On-set Performance Capture of Multiple Actors With A Stereo Camera

Chenglei Wu[1,2] * Carsten Stoll[1] † Levi Valgaerts[1] ‡ Christian Theobalt[1] §
[1]Max Planck Institute for Informatics [2]Intel Visual Computing Institute

Figure 1: *One of our performance capture results. Left to right: input video, reconstructed geometry, edited video with virtual logo added.*

Abstract

State-of-the-art marker-less performance capture algorithms reconstruct detailed human skeletal motion and space-time coherent surface geometry. Despite being a big improvement over marker-based motion capture methods, they are still rarely applied in practical VFX productions as they require ten or more cameras and a studio with controlled lighting or a green screen background. If one was able to capture performances directly on a general set using only the primary stereo camera used for principal photography, many possibilities would open up in virtual production and previsualization, the creation of virtual actors, and video editing during post-production. We describe a new algorithm which works towards this goal. It is able to track skeletal motion and detailed surface geometry of one or more actors from footage recorded with a stereo rig that is allowed to move. It succeeds in general sets with uncontrolled background and uncontrolled illumination, and scenes in which actors strike non-frontal poses. It is one of the first performance capture methods to exploit detailed BRDF information and scene illumination for accurate pose tracking and surface refinement in general scenes. It also relies on a new foreground segmentation approach that combines appearance, stereo, and pose tracking results to segment out actors from the background. Appearance, segmentation, and motion cues are combined in a new pose optimization framework that is robust under uncontrolled lighting, uncontrolled background and very sparse camera views.

CR Categories: I.3.7 [COMPUTER GRAPHICS]: Three-Dimensional Graphics and Realism; I.4.1 [IMAGE PROCESSING]: Digitization and Image Capture—Scanning; I.4.8 [IMAGE PROCESSING]: Scene Analysis;

Keywords: Performance Capture, Skeletal Motion Estimation, Shape Refinement, Bidirectional Reflectance Distribution Function

*e-mail: chenglei@mpi-inf.mpg.de
†e-mail: stoll@mpi-inf.mpg.de
‡e-mail: valgaerts@mpi-inf.mpg.de
§e-mail: theobalt@mpi-inf.mpg.de

Links: DL PDF WEB VIDEO

1 Introduction

Marker-less performance capture methods enable the reconstruction of detailed motion, dynamic geometry, and the appearance of real world scenes from multiple video recordings, for instance, reconstructing the full body or face of an actor, [Bradley et al. 2010; de Aguiar et al. 2008; Vlasic et al. 2008; Gall et al. 2009]. Despite the ability to capture richer and more expressive models than marker-based capture methods, marker-less methods are yet to be found in many practical feature film productions. One of the reasons for this is that most existing marker-less methods require studios with controlled lighting, controlled background, and a multitude of cameras. The benefit of being able to capture detailed models of actors in natural motion and natural apparel without markers is constrained in application by the remaining requirement to capture the actors in a separate green-screen controlled stage and not on set or on location. The ability to capture detailed moving 3D models of actors on the actual production set rather than a separate stage would broadly benefit movie and VFX production.

Currently, performances of real actors in a scene are frequently composited with virtual renditions of actors during post-processing. One example is the movie *Pirates of the Caribbean*, where real actors in a scene wear marker suits. The Imocap system is used to track their skeletal motion from the primary camera and a few satellite cameras and, in post-production, the actors in marker suits are replaced with virtual renditions. This common example shows the importance and tremendous difficulty of the task, since even the skeletal tracking alone required substantial manual marker labeling by an operator. On a real production set, it is difficult to effectively place additional satellite cameras for tracking as the environment

ACM Reference Format
Wu, C., Stoll, C., Valgaerts, L., Theobalt, C. 2013. On-set Performance Capture of Multiple Actors With A Stereo Camera. ACM Trans. Graph. 32, 6, Article 161 (November 2013), 11 pages.
DOI = 10.1145/2508363.2508418 http://doi.acm.org/10.1145/2508363.2508418.

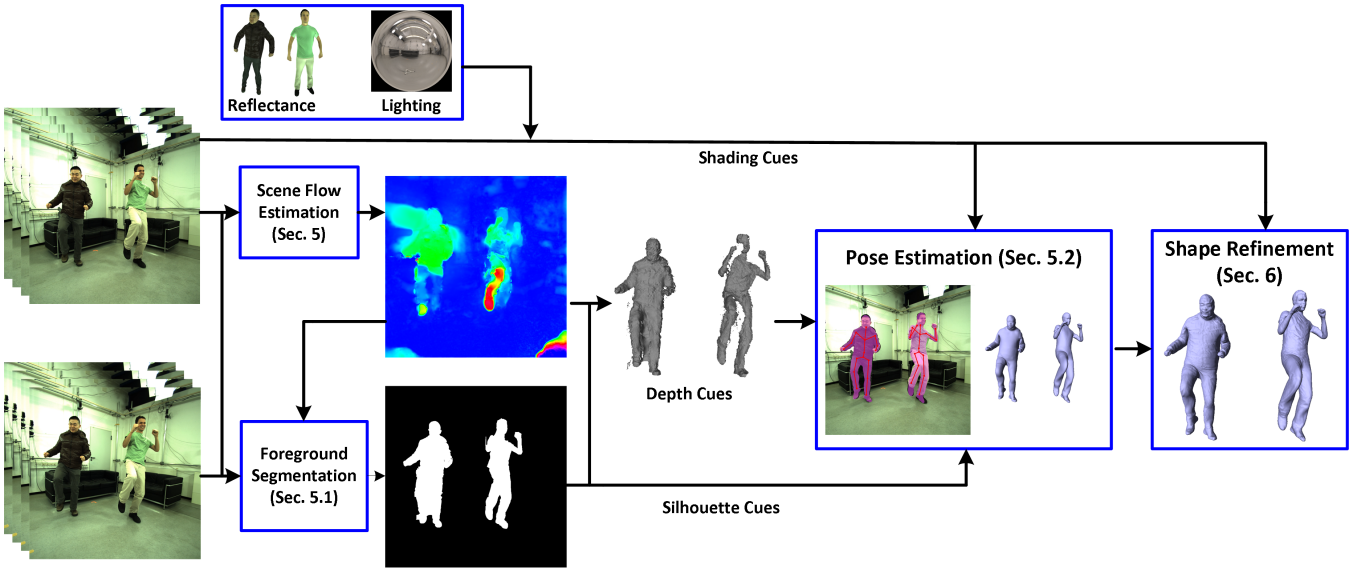

Figure 2: *Overview of our performance capture method.*

and scene conditions are very general and chosen with the visual quality of the shot in mind. This is usually orthogonal to the requirements that vision-based tracking algorithms have to operate robustly. If we could capture detailed motion *and* surface geometry automatically under the more general lighting conditions and backgrounds of a production set, while using only production cameras, then actors would benefit by being able to work on the real set while being captured, more realistic overlays of virtual actors could be created, more detailed pre-visualizations of CG augmented actors on set could be created, and the recovery of a 3D model underlying each actor in the scene would enable novel editing possibilities such as appearance modifications.

We describe a novel performance capture algorithm which works towards this goal. It enables us to capture the full body skeletal motion and detailed surface geometry of one or more actors using only a single stereo pair of video cameras. It is designed to work without additional depth sensors, such as Kinect, which only work indoors, have a limited range and accuracy, are not part of standard production cameras, and may interfere with other on-set equipment. The low-baseline stereo camera rig is permitted to move during recording. This setting is akin to modern movie production sets with a primary stereo camera. Our algorithm succeeds under uncontrolled lighting, non-frontal body poses of the actors, and scenes in which actors wear general apparel with non-Lambertian reflectance. It also succeeds in front of general scene backdrops where classical background subtraction would be infeasible.

1.1 Method Overview

Our algorithm tracks and deforms a template model for each actor in the scene such that it optimally aligns with stereo input images. In our setting, optimizing for pose and shape based on image features and silhouettes, as was applied in previous multi-view approaches with green screen backgrounds, would fail as now we have feature-full set backgrounds. Therefore, we rethink how illumination and reflectance are used for dynamic shape reconstruction. An overview is shown in Fig. 2.

