Garment Replacement in Monocular Video Sequences

Article *in* ACM Transactions on Graphics · December 2014

DOI: 10.1145/2634212

CITATIONS
31

READS
606

5 authors, including:

Lorenz Branz
GOM GmbH
7 PUBLICATIONS **46** CITATIONS

SEE PROFILE

Garment Replacement in Monocular Video Sequences

Lorenz Rogge, Felix Klose, Michael Stengel, Martin Eisemann and Marcus Magnor
Institut für Computergraphik, Technische Universität Braunschweig

We present a semi-automatic approach to exchange the clothes of an actor for arbitrary virtual garments in conventional monocular video footage as a post-process. We reconstruct the actor's body shape and motion from the input video using a parameterized body model. The reconstructed dynamic 3D geometry of the actor serves as an animated mannequin for simulating the virtual garment. It also aids in scene illumination estimation, necessary to realistically light the virtual garment. An image-based warping technique ensures realistic compositing of the rendered virtual garment and the original video. We present results for eight real-world video sequences featuring complex test cases to evaluate performance for different types of motion, camera settings, and illumination conditions.

Categories and Subject Descriptors: I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Virtual Reality*; I.3.8 [**Computer Graphics**]: Applications; I.4.8 [**Computer Graphics**]: Scene Analyis; I.6.3 [**Computer Graphics**]: Simulation—*Applications*

General Terms: Animation

Additional Key Words and Phrases: image based techniques, garment simulation, lighting reconstruction, video augmentation

ACM Reference Format:

Rogge, L., Klose, F., Stengel, M., Eisemann, M. and Magnor, M. 2014. Garment Replacement in Monocular Video Sequences. ACM Trans. Graph. 32, 6, Article xx (August 2014), 10 pages.
DOI = 10.1145/2508363.2508414
http://doi.acm.org/10.1145/2508363.2508414

1. INTRODUCTION

Digital workflows have greatly advanced video post-processing capabilities. However, photo-realistically modifying recorded real-world scene content is still a time-consuming, highly labor-intensive, and monetary costly process. In this work we address the challenge of realistically exchanging the attire of human actors in already captured, conventional, uncalibrated monocular video footage. Previous approaches addressed this long standing problem by rerendering the entire human actor by a virtual dummy [Divivier et al. 2004] or by altering only the texture print on the garment leaving the garment itself unchanged [Scholz and Magnor 2006]. To go beyond these approaches, a precise dynamic 3D body model reconstruction is usually required. This is the approach most often pursued in contemporary movie productions [Landgrebe 2012] but comes at the expense of a costly hardware setup and extensive manual labor. Earlier approaches in research have achieved impressive results based on multi-view recordings to estimate the necessary 3D body information from either a laser scanned model [de Aguiar et al. 2008] or using a statistical body model [Hasler et al. 2009]. The need for multi-video recordings, however, complicates acquisition, and such methods are not applicable to already recorded, conventional video footage. The input video and reconstructed body model generally do not match identically because of the limited fitting precision of the statistical body model and can therefore be used only as approximate guidance for video editing [Jain et al. 2010]. For virtual garment replacements, unfortunately, such imprecisions are unacceptable as they will inevitably produce noticeable visual artifacts.

In the following, we present a semi-automatic approach for perceptually convincing garment replacement in real-world video footage. In contrast to previous work, we intentionally concentrate on the difficult case of conventional *monocular* video recordings of actors wearing normal clothes. We do not use any additional input information besides the plain RGB images. While the problem is ill-posed, we are able to achieve high-quality results excelling in quality and general applicability over previous approaches that rely on additional information from depth cameras or multiview settings. Inspired by state-of-the-art literature [Jain et al. 2012], we aimed at achieving best-possible realism with user interaction times of approximately one minute per edited frame. This paper makes the following contributions:

(1) a complete and flexible pipeline for garment replacement in mononcular video sequences,
(2) a novel set of error terms for human pose estimation tailored specifically to track human motion in monocular video sequences while offering the possibility to interactively add soft pose constraints to resolve ambiguities,
(3) an approach to reconstruct dynamic scene illumination from a person's video recording alone, and
(4) an image-based body and silhouette matching algorithm for compositing and alignment of the virtual garment with the recorded actor based on an imprecise body model.

As potential applications of our approach we envision movie post-production and youth protection by sanitizing nudity in movie scenes.

2. RELATED WORK

Virtual clothing. Typically, virtual clothing systems follow one of three principles: they render a virtual avatar that is dressed with the virtual garment [Divivier et al. 2004; Hauswiesner et al. 2011] avoiding the problem of augmenting the real person with the virtual garment; the system changes only the texture of the garment [Scholz and Magnor 2006; Scholz et al. 2005; Pritchard and Heidrich 2003; Guskov et al. 2003]; or additional sensor data from multi view recordings [Hauswiesner et al. 2013] or depth sensors [Fitnect 2012; Giovanni et al. 2012] is used to render the garment on top of an actor in a video stream. To simulate even small wrinkles or wide apparel image-based clothing simulations provide a powerful tool [Xu et al. 2011; Hilsmann and Eisert 2012; Hilsmann et al. 2013]. For garment replacement in still images Yoon *et al.* [2011] proposed a semi-automatic system using a priori skeletal and silhouette information to properly drape a character. In this paper we propose a novel approach that tackles the problem of *realistically* augmenting a standard, *monocular* video of a human actor with virtual clothing requiring only minimal user interaction.

Shape and Pose Reconstruction. The simulation of virtual garments requires a proxy for the inherent collision detection and

Fig. 1: Our approach enables digitally changing people's clothes in conventional monocular video sequences for virtual garments.

plausible reconstruction of garment motion. In recent years, statistical deformable body models have received a lot of attention due to their flexibility and robustness [Anguelov et al. 2005; Hasler et al. 2009]. With these models, human shape and motion can be reconstructed from multiview video data [Hasler et al. 2009; Balan et al. 2007; Guan et al. 2010; Hasler et al. 2009]. Recent work shows that it is possible to solve the underconstrained problem of reconstructing complex 3D human poses [Agarwal and Triggs 2004; Guan et al. 2009] or approximate 3D geometry [Töppe et al. 2011] even from single images. Overcoming the problem of self-occlusions and ambiguities in monocular videos, however, requires additional information in the form of body markers on the actor [Rogge et al. 2011], a proper annotation of body joints [Ramakrishna et al. 2012] or by obeying to the laws of physics [Wei and Chai 2010; Vondrak et al. 2012]. Similar to [Zhou et al. 2010] reshaping bodies in still images, *MovieReshape* [Jain et al. 2010] manipulates the body shape of a human actor by transforming the deformation of a body model to a smooth deformation of the input video to prevent complex matting, segmentation, and occlusion problems. In a semi-automatic approach the required body shape and motion is reconstructed from silhouette comparisons and tracked features on the body surface. However, solving the problem of robust shape and motion reconstruction with the precision required for realistic garment replacement in arbitrary monocular videos seems currently still infeasible. In this work, we therefore relax the requirement of precise pose reconstruction from monocular video and instead fine-tune the garment simulation result in image space during compositing.

Lighting Reconstruction. For convincing realism of the final video augmentation, it is necessary to reconstruct not only the body model, but also scene illumination. In computer graphics the conventional approach is to estimate a radiance map from a light probe placed in the scene [Debevec 1998] or a fisheye lens recording [Frahm et al. 2005], but both require special preparations during recording and hence are not applicable to already captured video footage. However, given a reference geometry and input video allows to draw inferences from shading differences within a surface about the reflectance properties, normals and approximate lighting conditions [Gibson et al. 2001; Chen et al. 2011]. We extend the approach of Gibson *et al.* [2001] to estimate the albedo distribution function for a person's surface in the input video. We make use of the 3D information from the estimated parameterized body model to formulate the lighting reconstruction problem as a linear system that can be solved in a physically plausible way using a non-negative least-squares solver.

