MovieReshape: Tracking and Reshaping of Humans in Videos

Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel and Christian Theobalt
Max-Planck-Institut Informatik, Saarbrücken, Germany
{ajain, thormae, hpseidel, theobalt}@mpi-inf.mpg.de

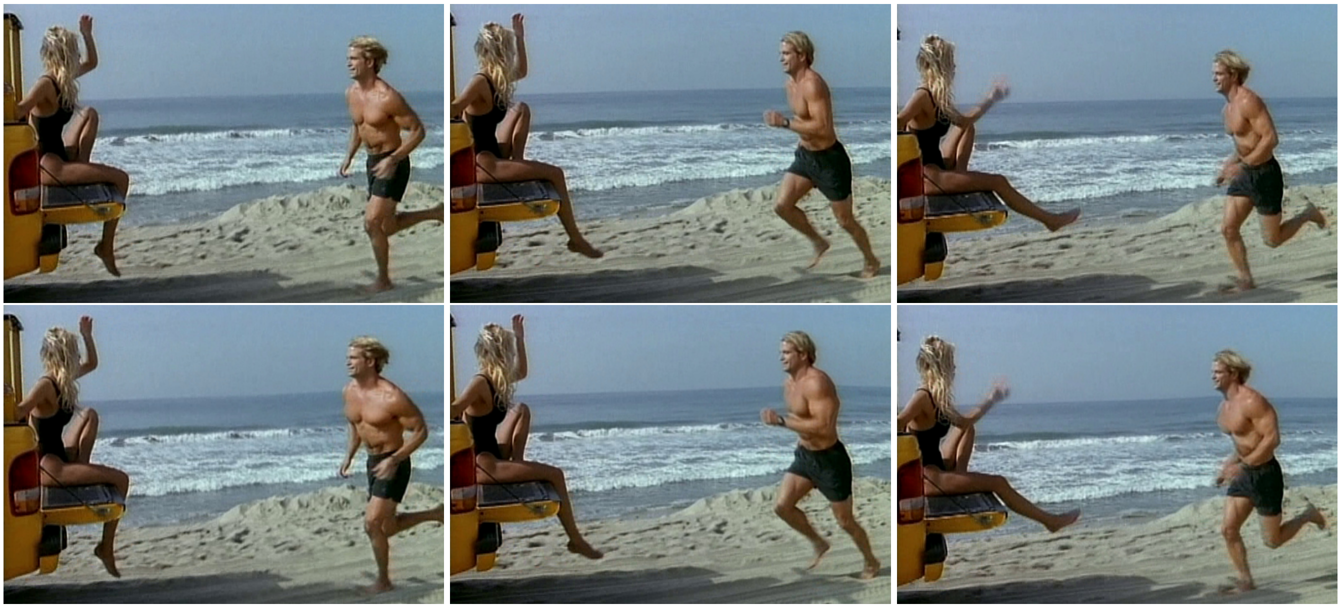


Figure 1: *In this sequence from the TV series Baywatch, we modified the original appearance of the actor (top row) such that he appears more muscular (bottom row). The edit was performed with our system by simply increasing the value on the muscularity control slider.*

Abstract

We present a system for quick and easy manipulation of the body shape and proportions of a human actor in arbitrary video footage. The approach is based on a morphable model of 3D human shape and pose that was learned from laser scans of real people. The algorithm commences by spatio-temporally fitting the pose and shape of this model to the actor in either single-view or multi-view video footage. Once the model has been fitted, semantically meaningful attributes of body shape, such as height, weight or waist girth, can be interactively modified by the user. The changed proportions of the virtual human model are then applied to the actor in all video frames by performing an image-based warping. By this means, we can now conveniently perform spatio-temporal reshaping of human actors in video footage which we show on a variety of video sequences.

Keywords: video editing, video retouching, reshaping of actors, morphable body model

1 Introduction

Digital retouching of photographs is an essential operation in commercial photography for advertisements or magazines, but is also increasingly popular among hobby photographers. Typical retouching operations aim for visual perfection, for instance by removing scars or birthmarks, adjusting lighting, changing scene backgrounds, or adjusting body proportions. Unfortunately, even commercial-grade image editing tools often only provide very basic manipulation functionality. Therefore, many advanced retouching operations, such as changing the appearance or proportions of the body, often require hours of manual work. To facilitate such advanced editing operations, researchers developed semantically-based retouching tools that employ parametric models of faces and human bodies in order to perform complicated edits more easily. Examples are algorithms to increase the attractiveness of a face [Leyvand et al. 2008], or to semi-automatically change the shape of a person in a photograph [Zhou et al. 2010].

While such semantically-based retouching of photographs is already very challenging, performing similar edits on video streams has almost been impossible up to now. Existing commercial video editing tools (Sec. 2) only provide comparatively basic manipulation functions, such as video object segmentation or video retargeting, and already these operations are computationally very demanding. Only a few object-based video manipulation approaches go slightly beyond these limits, for instance by allowing facial expression change [Vlasic et al. 2005], modification of clothing texture [Scholz and Magnor 2006], or by enabling simple motion edits of video objects [Scholz et al. 2009]. The possibility to easily manipulate attributes of human body shape, such as *weight*, *height* or *muscularity*, would have many immediate applications in movie and video post-production. Unfortunately, even with the most advanced object-based video manipulation tools, such retouching would take even skilled video professionals several hours of work. The primary challenge is that body shape manipulation, even in a single video frame, has to be performed in a *holistic* way. Since the appearance of the entire body is strongly correlated, body reshaping solely based on local operations is very hard. As an additional difficulty, body reshaping in video has to be done in a spatio-temporally coherent manner.

We therefore propose in this paper one of the first systems in the literature to easily perform holistic manipulation of body attributes of human actors in video. Our algorithm is based on a 3D morphable model of human shape and pose that has been learned from full body laser scans of real individuals. This model comprises a skeleton and a surface mesh. Pose variation of the model is described via a standard surface skinning approach. The variation of the body shape across age, gender and personal constitution is modeled in a low-dimensional principal-component-analysis (PCA) parameter space. A regression scheme enables us to map the PCA parameters of human shape onto semantically meaningful scalar attributes that can be modified by the user, such as: *height*, *waist girth*, *breast girth*, *muscularity*, etc. In a first step, a marker-less motion estimation approach spatio-temporally optimizes both the pose and the shape parameters of the model to fit the actor in each video frame. In difficult poses, the user can support the algorithm with manual constraint placement. Once the 3D model is tracked, the user can interactively modify its shape attributes. By means of an image-based warping approach, the modified shape parameters of the model are applied to the actor in each video frame in a spatio-temporally coherent fashion.

We illustrate the usefulness of our approach on *single-view* and *multi-view video* sequences. For instance, we can quickly and easily alter the appearance of actors in existing movie and video footage. Further on, we can alter the physical attributes of actors captured in a controlled multi-view video studio. This allows us to carefully plan desired camera viewpoints for proper compositing with a virtual background, while giving us the ability to arbitrarily retouch the shape of the actor during post-processing. We also confirmed the high visual fidelity of our results in a user study.