Input to our algorithm is a stereo video sequence of a scene filmed with a camera rig that can freely move, as well as a light probe image of the set without actors. We also expect for each actor a static triangle mesh shape template with an embedded kinematic skeleton that can be obtained from a laser scan or from image-based reconstruction. Instead of relying on simple light transport assumptions, and instead of assuming Lambertian surface reflectance, our performance capture method is the first to apply knowledge about the incident illumination and a detailed spatially-varying BRDF of every actor in a scene for both skeletal pose estimation and for reconstruction of detailed surface geometry. Therefore, we expect as additional input a spatially-varying parametric BRDF surface model for every actor template, captured prior to stereo recording. In practical productions, reconstruction of such a reflectance model for each actor is becoming standard and can be performed with a light stage [Vlasic et al. 2009]. However, inspired by previous capture methods under general illumination [Li et al. 2013], we describe in this paper a lightweight method to estimate the BRDF based on multi-view video footage of a moving actor recorded under standard studio lighting (Sec. 4).

Our main contribution is a new skeletal pose estimation approach. It relies on a new stereo-based foreground segmentation algorithm that employs appearance cues, scene flow, pose reconstruction results from previous frames, and stereo coherence to reliably segment out actors in front of general backgrounds (Sec. 5.1). Pose estimation is based on minimizing a new energy that measures the model-to-image consistency based on the segmented silhouettes, the depth map given by scene flow, and the shading consistency based on a full diffuse and specular surface BRDF model (Sec. 5.2).

First, our algorithm captures the skeletal pose of each actor together with surface geometry which lacks high frequency shape detail such as cloth folds (Sec. 5). Second, this detail is reconstructed by a new inverse rendering approach that refines the coarse geometry using shading-based dynamic scene refinement based on the scene illumination and the full surface BRDF (Sec. 6).

We demonstrate the performance of our algorithm on a variety of scenes with general uncontrolled lighting, and scenes showing several actors performing motions with difficult occlusions and out-of-plane motions. We also show results on footage with apparel with challenging non-Lambertian appearance, and scenes filmed with a moving camera rig. We qualitatively and quantitatively demonstrate the accuracy of our method and the importance of each step,

and showcase that the quality of our reconstructions enables appearance editing of actors in video (Sec. 7).

2 Related Work

Marker-less motion capture approaches reconstruct human skeletal motion and have been developed in vision and graphics for many years [Moeslund et al. 2006; Poppe 2007; Sigal et al. 2010]. Most of them rely on a template skeleton with simple attached shape primitives, to then minimize some form of model-to-image consistency, e.g., edge or silhouette features using local or global optimization methods [Deutscher et al. 2000; Bregler et al. 2004; Gall et al. 2008; Sigal et al. 2010]. Even though recent techniques approach real-time performance and capture complex motion [Stoll et al. 2011], these methods expect more cameras than a single stereo pair, and usually require recording in a studio environment with controlled lighting. Further, motion capture algorithms do not reconstruct detailed surface models as we do in our approach. Monocular approaches for skeletal motion estimation have also been proposed, but usually require manual interaction and do not reconstruct detailed 3D surface geometry, e.g. [Wei and Chai 2010]. To move towards combined skeleton and surface capture, researchers experimented with coarse 3D shape models in multi-view motion capture [Balan et al. 2007], e.g., with parametric human templates. However, they deliver very coarse geometry and expect actors to wear skin-tight clothing.

Marker-less performance capture approaches go beyond motion capture and reconstruct dynamic geometry, possibly with skeletal motion, of people in more general clothing. Some techniques rely on shape-from-silhouette or active or passive stereo [Zitnick et al. 2004; Matusik et al. 2000; Starck and Hilton 2007; Waschbüsch et al. 2005]. Vlasic et al. [2009] record a person with multiple cameras in a dense controlled light stage and perform photometric stereo for capturing space-time-incoherent shapes. Model-based approaches deform a shape template such that it resembles a person [de Aguiar et al. 2008; Vlasic et al. 2008; Gall et al. 2009] or a person's apparel [Bradley et al. 2008] alone in multi-view video, which yields spatio-temporally coherent reconstructions. Mesh-based tracking approaches, as proposed by [de Aguiar et al. 2008], provide frame-to-frame correspondences with a consistent topology. The approach by Cagniart et al. [2010] makes a weaker *a priori* assumption by modeling the scene as a set of moving patches that are tracked over time. Another set of model-based approaches combine skeleton tracking with deformable surface tracking to capture people in more general apparel [Vlasic et al. 2008; Gall et al. 2009; Liu et al. 2011]. Some of these methods combine pose estimation with image segmentation and optical flow [Bray et al. 2006; Brox et al. 2006; Brox et al. 2010], and by this means also capture more than one person in a scene [Liu et al. 2011]. However, most methods are still restricted to controlled studios with green screen background, and usually expect ten or more cameras. Moreover, the amount of surface detail captured by these approaches is limited.

When using a complex controlled light stage indoors, more surface detail can be reconstructed by exploiting visible shading information [Vlasic et al. 2009]. Wu et al. [2011] extract more detail on performance captured models recorded in uncontrolled indoor lighting conditions. They solve a sequence of inverse rendering problems and estimate the incident illumination in the scene, which they subsequently employ for shading-based refinement of the dynamic scene geometry. Shading information was also used for detailed face capture, e.g., Beeler et al. [2012] use ambient occlusion to improve the dynamic reconstruction of the face from multi-view video. Wu et al. [2012] show that shading cues can be used for more reliable skeletal pose tracking from multi-view video in indoor scenes with more general backgrounds, where classical features such as edges and silhouettes do not provide sufficiently reliable evidence. Li et al. [2013] not only capture the dynamic geometry, but also make use of the geometry to estimate the surface BRDF to obtain a relightable performance. Hasler et al. [2009] jointly employ feature-based performance capture and structure-from-motion of the background for outdoor motion capture with multiple cameras. Their approach requires manual interaction and does not produce detailed dynamic surface geometry.

Unlike our approach, none of the methods mentioned so far succeeds with only a single mobile stereo pair of input cameras and in general scenes with uncontrolled background and lighting. Further, no existing techniques deliver reliable skeletal pose and detailed 3D surface geometry in such conditions.

Instead of using a priori templates, some approaches build up a spatio-temporally coherent shape model by space-time-analysis of partial scanner data [Liao et al. 2009; Popa et al. 2010; Tevs et al. 2012]. For single objects in a scene, these approaches also succeed with sparse depth camera or scanner systems but, due to strong regularization, they often capture geometry lacking high-frequency detail. Further, due to drift, often the result becomes increasingly smooth as time goes on. Unlike these methods, our approach uses an a priori template, and captures highly detailed surface geometry plus skeletal motion parameters, of more than one person, from a stereo pair of cameras alone. Thus, it addresses the full pipeline from the image data to geometry, succeeds on long sequences without loss of detail, and delivers results in a parameterization that directly feeds into the processing pipeline known to animation artists.

Some earlier vision methods attempted to capture human skeletal motion from stereo footage, e.g., [Plankers and Fua 2001], but did not achieve a pose and reconstructions as detailed and reliable in similarly general scenes as our method. Our work is also related to recent works on skeletal pose estimation from depth cameras, such as the Kinect, e.g., [Shotton et al. 2011; Ganapathi et al. 2010; Wei et al. 2012]. These approaches are designed for real-time use and reconstruct coarse skeletal motion and coarse surface geometry [Taylor et al. 2012]. High-quality pose and shape reconstruction is not their goal. In addition, most depth cameras only work indoors, and have a very limited range and accuracy. In contrast, our approach is designed to work directly from the primary stereo rig used for filming, yet delivers results of higher detail and does not require any specialized hardware that would limit the application range. Conceptually related to our approach is the binocular face capture method by Valgaerts et al. [2012]. Their approach also succeeds in more general scenes, but our setting is even more challenging, and specific new algorithmic segmentation and tracking solutions are needed to capture the full body motion, and in particular the motion of several people when they are not clearly the largest and most frontal object in a scene.