3. OVERVIEW

We propose a semi-automatic approach to solve the highly underconstrained problem of augmenting *monocular* videos of a human actor with virtual garments by combining coarse pose estimation with an image-based refinement algorithm for more accurate garment alignment. We first fit a parameterized body model to the silhouettes of a recorded human actor (Sect. 4). This serves as a crude approximation to the actor's shape and motion. We propagate the model through the video by optimizing shape and pose parameters with a specifically designed error functional. Ambiguities are solved by allowing for soft user constraints. Based on the body model we reconstruct the BRDF of the actor and estimate the scene illumination (Sect. 5). The model serves as a proxy to simulate and render the virtual garment (Sect. 6). As the body model and the actor's shape only roughly match, we first apply a global image-based refinement by tracking the motion of both, model and actor, and using the discrepancy between these to fit the virtual garment to the human actor in image space (Sect. 7). This prevents a "floating" of the garment on the human actor. Small mismatches along the silhouettes are then corrected by a line matching algorithm. Both correction steps are combined in a single image-space warping procedure. The warped depth values of the body model handle occlusions during the final compositing step. Fig. 2 gives an overview of the proposed algorithm.

Input and Assumptions. The input to our algorithm is an arbitrary uncalibrated, monocular video sequence depicting a human actor. We assume that a matte $\mathbf{M}_t$ of the actor can be extracted from the video for each frame at time t. For our experiments we made use of the Grab-Cut algorithm by Rother *et al.* [2004] and Video SnapCut by Bai *et al.* [2009] available in Adobe AfterEffects™. Hair occluding any body parts must be removed from the matte.

Our work focuses on the garment simulation and compositing and we require the actor not to wear loosely fitting clothes. Estimating body poses from actors wearing loose clothes has been done in [Bălan and Black 2008]. We further assume that the actor's surface is largely diffuse.

4. SHAPE AND POSE RECONSTRUCTION

In the following we propose specifically tailored error functions to track an actor and estimate his/her shape in a monocular video sequence.

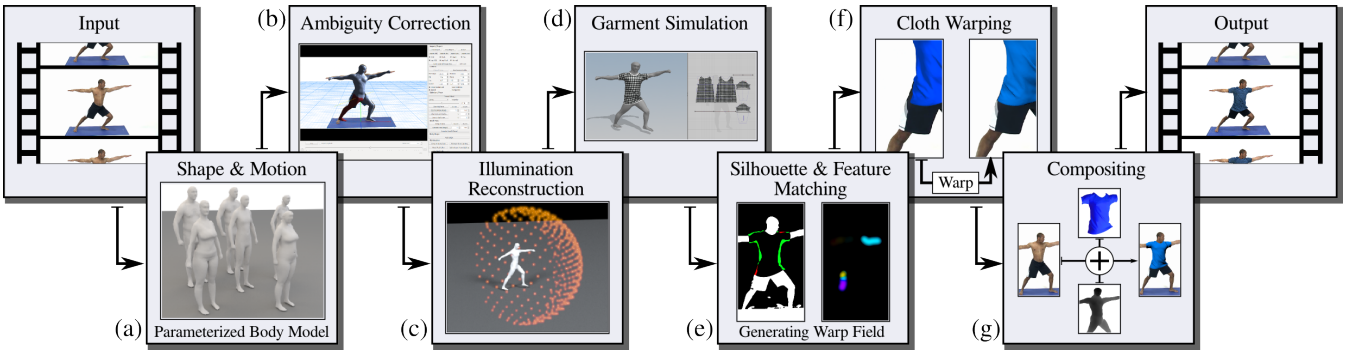

Fig. 2: **Overview of our approach.** Given an uncalibrated monocular input video (© Stock Footage), (a) we extract alpha mattes of the actor and make use of a parameterized body mesh model to estimate shape and motion. (b) If needed, ambiguous cases are solved by user interaction. (c) Reflectance parameters are computed for the body model, and scene illumination is reconstructed. (d) The body model is used to simulate a realistic motion of the virtual garment. (e) Interactive image-based matching and (f) image-warping are used to resolve remaining misalignments between the virtual garment and the actor. (g) Finally we create a composite of the virtual garment and the original video.

4.1 Camera Parameter Estimation

We estimate the extrinsic and intrinsic camera parameters using off-the-shelf bundle adjustment tools [Snavely et al. 2006] in case of a moving camera. In case of a static camera, we let the user fit a ground plane into the video by manually adjusting the distance, tilt and aperture angle of the camera from which the required projection matrix is then derived.

4.2 Morphable Body Model

For shape and pose estimation we require a parameterized body model. To this end, we tested the statistical body model from Jain *et al.* [2010] which is based on the work of Hasler *et al.* [2009], and the body model provided by the *MakeHuman*™ project [MakeHuman 2012]. The algorithm itself, however, is not model specific. Both tested models are fully skinned and are animated using shape $\Lambda = (\lambda_1, \ldots, \lambda_M)$ and pose parameters $\Phi = (\phi_1, \ldots, \phi_N)$. For the shape optimization it is sufficient to concentrate on optimizing the $M = 30$ most important shape parameters in the statistical model. For the MakeHuman model we can reduce the number of shape parameters from 50 to 35 by using the same parameter values for shaping symmetrical body parts, such as the left and right arm or the left and right leg. During pose estimation, we optimize all available parameters in both models. We proceed by first optimizing the shape for a single reference frame (Sect. 4.3) followed by optimizing the pose parameters $\Phi_t = (\phi_{1,t}, \ldots, \phi_{N,t})$ for each frame t in the video (Sect. 4.4).

4.3 Body Shape Estimation

We pose our shape estimation problem as an optimization of global shape parameters Λ based on an image-based energy function $E_{\text{Shape}}(\Lambda; \Phi_t, \mathbf{M}_t)$ for a specific frame time t:

$$\underset{\Lambda}{\operatorname{argmin}}\, E_{\text{Shape}}(\Lambda; \Phi_t, \mathbf{M}_t) = E_{\text{s}} + \alpha E_{\text{h}} \qquad (1)$$

We set $\alpha = 30$ in all our experiments.

The first term E_{s} measures the misalignment of the segmented actor's silhouette with the silhouette of the re-projected body model $\mathbf{B}$ using the camera projection P estimated in Sect. 4.1, similar to [Jain et al. 2010].

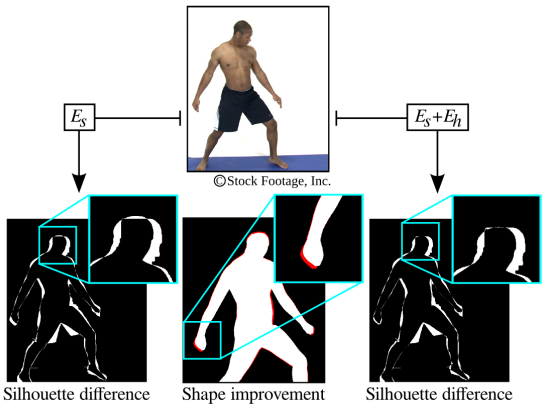

Fig. 3: **Left:** Shape fitting using only the silhouette term E_{s}. **Center:** Regions improved after including the term E_{h} (in red). **Right:** The error term E_{h} helps to determine the scale of the model, resulting in a better overall fit.

$$E_{\text{s}}(\Phi_t, \Lambda, \mathbf{M}_t) = \sum_{p=1}^{\#\text{pixels}} |\mathbf{M}_t(p) - P(\mathbf{B}(\Phi_t, \Lambda))(p)| \qquad (2)$$

The parameterized body model $\mathbf{B}(\Phi_t, \Lambda)$ or in short form $\mathbf{B}_t$ represents the template body model $\mathbf{B}$ deformed by the pose and shape parameters Φ_t, Λ.