2 Previous Work

In our work we can capitalize on previous research from a variety of areas. Exemplary work from the most important areas is briefly reviewed in the following.

Video Retouching Several commercial-grade image manipulation tools exist[1] that enable a variety of basic retouching operations, such as segmentation, local shape editing, or compositing. The research community also worked on object-based manipulation approaches that broaden the scope of the above basic tools, e.g., [Barrett and Cheney 2002]. Unfortunately, more advanced image edits are very cumbersome with the aforementioned approaches. A solution is offered by semantically-guided image operations, in which some form of scene model represents and constrains the space of permitted edits, such as a face model for automatic face beautification [Leyvand et al. 2008], or a body model for altering body attributes in photographs [Zhou et al. 2010].

Applying similarly complex edits to entire video streams is still a major challenge. The *Proscenium* system by Bennett et al. [2003] allows the user to shear and warp the video volumes, for instance to stabilize the camera or remove certain objects. [Liu et al. 2005] describe an algorithm for amplification of apparent motions in image sequences captured by a static camera. Wang et al. [2006] present the cartoon animation filter that can alter motions in existing video footage such that it appears more exaggerated or animated. Spatio-temporal gradient domain editing enables several advanced video effects, such as re-compositing or face replacement, at least if the faces remain static [Wang et al. 2007]. Spatio-temporal segmentation of certain foreground objects in video streams also paves the trail for some more advanced edits, such as repositioning of the object in the field of view [Wang et al. 2005; Li et al. 2005]. However, none of these methods enables easy complete reshaping of human actors in a way similar to the algorithm presented in this paper.

Our system has parallels to video retargeting algorithms that allow, for instance, to resize video while keeping the proportions of visually salient scene elements intact. Two representative video retargeting works are [Krähenbühl et al. 2009; Rubinstein et al. 2008]. However, complex plausible reshaping of humans in video is not feasible with these approaches.

Our approach employs a morphable model of human shape and pose to guide the reshaping of the actor in the video sequence. Conceptually related is the work by Scholz et al. who use a model of moving garment to replace clothing textures in monocular video [Scholz et al. 2009]. Vlasic et al. [2005] employ a morphable 3D face model to transfer facial expressions between two video sequences, where each one is showing a different individual. Finally, [Scholz and Magnor 2006] describe an algorithm to segment video objects and modify their motion within certain bounds by editing some key-frames. The algorithm by Hornung et al. [2007] solves a problem that is kind of opposite to what we aim for. They describe a semi-automatic method for animation of still images that is based on image warping under the control of projected 3D motion capture data. None of the aforementioned approaches could perform semantically plausible reshaping of actors in video footage in a similar manner as our approach.

Morphable 3D Body Models Our approach is based on a morphable model of human shape and pose similar to [Allen et al. 2003; Seo and Magnenat-Thalmann 2004; Anguelov et al. 2005; Allen et al. 2006; Hasler et al. 2009]. This model has been learned from a publicly available database of human body scans in different poses that is kindly provided by [Hasler et al. 2009]. Our body model is a variant of the SCAPE model by Anguelov et al. [2005] that describes body shape variations with a linear PCA model. Since SCAPE's shape PCA dimensions do not correspond to semantically meaningful dimensions, we remap the body parameters to semantically meaningful attributes through a linear regression similar to [Allen et al. 2003].

Marker-less Pose and Motion Estimation Monocular pose estimation from images and video streams is a highly challenging and fundamentally ill-posed problem. A few automatic approaches exist that attack the problem in the monocular case [Agarwal and Triggs 2006]. However, they often deliver very crude pose estimates and manual user guidance is required to obtain better quality results, e.g., [Davis et al. 2003; Parameswaran and Chellappa 2004; Hornung et al. 2007]. Recently, Wei and Chai [2010] presented an approach for interactive 3D pose estimation from monocular video. Similar, as with our approach in the monocular video case, manual intervention in a few keyframes is required.

In our research, we apply a variant of the marker-less pose estimation algorithm by [Gall et al. 2009] for pose inference in video. Our approach is suitable for both monocular and multi-view pose inference. A variety of marker-less motion estimation algorithms

[1] e.g. Adobe Photoshop$^{\text{TM}}$, GIMP, etc.

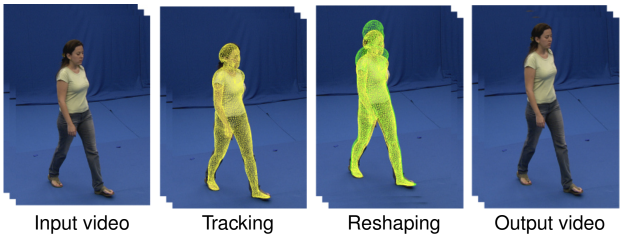


Figure 2: *The two central processing steps of our system are* **tracking** *and* **reshaping** *of a morphable 3D human model.*

for single and multi-view video have been proposed in the literature, see [Poppe 2007] for an extensive review. Many of them use rather crude body models comprising skeletons and simple shape proxies that would not be detailed enough for our purpose. At the other end of the spectrum, there are performance capture algorithms that reconstruct detailed models of dynamic scene geometry from multi-view video [de Aguiar et al. 2008; Vlasic et al. 2008]. However, they solely succeed on multi-view data, often require a full-body scan of the tracked individual as input, and do not provide a plausible parameter space for shape manipulation.

Therefore, our algorithm is based on a morphable human body model as described in the previous paragraph. Only a few other papers have employed such a model for full-body pose capture. Balan et al. [2007] track the pose and shape parameters of the SCAPE model from multi-view video footage. So far, monocular pose inference with morphable models has merely been shown for single images, [Guan et al. 2009; Hasler et al. 2010; Zhou et al. 2010; Sigal et al. 2007; Rosales and Sclaroff 2006], where manual intervention by the user user is often an integral part of the pipeline. In contrast, in our video retouching algorithm we estimate time-varying body shape and pose parameters from both single and multi-view footage, with only a small amount of user intervention needed in the monocular video case.

3 Overview

Our system takes as input a *single-view* or *multi-view* video sequence with footage of a human actor to be spatio-temporally reshaped (Fig. 2). There is no specific requirement on the type of scene, type of camera, or appearance of the background. As a first step, the silhouette of the actor in the video footage is segmented using off-the-shelf video processing tools. The second step in the pipeline is marker-less model fitting. There, both the shape and the pose parameters of the 3D model are optimized such that it re-projects optimally into the silhouette of the actor in each video frame (Sec. 4). Once the model is tracked, the shape parameters of the actor can be modified by simply tweaking a set of sliders corresponding to individual semantic shape attributes. Since the original PCA parameter dimensions of the morphable shape model do not directly correspond to plausible shape attributes, we learn a mapping from intuitive attributes, such as muscularity or weight, to the underlying PCA space (Sec. 5.1). Now reshaping can be performed by adjusting plausible parameter values. Once the target set of shape attributes has been decided on, they are applied to the actor in all frames of the video input by performing image-based warping under the influence of constraints that are derived from the re-projected modified body model (Sec. 5.2).

