = \exp\left(\frac{\|I(x_i) - I(x_j)\|_2^2}{2}\right)\,. \qquad (10)$$

The purpose of the weight ω_{ij} is to increase the cost for assigning different labels to neighboring pixels that have similar color appearance, and to decrease the cost if the color appearance is different.

3.2.3 Minimization of the MRF energy (1)

In order to minimize (1) we use the alpha-expansion algorithm [7] that has appealing properties both from a theoretical and from a practical point-of-view. On the one hand, when minimizing energy (1) with a generalized Potts model as smoothness term as in (9), the alpha-expansion algorithm has the guarantee that the so-obtained local optimum lies within a factor of the global optimum [41]. Moreover, the alpha-expansion algorithm is very efficient and is known to produce good solutions in practice.

3.3. Pose Tracking

Given the material segmentation of the multiview images in the current frame, our task is now to estimate the pose parameters $\mathcal{S}$.

Gaussian blob tracker: In order to perform tracking, we adopt the SoG-based skeleton tracking method of [38]. To be more specific, we attach a sum of 3D spatial Gaussians (SoG) to the skeleton in order to approximate the actor's body model (cf. Fig. 2). The 3D SoG body model can be expressed as:

$$\mathcal{M}(x) = \sum_{i=1}^{N} \exp(-\frac{\|x - \bar{x}_i\|_2^2}{2\sigma_i^2}), \qquad (11)$$

where x is a 3D coordinate, N is the number of spatial Gaussians, and $\bar{x}_i$ and σ_i^2 are the mean and the variance of the i-th Gaussian blob, respectively. Then, the 2D images and the 3D body model are approximated as 2D and 3D SoG respectively, and the skeleton parameters are obtained by maximizing the overlap of the 2D image SoG and the projected 3D body model SoG. For further details on this approach we refer the reader to [38].

In addition to the blobs that represent the human body shape, we create 14 special detection Gaussians placed at the 14 most prominent joints (cf. Fig. 2), which are used to match the skeleton pose with CNN-based 2D detections obtained by the *convolutional pose machine* approach [42].

In contrast to the original tracking method [38], instead of the raw images we use the segmented material images (as described in section 3.2) in combination with the heatmaps of the 14 CNN-based joint position detections.

Adaptive weighting: In order to improve upon the robustness and to prevent that wrong segmentations lead to error propagation, we employ an adaptive weighting strategy to set the relative importance between the material segmentation image and the CNN-based joint detections. Let w_s and w_d be the weights of the segmentation and joint detections that are used in the blob tracker. Initially, the weights are set to w_s=0.8 and w_d=0.2. In order to check for unreliable segmentations, we compute for each material $\ell_1, \ldots \ell_L$ the norm of the difference s_ℓ of the geometric median μ_ℓ and its value of the previous frame μ_ℓ^{t-1}, i.e. $s_\ell = \|\mu_\ell - \mu_\ell^{t-1}\|_2$. If any of the s_ℓ for $\ell = \ell_1, \ldots \ell_L$ is larger than the threshold θ, where we used $\theta \approx 0.08$, we consider the segmentation as failure. In the case of a failure, we update $w_s = \frac{w_s}{2}$ to decrease the relative importance of the segmentation, otherwise we set $w_s = \max(0.8, w_s + 0.1)$. The weight w_d is obtained as $w_d = 1 - w_s$.

4. Experimental Results

We evaluated our proposed approach on 5 outdoor and 2 indoor sequences. The outdoor sequences include harsh (e.g. *walk1_outdoor*) and soft shadow, while the indoor light

Figure 4. Qualitative results. The figure shows tracking results obtained with our approach on both indoor and outdoor sequences (columns) and different frames (rows). From the top to the bottom: $girl\_indoor$, $boy\_outdoor$, $girl1\_outdoor$, boy_indoor, $walk1\_outdoor$.

changes are simulated by randomly switching on and off a subset of the studio lights.