The second component E_{h} penalizes differences in height between the actor silhouette and the projected body model in image space:

$$E_{\text{h}}(\Phi_t, \Lambda, \mathbf{M}_t) = |y_{\min,\mathbf{M}_t} - y_{\min,\mathbf{B}_t}| + |y_{\max,\mathbf{M}_t} - y_{\max,\mathbf{B}_t}| \qquad (3)$$

$y_{\min,\mathbf{M}_t}, y_{\min,\mathbf{B}_t}$ and $y_{\max,\mathbf{M}_t}, y_{\max,\mathbf{B}_t}$ are the minimum and maximum y-values of non-zero pixels in the respective mattes $\mathbf{M}_t$ of the actor and of the body model $\mathbf{B}_t$ at frame time t. Fig. 3 shows a comparison with and without the error term E_{h}.

The pose parameters Φ_t for the shape estimation are set by the user in a simple click-and-drag fashion for a single frame t, having

a matte $\mathbf{M}_t$ describing body shape and joint length properly. We then keep Φ_t fixed for the optimization of Λ in Eq. (1) for which we use a multidimensional linear optimizer [Nelder and Mead 1965]. Once Λ has been computed, it is kept fixed throughout the video sequence. For efficiency reasons when using the MakeHuman model, we first estimate the so-called macro parameters for height, weight, gender, and tone as these are linear projections of the other parameters and therefore influence them. We then use the result as an initialization to optimize all 35 parameters. We explicitly decouple shape and pose estimation, as the size and orientation of the actor are unknown a priori and have to be initialized by hand. Manually setting the pose parameters for a reference frame during this initialization step is more efficient and less time consuming than optimizing pose and shape in a coupled manner for every video frame. If required for more precise shape estimation one could easily define several keyframes for a joint shape optimization or optimize shape and pose together using keyframes as soft constraints, as described in Sect. 4.4.

4.4 Pose and Motion Estimation

Estimating the pose for each frame t is more intricate as the monocular projection is ambiguous with respect to the pose parameters Φ_t. We solve ambiguities and physical implausibilities by enforcing temporal coherence of the pose from one frame to the next, incorporating an interpenetration check as a separate error term and allowing for soft user constraints at keyframes which we describe in detail in the following.

Error Term. We formulate the error term for pose estimation as:

$$\underset{\Phi_t}{\operatorname{argmin}} E_{\text{Pose}}(\Phi_t; \Lambda, \mathbf{M}_t) = E_{\text{s}} + \beta E_{\text{t}} + \gamma E_{\text{i}} \,, \tag{4}$$

We set $\beta = 2.5$ and $\gamma = 500$. Λ is known from the previous shape optimization, Sect. 4.3. The first term E_{s} is equivalent to the silhouette error term in Eq. (1). The second term E_{t} penalizes strong temporal deformations by comparing the joint angles between frames:

$$E_{\text{t}}(\Phi_{t-1}, \Phi_t) = \sum_{i=1}^{N} e^{\delta|\phi_{i,t} - \phi_{i,t-1}|} - 1 \tag{5}$$

$\phi_{i,t}$ is the i^{th} pose parameter in frame t. The third term E_{i} penalizes self-interpenetrations of the body mesh. For every vertex $\mathbf{v}_j \in \mathbf{B}_t$ the penetration depth $d_{\text{p}}(\mathbf{v}_j, \Phi_t, \Lambda)$ is accumulated:

$$E_{\text{i}}(\Phi_t, \Lambda) = \sum_{j=1}^{|\mathbf{V}|} d_{\text{p}}(\mathbf{v}_j, \Phi_t, \Lambda) \,, \tag{6}$$

where $d_{\text{p}}(\mathbf{v}_j, \Phi_t, \Lambda) = 0$ for all vertices $\mathbf{v}_j \in \mathbf{B}_t$ that are not penetrating any body surface. Otherwise, d_{p} is the Euclidean distance to the closest surface point of the penetrated body part.

The algorithm is initialized with the shape and pose $\mathbf{B}_t$ estimated during shape optimization, Sect. 4.3. The optimizer then proceeds to estimate each individual succeeding frame in sequential order. To optimize preceeding frames, Φ_{t-1} is exchanged with Φ_{t+1} in E_{t}, Eq. 5.

Soft user constraints. To solve ambiguities resulting from the monocular input, we allow for additional soft constraints at each joint. In any frame where errors occur, the user may mark interactively invalid pose parameters, either for single joints or whole kinematic chains and specify a range of frames R that are to be re-optimized, Fig. 4.

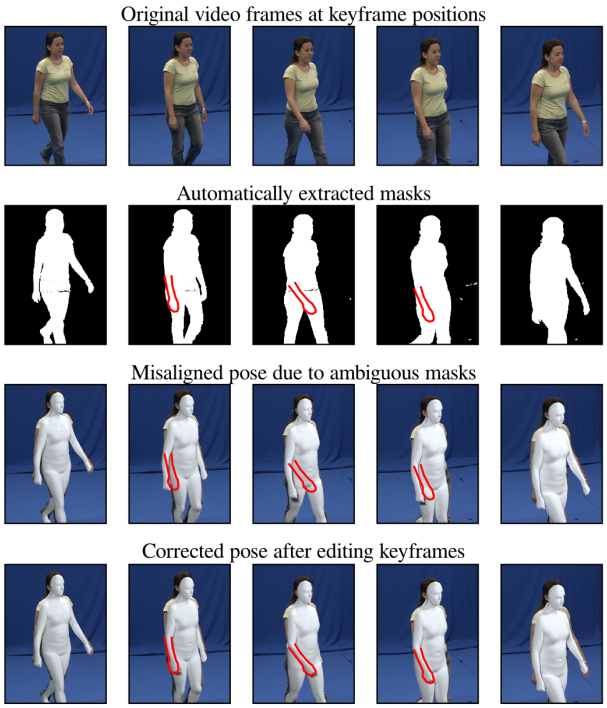

Fig. 4: **Pose estimation correction.** In this example, ambiguous silhouette information (2^{nd} row) caused the arm pose to be reconstructed incorrectly (3^{rd} row). The correct pose is marked in red. The user selects a keyframe (3^{rd} column) and manually adjusts the joints of the lower right arm of the body model. The pose parameters in a user-specified range are interpolated from the corrected keyframe and the previous as well as later correct poses outside the range. These are then used as initialization parameters during a second run of the automatic pose estimator.

This is done using an interactive 3D editor allowing the user to select the invalid body parts at a specific frame and to move them towards the correct position, overwriting the falsely reconstructed pose parameters for this frame. For re-initialization of the optimizer, we linearly interpolate the pose parameters of the body model in R between the user-specified keyframe and the already well-optimized frames at the boundaries of the range R. We additionally remove small jittering artifacts in the joints by temporally smoothing the pose parameters. Using a simple boxfilter with a one-frame radius in temporal direction proved to be sufficient in our cases.

5. SCENE RECONSTRUCTION

Once the actor's shape and pose are reconstructed from the monocular input video, we can use the 3D information to reconstruct approximate scene lighting. Our approach consists of two steps. In the first step, we determine diffuse BRDF parameter values for all pixels within the actor's silhouette $\mathbf{M}$ from the monocular input video (Sect. 5.1). In the second step, we make use of the reconstructed BRDFs to estimate scene illumination on a per frame basis, allowing for changing lighting conditions (Sect. 5.2).

5.1 Albedo Estimation

Inspired by [Ziegler et al. 2004], we derive a time-dependent representation of the actor's surface shading by reprojecting each visible vertex of the statistical body model $\mathbf{B}$ into each frame of the input video. We then estimate the diffuse component of a parametric BRDF model by averaging each vertex color over time. Our basic assumption is that for a fixed surface point only its intensity values should change over time due to changes in surface orientation but not its chromaticity or saturation. Thus, averaging colors should properly describe the saturation and chromaticity of the diffuse surface and reduce the influence of shadows affecting color brightness. We are aware that more sophisticated approaches for BRDF reconstruction exist [Theobalt et al. 2007], but these generally require more than one viewpoint or known scene lighting.