3 Preliminaries

In this paper we solve a variety of inverse rendering problems based on the following light transport model. Given a model of shape, illumination, and surface reflectance, the reflectance equation at a surface point is defined as [Kajiya 1986]:

$$B\left(x,\omega_o\right) = \int_{\Omega} L(\omega_i)\, V(x,\omega_i)\, \rho\left(\omega_i,\omega_o\right) \max(\omega_i \cdot \boldsymbol{n}(x), 0)\, d\omega_i \ , \quad (1)$$

where $B\left(x,\omega_o\right)$ is the reflected radiance, and the variables x, $\boldsymbol{n}$, ω_i, and ω_o are the surface location, the surface normal, and the incident and outgoing directions. The symbol Ω represents the domain of all possible directions, $L(\omega_i)$ represents the incident lighting, $V(x,\omega_i)$ is a binary visibility function, and $\rho\left(\omega_i,\omega_o\right)$ is the

bidirectional reflectance distribution function (BRDF). By defining $L_v(\omega_i) = L(\omega_i)\, V(x, \omega_i)$ as the visible lighting, $\hat{\rho}\,(\omega_i, \omega_o) = \rho\,(\omega_i, \omega_o)\max(\omega_i \cdot \boldsymbol{n}(x), 0)$, and parameterizing them using Spherical Harmonics (SH) [Ramamoorthi and Hanrahan 2001], the reflectance equation can be rephrased in the frequency domain:

$$B\,(\alpha, \beta, \theta_o, \phi_o) = \sum_{l=0}^{F_B} \sum_{m=-l}^{l} \sum_{p=0}^{P_B} \sum_{q=-p}^{p} L_{lm}\, \hat{\rho}_{lpq}\, D^l_{mq}(\alpha)\, e^{Im\beta}\, Y_{pq}(\theta_o, \phi_o) \ , \quad (2)$$

where (α, β) and (θ_o, ϕ_o) are the spherical angular parameters of $\boldsymbol{n}$ and ω_o, F_B and P_B are the SH orders, and L_{lm} and $\hat{\rho}_{lpq}$ are the SH coefficients of $L_v(\omega_i)$ and $\hat{\rho}\,(\omega_i, \omega_o)$. $D^l_{mq}(\alpha)$ is a matrix modeling how a spherical harmonic transforms under rotation into direction α, and $Y_{pq}(\theta_o, \phi_o)$ is the SH function. While (α, β) are defined in global coordinates, (θ_o, ϕ_o) are defined in local surface coordinates with the normal direction as north pole. We assume the BRDF to be isotropic. In the case of Lambertian reflectance, the image irradiance equation further simplifies to [Ramamoorthi and Hanrahan 2001]:

$$B_d(\alpha, \beta) = \sum_{l=0}^{F_D} \sum_{m=-l}^{l} \Lambda_l\, L_{lm}\, \hat{\rho}_{dl}\, Y_{pq}(\alpha, \beta) \ , \quad (3)$$

where $\hat{\rho}_{dl}$ are the SH coefficients for the clamped cosine function, Λ_l is a constant scalar for normalization, and F_D is the SH order, which is taken to be $F_D = 4$ in our experiment.

As can be seen in Eq. (2), parameterizing the reflectance equation in SHs for general BRDFs is much more complex than the Lambertian case. However, Ramamoorthi and Hanrahan [2001] showed that when the central direction of the BRDF is available, like in the Phong or Torrance-Sparrow model, a simple reparameterization in the SH domain similar to the Lambertian case can be obtained:

$$B_s(\alpha, \beta) = \sum_{l=0}^{F_S} \sum_{m=-l}^{l} \Lambda_l\, L_{lm}\, \hat{\rho}_{sl}\, Y_{pq}(\alpha, \beta) \ , \quad (4)$$

where $\hat{\rho}_{sl}$ are the SH coefficients of the properly reparameterized BRDF, and F_S is the order of SH, which is generally higher than for the Lambertian case. In our paper, we take $F_S = 10$ and will reduce it accordingly when BRDF parameters are obtained.

Similar to [Li et al. 2013], we estimate the BRDF as consisting of a diffuse part with Lambertian reflectance and a specular part under general illumination. However, instead of using a Phong reflectance model as in their approach, we use a simplified Torrance-Sparrow model [Torrance and Sparrow 1967] for the specular component. Thus, for our setting, the BRDF can be defined as:

$$\rho\,(\omega_i, \omega_o) = k_d + \frac{k_s}{4\pi\sigma_b^2 \cos\theta_i \cos\theta_o} \exp\left(-(\theta_h/\sigma_b)^2\right) \ , \quad (5)$$

where k_d and k_s are the diffuse and specular albedos, θ_i, θ_o, and θ_h are the incoming light direction, the viewing direction, and the half angle, all defined with respect to the surface normal, and σ_b is the surface roughness. The Torrance-Sparrow can be parameterized using a model based on the viewing vector mirrored at the surface normal in the SH domain by using coefficients of the form $\Lambda_l\, \hat{\rho}_{sl} \approx \exp\left(-(\sigma_b\, l)^2\right)$. Therefore, our final SH-parameterized reflectance equation takes the form:

$$B(\alpha, \beta) = k_d B_d(\alpha, \beta) + k_s B_s(\alpha, \beta) \ , \quad (6)$$

Based on the image irradiance equation described above, we first introduce how to inversely estimate the BRDF function from a multi-view and multi-lighting image sequence in Sec. 4. Afterwards, an analysis-through-synthesis pose estimation method is explained in Sec. 5 and a new shape refinement method based on Eq. (6) is introduced in Sec. 6.

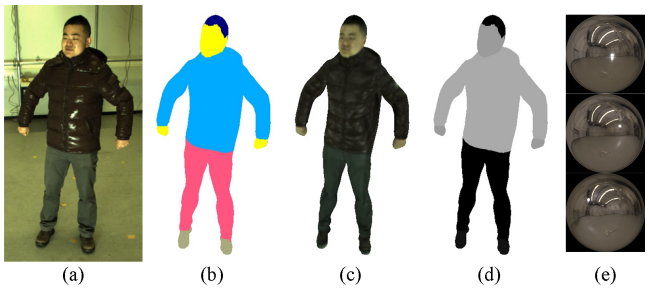

Figure 3: *Reflectance estimation: (a) input image, (b) material segmentation, (c) estimated spatially varying diffuse albedo, (d) estimated per-segment specular albedo, (e) the light probe images.*

4 Template and Reflectance Reconstruction

A first input to our algorithm is a static triangle mesh shape template M for each tracked actor. We use a laser scanner, but M could also be obtained via image-based reconstruction. The template is purposefully smoothed to remove static high-frequency shape detail. A bone skeleton with 20 joints and 37 degrees of freedom controls the motion of the shape template via skinning.

A model of surface reflectance for every actor is a second important prerequisite enabling stable binocular performance capture in general scenes (see Sec. 5). Such a model can be captured using a light-stage [Vlasic et al. 2009]. However, with wider applicability in mind, we employ an alternative solution that is based on a simpler studio setup and is inspired by methods that are able to capture the BRDF under general illumination [Li et al. 2013]. Our approach consists of three light sources placed vertically at different heights and a calibrated multi-view camera system (see Fig. 3), a set-up that is close in spirit to [Theobalt et al. 2007]. The actor is recorded performing a simple rotational motion with all three light sources turned on sequentially. Prior to recording, a ground truth environment map is captured for each such lighting condition and projected into spherical harmonics space. Then, we use a state-of-the-art performance capture algorithm [Wu et al. 2012] to track this simple sequence with our shape template.

The rotating motion of the performer allows us to collect reflectance samples of visible surface locations in the camera views. These samples are captured under different illumination and viewing conditions covering a range of azimuthal angles, while the vertical displacement of the light sources gives us measurements on different elevation angles. From these samples, we estimate the BRDF parameters k_d, k_s and σ_b (see Sec. 3). Although it is desirable to estimate these parameters for each point on the template mesh, the high-frequency reflectance component is particularly hard to estimate from a relative sparse set of samples. To make the calculation tractable, we assume the surface to comprise a discrete set of B materials K_b (such as skin), each with a constant specular reflectance. This discrete set of materials is manually segmented in the first frame (see Fig. 3), but could also be found via color clustering [Wu et al. 2011]. Another simplifying observation is that many common materials are dielectric, i.e., the generated highlights are of the same color as the light source. Following this assumption, we can represent the specular albedo k_s as a scalar value. Thus, we solve for a per-vertex k_d, a per-patch specular albedo k_s, and a per-patch surface roughness parameter σ_b.

BRDF estimation is performed in an iterative coarse-to-fine way. In a first iteration, we assume that all BRDF parameters are constant for all vertices of a material K_b. Then, we minimize the error between the rendered model under the calibrated lighting and the input frames:

$$E_K^B = \sum_f \sum_{v \in K, c \in N_c} w_{x,c} \, \| k_d B_d + k_s B_s(\sigma_b) - I_c(x,f) \|, \quad (7)$$

where f is the frame index, v is the vertex index, c is the camera index, and $w_{x,c}$ is a weighting factor. The surface normals of the coarse model reported by performance capture are too coarse to estimate the reflectance reliably. Therefore, we interleave a refinement of the surface normal orientations using a shape-from-shading approach similar to [Zhang et al. 1999] with the estimation of the reflectance. We perform normal refinement for each camera view. We iterate normal refinement and reflectance estimation, typically twice. After the first iteration of BRDF and normal estimation, we allow the diffuse albedo k_d to vary for every vertex, while keeping k_s and σ_b fixed per material. To prevent k_d and k_s from being negative in the optimization, we reparameterize them as $k_d = r_d^2$ and $k_s = r_s^2$, and optimize r_d and r_s instead. All optimization steps are performed with a conjugate gradient solver. We start by setting F_S=10, and when the BRDF parameters are obtained, we adaptively reduce F_S for each material segment using a strategy similar to [Ramamoorthi and Hanrahan 2002] to reduce processing time.