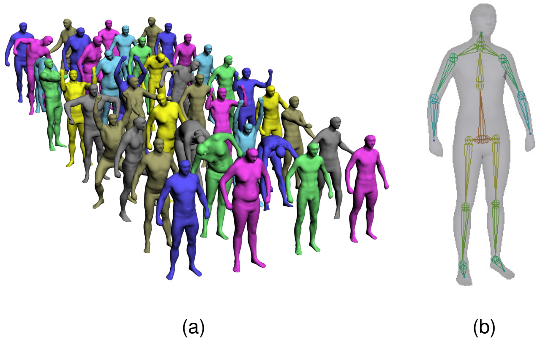


Figure 3: *Morphable body model - (a) Samples of the pose and shape parameter space that is spanned by the model. (b) The average human shape with the embedded kinematic skeleton.*

4 Tracking with a Statistical Model of Pose and Shape

In the following, we review the details of the 3D human shape model, and explain how it is used for tracking the actor in a video.

4.1 3D Morphable Body Model

We employ a variant of the SCAPE model [Anguelov et al. 2005] to represent the pose and the body proportions of an actor in 3D. We learned this model from a publicly available database of 550 registered body scans of over 100 people (roughly 50% male subjects, and 50% female subjects, aged 17 to 61) in different poses (Fig. 3(a)). The motion of the model is represented via a kinematic skeleton comprising of 15 joints. The surface of the model consists of a triangle mesh with roughly 6500 3D vertices $\mathbf{v_i}$. As opposed to the original SCAPE model, we do not learn per-triangle transformation matrices to represent subject-specific models of pose-dependent surface deformation. In our application, this level of detail is not required to obtain realistic reshaping results. Further on, the omission of this per-triangle model component prevents us form having to solve a large linear system to reconstruct the model surface, every time the model parameters have changed. This, in turn, makes pose estimation orders of magnitude faster. Instead of per-triangle transformations, we use a normal skinning approach for modeling pose-dependent surface adaptation. To this end, the skeleton has been rigged into the average shape human shape model by a professional animation artist (Fig. 3(b)).

Similar to the original SCAPE model, we represent shape variation across individuals via principal component analysis (PCA). We employ the first 20 PCA components which capture 97% of the body shape variation. In total, our model thus has $N = 28$ pose parameters $\mathbf{\Phi} = (\phi_1, \ldots, \phi_N)$ and $M = 20$ parameters $\mathbf{\Lambda} = (\lambda_1, \ldots, \lambda_M)$ to represent the body shape variation.

4.2 Marker-less Tracking

We use a marker-less motion capture approach to fit the pose and shape of the body model to a human actor in each frame of a single-view or multi-view video sequence. In case the input is an arbitrary monocular video sequence, we make the simplifying assumption that the recording camera is faithfully modeled by a scaled orthographic projection. In the multi-view video case we expect fully-calibrated frame-synchronized cameras, which is a reasonable assumption to make as most of these sequences are captured under controlled studio conditions.

Henceforth, we denote a video frame at time stamp t seen from

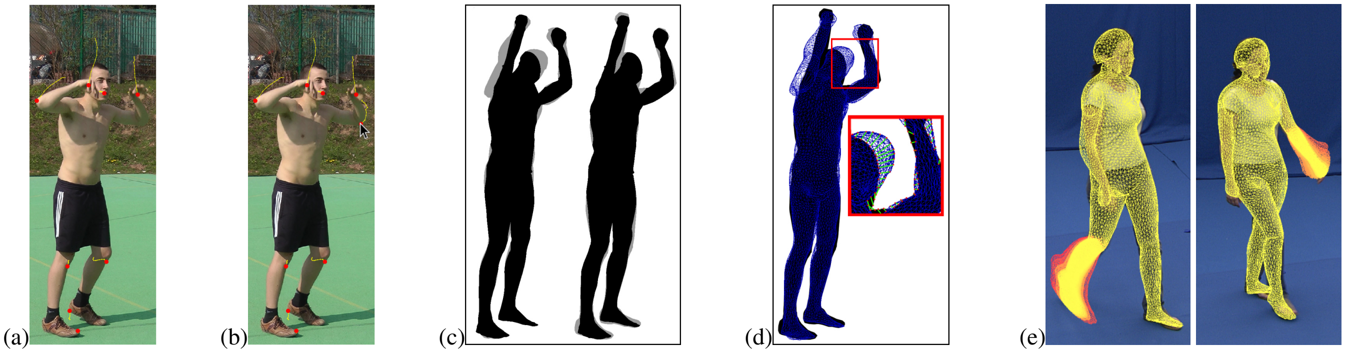


Figure 4: (a)-(d) Components of the pose error function: *(a) KLT features and their trajectories (yellow) over several frames; (b) in the monocular video case, additional feature point tracks can be manually generated or broken trajectories can be linked; (c) silhouette error term used during global optimization; a sum of image silhouette pixels not covered by the model, and vice versa (erroneous pixels in dark grey), (d) silhouette error term used during local optimization - corresponding points between image and model silhouettes and their distances are shown;* **(e) Global pose optimization:** *sampled particles (model pose hypotheses) are overlaid for the leg and the arm.*

camera c ($c = 1, \dots, C$) with $I_{t,c}$. Before tracking commences, the person is segmented from the background in each video frame, yielding a foreground silhouette. To serve this purpose, we rely on standard video processing tools[2] if chroma-keying is not possible, but note that alternative video object segmentation approaches, such as [Wang et al. 2005; Li et al. 2005], would be equally applicable.

Our motion capture scheme infers pose and shape parameters by minimizing an image-based error function $E(\mathbf{\Phi}, \mathbf{\Lambda}, t)$ that, at each time step of video t, penalizes misalignment between the 3D body model and its projection into each frame:

$$E(\mathbf{\Phi}_t, \mathbf{\Lambda}_t) = \sum_{c=1}^{C} E_s(\mathbf{\Phi}, \mathbf{\Lambda}_t, I_{t,c}) + E_f(\mathbf{\Phi}_t, \mathbf{\Lambda}_t, I_{t,c}) \; . \qquad (1)$$

The first component E_s measures the misalignment of the silhouette boundary of the re-projected model with the silhouette boundary of the segmented person. The second component E_f measures the sum of distances in the image plane between feature points of the person tracked over time, and the re-projected 3D vertex locations of the model that - in the previous frame of video - corresponded to the respective feature point. Feature trajectories are computed for the entire set of video frames before tracking commences (Fig. 4(a)). To this end, an automatic Kanade-Lucas-Tomasi (KLT) feature point detector and tracker is applied to each video frame. Automatic feature detection alone is often not sufficient, in particular if the input is a monocular video: Trajectories easily break due to self-occlusion, or feature points may not have been automatically found for body parts that are important but contain only moderate amounts of texture. We therefore provide an interface in which the user can explicitly mark additional image points to be tracked, and in which broken trajectories can be linked (Fig. 4(b)).