5.2 Scene Illumination Estimation

We estimate scene illumination on a per-frame basis. The idea is to distribute virtual point lights in the scene and adjust their intensities so that the rendered image of the body model using the reconstructed diffuse BRDF closely matches the appearance in the original video image $\mathbf{I}_t$, Fig. 5. The problem can be formulated as a linear system

$$\underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \tag{7}$$

describing the shading equations for a set of L light sources which can be solved in a least-squares sense. Here, $\mathbf{x}$ is a column vector of size $3L \times 1$ that describes the RGB color components for all L virtual light sources and $\mathbf{b}$ is the vectorized version of the reference image $\mathbf{I}_t$. Each row of the matrix $\mathbf{A}$ represents the influence of each single light source on a single sample pixel according to the utilized shading model. Although this approach is capable of reconstructing the illumination using more complex BRDF models, we found it sufficient to use the albedo estimate of Sect. 5.1 as this reduces the complexity of the linear system and yields plausible results in all of our test scenes.

In contrast to the work in [Gibson et al. 2001], we initially distribute the light sources uniformly over a hemisphere using a deterministic golden spiral point distribution [Swinbank and Purser 2006], Fig. 5.

The hemisphere is aligned with the camera orientation surrounding the reconstructed body model at an infinitely far distance. The positions are kept fixed and we apply a non-negative least squares solver [Lawson and Hanson 1995] to solve the linear system in Eq. (7). We use $L = 400$ initial light source positions in all our test scenes. While the solver avoids negative light emission it also minimizes the amount of non-zero lights during optimization. Therefore, only a minimum amount of light sources remains active to illuminate the scene properly. For our test sequences, this varied between one to six light sources, depending on scene and frame.

The visual appearance of the actor model compared to the input image is mainly influenced by the light sources on the frontal hemisphere. We found it sufficient in terms of quality of the results to limit the reconstruction to the frontal light source positions, as this reduces the complexity of the linear system.

The resulting light source positions can be interpreted as an environment map or used directly as point light sources when rendering the virtual garment. Examples can be seen in Fig. 6.

6. CLOTH ANIMATION AND RENDERING

The reconstructed body model is used to run a physical cloth simulation using the commercially available *Marvelous Designer™*. The body model assures that the garment interacts plausibly with the human actor. We export the animated garment model, the predefined camera configuration, and the reconstructed virtual point lights into *Mental Ray™* for rendering. Along with the pixel color, we also render into a separate depth buffer which we later use for depth-based compositing with the input video. Finally, we make use of differential rendering [Debevec 1998] to transfer all illumination effects caused by the artificial garment into the original scene. For this we use the shading difference of the reconstructed body model rendered with and without the virtual garment to extract information about shadows and ambient occlusion caused by the garment and apply them to the original actor during the compositing step (Sect. 7.3).

7. IMAGE-BASED REFINEMENT

To obtain convincing compositing results, we propose to refine the rendering result of the virtual garment in image space for pixel accurate alignment with the actor in the input video. We first correct for remaining differences in body shape between the actor and the body model (Sect. 7.1), then align the silhouettes between the rendered virtual garment $\mathbf{G}$ and the silhouette of the actor $\mathbf{M}$ (Sect. 7.2), and, finally, warp and composite $\mathbf{G}$ into the input video (Sect. 7.3). This approach follows the intuition that similar body shapes create similar wrinkles and deformations in garment simulations [Guan et al. 2012].

7.1 Image-based Garment Motion Correction

The optimized body model (Sect. 4) provides a robust but usually not pixel-precise representation of the human actor. To create a convincing illusion of the actor wearing the virtual garment, the garment needs to move precisely according to the actor in the video. A mismatch in the motion between the reconstructed body model and the actor otherwise results in an unnatural "floating" of the gar-

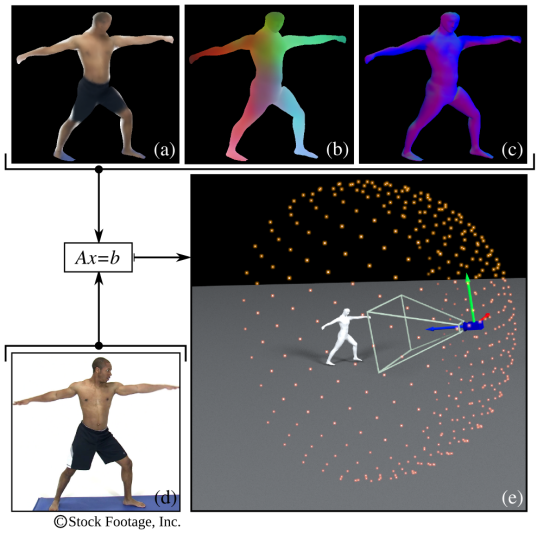

Fig. 5: **Scene illumination estimation.** We use the estimated diffuse surface color on the body model (a), 3D surface position (b) and orientation (c) and the original video (d) to reconstruct scene illumination for each video frame. Virtual point light sources are distributed across the frontal hemisphere (e). For each light source, intensity and color is determined to match the actor's illumination in the video frame.

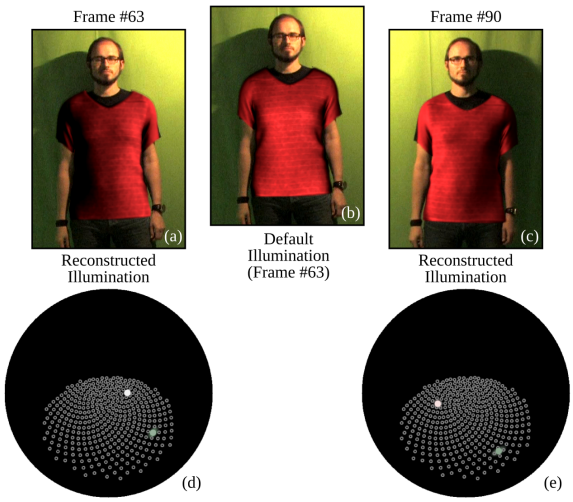

Fig. 6: **Reconstructed scene illumination.** Reconstructed illumination of the *Dyn. Light* sequence depicting dynamic illumination with a moving light source. **Top:** The illumination estimate improves the realism of the composite, especially in the case of an animated light source, as the reconstructed light positions illuminate the shirt from the correct direction ((a) and (c)), while the default illumination of the used render tool (*Mental Ray*™) introduces inconsistencies between background shadows and shading of the virtual garment (b), as directions of the incident light do not match. **Bottom:** Visualization of the reconstructed light positions as environment maps ((d) and (e)). Grey cirles mark all L possible light source positions used for optimization. The bigger dots mark color and position of light sources of the illumination reconstruction result. A greenish light source has been reconstructed at similar positions for both sample frames, recreating the indirect illumination from the floor and static illumination. A brighter white light source was reconstructed at positions resembling the direct illumination of the moving light source in the *Dyn. Light* sequence.

ment on the actor. In our approach we establish a correspondence field between the body model and the actor that we use to correct for the differences in motion between both in image space.

The user starts by selecting keyframes in the video where the rendered clothes visually fit the actor as desired. Less than ten key frames have proven to be sufficient for all our test sequences. From there all further correction is done automatically. The motion of the rendered body model in image space is known. For a continuous motion estimation of the original actor, we compute the optical flow inbetween adjacent frames of the input video. To this end we make use of the long-range optical by [Lipski et al. 2010] which uses a robust belief propagation algorithm and SIFT-feature descriptors per pixel, extended with RGB information, to compute a pixel precise matching. We compute trajectories for each pixel (x, y) in each keyframe by concatenating the frame-to-frame flow fields for the input video $\mathbf{T_I}(x, y, t)$ and the rendered body model $\mathbf{T_B}(x, y, t)$, respectively. $\mathbf{T}(x, y, t)$ tells for each intermediate frame t where a certain pixel (x, y) from the previous keyframe moved during the animation. For readability reasons we omit the x, y, t parameters of the trajectories, when they are not required. We discard a pixel trajectory in $\mathbf{T_B}$ if it crosses a strong discontinuity in the depth map of the body model and for $\mathbf{T_I}$ if it crosses a silhouette boundary as the correspondences and therefore $\mathbf{T}$ is not reliable anymore.

