Performance Capture from Multi-View Video

Christian Theobalt, Edilson de Aguiar, Carsten Stoll, Hans-Peter Seidel, and Sebastian Thrun

Abstract Nowadays, increasing performance of computing hardware makes it feasible to simulate ever more realistic humans even in real-time applications for the end-user. To fully capitalize on these computational resources, all aspects of the human, including textural appearance and lighting, and, most importantly, dynamic shape and motion have to be simulated at high fidelity in order to convey the impression of a realistic human being. In consequence, the increase in computing power is flanked by increasing requirements to the skills of the animators. In this chapter, we describe several recently developed performance capture techniques that enable animators to measure detailed animations from real world subjects recorded on multi-view video. In contrast to classical motion capture, performance capture approaches don't only measure motion parameters without the use of optical markers, but also measure detailed spatio-temporally coherent dynamic geometry and surface texture of a performing subject. This chapter gives an overview of recent state-of-the-art performance capture approaches from the literature. The core of the chapter describes a new mesh-based performance capture algorithm that uses a combination of deformable surface and volume models for high-quality reconstruction of people in general apparel, i.e. also wide dresses and skirts. The chapter concludes with a discussion of the different approaches, pointers to additional literature and a brief outline of open research questions for the future.

1 Introduction

Today, photo-realistically rendered virtual humans are becoming ever more important elements of feature films. They can perform almost any type of action or stunt at no risk of fatality, as long as an animator is capable of creating the desired effect.

C. Theobalt (✉), E. de Aguiar, C. Stoll, and H.-P. Seidel
MPI Informatik, Saarbruecken, Germany
e-mail: theobalt@mpii.de, theobalt@cs.stanford.edu,edeaguia@mpi-inf.mpg.de,stoll@mpi-inf.mpg.de,hpseidel@mpi-inf.mpg.de

C. Theobalt, S. Thrun
Stanford University, Stanford, CA, USA
e-mail: thrun@stanford.edu

R. Ronfard and G. Taubin (eds.), *Image and Geometry Processing for 3-D Cinematography*, Geometry and Computing 5, DOI 10.1007/978-3-642-12392-4_6,

In recent years, ever more powerful computing hardware and rendering algorithms have made it feasible to display detailed realistic humans not only in big-budget feature films, but even in real-time applications available to the end-user at home. For instance, it is foreseeable that in the near future computer game engines will be able to display characters with detailed texture and dynamic geometry, such as correctly deforming cloth. Another application that will gain increasing importance is 3-D video, a new form of media where either the user or the broadcasting company can instantaneously change the viewpoint on a displayed scene. In both cases, it will be important to be able to capture detailed time-varying 3-D models of humans.

Unfortunately, currently available acquisition technology frequently falls short of capturing such rich 3-D scene descriptions that would be directly applicable in the application scenarios mentioned above. Motion capture systems have been around for many years, but they are merely able to measure skeletal motion under controlled conditions. Currently, they are unable to capture shape, motion and appearance of actors in general everyday apparel. Image-based rendering techniques have been proposed to create novel view points of scenes by computationally combining views taken from a few input video streams. However, as we will see later in this chapter, many of these approaches fail to fulfill the visual quality requirements that most professional productions have.

This chapter therefore describes a new category of algorithms, performance capture methods, which are able to fulfill these requirements. Performance capture methods retrieve highly-detailed dynamic shape and motion of moving subjects from (usually) only a handful of unmodified video recordings, i.e. actively placed visual markers are not required. In contrast to previous methods from the literature they are able to handle people in general everyday apparel, such as a skirt or a dress. Also, they are able to capture spatio-temporally coherent geometry, a characteristic that sets them apart from many previous methods from the literature, in particular image-based rendering approaches. Spatio-temporal coherence is an important feature since only if correspondences between reconstructed poses over time are known it is easy to post-process, store and modify the captured data.