5 Skeletal Motion Estimation

It is our goal to estimate detailed surface and skeleton motion of actors in general clothing, who perform general motion in sets with no controlled background, merely from the video footage of a possibly moving stereo camera rig. Compared to previous multi-view performance capture algorithms that operate with tens of cameras and in front of a green screen for easier background subtraction, the drastically reduced set of views and the uncontrolled environment represent a previously unseen challenge. Thus, we need to fundamentally rethink which data cues to use for tracking, how to measure the model-to-image data consistency, and how to optimize the pose and shape parameters of the template model.

To meet this challenge, our method is the first to jointly employ shading cues from a full BRDF model, from depth information, and from motion information extracted from binocular views, and to robustly extracted foreground regions representing actors, all from binocular footage in general scenes. First, a light probe image of the empty set is captured, assuming that the lighting is constant for the duration of the recording. Then, we employ the variational approach of Valgaerts et al. [2010] to compute the 3D scene flow between each consecutive pair of frames. This approach computes optical flows in each camera view and 3D stereo geometry for each time step, both of which are used by our algorithm.

Performance capture now subsequently processes pairs of stereo video frames, by alternating the following two steps:

1. A new segmentation method is applied to robustly segment out the regions in the depth maps corresponding to persons in the foreground, even if the stereo rig is moving and the background has a general appearance and shape (Sec. 5.1). To succeed in this challenging setting, the segmentation method jointly relies on color information, a scene flow-induced body shape prior derived from previous body poses, and stereo constraints between input image pairs. Segmentation produces a depth region of the person to be tracked whose outlines provide additional silhouette cues for performance capture.
2. The current pose and shape of the actor are found by optimizing a pose error (Sec. 5.2). To this end, we employ a tracking algorithm that, for the first time, jointly relies on appearance cues from a full BRDF with diffuse and specular component, silhouette cues, and scene flow information.

5.1 Foreground Segmentation

Automatically obtaining clean segmented regions of depth belonging to persons in the foreground is a prerequisite for reliable binocular full body performance capture. Many previous segmentation approaches used color alone for segmenting foreground objects in video. Unfortunately, colors of foreground objects in general scenes can be very similar, leading to segmentation errors. Often only manual intervention can resolve these problems [Rother et al. 2004]. However, even for multi-view performance capture of interacting persons in front of a green screen, color information alone was found to be insufficient for labeling persons in video [Liu et al. 2011]. Depth thresholding alone is also not a reliable cue to segment out the person in a scene since, depending on the surrounding geometry, the person may not be the closest object to the camera. Finally, depth or image differencing alone is not suitable, since with a moving camera rig the background model would need to be permanently updated and possibly tracked with a structure-from-motion approach, which is error-prone with a dynamic foreground.

To succeed with a sparse set of binocular views, a general background, and a possibly moving rig, we employ a Markov-Random-Field (MRF)-based segmentation approach that combines evidence from a variety of scene cues to obtain a reliable segmentation of the persons in the foreground in both input views, and thus in the stereo depth. Foreground segmentation was also employed for motion tracking by Brox et al. [2006; 2010] in a multi-view setting by combining appearance cues, modeled by a Gaussian distribution, with a shape prior, provided by the object contour at the current pose. They evolve the object contour by minimizing a non-linear energy, which is sensitive to local minima. Here, we formulate the segmentation as a labeling problem which can be solved efficiently by a graph cut algorithm, and we model the appearance by a Gaussian mixture model (GMM) which enables the segmentation to work for textured objects. Further, we include a shape prediction by the estimated scene flow to obtain a more accurate shape prior and add a new stereo constraint as a consistency check between both cameras.

For every time step, segmentation is performed in two stages: 1) In a first stage, pixels in the left and right images are labeled separately as *person in the foreground* or as *background*. In case of multiple persons in the scene, a separate two-label segmentation is solved for each person. 2) In a second stage, the segmentations of each person from both views are fused.

Stage I: In the first stage, the segmentation finds the least energy (maximum likelihood) configuration $L = \{l_p \mid p \in 1, \ldots, N_p; l_p \in \{0,1\}\}$ of the MRF, assigning binary labels l_p to each of the N_p pixels. The energy is defined as follows:

$$D_1^S(L) = \sum_{p \in P} \lambda_A D_p^A(l_p) + \lambda_G D_p^G(l_p) + \lambda_S D_{pq}(l_p, l_q) \ . \quad (8)$$

$D_p^A(l_p)$ is a likelihood term penalizing the assignment of label l_p to pixel p based on its color, $D_p^G(l_p)$ is a shape prior exploiting that the body model pose in the previous frame is known and the scene flow between the previous and the current frame is available, and $D_{pq}(l_p, l_q)$ is a regularizing contrast term which favors that pixels have the same label when their color is similar. The weighting factors are experimentally set to $\lambda_A = 2$, $\lambda_G = 10$ and $\lambda_S = 50$.

Figure 4: *Foreground segmentation: (a) input image, (b)+(c) color likelihood for background and foreground (white: high, black: low), (d) segmentation result using only color likelihood, (e) color term and shape prior likelihood, (f) final segmentation result using all components.*

Color Likelihood A separate color model is used for foreground ($l = 0$) and background ($l = 1$). For both, it is implemented as a GMM of RGB colors with $K = 6$ mixture components. The k-th Gaussian $N\big(I(p) \,|\, \mu^l, \Sigma^l\big)$ in the GMM corresponding to label l is parameterized by a mean μ_k^l, a covariance Σ_k^l, and a weight ω_k^l. Given the pixel color $I(p)$, the appearance cost $D_p^A(l)$ for assigning to it a label l is defined as the negative log-likelihood of:

$$p\big(I(p) \,|\, \mu_1^l, \Sigma_1^l, \omega_1^l, \ldots, \mu_K^l, \Sigma_K^l, \omega_K^l\big) = \sum_{k=1}^{K} \omega_k N\left(I(p) \,|\, \mu_k^l, \Sigma_k^l\right) \quad . \quad (9)$$

The GMMs for foreground and background are continuously retrained over time to increase robustness under lighting and appearance changes. To train the GMMs for the current frame, we take the foreground and background regions of the previous frame and warp them to the current frame by means of the optical flow computed as part of the scene flow estimation. The colors of the warped regions are used for training the GMM models of the current frame. Fig. 4 (b) and (c) show the results of assigning pixels to foreground or background using the color term for the input image of Fig. 4 (a). As shown, the color term is able to distinguish most of the foreground. However, it may not be sufficient when the foreground color is similar to the background, e.g., the lower foot in Fig. 4 (a), which leads to an incorrect segmentation result as seen in Fig. 4 (d).

Shape Prior The shape prior measures the label assignment cost based on a prediction of the pose of the model in the current time step given that its pose in the previous time step is known and motion is smooth. We model this term by warping the previous pose of the shape model onto the current frame via scene flow. The warped model is projected onto the image and we build a heat map H^G based on the pixel's distance to the outer contour of the projected model. The shape prior cost is defined as the negative logarithm of:

$$H_p^G(l_p) = \begin{cases} \frac{1}{1+\exp\left(-d_p^2/(2\sigma_p^2)\right)} & l_p = \hat{l_p} \\ \frac{1}{1+1/\exp\left(-d_p^2/(2\sigma_p^2)\right)} & l_p \neq \hat{l_p} \end{cases} , \quad (10)$$

where d_p is the distance from pixel p to the nearest contour point, $\hat{l_p}$ is the pixel label given by the warped projected model, and σ_p is experimentally set to 5 for all experiments. Fig. 4 (e) shows the cost function of assigning pixels to the background, which helps to correctly segment the foot part to the foreground in Fig. 4 (f).

Smoothness Term The contrast term D_{pq} takes the same form as described in [Liu et al. 2011] and is defined as:

$$D_{pq}(l_p, l_q) = \begin{cases} \frac{\gamma}{s(p,q)} \exp\left(\frac{-\|I_p - I_q\|^2}{2\sigma_c^2}\right) & l_q \neq l_q \\ 0 & l_p = l_q \end{cases} , \quad (11)$$

where $s(p, q)$ is the spatial distance between the pixels.