Given $\mathbf{T_B}$ and $\mathbf{T_I}$ for the body model and the input video, we create a difference map $\mathbf{D}$ that describes for each frame how the motion of the actor and the body mesh differ. For this we subtract $\mathbf{T_B}$ and $\mathbf{T_I}$ from each other and write the result to the warped pixel position according to $\mathbf{T_I}$ in each frame:

$$\mathbf{D}(x', y', t) = \mathbf{T_I}(x, y, t) - \mathbf{T_B}(x, y, t) \quad (8)$$

Where (x, y) is a pixel position in a keyframe and (x', y') is the pixel position according to $\mathbf{T_B}(x, y, t)$, cf. Fig. 7.

Assuming that the motion of the actor is locally smooth, we can safely apply an outlier removal to all $\mathbf{D}$ using a small median filter. $\mathbf{D}$ is only sparsely populated, due to discarded trajectories, occlusion and disocclusion. To keep the change subtle and to avoid visible artifacts we interpolate and smooth the values of $\mathbf{D}(x, y, t)$ by applying a domain transform filter [Gastal and Oliveira 2011], which is a versatile and edge-preserving filter function capable of processing high resolution images in real time. The depth of the rendered garment serves as the edge function for the filter. Applying the resulting smooth warp field to the rendered garment $\mathbf{G}$ for each frame nicely adapts the motion of the garment from the body model to the real actor. To avoid drifts due to imprecise optical flow computations, the same procedure is applied in backward direction between two keyframes and the warps are averaged.

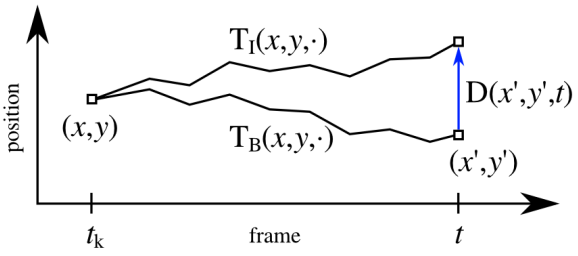

Fig. 7: **Image-space garment motion correction.** The difference vector $\mathbf{D}(x', y', t)$ describes the overall accumulated motion difference of pixel (x, y) at frame t. It is used to correct the misaligned pixel position (x', y') of the rendered garment stemming from motion differences between body model and actor. $\mathbf{D}(x', y', t)$ is computed from the trajectories $\mathbf{T_B}(x, y, \cdot)$ and $\mathbf{T_I}(x, y, \cdot)$ at frame t, which represent the motion of corresponding pixels of body model resp. actor starting from a keyframe t_k up to frame t.

7.2 Silhouette Matching

While the last step corrected for general screen-space motion differences stemming from shape deviations between the body model and the actor, we now need to track non-matching silhouettes over time and correct them so that the garment correctly overlaps the actor in the video. Detecting semantically meaningful silhouettes for matching is an ill-posed problem as the cause for mismatching silhouettes between garment and actor may also be intended for wider apparel. We, therefore, opted for a semi-automatic approach. For preview and misalignment detection, we warp the garment according to the warp field from the last step.

We track a set of silhouettes as follows: The user begins by selecting an input frame at a time t_0 in the video sequence and specifies two points s_{start} and s_{end} along the silhouettes of the garment and also along the silhouettes of the actor that are to be matched, Fig. 8. Since the silhouettes of both $\mathbf{G}_{t_0}$ and $\mathbf{M}_{t_0}$ are known, the segments $S_{\mathbf{G}_{t_0}}$ and $S_{\mathbf{M}_{t_0}}$ connecting start and endpoints in the respective mattes can be found by tracing both silhouette contours from s_{start} to s_{end}. For the matching, we treat both lines $S_{\mathbf{G}_{t_0}}$ and

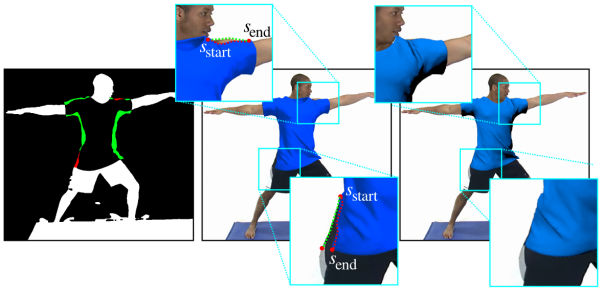

Fig. 8: **Image-space refinement.** To align the rendered garment with the actor in the video (© Stock Footage) more accurately, we warp silhouette mismatches in image space. (a) Video frame and rendered garment with mismatching silhouettes (in green and red). (b) Composite without silhouette warp. (c) Composite after image-space refinement and depth-based compositing.

$S_{\mathbf{M}_{t_0}}$ as parametric curves and match points with the same interpolation parameter.

The user then selects a later frame at time t_1 and specifies the start and end points along the same silhouette segments which do *not* have to be exactly the same position along the silhouette as in frame t_0. To solve for intermediate frames, we linearly interpolate the positions of s_{start} and s_{end} for each silhouette and snap them to the position of the closest silhouette pixel in each frame. We extract the silhouette segments $S_{\mathbf{G}_t}$ and $S_{\mathbf{M}_t}$ given s_{start} and s_{end} for all frames at time t with $t_0 \leq t \leq t_1$ and match them again as described above.

7.3 Warping and Compositing

We use radial basis functions to establish a smooth warp field that matches the silhouettes of $\mathbf{G}$ and $\mathbf{M}$ for all frames between time t_0 and t_1. Let $s_i,\ i \in 1,..,K$ be the i^{th} pixel along $S_{\mathbf{G}_t}$ of length K and $w_s(s_i)$ be the associated warp vectors to match s_i to its correspondence on $S_{\mathbf{M}_t}$. The warp $w(p)$ for the position p of each pixel in the image of the rendered garment $\mathbf{G}$ is estimated as an inverse multi quadric weighting of all warps along the silhouette segment with a Gaussian falloff:

$$w(p) = \frac{\sum_i^K \frac{1}{\sqrt{1+\|p-s_i\|}} \cdot w_s(s_i)}{\sum_i^K \frac{1}{\sqrt{1+\|p-s_i\|}}} \cdot e^{-\frac{1}{2}\frac{\|p-s_c\|}{\sigma^2}} \qquad (9)$$

where $\sigma = 8.3$ and s_c is the closest point to p along any silhouette segment $S_{\mathbf{G}_t}$.

This warp field is then concatenated with the warp field from the image-based body model correction (Sect. 7.1) and applied to the original rendering of $\mathbf{G}$ and the depth map of the rendered body model. Finally, the virtual garment is composited into the input video by comparison of the depth values of the warped garment $\mathbf{G}$ and warped body model $\mathbf{B}$.

Further, we use differential rendering [Debevec 1998] to darken those regions of the input video that are in shadow of the garment using the shadow matte generated during the garment rendering step (Sect. 6).