Runtime: Our current implementation typically takes around half a minute per multiview frame when using 8 views. The biggest overhead is the estimation of color and pose costs for each camera and pixel. While runtime was not a key concern for us, we believe that it can be drastically improved using parallel computing techniques.

Qualitative results: In Fig. 4, we provide some examples of the motion capture results obtained with our method. We can see from these results that our method accurately tracked the skeletal motion of the actors in both outdoor and indoor scenarios. Note that, in the outdoor sequences of row 2 and 3, although the illumination changes significantly as the actors walk into or out of the shadow, our method is still able to track the motion stably. In the sequence of row 5, there is a harsh shadow casted on the actor, but our method still yields successful tracking results. The benefit of our approach is further evidenced in the simulated varying illumination scenarios of row 1 and 4. Even in the extreme case of row 1, where the actor is hardly visible for a human observer, our method still works in such a challenging scenario. The complete results on all sequences are provided in our supplementary video.

Furthermore, we compare our approach with the method by Rhodin et al. [31] and the Gaussian blob tracker [38] on the *boy_outdoor* sequence, for which we provide the ground truth. For the latter, we evaluated two different scenar-

Figure 5. Comparison of estimated poses in the *boy_outdoor* sequence for different methods. From left to right: ours, *BT+Det*, *Rhodin et al.* and *BT*. The circles point to tracking failures.

ios. In the first scenario, we apply the blob tracker to the raw RGB images, which we denote as *BT*. In the second scenario, we use the CNN-based joint detection blobs only (cf. section 3.3) without considering the material segmentation, which we denote by *BT+Det* in the results. The qualitative comparison is shown in Fig. 5. From these results, we can see that our method yields more accurate tracking results than the existing methods.

Quantitative results: Quantitative results evaluating those methods on the *boy_outdoor* sequence are shown in Fig. 6, as well as in Tables 1 and 2. In Fig. 6 we show the percentage of correct keypoints (3DPCK [25]) for four end-effector joints across the entire sequence comprising 400 frames from 8 different views, with Table 1 summarizing the corresponding area under the curve (AUC). In Table 2 we summarise the 3D ground truth error (in cm) for the four methods for five end-effector joints. Overall, our method and the *BT+Det* method significantly outperform the other two approaches. While results of our method and

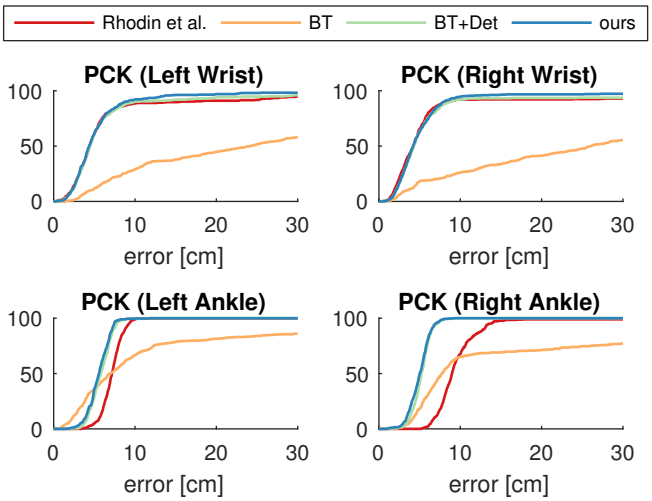

Figure 6. 3DPCK values of end-effector joint positions for *boy_outdoor* sequence when using 8 cameras. The value on the vertical axis shows the percentage of frames where the error is smaller than or equal to the value on the horizontal axis. The corresponding AUC values are shown in Table 1.

Table 1. Area under curve (AUC) values of 3DPCK curves in Fig. 6.