The minimum energy (8) is found via graph cuts [Boykov and Funka-Lea 2006]. For efficiency, segmentation is performed for a conservative extended bounding box around the foreground actor, centered at the location from the previous frame warped by the scene flow. Pixels outside the box are labeled as background.

Stage II: In the second stage, we perform another segmentation of each image by taking into account information from the other camera. Specifically, we augment the MRF energy such that for each pixel in the current view, we check the consistency with the segmentation in the other view. We derive a stereo-based confidence measure by warping the segmentation of the other view into the current view using scene flow. If the warped segmentation assigns the same label to a pixel in the current view, the pixel is marked as trusted. Then, we retrain the color GMMs for the foreground and background using only trusted pixels in both views. Finally, another graph cut segmentation is performed by minimizing:

$$D_2^S(L) = \sum_{p \in P} \lambda_A D_p^A(l_p) + \lambda_G D_p^G(l_p) + \lambda_S D_{pq}(l_p, l_q) + \lambda_O D_p^O(l_p). \quad (12)$$

The main extension is the added stereo constraint D_p^O. It assumes the value 1 if trusted pixels are assigned a different label than in Stage I, 0 otherwise. For untrusted pixels, D_p^O is set to 0.5 for both labels. The weighting factor λ_O is experimentally set to 100.

5.2 Pose Estimation

Given a template model of the actor, including a rigged and skinned 3D mesh with reflectance information for each vertex, we track the

motion of actors in a binocular input video recorded in an arbitrary uncontrolled environment. As common in related work, we formulate this as a sequential problem. Given the pose at time $t-1$, the geometry M_{t-1} at time $t-1$, and two pairs of images at times $t-1$ and t respectively, we want to estimate the skeletal pose at time t. We formulate this as an energy minimization based on the constraints coming from the cues obtained in the previous steps. Li et al. [2013] employ the silhouette and feature constraints for pose estimation in a multi-view setup, which is not enough for our setup (see the comparison in Sec. 7). Our energy for pose estimation takes three terms. The first term E^S encodes information from shading cues and measures the difference between the captured images and a rendered version of the character based on the reflectance and the captured environment map. The second term E^G comes from the depth cues, which measures the difference between our current pose and a depth map of the current image pair calculated as a by-product of the scene flow method. The third term E^H contains the silhouette cues and measures the difference between the projected contour of the mesh at the current pose and the segmented silhouette. The three terms are combined into a single total energy term:

$$E^T = \beta_S\, E^S + \beta_G\, E^G + \beta_L\, E^H \quad , \tag{13}$$

where β_S, β_G, and β_L are weighting factors. We optimize this energy in as a function of the skeletal joint angle parameters using a simple conditioned gradient descent method similar to [Stoll et al. 2011]. The weighting factors are experimentally set to $\beta_S = 1$ and $\beta_L = 10$ for all sequences, while $\beta_G = 20$ for sequences with moving cameras and $\beta_G = 10$ for all other sequences.

Shading Term Similar to [Wu et al. 2012], the shading energy E^S measures the similarity between a rendered image of the current pose of the actor under the known lighting and reflectance and the captured images. In contrast to previous work, we do not assume Lambertian reflectance, but propose one of the first methods to employ a full BRDF model with diffuse and specular reflectance as cues in a 3D pose tracking framework. We demonstrate in the experimental section Sec. 7 that by relying on this more advanced light transport model, we can obtain more accurate and more robust tracking results even with sparse input data captured in general environments. For a single camera c, we write:

$$E_c^S = \frac{1}{N_c^s} \sum_i \left(B\left(c, v_i^t, \boldsymbol{n}_i^t\right) - I_c^t(x_i^t, y_i^t) \right)^2 \quad , \tag{14}$$

where N_c^s is the number of visible vertices in camera c, (x_i^t, y_i^t) is the projection of the surface vertex v_i^t, $\boldsymbol{n}_i^t$ is the corresponding surface normal, and B is the radiance calculated according to Eq. (3) and Eq. (4). While lighting and reflectance functions are constant, the vertex positions v_i^t, the projections (x_i^t, y_i^t), and the normals $\boldsymbol{n}_i^t$ depend on the pose parameters of the model. If we ignore potential visibility changes in the vertices, we can calculate analytical derivatives of this function using a Taylor expansion.

Depth Term We estimate per-camera depth maps as part of the scene flow computation. Using the segmentation obtained in Sec. 5.1, we remove the background from the depth map. The segmented foreground depth is then refined by removing interpolated depth values at occlusion boundaries via triangle normal orientation thresholding relative to the viewing direction. Based on the filtered foreground depth map, the second component of the pose energy encodes iterative-closest-point-like constraints:

$$E_c^G = \frac{1}{N_c^g} \sum_i \left(v_i^t - c\left(v_i^t\right) \right)^2 \quad , \tag{15}$$

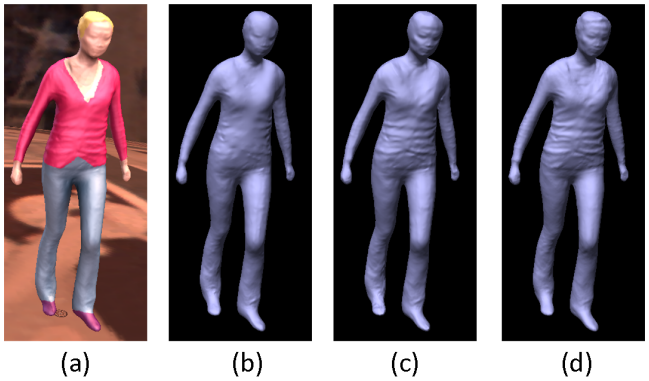

Figure 5: *Surface refinement on a synthetic test sequence: (a) one of two input images, (b) refined shape using [Wu et al. 2011], (c) refined shape using our method, (d) ground truth shape.*

where $c\left(v_i^t\right)$ is the corresponding 3D point for vertex v_i^t in the re-projected depth map of camera c based on an approximate nearest neighbor search.

Silhouette Term Following our segmentation, the contour pixels of an actor in the foreground can be conveniently detected in each camera view, enabling us to define a silhouette consistency term. For each of the N_c^h contour pixels, we define a projection ray that can be parameterized as a Plücker line $H_i = (n_i, m_i)$ [Gall et al. 2009]. The silhouette consistency term sums the distance between each line H_i and its closest vertex $v(i)^t$ on the body model:

$$E_c^H = \frac{1}{N_c^h} \sum_i \left(v(i)^t \times n_i - m_i \right)^2 \quad , \tag{16}$$

If more than one person is present in the scene, the steps in this section are run for each person separately.

6 Shape Refinement

Skeletal tracking yields the coarse shape of each actor in the scene at every time step. However, fine scale surface detail visible in the images is missing. We recover this with an extended version of the photometric refinement process described in [Wu et al. 2011]. We formulate this problem as a spatio-temporal MAP inference problem, where the cost function takes the form:

$$\psi\left(\boldsymbol{g}^{\boldsymbol{t}}\right) = \phi\left(I^t \,|\, \boldsymbol{g}^{\boldsymbol{t}}\right) + \phi\left(\boldsymbol{g}^{\boldsymbol{t}} \,|\, \boldsymbol{g}^{\boldsymbol{t-1}}\right) \quad , \tag{17}$$

where $\phi\left(I^t \,|\, \boldsymbol{g}^{\boldsymbol{t}}\right)$ is the shading error that measures the similarity of the image gradients in the input image I^t to the predicted rendered shading gradients according to the image reflectance equation described in Sec. 3. The unknown $\boldsymbol{g}^{\boldsymbol{t}}$ represents the refined surface geometry for every vertex as a displacement from M_c^t in the local normal direction. The term $\phi\left(\boldsymbol{g}^{\boldsymbol{t}} \,|\, \boldsymbol{g}^{\boldsymbol{t-1}}\right)$ is a prior that requires the current refined surface geometry to be similar to the refined surface geometry of the previous time-step, transformed to the current time-step via skeleton-based deformation and surface skinning using the pose parameters obtained in Sec. 5.