8. RESULTS

We tested a variety of video sequences in our experiments, Tab. I, which differ in resolution, duration, motion complexity of the actor, scene illumination, and camera motion. The sequence *Ballet* depicts a ballet dancer performing a fast pirouette motion, while the *Dancer* scene contains a combination of slow and fast movements. The *Hulk* sequence allows evaluating complex interactions of the actor and virtual garment, while the *Yoga* sequence depicts time-coherent realistic folding of clothes. We use *Dyn. Light* as an example for a scene with a moving light source. While the scenes *Ballet* and *Yoga* are taken from videos available online [Howcast Media, Inc. ; Stock Footage, Inc.], *Dancer*, *Dyn. Light*, and *Hulk* are own recordings. The *Haidi* sequence taken from the *i3DPost* dataset [Starck and Hilton 2007; Gkalelis et al. 2009] comprises a straight walking motion. The sequence *Into The Blue* is a short clip taken from the eponymous movie [MGM, Inc. 2005] while *Parkour* is taken from the short film *Aaron Martin - Parkour* provided by Yannick Wolff [Wolff]. Both clips demonstrate the applicability of our approach to professional shots with strong camera motion. All scenes have a resolution of 1080p except for *Ballet* (720p) and *Dancer* (4k).

Scene	Frames	Edited frames		Editing time (in min)		Processing time (in min)
		pose	warp	pose	warp	
Ballet	93	25	81	20	60	360
Dancer	140	27	89	20	50	1200
Dyn. Light	240	7	66	5	82	820
Haidi	110	20	79	20	40	420
Hulk	150	9	0	6	0	760
Yoga	299	12	139	15	48	1380
Into The Blue	149	22	7	25	3	250
Parkour	256	35	12	30	10	420

Table I. : **Test sequences.** The first two columns (1, 2) state the test sequence name resp. duration in number of frames. Column (3) shows the number of manually edited frames for body pose and image warp corrections. The number of edited frames regarding pose corrections, cf. Sect. 4.4, depends on the complexity of the motion performed by the actor. The number of frames with image-based corrections is the total of all keyframe editing operations required for the warp computation in Sect. 7. Column (4) shows the corresponding total manual editing time. Column (5) presents the computer processing time of a CPU implementation on a commodity PC (Intel Core i7, with 3.2 GHz and 12GB RAM).

The MakeHuman body model was used in the scenes *Into The Blue* and *Parkour*. In all other scenes we used the body model of [Hasler et al. 2009]. We found the MakeHuman model allowed for a slightly more precise adjustment of gender specific body proportions.

The number of manually edited frames during shape and pose optimization, cf. Tab. I, includes the initial positioning of the body model for shape reconstruction as well as the corrections of falsely reconstructed joint orientations after automatic pose reconstruction (Sect. 4). The amount of required manual guidance of the body and silhouette matching algorithm (Sect. 7), is included in Tab. I, 3rd column. Scenes with more complex or fast motions require more corrections of the pose estimation (*Ballet*, *Dancer*) than scenes with slower motions (*Dyn. Light*, *Yoga*). Pose correction was necessary in frames with ambiguous silhouettes resulting from body self-occlusions in the scenes *Haidi*, *Hulk*, *Into The Blue*, and *Parkour*. Editing pose keyframes, however, is not overly time consuming, taking less than a minute per edited frame in general.

User guidance for image-based refinements is needed in frames with strong silhouette deformations. Correcting a silhouette mismatch in a single frame only takes a few seconds, and since the silhouette start and end position are interpolated linearly, corrections for slow moving objects are completed quickly. However,

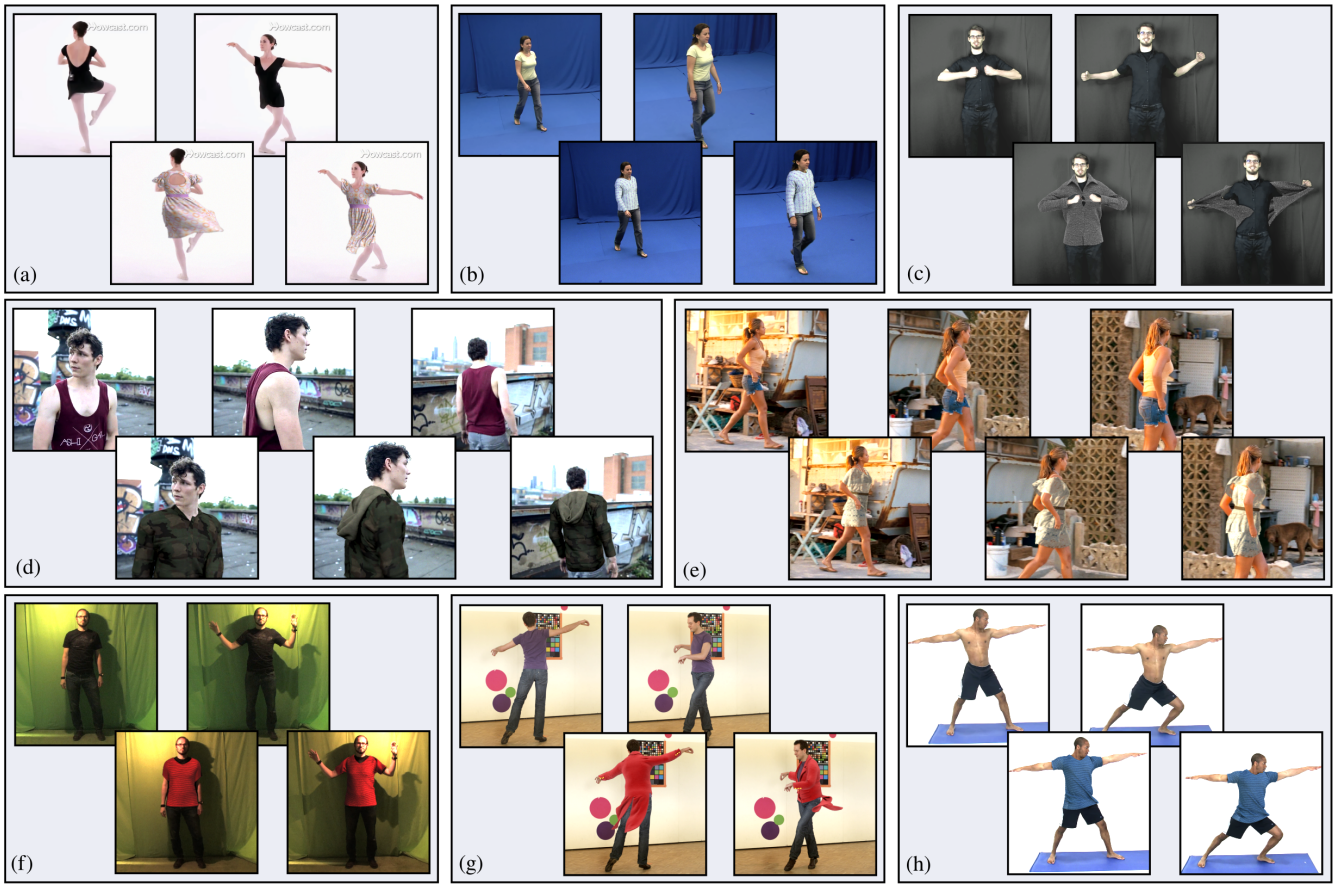

Fig. 9: **Results.** Original (top) and augmented frames (bottom) of our test sequences (a) *Ballet* © Howcast Media, (b) *Haidi*, (c) *Hulk*, (d) *Parkour* Yannick Wolff, (e) *Into The Blue* © MGM, (f) *Dyn. Light*, (g) *Dancer*, and (h) *Yoga* © Stock Footage.

more editing time is necessary for fast non-linear motions. By only applying a few corrections the overall quality of most sequences increases considerably and the majority of the time listed in the fourth column of Tab. I was spent on correcting details. This is especially the case in the *Ballet* scene due to the fast, motion blurred rotation of the dancer.

On average, manual interaction using the provided tools takes approximately one minute per edited frame. Total editing time for an entire sequence varies between 60 and 90 minutes. We think this amount of manual interaction is permissable to achieve best-possible results, as state-of-the-art approaches allow comparable amounts of user guidance in the range of one to ten minutes per edited frame [Jain et al. 2012].