	Rhodin et al.	BT	BT+Det	ours
LW	0.9249	0.6858	0.9295	**0.9428**
RW	0.9298	0.6023	0.9326	**0.9451**
LA	0.8839	0.7979	0.9075	**0.9114**
RA	0.8279	0.7003	0.9061	**0.9105**

Table 2. Numerical summary of average ground truth errors (in cm) and standard deviation for five different joints (LW: left wrist, RW: right wrist, LA: left ankle, RA: right ankle, N: neck).

	Rhodin et al.	BT	BT+Det	ours
LW	7.35±9.68	30.92±26.18	6.89±8.77	**5.59**±4.91
RW	7.20±11.39	41.02±35.09	6.91±9.73	**5.61**±6.32
LA	7.28±3.05	12.69±14.35	5.79±1.28	**5.55**±1.33
RA	9.60±3.34	16.73±16.82	5.22±1.14	**4.98**±1.25
N	3.23±1.45	8.34±5.71	2.91±1.18	**2.86**±1.26

BT+Det are comparable in Fig. 6, Tables 1 and 2 reveal that our method results in a higher AUC and reduced errors, respectively. In addition, in Fig. 7 we also evaluate the performance of our method and *BT+Det* depending on the number of views.

5. Discussion & Limitations

Our segmentation method is robust thanks to the dynamic appearance term. Moreover, since the blob tracker is based on (smooth) Gaussian blobs, it does not consider sharp object boundaries/edges, so that non-perfect boundary segmentations are very likely to still lead to good tracking results. In addition, due to the adaptive weighting

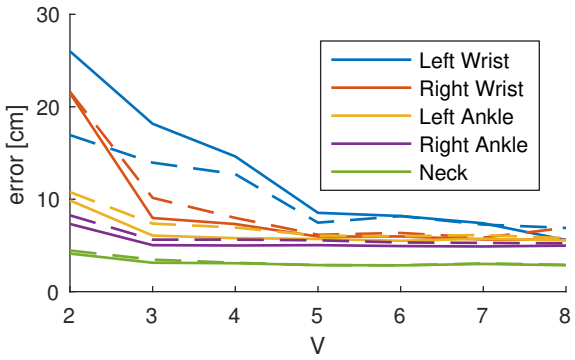

Figure 7. Ground truth error (vertical axis) of our method (solid lines) and *BT+Det* (dashed lines) depending on the used number of views (horizontal axis).

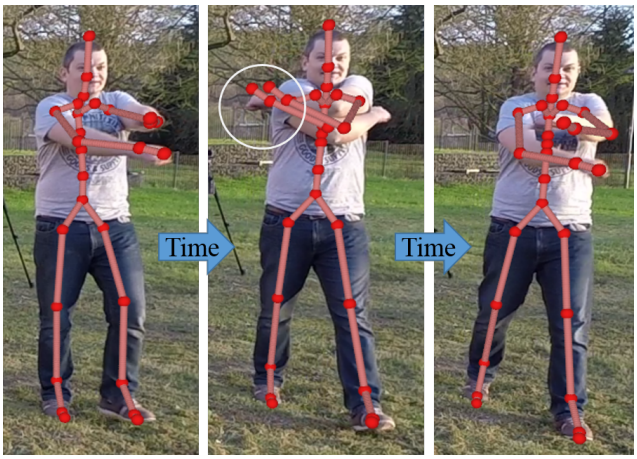

Figure 8. The image shows a tracking failure during a challenging motion in the *boy_outdoor* sequence. Our method recovers in the next frames. The white circle points to tracking failures.

scheme our method is designed such that it can recover from segmentation errors, even in presence of strong initial skeleton pose misalignments, as shown in Fig. 9. Please refer to the supplementary video for animated results.