Pose inference at each time step t of a video is initialized with the pose parameters $\mathbf{\Phi}_{t-1}$ and shape parameters $\mathbf{\Lambda}_{t-1}$ determined in the preceding time step. For finding $\mathbf{\Phi}_t$ and $\mathbf{\Lambda}_t$ we adapt the combined local and global pose optimization scheme by [Gall et al. 2009].

Given a set of K 3D points $\mathbf{v_i}$ on the model surface and their corresponding locations in the video frame $\mathbf{u}_{i,c}$ at time t in camera c (these pairs are determined during evaluation of the silhouette and feature point error), a fast local optimization is first performed to determine the pose parameters of each body part. During local optimization, E_s in Eq. (1) is computed by assigning a set of points on the model silhouette to the corresponding closest points on the image silhouette, and summing up the 2D distances (Fig. 4(c)).

Each 2D point $\mathbf{u}_{i,c}$ defines a projection ray that can be represented as a Plücker line $L_{i,c} = (n_{i,c}, m_{i,c})$ [Stolfi 1991]. The error of pair $(\mathcal{T}(\mathbf{\Phi}_t, \mathbf{\Lambda}_t)\mathbf{v}_i, \mathbf{u}_{i,c})$ is given by the norm of the perpendicular vector between the line L_i and the 3D point $\mathbf{v}_i$ from the body models standard pose, transformed by transformation $\mathcal{T}(\mathbf{\Phi}_t, \mathbf{\Lambda}_t)$ that concatenates the pose, shape, and skinning transforms. Finding the nearest local pose and shape optimum of Eq. (1) therefore corresponds to solving

$$\underset{(\mathbf{\Phi}_t, \mathbf{\Lambda}_t)}{\operatorname{argmin}} \sum_{c}^{C} \sum_{i}^{K} w_i \| \Pi(\mathcal{T}(\mathbf{\Phi}_t, \mathbf{\Lambda}_t)\mathbf{v}_{i,c}) \times n_{i,c} - m_{i,c} \|_2^2 \qquad (2)$$

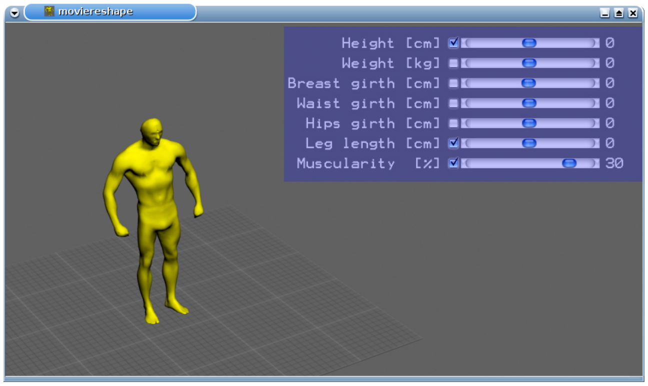
which is linearized using Taylor approximation and solved iteratively. Π is the projection from homogeneous to non-homogeneous coordinates.

Local pose optimization is extremely fast but may in some cases get stuck in incorrect local minima. Such pose errors could be prevented by running a full global pose optimization. However, global pose inference is prohibitively slow when performed on the entire pose and shape space. We therefore perform global pose optimization only for those sub-chains of the kinematic model, which are incorrectly fitted. Errors in the local optimization result manifest through a limb-specific fitting error $E(\mathbf{\Phi}_t, \mathbf{\Lambda}_t)$ that lies above a threshold. For global optimization, we utilize a particle filter. Fig. 4(d) overlays the sampled particles (pose hypotheses) for the leg and the arm.

In practice, we solve for pose and shape parameters in a hierarchical way. First, we solve for both shape and pose using only a subset of key frames of the video in which the actor shows a sufficient range pose and shape deformation. It turned out that in all our test sequences the first 20 frames form a suitable subset of frames. In this first optimization stage, we solely perform global pose and shape optimization and no local optimization. Thereafter, we keep the shape parameters fixed, and subsequently solve for the pose in *all* frame using the combined local and global optimization scheme.

We employ the same tracking framework for both multi-view ($C > 1$) and single view video sequences ($C = 1$). While multi-view data can be tracked fully-automatically, single view data may need more frequent manual intervention. In all our monocular test sequences,

[2] Mocha$^{\text{TM}}$, Adobe AfterEffects$^{\text{TM}}$



Height [cm] ☑ 0
Weight [kg] □ 0
Breast girth [cm] □ 0
Waist girth [cm] □ 0
Hips girth [cm] □ 0
Leg length [cm] ☑ 0
Muscularity [%] ☑ 30

Figure 5: *The reshaping interface allows the user to modify semantic shape attributes of a person.*

though, only a few minutes of manual user interaction were needed. Please note that monocular pose tracking is ill-posed, and therefore we cannot guarantee that the reconstructed model pose and shape are correct in a metric sense. However, in our retouching application such 3D pose errors can be tolerated as long as the re-projected model consistently overlaps with the person in all video frames. Also, for our purpose it is not essential that the re-projected model aligns exactly with the contours of the actor. The image-based warping deformation described in the following also succeeds in the presence of small misalignments.

5 Reshaping Interface

Once tracking information for shape and pose has been obtained, the body shape of the actor can be changed with our interactive reshaping interface (see Fig. 5).

5.1 Deformation of Human Shape

The PCA shape space parameters $\mathbf{\Lambda}$ do not correspond to semantically meaningful dimensions of human constitution. The modification of a single PCA parameter λ_k will simultaneously modify a combination of shape aspects that we find intuitively plausible, such as *weight* or *strength of muscles*. We therefore remap the PCA parameters onto meaningful scalar dimensions. Fortunately, the scan database from which we learn the PCA model contains for each test subject a set of semantically meaningful attributes, including: *height*, *weight*, *breast girth*, *waist girth*, *hips girth*, *leg length*, and *muscularity*. All attributes are given in their respective measurement units, as shown in Fig. 5.

Similar to [Allen et al. 2003] we project the $Q = 7$ semantic dimensions onto the M PCA space dimensions by constructing a linear mapping $\mathbf{S} \in \mathcal{M}((M-1) \times (Q+1))$ between these two spaces:

$$\mathbf{S}\,[f_1\ \ldots\ f_Q\ 1]^T = \mathbf{\Lambda}\,, \tag{3}$$

where f_i are the semantic attribute values of an individual, and $\mathbf{\Lambda}$ are the corresponding PCA coefficients. This mapping enables us to specify offset values for each semantic attribute $\Delta\mathbf{f} = [\Delta f_1\ \ldots\ \Delta f_Q\ 0]^T$. By this means we can prescribe by how much each attribute value of a specific person we tracked should be altered. For instance, one can specify that the weight of the person shall increase by a certain amount of kilograms. The offset feature values translate into offset PCA parameters $\Delta\mathbf{\Lambda} = \mathbf{S}\Delta\mathbf{f}$ that must be added to the original PCA coefficients of the person to complete the edit.