The results, cf. Fig. 9, as well as the accompanying video show that our approach allows augmenting an uncalibrated, monocular video with virtually rendered garments in a variety of different settings. The selected scenes illustrate nicely that our approach can handle tight as well as loose-fitting virtual garments sufficiently. As we reconstruct the scene illumination on a per frame basis, changing directions of incident light are modeled properly, cf. Fig. 9(f) and our accompanying video. Since we use the reconstructed body model as collision proxy in the garment simulation, complex interactions of body and garment can be generated, and convincingly composed into the original video, cf. Fig. 9(c). In the case of the *Parkour* sequence, we removed parts of the actors loose fitting shirt by applying a fast diffusion based inpainting technique [Oliveira et al. 2001] and adjusting the actor's silhouette accordingly.

9. DISCUSSION AND FUTURE WORK

The results demonstrate that the hybrid 3D/image-based technique provides an easy-to-use and versatile way to convincingly augment general monocular videos of human actors with virtual garments. While our approach presents only a first step towards realistic video editing and augmentation, our solution provides a generally applicable framework that enables to solve the problem of garment replacement for a variety of different settings. Given the methods and techniques available today, we are convinced that the only way to attain even more realistic results would come at the price of investing disproportionately more manual work. In this work we tried to find an optimal balance between automatic processing and manual guidance to obtain highest-possible visual quality for still passable manual labor costs. The problem tackled in this paper is ill-posed, and many opportunities for future improvements exist.

Body and Shape Reconstruction. Shape optimization requires a "cooperative" silhouette. In cases where only a frontal view is provided, gender-specific shape characteristics can not be derived adequately. This problem is currently handled by allowing the user to adjust or constrain individual shape characteristics.

Scene Reconstruction. BRDF reconstruction from monocular video sequences with unknown scene lighting is an ill-posed problem. In the worst case, some color channels cannot be reconstructed at all, e.g., a white dress under red illumination. This is no limitation of our algorithm but rather a physical limitation and requires either manual adjustment of the BRDF, additional input recordings, or manual illumination modeling. Extending the model to handle more complex BRDFs would help to reconstruct more complex lighting conditions.

Image-based Refinement. We found that our user-guided image-based matching algorithms provide a powerful tool to correct small misalignments between the actor and the virtual garment in an unobstrusive way. The image-based model correction is able to handle inaccuracies of the body model. The silhouette matching nicely corrects small misalignments without the need for the user to explicitely track specific points on the silhouette which makes the correction tool easy and efficient to use. Although we do not explicitely enforce temporal consistency in the warping step, we did not encounter visually disturbing artifacts in our examples. However, in cases where the mismatch is too large the warping will become visible.

Clothing and Rendering. Our algorithm demands close-fitting clothes in the original video. Supporting arbitrary apparel in the input video requires more robust pose estimators that can deal with the additional uncertainties due to loose-fitting clothes [Hasler et al. 2009]. It also requires more sophisticated segmentation and inpainting techniques to remove the clothing from the input video [Granados et al. 2012]. Besides these limitations to the input video, certain pose parameters of the body model strongly influence the result of the garment simulation, as garment geometry might get stuck inbetween parts of the body geometry. The physics simulation is not able to resolve disadvantageous geometry collisions, resulting in a jittering mesh animation. This can be seen in the arm pit regions of the *Parkour* sequence.

Real-Time Implementation. Our approach is currently limited to off-line processing due to the high computational demands for fitting the body model parameter, manual interaction and garment rendering. While rendering can be accelerated, precise real-time pose estimation is not yet possible. Still, porting the solver to the GPU or making use of temporal predictions to find better initial poses will let the solver converge more quickly. Real-time implementations of our approach can open up other fields of applications, such as virtual try-on systems, e.g. for internet shopping using only a simple webcam, or visualizing people in different apparel using augmented reality eye wear.

10. CONCLUSION

We have presented a hybrid 3D/image-based approach to augment actors in general monocular video recordings with arbitrary garments without any additional input information or intrusive markers on the actor. The approach requires considerably less user interaction than state-of-the-art approaches. Our specifically tailored error function takes the silhouette, height, and information on self-intersection into account. Our matching and warping technique robustly removes visual artifacts caused by inaccuracies of the body model. A high degree of realism is achieved by our automatic per-frame scene illumination reconstruction. The 3D/image-based approach enables us to create realistically looking results in a flexible and versatile way. We applied our approach to scenes containing a variety of different motions, resolutions, video qualities and lighting conditions. The algorithm forms a solid basis for further research including partial occlusions, real-time implementations for virtual try-on systems, and realistically editing and augmenting real-world footage.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Unions Seventh Framework Programme FP7/2007-2013 under grant agreement no. 256941, Reality CG.

We would also like to thank The Foundry Ltd. for providing Nuke software licenses.

REFERENCES

AGARWAL, A. AND TRIGGS, B. 2004. 3D human pose from silhouettes by relevance vector regression. In *Computer Vision and Pattern Recognition*. Vol. 2. II–882–II–888.

ANGUELOV, D., SRINIVASAN, P., KOLLER, D., THRUN, S., RODGERS, J., AND DAVIS, J. 2005. SCAPE: shape completion and animation of people. *ACM Trans. Graph. 24,* 3, 408–416.

BAI, X., WANG, J., SIMONS, D., AND SAPIRO, G. 2009. Video SnapCut: robust video object cutout using localized classifiers. *ACM Trans. Graph. 28,* 3, 70:1–70:11.

BALAN, A., SIGAL, L., BLACK, M., DAVIS, J., AND HAUSSECKER, H. 2007. Detailed human shape and pose from images. In *Computer Vision and Pattern Recognition*. 1–8.

BĂLAN, A. O. AND BLACK, M. J. 2008. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision: Part II*. 15–29.

CHEN, X., WANG, K., AND JIN, X. 2011. Single image based illumination estimation for lighting virtual object in real scene. In *Comp.-Aided Design and Comp. Graph.* 450 –455.

DE AGUIAR, E., STOLL, C., THEOBALT, C., AHMED, N., SEIDEL, H.-P., AND THRUN, S. 2008. Performance capture from sparse multi-view video. *ACM Trans.Graph. 27,* 3, 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 *Conference on Computer graphics and interactive techniques*. SIGGRAPH '98. 189–198.

DIVIVIER, A., TRIEB, R., EBERT, A., HAGEN, H., GROSS, C., FUHRMANN, A., LUCKAS, V., ENCARNAO, J. L., KIRCHDÖRFER, E., RUPP, M., VIETH, S., KIMMERLE, S., KECKEISEN, M., WACKER, M., STRASSER, W., SATTLER, M., AND SAR, R. 2004. Virtual Try-On: Topics in realistic, individualized dressing in virtual reality. In *Virtual and Augmented Reality Status*. 1–17.

FITNECT. 2012. Fitnect, Interactive Kft. Website. Available online at http://www.fitnect.hu/, visited in Jan. 2013.

FRAHM, J., KOESER, K., GREST, D., AND KOCH, R. 2005. Markerless augmented reality with light source estimation for direct illumination. In *Conference on Visual Media Production*. 211–220.

GASTAL, E. S. L. AND OLIVEIRA, M. M. 2011. Domain transform for edge-aware image and video processing. *ACM Trans. Graph. 30,* 4, 69:1–69:12.

GIBSON, S., HOWARD, T., AND HUBBOLD, R. 2001. Flexible image-based photometric reconstruction using virtual light sources. *Computer Graphics Forum 20,* 3, 203–214.

GIOVANNI, S., CHOI, Y., HUANG, J., KHOO, E., AND YIN, K. 2012. Virtual try-on using kinect and hd camera. In *Motion in Games*, M. Kallmann and K. Bekris, Eds. Lecture Notes in Computer Science, vol. 7660. Springer Berlin Heidelberg, 55–65.