The initial model acquisition and material segmentation have a direct impact on the quality of the results, as poorly estimated texture information results in tracking failures. A way to further automatize this required pre-processing step is to employ existing methods (*e.g.* Rhodin et al. [32]) for the shape approximation. The corresponding texture can be back-projected from the multi-camera images accounting for self-occlusions and coherence. A simple color clustering might suffice to automatically identify the actor materials. Automatic identification of the materials could however produce poor quality segmentations, *e.g.* in presence of highly textured apparel. Our method cannot directly handle non-homogeneous as well as highly specular foreground materials. A multi-modal color term (*e.g.* Gaussian mixture models) could help in improving segmentation of such materials. Our method is in general robust in cases where the foreground and the background appearance coincide,

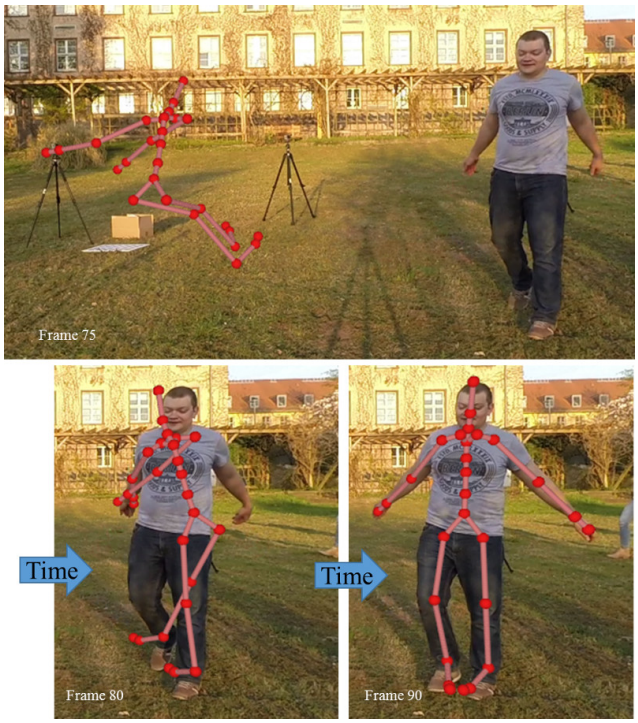

Figure 9. The image shows skeleton tracking resulting from bad pose initialization in the *boy_outdoor* sequence. Our method is able to successfully recover the correct pose after few frames.

thanks to the pose prediction term accounting for plausible motions. Typically, in a multiview setting the background varies a lot and the combination of the segmentations of all the views suffices to converge to the right pose.

6. Conclusion

In this work we have presented a novel approach for illumination-invariant human motion capture in a multiview setup. Our key idea is to employ an intermediate image representation that factors out variations in lightings across the sequence in time, or variations in appearance across the views. In order to obtain this invariant representation, for each frame and each view we solve a segmentation problem that uses the previous tracking result in order to infer cues about the individual materials' appearances in the current frame. By fusing this approach with CNN-based joint detectors as well as with a model-based tracker we are able to demonstrate superior performance compared to other methods, even under difficult conditions. Using a dynamical weighting strategy for determining the relative importance between the image segmentation and the joint detections, the method is less prone to errors due to bad segmentations and is able to recover from tracking failures (cf. Fig. 8).

Acknowledgements: We thank Helge Rhodin for his comparative results, Alkhazur Manakov and Hyeongwoo Kim for helping with the recordings. This work was funded by the ERC Starting Grant CapReal (335545).

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *TPAMI*, 2006.

[2] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view Pictorial Structures for 3D Human Pose Estimation. In *BMVC*, 2013.

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

[4] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *IJCV*, 2010.

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016.

[6] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *CVPR*, 1998.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *TPAMI*, 2001.

[8] M. Bray, P. Kohli, and P. Torr. PoseCut: simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In *ECCV*, 2006.

[9] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers. Combined Region and Motion-Based 3D Tracking of Rigid and Articulated Objects. *TPAMI*, 2010.

[10] M. Burenius, J. Sullivan, and S. Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013.

[11] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. Marker-less 3D human motion capture with monocular image sequence and height-maps. In *ECCV*, 2016.