Please note that certain semantic attributes are implicitly correlated to each other. For instance, increasing a woman's height may also lead to a gradual gender change since men are typically taller than women. In an editing scenario, such side-effects may be undesirable, even if they would be considered as generally plausible. In the end, it is a question of personal taste which correlations should be allowed to manifest and which ones should be explicitly suppressed. We give the user control over this decision and give him the possibility to explicitly fix or let free certain attribute dimensions when performing an edit. To start with, for any attribute value our reshaping interface provides reasonable suggestions of what parameters to fix when modifying certain attributes individually. For instance, one suggestion is that when editing the height, the waist girth should be preserved.

5.2 Consistent Video Deformation

Our reshaping interface allows the user to generate a desired 3D target shape $\mathbf{\Lambda}' = \Delta\mathbf{\Lambda} + \mathbf{\Lambda}$ from the estimated 3D source shape $\mathbf{\Lambda}$ (remember that $\mathbf{\Lambda}$ is constant in all frames after tracking has terminated). This change can be applied automatically to all the images of the sequence. In our system the user-selected 3D shape change provides the input for a meshless moving least squares (MLS) image deformation, which was introduced by [Müller et al. 2005; Schaefer et al. 2006] (see Sec.7 for a discussion on why we selected this approach).

The 2D deformation constraints for MLS image deformation are generated by employing a sparse subset $\mathbb{S}$ of all surface vertices $\mathbf{v}_i$ of the body model. This set $\mathbb{S}$ is defined once manually for our morphable body model. We selected approx. 5 to 10 vertices per body part making sure that the resulting 2D MLS constraints are well distributed from all possible camera perspectives. This selection of a subset of vertices is done only once and then kept unchanged for all scenes. In the following, we illustrate the warping process using a single frame of video (Fig. 6). To start with, each vertex in $\mathbb{S}$ is transformed from the standard model pose into the pose and shape of the *source body*, i.e., the model in the pose and shape as it was found by our tracking approach. Afterwards, the vertex is projected into the current camera image, resulting in the source 2D deformation point $\mathbf{s}_i$. Then, each subset vertex is transformed into the pose and shape of the *target body* - i.e., the body with the altered shape attributes - and projected in the camera image to obtain the target

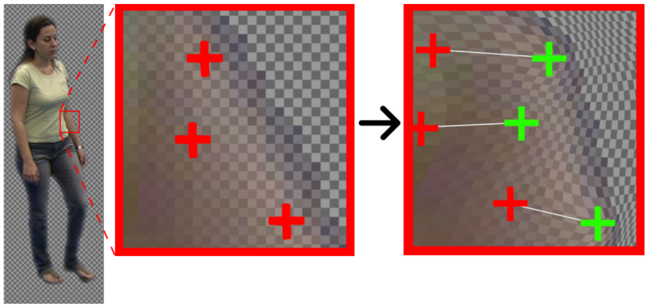


Figure 6: *Illustration of the MLS-based warping of the actor's shape. The zoomed in region shows the projected deformation constraints in the source model configuration (left), and in the target model configuration (right). The red points show the source constraint positions, the green points the target positions. The image is warped to fulfill the target constraints.*

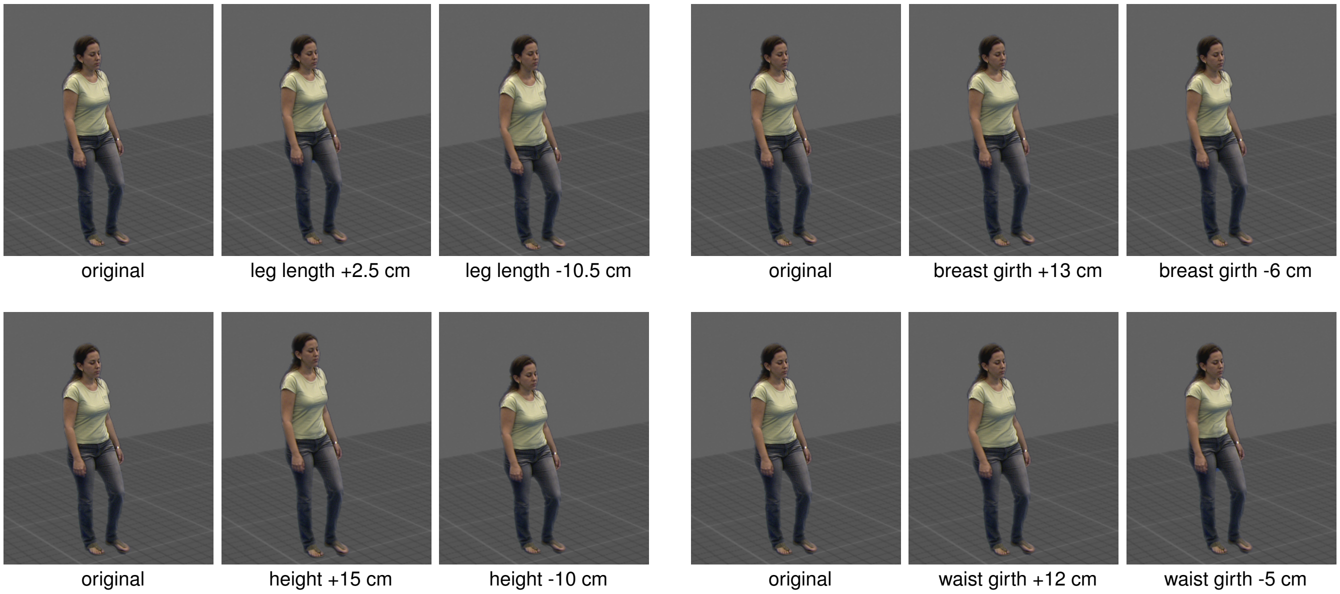


Figure 7: *A variety of reshaping results obtained by modifying several shape attributes of the same actor.*

2D deformation points $\mathbf{t}_i$:

$$\begin{aligned}\mathbf{s}_i &= \mathcal{P}_t\left(\mathcal{T}(\mathbf{\Phi}_t, \mathbf{\Lambda})\mathbf{v}_i\right) \\ \mathbf{t}_i &= \mathcal{P}_t\left(\mathcal{T}(\mathbf{\Phi}_t, \mathbf{\Lambda}')\mathbf{v}_i\right) \quad ,\end{aligned} \tag{4}$$

where $\mathcal{P}_t$ denotes the projection in the current camera image at time t.

Given the deformation constraints $\mathbf{s}_i \rightarrow \mathbf{t}_i$, MLS deformation finds for each pixel $\mathbf{x}$ in the image the optimal 2D transformation $\mathcal{M}_\mathbf{x}$ to transform the pixel to its new location $\mathbf{x}' = \mathcal{M}_\mathbf{x}(\mathbf{x})$. Thereby, the following cost function is minimized:

$$\underset{\mathcal{M}_\mathbf{x}}{\arg\min} \quad \sum_{\mathbf{s}_i, \mathbf{t}_i \in \mathbb{S}} \frac{1}{|\mathbf{x} - \mathbf{s}_i|^2} \left(\mathcal{M}_\mathbf{x}(\mathbf{s}_i) - \mathbf{t}_i\right)^2 \quad . \tag{5}$$

The closed-form solution to this minimization problem is given in [Müller et al. 2005]. Similar as in [Ritschel et al. 2009], our system calculates the optimal 2D deformation in parallel for all pixels of the image using a fragment shader on the GPU. This allows the user of the reshaping interface to have an immediate *What You See Is What You Get*-feedback when a semantic shape attribute is changed. In practice, the user decides on the appropriate reshaping parameters by inspecting a single frame of video (typically the first one) in our interface. Fig. 7 shows a variety of attribute modifications on the same actor. Once the user is satisfied with the new shape, the warping procedure for the entire sequence is started with a click of a button.