GKALELIS, N., KIM, H., HILTON, A., NIKOLAIDIS, N., AND PITAS, I. 2009. The i3dpost multi-view and 3d human action/interaction database. In *Conference for Visual Media Production*. 159 –168.

GRANADOS, M., KIM, K. I., TOMPKIN, J., KAUTZ, J., AND THEOBALT, C. 2012. Background inpainting for videos with dynamic objects and a free-moving camera. In *Proceedings of the 12th European conference on Computer Vision - Volume Part I*. ECCV'12. 682–695.

GUAN, P., FREIFELD, O., AND BLACK, M. 2010. A 2D human body model dressed in eigen clothing. In *European Conference on Computer Vision*. Lecture Notes in Computer Science, vol. 6311. 285–298.

GUAN, P., REISS, L., HIRSHBERG, D., WEISS, A., AND BLACK, M. J. 2012. Drape: Dressing any person. *ACM Trans. on Graph. 31,* 4, 35:1–35:10.

GUAN, P., WEISS, A., BALAN, A., AND BLACK, M. 2009. Estimating human shape and pose from a single image. In *IEEE Conference on Computer Vision*. 1381–1388.

GUSKOV, I., KLIBANOV, S., AND BRYANT, B. 2003. Trackable surfaces. In *ACM SIGGRAPH/Eurographics symposium on Computer animation*. 251–257.

HASLER, N., ROSENHAHN, B., THORMAHLEN, T., WAND, M., GALL, J., AND SEIDEL, H.-P. 2009. Markerless motion capture with unsynchronized moving cameras. In *Computer Vision and Pattern Recognition*. 224 –231.

HASLER, N., STOLL, C., ROSENHAHN, B., THORMÄHLEN, T., AND SEIDEL, H.-P. 2009. Estimating body shape of dressed humans. *Computers & Graphics 33,* 3, 211–216.

HASLER, N., STOLL, C., SUNKEL, M., ROSENHAHN, B., AND SEIDEL, H.-P. 2009. A statistical model of human pose and body shape. *Computer Graphics Forum 28,* 2, 337–346.

HAUSWIESNER, S., STRAKA, M., AND REITMAYR, G. 2011. Free viewpoint virtual try-on with commodity depth cameras. In *Virtual Reality Continuum and Its Applications in Industry*. 23–30.

HAUSWIESNER, S., STRAKA, M., AND REITMAYR, G. 2013. Virtual try-on through image-based rendering. *Visualization and Computer Graphics, IEEE Transactions on 19,* 9, 1552–1565.

HILSMANN, A. AND EISERT, P. 2012. Image-based animation of clothes. In *Eurographics (Short Papers)*. Eurographics Association, 69–72.

HILSMANN, A., FECHTELER, P., AND EISERT, P. 2013. Pose space image based rendering. *Computer Graphics Forum 32,* 2pt3, 265–274.

HOWCAST MEDIA, INC. Ballet dancing: How to do a pirouette. http://www.howcast.com/videos/497190-How-to-Do-a-Pirouette-Ballet-Dance.

JAIN, A., THORMÄHLEN, T., SEIDEL, H.-P., AND THEOBALT, C. 2010. Moviereshape: Tracking and reshaping of humans in videos. *ACM Trans. Graph. 29,* 5.

JAIN, E., SHEIKH, Y., MAHLER, M., AND HODGINS, J. 2012. Three-dimensional proxies for hand-drawn characters. *ACM Trans. Graph. 31,* 1 (Feb.), 8:1–8:16.

LANDGREBE, M. 2012. Underworld: Awakening. *Digital Production* 3.

LAWSON, C. L. AND HANSON, R. J. 1995. *Solving least squares problems*. Society for Ind. and Appl. Math.

LIPSKI, C., LINZ, C., NEUMANN, T., WACKER, M., AND MAGNOR, M. 2010. High resolution image correspondences for video post-production. In *Proc. European Conference on Visual Media Production (CVMP) 2010*. Vol. 7. 33–39.

MAKEHUMAN. 2012. Make human - open source tool for making 3d characters. http://www.makehuman.org, visited in April 2013.

MGM, INC. 2005. Into the blue.

NELDER, J. A. AND MEAD, R. 1965. A simplex method for function minimization. *The Computer Journal 7,* 4, 308–313.

OLIVEIRA, M. M., BOWEN, B., MCKENNA, R., AND CHANG, Y.-S. 2001. Fast digital image inpainting. In *Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP 2001), Marbella, Spain*. 106–107.

PRITCHARD, D. AND HEIDRICH, W. 2003. Cloth motion capture. *Comput. Graph. Forum 22*, 263–272.

RAMAKRISHNA, V., KANADE, T., AND SHEIKH, Y. 2012. Reconstructing 3d human pose from 2d image landmarks. In *European conference on Computer Vision - Part IV*. Lecture Notes in Computer Science, vol. 7575. 573–586.

ROGGE, L., NEUMANN, T., WACKER, M., AND MAGNOR, M. 2011. Monocular pose reconstruction for an augmented reality clothing system. In *Vision, Modeling and Visualization*. 339–346.

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

SCHOLZ, V. AND MAGNOR, M. 2006. Texture replacement of garments in monocular video sequences. In *Eurographics Symposium on Rendering*. 305–312.

SCHOLZ, V., STICH, T., KECKEISEN, M., WACKER, M., AND MAGNOR, M. 2005. Garment motion capture using color-coded patterns. *Computer Graphics Forum 24,* 3, 439–448.

SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. 2006. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph. 25,* 3, 835–846.

STARCK, J. AND HILTON, A. 2007. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications 27,* 3, 21 –31.

STOCK FOOTAGE, INC. Man doing yoga on a white background. http://www.stockfootage.com/shop/man-doing-yoga-on-a-white-background.

SWINBANK, R. AND PURSER, R. J. 2006. Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society 132,* 619, 1769–1793.

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 Transactions on Visualization and Computer Graphics 13,* 4, 663–674.

TÖPPE, E., OSWALD, M., CREMERS, D., AND ROTHER, C. 2011. Silhouette-based variational methods for single view reconstruction. In *Video Processing and Computational Video*, D. Cremers, M. Magnor, M. Oswald, and L. Zelnik-Manor, Eds. Lecture Notes in Computer Science, vol. 7082. 104–123.

VONDRAK, M., SIGAL, L., HODGINS, J. K., AND JENKINS, O. C. 2012. Video-based 3d motion capture through biped control. *ACM Trans. Graph. 31,* 4, 27.

WEI, X. AND CHAI, J. 2010. Videomocap: modeling physically realistic human motion from monocular video sequences. *ACM Trans. Graph. 29,* 4, 42:1–42:10.

WOLFF, Y. Parkour. http://vimeo.com/68317895.

XU, F., LIU, Y., STOLL, C., TOMPKIN, J., BHARAJ, G., DAI, Q., SEIDEL, H.-P., KAUTZ, J., AND THEOBALT, C. 2011. Video-based characters: creating new human performances from a multi-view video database. *ACM Trans. Graph. 30,* 4, 32:1–32:10.

YOON, J.-C., LEE, I.-K., AND KANG, H. 2011. Image-based dress-up system. In *Ubiquitous Information Management and Communication*. ICUIMC '11. 52:1–52:9.

ZHOU, S., FU, H., LIU, L., COHEN-OR, D., AND HAN, X. 2010. Parametric reshaping of human bodies in images. In *ACM SIGGRAPH 2010 papers*. SIGGRAPH '10. ACM, New York, NY, USA, 126:1–126:10.

ZIEGLER, G., LENSCH, H., AHMED, N., MAGNOR, M., AND SEIDEL, H.-P. 2004. Multivideo compression in texture space. In *International Conference on Image Processing*. Vol. 4. 2467 – 2470 Vol. 4.

Received March 2014; accepted April 2014