Unlike [Wu et al. 2011], we adapt the geometry refinement approach to explicitly consider a full diffuse and specular BRDF, rather than just diffuse reflectance. Our method is related to previous stereo methods that phrase multi-view consistency under general surface BRDFs, e.g., [Davis et al. 2005], but unlike these we do not require images under multiple and often calibrated lighting

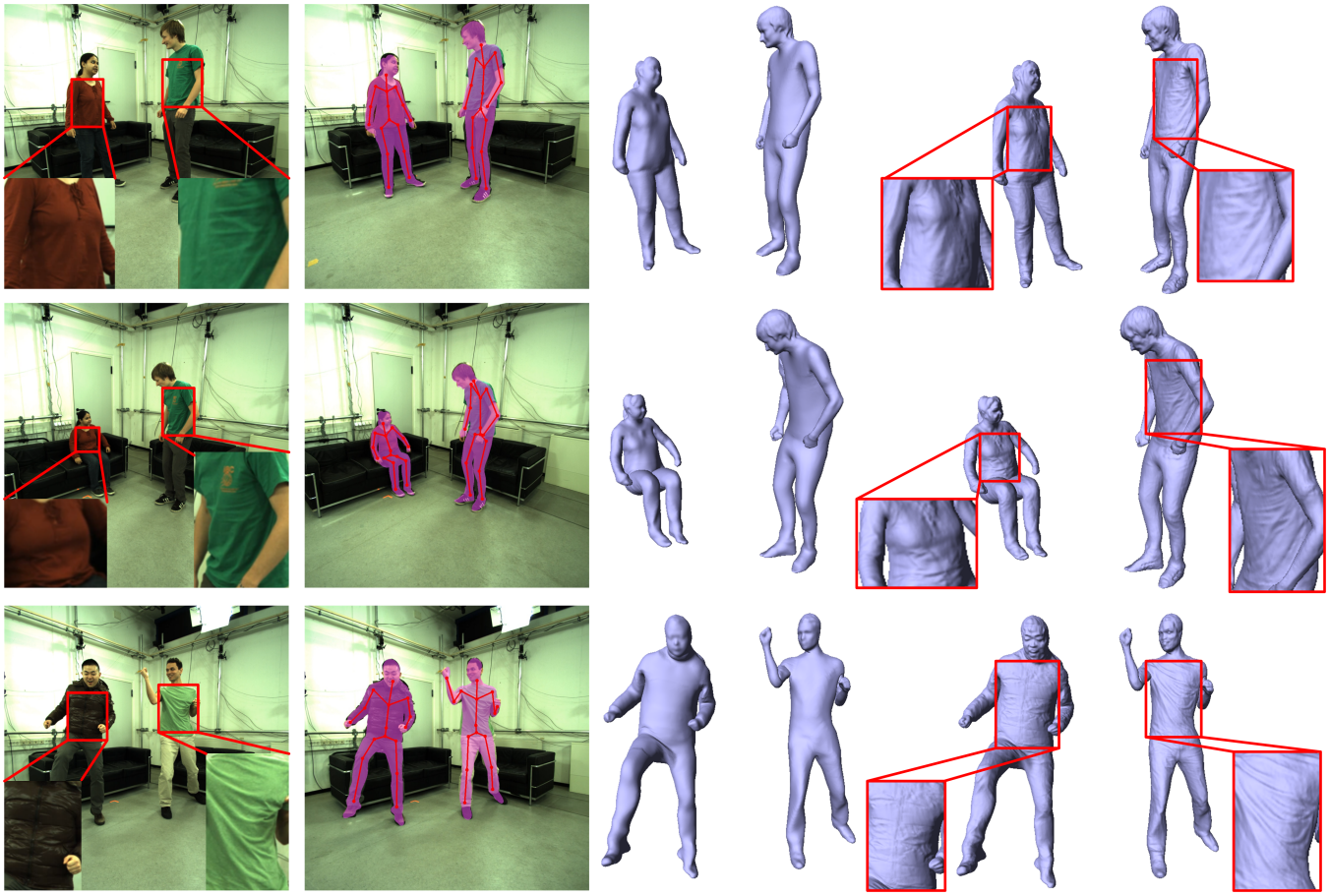

Figure 6: *Performance capture results of our algorithm on real world sequences. Left to right: one of the two input images, segmentation and tracked skeleton as an overlay, 3D geometry after skeletal pose estimation, 3D geometry after surface refinement.*

conditions. Since we are able to exploit the information in the full BRDF, our method not only works as well as [Wu et al. 2011] on diffuse surfaces with sparse binocular input data, it also successfully recovers surface detail on very specular surfaces where the previous method would fail, e.g., the specular jacket in the bottom row in Fig. 6. As a final result, high frequency shape detail on the surface, such as fine folds and creases, are recovered in a spatio-temporally coherent way. To optimize this energy function, we employ the Levenberg-Marquardt algorithm, which is similar to [Wu et al. 2011]. Fig. 5 shows a comparison of our refinement method with [Wu et al. 2011] on a specular surface.

7 Results

We recorded 3 test sequences consisting of over 1300 frames. The data was recorded with a stereo rig with a baseline of ≈ 22 cm at a resolution of 1024×1024 pixels and at a frame rate of 45 fps. Each sequence shows two people wearing casual clothing performing a variety of different motions in front of a general background. The scenes provide various challenges, such as moving cameras, specular apparel, close contact with background objects, and partial occlusions (see Fig. 6 and supplementary video), which would make tracking these sequences with previous approaches challenging. We also evaluate our method on a synthetic data set. The pose for the first frame is initialized manually, followed by the local optimization described in Sec. 5.2. The mask image for the first frame is generated using a segmentation tool [Rother et al. 2004].

The first sequence (Fig. 6, top) contains two people who initially are standing and talking, and then start to dance. Our algorithm successfully evaluates the pose and reconstructs small details such as the folds in the shirts accurately from the stereo images. The second sequence (Fig. 6, middle) shows two actors in the process of sitting down on a couch, and is recorded with a moving camera. Even though the actors are in contact with the couch in the background and some partial occlusions take place, the motion and surface detail is reconstructed accurately. As the camera is moving, we only reconstruct the relative pose of the actors with respect to the camera (i.e., we do not distinguish between camera motion and actor motion). The third sequence (Fig. 6, bottom) shows two actors jumping and kicking. Even though the motions are very fast, the pose estimation is successful. Further, even though the left actor wears a highly specular jacket, the surface detail is reconstructed accurately by our method. This highlights again the importance of using a non-Lambertian BRDF, since a method based on Lambertian shading would fail in estimating surface detail accurately.

Scene Enhancement We use the tracked motion and refined surface of the scenes to modify the original footage from the stereo camera. As our geometry is spatio-temporally coherent, it is easy to add new textures on top of the original footage or peform other modifications (see Fig. 1 and supplementary material).

Quantitative Evaluation To evaluate our method quantitatively, we generated a synthetic data set consisting of 100 frames by ren-

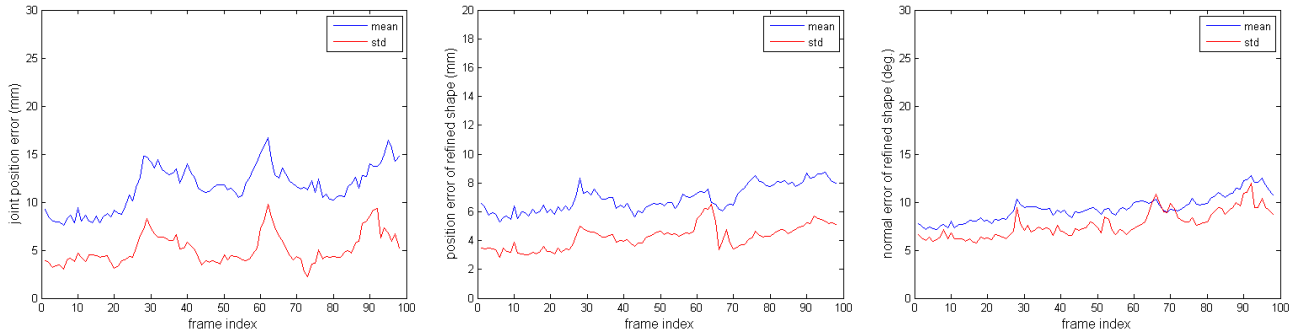

Figure 7: *Quantitative evaluation on a synthetic sequence, showing mean and standard deviation for each frame: (a) joint position error, (b) vertex position error for refined shape, (c) normal direction error for refined shape.*

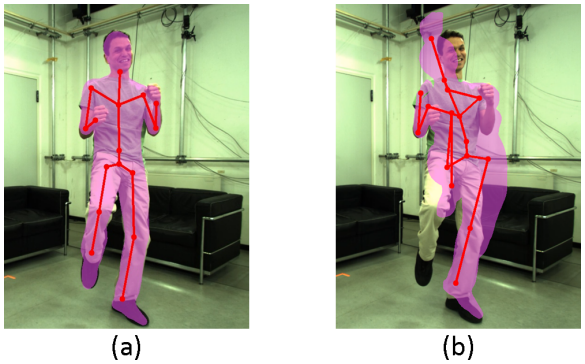

Figure 8: *Comparison of our method with [Li et al. 2013]: (a) our tracked skeleton, (b) tracked skeleton of [Li et al. 2013].*