6 Results

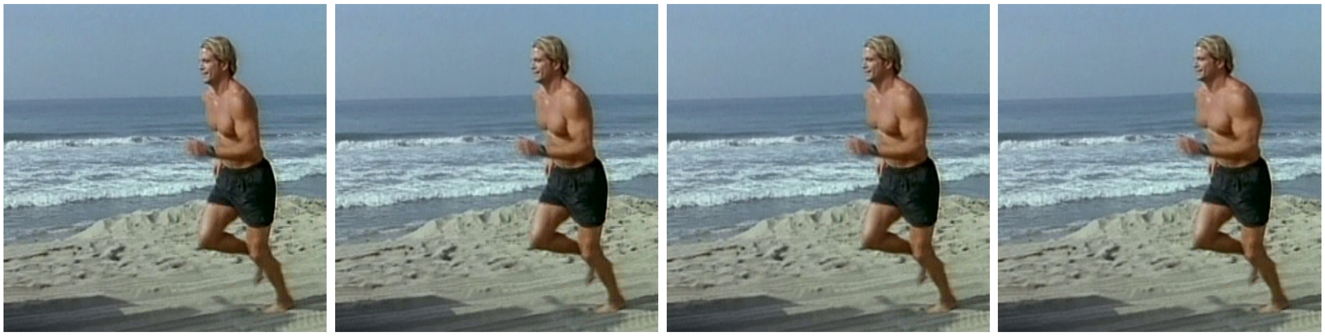
We performed a wide variety of shape edits on actors from three different video sequences: **1)** a monocular sequence from the TV series Baywatch showing a man jogging on the beach (DVD quality, resolution: 720 × 576, 25 fps, duration 7 s), Fig. 1; **2)** a monocular sequence showing a male basketball player (resolution: 1920 × 1080, 50 fps, duration 8 s), Fig. 9; **3)** a multi-view video sequence kindly provided by the University of Surrey[3] showing a female actor walking/sitting down in a studio (8 HD video cameras, 25 fps, blue screen background, duration 5 s), Fig. 7.

The sequences thus cover a wide range of motions, camera angles, picture formats, and real and synthetic backgrounds. The multi-view video sequence was tracked fully-automatically. In the monocular sequences, on average 1 in 39 frames needed manual user intervention, for instance the specification of some additional locations to be tracked. In neither case more than 5 minutes of user interaction were necessary. In the single-view sequences, the actor is segmented from the background using off-the-shelf tools, which takes on average 20 s per frame. All camera views in the multi-view sequence are chroma-keyed automatically.

The result figures, as well as the accompanying video show that we are able to perform a large range of semantically guided body reshaping operations on video data of many different formats that are typical in movie and video production. Fig. 7 illustrates nicely the effect of the modification of individual shape attributes of the same individual. In all cases, the resulting edits are highly realistic. In the Baywatch sequence in Fig. 1 we increased the muscularity of the actor by a significant amount. The final result looks highly convincing and consistent throughout the sequence. Fig. 8 shows that gradual changes of the muscularity can be easily achieved. Fig. 9 shows a basketball player filmed from a lateral angle. Our modification of the actor's waist girth looks very natural throughout the sequence, even for extreme edits that already lie beyond shape variations observed in reality. Overall, the modified actors look highly plausible and it is extremely hard to unveil them as video retouching results. Note that our edits are not only consistent over time, but also perspectively correct. Without an underlying 3D model such results would be hard to achieve.

Our results on the multi-view data (Fig. 7 and supplemental video) illustrate that the system is also useful when applied to footage that has been captured under very controlled studio conditions. For instance, if scene compositing is the goal, an actor can be captured on set from a variety of pre-planned camera positions in front of a blue screen. Now, with our system the shape of the actor can be arbitrarily modified in any of the camera views, such that the director can decide during compositing if any shape edit is necessary. As

[3] `http://kahlan.eps.surrey.ac.uk/i3dpost_action/`



original | muscularity +10% | muscularity +20% | muscularity +30%

Figure 8: *Gradual increase of the muscularity of the Baywatch actor from his original shape (shown at the left).*

an additional benefit, on multi-view data no manual intervention is needed, except the user input defining the edit. The accompanying video shows a few examples of combined shape editing and compositing with a rendered backdrop.

Using an unoptimized implementation on an Intel Core 2 Duo CPU, @3.0 GHz it takes around 9 s per frame to track the pose of the actor in a monocular sequence, and 22 s to do the same in the multi-view case. Note that tracking is only performed once for each sequence. In our reshaping tool, shape attributes can be modified in real-time, with immediate visual feedback given for the initial frame of the video. Generating the video with the new shape parameters, i.e., applying image-based warping to the entire video, takes approx. 20 ms per frame.

6.1 User Study

We evaluated our system in a user study. The goal of the study was to find out if small artifacts that may be introduced by our algorithm are noticeable by a human observer. We presented 30 participants the *Baywatch* video (shown in Fig. 1 and in the supplemental video). Half of the participants were shown the original video and were asked to rate the amount of visible artifacts. The other half was shown our modified video, where the running man is rendered more muscular, and were asked the same question. The participants rated the amount of visible artifacts on a 7-point Likert scale, where 1 means no artifacts and 7 very disturbing artifacts. The first group, which watched the original video, rated the amount of visible artifacts on average with 2.733 ± 1.22, where $\pm$ denotes the standard deviation. Our modified video received only a slightly worse rating of 2.866 ± 1.414. This may indicate that slight artifacts are introduced by our method. We validated this assumption with a two-way analysis of variance (ANOVA). The null hypothesis that the means of the two groups are equal does results in a very high p-value of 0.709 and, consequently, such a null hypothesis should not be rejected. This leads us to the conclusion that the amount of artifacts introduced by our method is very low and, thus, the anova analysis does not show a significant effect to reject such a null hypothesis in our experiment (on the other hand, this does not show that such a null hypothesis is true and we have proven that there are no artifacts introduced by our method).

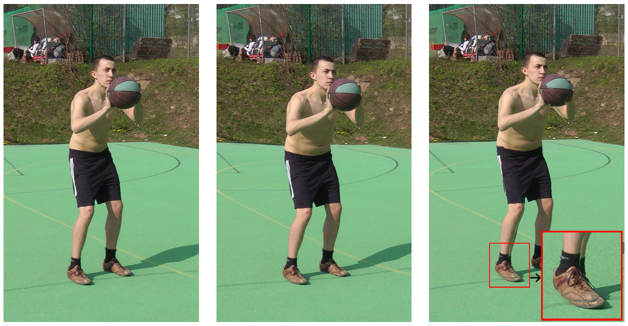
We then showed all 30 participants a side-by-side comparison of the original and the modified video and asked them if they could spot the difference. 28 out of 30 participants realized that we have made the running man more muscular, and only two participants thought that we changed something in the background. This indicates that our system is capable of achieving a noticeable reshaping result without introducing significant artifacts.