[12] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015.

[13] A. Elhayek, C. Stoll, K. I. Kim, and C. Theobalt. Outdoor human motion capture by simultaneous optimization of pose and camera parameters. In *CGF*, 2015.

[14] A. Elhayek, C. Stoll, K. I. Kim, and C. Theobalt. Outdoor Human Motion Capture by Simultaneous Optimization of Pose and Camera Parameters. In *CGF*, 2015.

[15] J. Gall, B. Rosenhahn, T. Brox, and H. P. Seidel. Optimization and filtering for human motion capture : A multi-layer framework. *IJCV*, 2010.

[16] J. Gall, B. Rosenhahn, and H. P. Seidel. Drift-free tracking of rigid and articulated objects. In *CVPR*, 2008.

[17] N. Hasler, B. Rosenhahn, T. Thormählen, M. Wand, J. Gall, and H. P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR Workshops*, 2009.

[18] O. Hössjer and C. Croux. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *Journal of Nonparametric Statistics*, 1995.

[19] C. Ionescu, L. Bo, and C. Sminchisescu. Structural SVM for visual localization and continuous state estimation. In *ICCV*, 2009.

[20] L. Kavan, S. Collins, J. Žára, and C. O'Sullivan. Skinning with dual quaternions. In *I3D*, 2007.

[21] C. S. Lee and A. Elgammal. Coupled visual and kinematic manifold models for tracking. *IJCV*, 2010.

[22] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH*, 2000.

[23] S. Li, A. B. Chan, and A. B. C. Sijin Li. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In *ACCV*, 2014.

[24] Y. Liu, J. Gall, C. Stoll, Q. Dai, H. P. Seidel, and C. Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *TPAMI*, 2013.

[25] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017.

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

[27] G. Mori and J. Malik. Recovering 3D human body configurations using shape contexts. *TPAMI*, 2006.

[28] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017.

[29] S. Prince. Computer vision: models, learning, and inference, 2012.

[30] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. EgoCap: Egocentric Marker-less Motion Capture with Two Fisheye Cameras. In *SIGGRAPH Asia*, 2016.

[31] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*, 2016.

[32] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *ICCV*, 2015.

[33] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt. Model-based outdoor performance capture. In *3DV*, 2016.

[34] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H.S. Torr. Randomized trees for human pose detection. In *CVPR*, 2008.

[35] C. Schmaltz, B. Rosenhahn, T. Brox, and J. Weickert. Region-based pose tracking with occlusions using 3D models. *Mach. Vision Appl.*, 2012.

[36] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, 2003.

[37] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *IJCV*, 2012.

[38] C. Stoll, N. Hasler, J. Gall, H. P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *ICCV*, 2011.

[39] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured Prediction of 3D Human Pose with Deep Neural Networks. In *BMVC*, 2016.

[40] M. Trumble, A. Gilbert, A. Hilton, and J. Collomosse. Deep convolutional networks for marker-less human pose estimation from multiple views. In *CVMP*, 2016.

[41] O. Veksler. *Efficient Graph-based Energy Minimization Methods in Computer Vision*. PhD Thesis, 1999.

[42] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[43] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donns est minimum. *Tohoku Mathematical Journal, First Series*, 1937.

[44] C. Wu, K. Varanasi, and C. Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *ECCV*, 2012.

[45] L. Xu, C. Lu, Y. Xu, and J. Jia. Image smoothing via L0 gradient minimization. In *SIGGRAPH*, 2011.

[46] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. MonoPerfCap: Human Performance Capture from Monocular Video. *arXiv:1708.02136*, 2017.

[47] F. Zhou and F. De La Torre. Spatio-temporal matching for human detection in video. In *ECCV*, 2014.

[48] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3D shape estimation from 2D landmarks: A convex relaxation approach. In *CVPR*, 2015.

[49] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCV*, 2016.

[50] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *CVPR*, 2016.