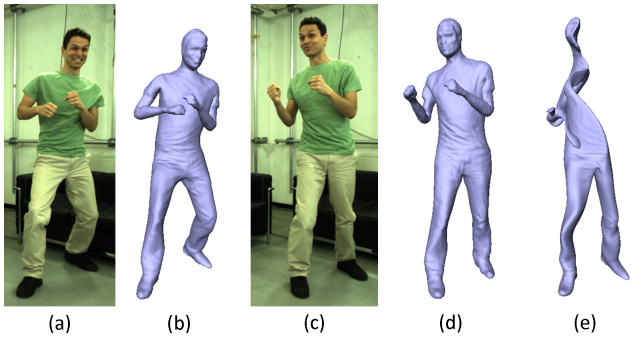

Figure 9: *Comparison of our method with [Valgaerts et al. 2012]: (a) first frame, (b) corresponding geometry, (c) 100th frame, (d) our reconstructed geometry, (e) reconstructed geometry of [Valgaerts et al. 2012].*

dering a captured sequences with a manually painted Phong-based material and texture onto a virtual stereo rig with a baseline of ≈ 4 cm under the environment lighting of St. Peter's Basilica [Debevec 1998]. Given the images, the initial model, and its BRDF, as well as the incident lighting, we ran our complete pipeline, including the scene flow estimation, foreground segmentation, motion tracking, and surface refinement. We then compared the results against the ground truth to quantify the accuracy of the skeletal motion and surface reconstruction (see Fig. 7). The evaluation shows that our algorithm is able to create a very accurate reconstruction of the synthetic scene, with an average joint position error of only 11.6 ± 5.09 mm, and average surface position and normal error of 6.92 ± 4.23 mm and 9.34 ± 7.7 degrees respectively.

To make sure that all parts of our pipeline are actually important, we also evaluated the approach on a real sequence of 500 frames by leaving out one or several stages of our pose estimation pipeline. Possible algorithmic components for the pose estimation pipeline are: (a) image segmentation, (b) scene flow constraints, (c) depth map constraints, (d) shading constraints, and (e) silhouette constraints. Using only (c), (c+d), or (a+b), the pose estimation fails to track the sequence completely. Using (a+c), or (a+c+d), the pose estimation is able to track the whole sequence, however some body parts get lost during tracking. Our pipeline, consisting of (a+c+d+e) is able to track the whole sequence correctly and performs best of all the combinations (see also the supplementary video).

Comparison with State-of-the-art We compared our tracking approach with the method described in [Li et al. 2013] for the real-world sequence shown in the bottom row of Fig. 6. As can be clearly seen in Fig. 8, the tracking method of Li et al. [2013], which employs the silhouette and feature constraints, fails on this binocular data, while our method successfully estimates the correct pose.

We also compared our method with a purely surface-based tracking method recently proposed for binocular facial performance capture [Valgaerts et al. 2012]. Fig. 9 shows the results of tracking the template mesh over ≈ 200 frames for the real-world sequence in the bottom row of Fig. 6. The method of [Valgaerts et al. 2012], which only propagates mesh vertices by means of scene flow, clearly suffers from self-occlusions, motion estimation errors near boundaries, and the inability of the applied Laplacian regularization to deal with rotating motion. Our method, on the other hand, builds on a model-based skeleton tracking that is much more robust to the articulated motion that is typical for full body tracking.

Run Time We ran our algorithm on a commodity PC with a dual-core 3GHz processor and 8GB RAM with a single threaded, non-optimized implementation. Scene flow calculation takes ≈ 3 min per frame. Motion tracking including foreground segmentation takes ≈ 2 min per frame. The final shape refinement step takes ≈ 1 min for a template mesh resolution of ≈ 80000 vertices. As these three steps are independent of each other, they can be pipelined into multiple threads or machines.

Discussion Our method succeeds to handle many challenging cases, including moving cameras, specular apparel, and partial occlusions. However, there are limitations to its use. As we use only

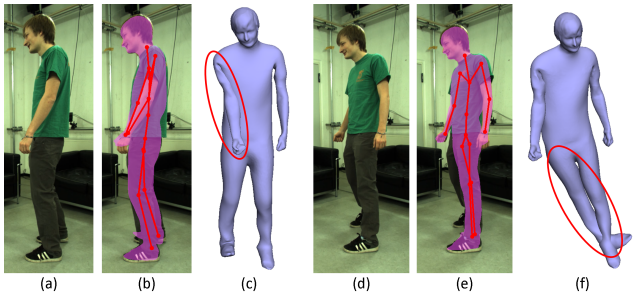

Figure 10: *A failure case of our tracking method: (a) input frame in which an actor is turning away from the camera, (b) corresponding tracked skeleton, (c) corresponding tracked mesh, (d) input frame in which the same actor is turning back towards the camera, (e) corresponding tracked skeleton, (f) corresponding tracked mesh.*

a small-baseline stereo rig, some body parts may be completely occluded in some frames. Our current local optimization scheme may fail to recover from these occlusions when the body parts appear again. Fig. 10 shows such a failure case for our method on a sequence with self-occlusions. Fig. 10 (a), (b) and (c) show one of the input images, the corresponding tracked skeleton from the camera view point, and the tracked mesh geometry for a frame in which an actor is turning away from the camera, thus occluding his right side (see also the supplementary video). The tracked mesh makes it clear that the occluded arm is not tracked correctly and intersects the torso. Fig. 10 (d), (e) and (f) show results from the same camera view point at a later point in time where the right leg starts to reappear again. Both the tracked skeleton and mesh show an incorrect pose for the leg that was occluded in the previous frames. For the same reason, multiple interacting actors currently cannot be handled by our method. Occlusions could be handled by first detecting them and then using a global optimization for the occluded parts to make sure they are recovered correctly. The fact that occluded body parts do not have a correct pose during occlusion is not a major concern since our primary interest lies in the geometry visible from the perspective of the stereo camera. Nevertheless, recovering a reliable pose for occluded parts is an important open problem and may be relevant for some applications.

Extending the current method to outdoor performance capture is another interesting direction for future work. While our shape refinement algorithm is able to generate detailed geometry for most surfaces, it may fail for saturated and over-exposed highlights where no information can be extracted. Topological changes can not be handled either as we assume a constant connectivity and topology. Even though the output of our algorithm is spatio-temporally coherent (i.e., it has a constant connectivity and mesh topology), the shape refinement currently does not account for minor motion of garments such as a shifting shirt, which may lead to slight swimming artifacts in the range of 1-2 cm when rendering virtual textures in the original video. This could be improved by performing an additional scene flow-based alignment between the virtual actor and the current input images and performing an additional adaptation of the actor to the foreground segmentation to capture cloth motion.

8 Conclusion

We have presented a novel performance capture algorithm that reconstructs detailed human skeletal motion and space-time coherent surface geometry from a potentially moving, low-baseline stereo camera rig. It is able to track skeletal motion and detailed surface geometry of one or more actors in uncontrolled environments by exploiting BRDF information, scene illumination, and background segmentation. With our approach we are able to produce high quality results from a simple stereo camera setup that approach the quality of results previously only achievable with complex setups containing 10 or more cameras. We believe that our method steps towards enabling the use of full-body performance capture for wider use, such as on-set performance capture without additional hardware, video editing, and the creation of virtual actors.

Acknowledgements

We gratefully acknowledge all our actors for their participation in the recordings and thank the reviewers for their helpful comments. We thank James Tompkin for his suggestions and corrections.

References

BALAN, A., SIGAL, L., BLACK, M., DAVIS, J., AND HAUSSECKER, H. 2007. Detailed human shape and pose from images. In *Proc. CVPR*.

BEELER, T., BRADLEY, D., ZIMMER, H., AND GROSS, M. 2012. Improved reconstruction of deforming surfaces by cancelling ambient occlusion. In *Proc. ECCV*, 30–43.

BOYKOV, Y., AND FUNKA-LEA, G. 2006. Graph cuts and efficient N-D image segmentation. *IJCV 70*, 2, 109–131.

BRADLEY, D., POPA, T., SHEFFER, A., HEIDRICH, W., AND BOUBEKEUR, T. 2008. Markerless garment capture. *ACM TOG (Proc. SIGGRAPH) 27*, 3, 99:1–99:9.

BRADLEY, D., HEIDRICH, W., POPA, T., AND SHEFFER, A. 2010. High resolution passive facial performance capture. *ACM TOG (Proc. SIGGRAPH) 29*, 3, 41:1–41:10.