7 Discussion

We demonstrated that our approach can modify the body shape of actors in videos extremely realistically.

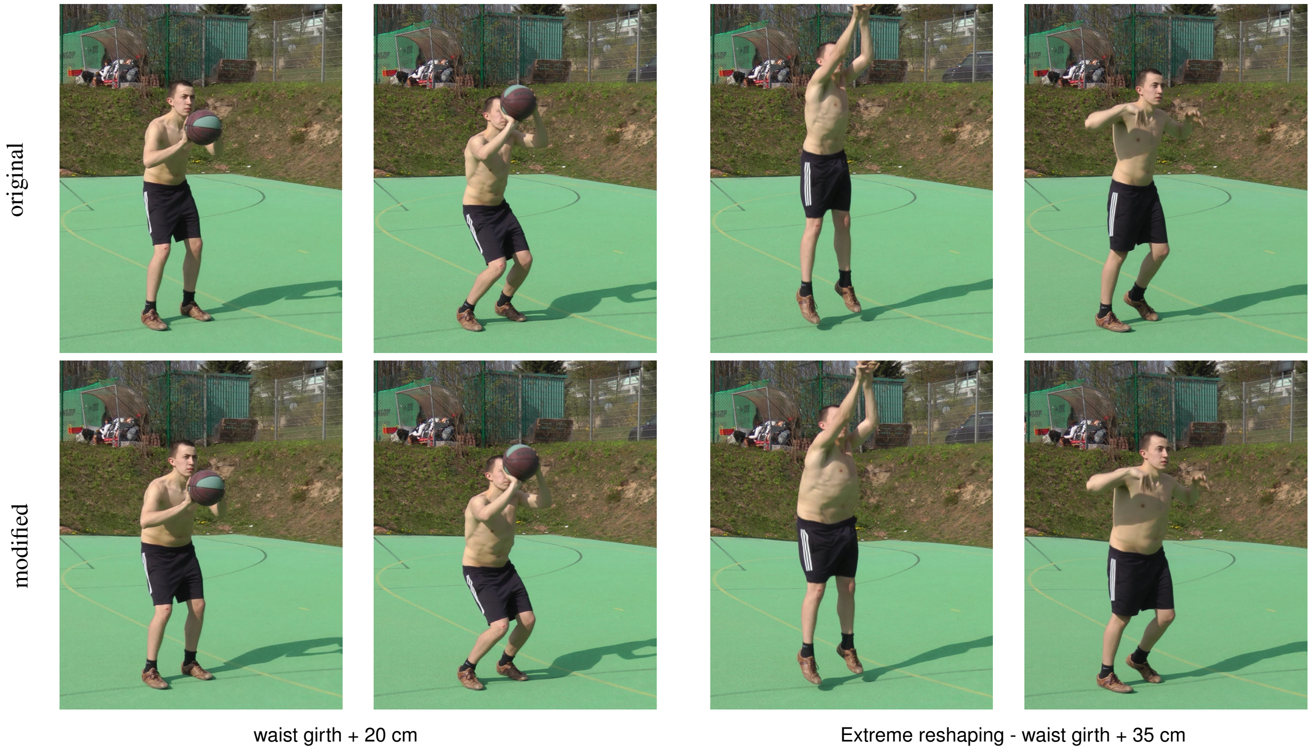
Pixel-accurate tracking is hard to achieve, especially in monocular sequences. Therefore, we refrain from using a 3D model, which could be textured with the original video frame, for rendering the reshaped human. This would inevitably lead to noticeable artifacts. In contrast, our 2D image deformation that is guided by the 3D model is robust against small tracking errors and still produces perspectively correct warps.

Nonetheless, our approach is subject to a few limitations. If the pose tracking was sub-optimal, deformation constraints may be placed very close to or in the scene background. In this case, the image deformation applied to the actor may propagate into the background leading to a halo-like warp. When the person's shape is extremely enlarged, distortions may become noticeable in the background (Fig. 10). Similarly, when the person's apparent size is strongly reduced, the background is warped to fill the whole, whereas another option would be a spatio-temporal inpainting of the disocclusions. However, as confirmed in the user study, we



(a) | (b) | (c)

Figure 10: *MLS-based image warping compared to segmentation-based deformation. (a) Original Image, (b) Deformation using MLS-based image warping. One can notice slight artifacts in the background when the human deformation is too strong, e.g. the straight edge of the basket ball court appears curved. (c) Covering the background with the modified image of the segmented human often produces more objectionable artifacts, such as a double arm, double legs or shoes.*



waist girth + 20 cm

Extreme reshaping - waist girth + 35 cm

Figure 9: *Change of waist girth of a basketball player recorded with a single video camera - on the left, the waist girth was increased moderately; on the right the waist girth was increased way beyond a natural range, but still the deformation looks coherent and plausible.*

found out that for a normal range of edits, these effects are hardly noticeable. In future, we plan to include inpainting functionality and apply a more advanced contour tracking and automatic segmentation approach. Fig. 10(c) shows an example, where the shape manipulation enlarges the silhouette of the person. In that case it would be feasible to segment the person in the foreground, deform it, and overlay it with the original frame. This way, background distortions could be prevented. However, this alternative method may lead to even more objectionable artifacts, in particular if the segmentation is not accurate since the model boundary did not exactly coincide with the person's silhouette. As a consequence, we currently always employ MLS-based global image warping.

Another problematic situation arises when limbs are occluding other parts of the body. In this case the deformation of the occluded body part is also applied to the limbs, which is an undesired artifact. In practice the effect is not very noticeable for shape modifications in a normal range.

While our system works for people dressed in normal apparel, our approach might face difficulties when people wear very wide clothing, such as a wavy skirt or long coat. In such cases, automatic pose tracking would fail. In addition, our warping scheme may not lead to plausible reshaping results that reflect the expected deformation of wide apparel. Also, shape edits often leads to corresponding changes in skeletal dimensions. When editing a video, this might make motion retargeting necessary in order to preserve a natural motion (e.g. to prevent foot skating). However, for most attribute dimensions this plays no strong role and even a modification of the leg length of an actor within certain bounds does not lead to noticeable gait errors.

Finally, our approach is currently not fully automatic. For segmentation of monocular video we heavily rely on commercial tools that may require manual intervention. However, we believe that the amount of user interaction required in order to make ill-posed monocular tracking feasible is acceptable, given the ability to perform previously unseen shape edits in videos.

8 Conclusion

We have presented *MovieReshape*, a system to perform realistic spatio-temporal reshaping of human actors in video sequences. Our approach is based on a statistical model of human shape and pose which is tracked to follow the motion of the actor. Spatio-temporally coherent shape edits can be performed efficiently by simply modifying a set of semantically meaningful shape attributes. We have demonstrated the high visual quality of our results on a variety of video sequences of different formats and origins, and validated our approach in a user study. Our system paves the trail for previously unseen post-processing applications in movie and video productions.

Acknowledgements

We would like thank Betafilm and FremantleMedia Ltd for the permission to use the Baywatch footage. We also thank the University of Surrey [Gkalelis et al. 2009] as well as [Hasler et al. 2009] for making their data available.

References

AGARWAL, A., AND TRIGGS, B. 2006. Recovering 3d human pose from monocular images. *IEEE Trans. PAMI 28*, 1, 44–58.

ALLEN, B., CURLESS, B., AND POPOVIĆ, Z. 2003. The space of human body shapes: reconstruction and parameterization from range scans. In *Proc. ACM SIGGRAPH '03*, 587–594.

ALLEN, B., CURLESS, B., POPOVIĆ, Z., AND HERTZMANN, A. 2006. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proc. SCA*, 147–156.

ANGUELOV, D., SRINIVASAN, P., KOLLER, D., THRUN, S., RODGERS, J., AND DAVIS, J. 2005. SCAPE: Shape completion and animation of people. In *ACM TOG (Proc. SIGGRAPH '05)*.

BARRETT, W. A., AND CHENEY, A. S. 2002. Object-based image editing. In *Proc. ACM SIGGRAPH '02*, ACM, 777–784.

BENNETT, E. P., AND MCMILLAN, L. 2003. Proscenium: a framework for spatio-temporal video editing. In *Proc. ACM MULTIMEDIA '03*, 177–184.

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

DAVIS, J., AGRAWALA, M., CHUANG, E., POPOVIĆ, Z., AND SALESIN, D. 2003. A sketching interface for articulated figure animation. In *Proc. SCA*, 320–328.

DE AGUIAR, E., STOLL, C., THEOBALT, C., AHMED, N., SEIDEL, H.-P., AND THRUN, S. 2008. Performance capture from sparse multi-view video. In *ACM TOG (Proc. SIGGRAPH '08)*.

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

GKALELIS, N., KIM, H., HILTON, A., NIKOLAIDIS, N., AND PITAS, I. 2009. The i3dpost multi-view and 3d human action/interaction database. In *Proc. CVMP 2009*.

GUAN, P., WEISS, A., BĂLAN, A. O., AND BLACK, M. J. 2009. Estimating human shape and pose from a single image. In *Proc. IEEE ICCV*.

HASLER, N., STOLL, C., SUNKEL, M., ROSENHAHN, B., AND SEIDEL, H.-P. 2009. A statistical model of human pose and body shape. In *CGF (Proc. Eurographics 2008)*, vol. 2.

HASLER, N., ACKERMANN, H., ROSENHAHN, B., THORMÄHLEN, T., AND SEIDEL, H.-P. 2010. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Proc. IEEE CVPR*.

HORNUNG, A., DEKKERS, E., AND KOBBELT, L. 2007. Character animation from 2d pictures and 3d motion data. *ACM TOG 26*, 1, 1.

KRÄHENBÜHL, P., LANG, M., HORNUNG, A., AND GROSS, M. 2009. A system for retargeting of streaming video. In *Proc. ACM SIGGRAPH Asia '09*, 1–10.

LEYVAND, T., COHEN-OR, D., DROR, G., AND LISCHINSKI, D. 2008. Data-driven enhancement of facial attractiveness. *ACM TOG 27*, 3, 1–9.

LI, Y., SUN, J., AND SHUM, H.-Y. 2005. Video object cut and paste. *ACM TOG 24*, 3, 595–600.

LIU, C., TORRALBA, A., FREEMAN, W. T., DURAND, F., AND ADELSON, E. H. 2005. Motion magnification. In *Proc. ACM SIGGRAPH '05*, 519–526.

MÜLLER, M., HEIDELBERGER, B., TESCHNER, M., AND GROSS, M. 2005. Meshless deformations based on shape matching. *ACM TOG 24*, 3, 471–478.

PARAMESWARAN, V., AND CHELLAPPA, R. 2004. View independent human body pose estimation from a single perspective image. In *Proc. IEEE CVPR*, II: 16–22.

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

RITSCHEL, T., OKABE, M., THORMÄHLEN, T., AND SEIDEL, H.-P. 2009. Interactive reflection editing. *ACM TOG (Proc. SIGGRAPH Asia '09) 28*, 5.

ROSALES, R., AND SCLAROFF, S. 2006. Combining generative and discriminative models in a framework for articulated pose estimation. *Int. J. Comput. Vision 67*, 3, 251–276.

RUBINSTEIN, M., SHAMIR, A., AND AVIDAN, S. 2008. Improved seam carving for video retargeting. *ACM TOG (Proc. SIGGRAPH '08) 27*, 3, 1–9.

SCHAEFER, S., MCPHAIL, T., AND WARREN, J. 2006. Image deformation using moving least squares. *ACM TOG 25*, 3, 533–540.

SCHOLZ, V., AND MAGNOR, M. 2006. Texture replacement of garments in monocular video sequences. In *Proc. EGSR*, 305–312.

SCHOLZ, V., EL-ABED, S., SEIDEL, H.-P., AND MAGNOR, M. A. 2009. Editing object behaviour in video sequences. *CGF 28*, 6, 1632–1643.

SEO, H., AND MAGNENAT-THALMANN, N. 2004. An example-based approach to human body manipulation. *Graph. Models 66*, 1, 1–23.

SIGAL, L., BALAN, A. O., AND BLACK, M. J. 2007. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Proc. NIPS*.

STOLFI, J. 1991. *Oriented Projective Geometry: A Framework for Geometric Computation.* Academic Press.

VLASIC, D., BRAND, M., PFISTER, H., AND POPOVIĆ, J. 2005. Face transfer with multilinear models. *ACM TOG 24*, 3, 426–433.

VLASIC, D., BARAN, I., MATUSIK, W., AND POPOVIĆ, J. 2008. Articulated mesh animation from multi-view silhouettes. *ACM TOG (Proc. SIGGRAPH '08)*.

WANG, J., BHAT, P., COLBURN, R. A., AGRAWALA, M., AND COHEN, M. F. 2005. Interactive video cutout. In *Proc. ACM SIGGRAPH '05*, ACM, 585–594.

WANG, J., DRUCKER, S. M., AGRAWALA, M., AND COHEN, M. F. 2006. The cartoon animation filter. *ACM TOG (Proc. SIGGRAPH '06)*, 1169–1173.

WANG, H., XU, N., RASKAR, R., AND AHUJA, N. 2007. Videoshop: A new framework for spatio-temporal video editing in gradient domain. *Graph. Models 69*, 1, 57–70.

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

ZHOU, S., FU, H., LIU, L., COHEN-OR, D., AND HAN, X. 2010. Parametric reshaping of human bodies in images. *ACM TOG (Proc. SIGGRAPH '10) 29*, 4.