BRAY, M., KOHLI, P., AND TORR, P. H. S. 2006. POSECUT: simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In *Proc. ECCV*, 642–655.

BREGLER, C., MALIK, J., AND PULLEN, K. 2004. Twist based acquisition and tracking of animal and human kinematics. *IJCV 56*, 3, 179–194.

BROX, T., ROSENHAHN, B., CREMERS, D., AND SEIDEL, H.-P. 2006. High accuracy optical flow serves 3-D pose tracking: exploiting contour and flow based constraints. In *Proc. ECCV*, 98–111.

BROX, T., ROSENHAHN, B., GALL, J., AND CREMERS, D. 2010. Combined region and motion-based 3D tracking of rigid and articulated objects. *IEEE TPAMI 32*, 3, 402–415.

CAGNIART, C., BOYER, E., AND ILIC, S. 2010. Free-form mesh tracking: a patch-based approach. In *Proc. CVPR*, 1339–1346.

DAVIS, J. E., YANG, R., AND WANG, L. 2005. BRDF invariant stereo using light transport constancy. In *Proc. ICCV*, 436–443.

DE AGUIAR, E., STOLL, C., THEOBALT, C., AHMED, N., SEIDEL, H.-P., AND THRUN, S. 2008. Performance capture from sparse multi-view video. *ACM TOG (Proc. of SIGGRAPH) 27*, 98:1–98:10.

DEBEVEC, P. 1998. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proc. SIGGRAPH*, 189–198.

DEUTSCHER, J., BLAKE, A., AND REID, I. 2000. Articulated body motion capture by annealed particle filtering. In *Proc. CVPR*, 1144–1149.

GALL, J., ROSENHAHN, B., AND SEIDEL, H.-P. 2008. *Human Motion: Understanding, Modelling, Capture and Animation*. ch. An Introduction to Interacting Simulated Annealing, 319–343.

GALL, J., STOLL, C., AGUIAR, E., THEOBALT, C., ROSENHAHN, B., AND SEIDEL, H.-P. 2009. Motion capture using joint skeleton tracking and surface estimation. In *Proc. CVPR*, 1746–1753.

GANAPATHI, V., PLAGEMANN, C., KOLLER, D., AND THRUN, S. 2010. Real time motion capture using a single time-of-flight camera. In *Proc. CVPR*, 755–762.

HASLER, N., ROSENHAHN, B., THORMÄHLEN, T., WAND, M., GALL, J., AND SEIDEL, H.-P. 2009. Markerless motion capture with unsynchronized moving cameras. In *Proc. CVPR*, 224–231.

KAJIYA, J. T. 1986. The rendering equation. In *Proc. SIGGRAPH*, 143–150.

LI, G., WU, C., STOLL, C., LIU, Y., VARANASI, K., DAI, Q., AND THEOBALT, C. 2013. Capturing relightable human performances under general uncontrolled illumination. *CGF (Proc. EUROGRAPHICS) 32*, 275–284.

LIAO, M., ZHANG, Q., WANG, H., YANG, R., AND GONG, M. 2009. Modeling deformable objects from a single depth camera. In *Proc. ICCV*, 167–174.

LIU, Y., STOLL, C., GALL, J., SEIDEL, H.-P., AND THEOBALT, C. 2011. Markerless motion capture of interacting characters using multi-view image segmentation. In *Proc. CVPR*, 1249–1256.

MATUSIK, W., BUEHLER, C., RASKAR, R., GORTLER, S. J., AND MCMILLAN, L. 2000. Image-based visual hulls. In *Proc. SIGGRAPH*, 369–374.

MOESLUND, T., HILTON, A., AND KRÜGER, V. 2006. A survey of advances in vision-based human motion capture and analysis. *CVIU 104*, 2, 90–126.

PLANKERS, R., AND FUA, P. 2001. Tracking and modeling people in video sequences. *CVIU 81*, 3, 285–302.

POPA, T., SOUTH-DICKINSON, I., BRADLEY, D., SHEFFER, A., AND HEIDRICH, W. 2010. Globally consistent space-time reconstruction. *CGF (Proc. SGP) 29*, 5, 1633–1642.

POPPE, R. 2007. Vision-based human motion analysis: An overview. *CVIU 108*, 1-2, 4–18.

RAMAMOORTHI, R., AND HANRAHAN, P. 2001. A signal-processing framework for inverse rendering. In *Proc. SIGGRAPH*, 117–128.

RAMAMOORTHI, R., AND HANRAHAN, P. 2002. Frequency space environment map rendering. *Proc SIGGRAPH 21*, 3, 517–526.

ROTHER, C., KOLMOGOROV, V., AND BLAKE, A. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM TOG 23*, 3, 309–314.

SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP, T., FINOCCHIO, M., MOORE, R., KIPMAN, A., AND BLAKE, A. 2011. Real-time human pose recognition in parts from single depth images. In *Proc. CVPR*, 1297–1304.

SIGAL, L., BALAN, A., AND BLACK, M. 2010. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV 87*, 4–27.

STARCK, J., AND HILTON, A. 2007. Surface capture for performance based animation. *IEEE CGA 27*, 3, 21–31.

STOLL, C., HASLER, N., GALL, J., SEIDEL, H.-P., AND THEOBALT, C. 2011. Fast articulated motion tracking using a sums of gaussians body model. In *Proc. ICCV*, 951–958.

TAYLOR, J., SHOTTON, J., SHARP, T., AND FITZGIBBON, A. W. 2012. The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proc. CVPR*, 103–110.

TEVS, A., BERNER, A., WAND, M., IHRKE, I., BOKELOH, M., KERBER, J., AND SEIDEL, H.-P. 2012. Animation cartography: intrinsic reconstruction of shape and motion. *ACM TOG 31*, 2, 12:1–12:15.

THEOBALT, C., AHMED, N., LENSCH, H., MAGNOR, M., AND SEIDEL, H. P. 2007. Seeing people in different light – joint shape, motion, and reflectance capture. *IEEE TVCG 13*, 3, 663–674.

TORRANCE, K. E., AND SPARROW, E. M. 1967. Theory for off-specular reflection from roughened surfaces. *J. Opt. Soc. Am. 57*, 9, 1105–1112.

VALGAERTS, L., BRUHN, A., ZIMMER, H., WEICKERT, J., STOLL, C., AND THEOBALT, C. 2010. Joint estimation of motion, structure and geometry from stereo sequences. In *Proc. ECCV*, 568–581.

VALGAERTS, L., WU, C., BRUHN, A., SEIDEL, H.-P., AND THEOBALT, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. In *ACM TOG (Proc. SIGGRAPH Asia)*, vol. 31, 187:1–187:11.

VLASIC, D., BARAN, I., MATUSIK, W., AND POPOVIĆ, J. 2008. Articulated mesh animation from multi-view silhouettes. *ACM TOG (Proc. SIGGRAPH) 27*, 3, 97:1–97:9.

VLASIC, D., PEERS, P., BARAN, I., DEBEVEC, P., POPOVIC, J., RUSINKIEWICZ, S., AND MATUSIK, W. 2009. Dynamic shape capture using multi-view photometric stereo. *ACM TOG (Proc. SIGGRAPH Asia) 28*, 5, 174:1–174:11.

WASCHBÜSCH, M., WÜRMLIN, S., COTTING, D., SADLO, F., AND GROSS, M., 2005. Scalable 3D video of dynamic scenes.

WEI, X., AND CHAI, J. 2010. Videomocap: modeling physically realistic human motion from monocular video sequences. *ACM TOG (Proc. SIGGRAPH) 29*, 4, 42:1–42:10.

WEI, X., ZHANG, P., AND CHAI, J. 2012. Accurate realtime full-body motion capture using a single depth camera. *ACM Trans. Graph. 31*, 6, 188:1–188:12.

WU, C., VARANASI, K., LIU, Y., SEIDEL, H.-P., AND THEOBALT, C. 2011. Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proc. IEEE ICCV*, 1108–1115.

WU, C., VARANASI, K., AND THEOBALT, C. 2012. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *Proc. ECCV*, 757–770.

ZHANG, R., TSAI, P., CRYER, J., AND SHAH, M. 1999. Shape from shading: A survey. *IEEE TPAMI 21*, 8, 690–706.

ZITNICK, C. L., KANG, S. B., UYTTENDAELE, M., WINDER, S., AND SZELISKI, R. 2004. High-quality video view interpolation using a layered representation. *ACM TOG (Proc. SIGGRAPH) 23*, 3, 600–608.