In the following chapter, we will first review general related work from the fields of motion capture and image-based rendering. Thereafter, we will discuss four representative, but conceptually different performance capture methods. The first method retrieves detailed time-varying geometry of pieces of apparel from multi-view video using a combination of stereo and cross-parameterization. Along a similar line of thinking, the second approach described employs a combination of visual hulls, multi-view stereo and spatio-temporal cross-parameterization to reconstruct complete performances of humans. The third method described differs from these two approaches in that it employs a template model and skeleton-based pose-fitting to visual hull sequences to measure full human performances. The core of the chapter is a new performance capture approach that takes an unconventional, yet very effective alternative route. Instead of relying on a classical skeleton-based representation of humans, it exploits deforming meshes to faithfully capture the dynamic appearance of actors in arbitrary general apparel. The paper concludes with a discussion and some pointers to additional reading.

2 Paving the Way for Performance Capture: Motion Capture, Image-Based Rendering and 3-D Video Approaches

Modern Performance Capture algorithms can capitalize on a body of related methods which focused on solving sub-problems of the overall performance capture problem. In the following we give a brief overview of important categories of such techniques.

Marker-based optical motion capture systems are the workhorses in many game and movie production companies for measuring motion of real performers [29]. Despite their high accuracy, their very restrictive capturing conditions (that often require the subjects to wear skin-tight body suits and reflective markings) make them incapable of capturing shape and texture simultaneously with motion. Park et al. [35] try to overcome part of this limitation by using several hundred markers to extract a model of human skin deformation. While their animation results are very convincing, manual mark-up and data cleanup times can be tremendous in such a setting and generalization to normally dressed subjects is difficult. In contrast, marker-free performance capture algorithm require a lot less setup time and enable *simultaneous* capture of shape, motion and texture of people wearing everyday apparel.

Marker-less motion capture approaches are designed to overcome some restrictions of marker-based techniques and enable performance recording without optical scene modification [31,39]. Although they are more flexible than intrusive methods, it remains difficult for them to achieve the same level of accuracy and the same application range. Furthermore, since most approaches employ kinematic body models, it is hard for them to capture motion, let alone detailed shape, of people in loose everyday apparel. Some methods, such as [42] and [4] try to capture more detailed body deformations in addition to skeletal joint parameters by adapting the models closer to the observed silhouettes, or by using captured range scan data [2]. But both algorithms require the subjects to wear tight clothes. Only few approaches, such as the work by [40], aim at capturing humans wearing more general attire, e.g. by jointly relying on kinematic body and cloth models. Unfortunately, these methods typically require hand-crafting of shape and dynamics for each individual piece of apparel. Also, they focus on joint parameter estimation under occlusion rather than accurate geometry capture, and therefore the shape quality of the captured performers is typically very crude.

Other related work explicitly reconstructs highly-accurate geometry of moving cloth from video [43, 56]. However, these methods require visual interference with the scene in the form of specially tailored color patterns on each piece of garment which renders simultaneous shape and texture acquisition infeasible.

A slightly more application-driven concept related to performance capture is put forward by *3-D video* methods which aim at rendering the appearance of reconstructed real-world scenes from new synthetic camera views never seen by any real camera. Early shape-from-silhouette methods reconstruct rather coarse approximate 3-D video geometry by intersecting multi-view silhouette cones [19, 28].

Despite their computational efficiency, the moderate quality of the textured coarse scene reconstructions often falls short of production standards in the movie and game industry. To boost 3-D video quality, researchers experimented with image-based methods [52], multi-view stereo [61], multi-view stereo with active illumination [55], or model-based free-viewpoint video capture [10]. In contrast to performance capture approaches, the first three methods do not deliver spatio-temporally coherent geometry or full 360 degree shape models, which are both essential prerequisites for animation post-processing. At the same time, previous kinematic model-based 3-D video methods were unable to capture performers in general clothing.

Data-driven 3-D video methods synthesize novel perspectives by a pixel-wise blending of densely sampled input viewpoints [57]. While even renderings under new lighting can be produced at high fidelity [15], the complex acquisition apparatus requiring hundreds of densely spaced cameras makes practical applications often difficult. Further on, the lack of geometry makes subsequent editing a major challenge.

3 Performance Capture Approaches

Performance capture approaches differ from the methods described in the previous section in a few key aspects. First, they aim at reconstruction of highly detailed dynamic scene geometry. By this we mean that the quality of the reconstructed shape should be of such high fidelity that it can even be used without original texture, e.g. for rendering under new artificial lighting and surface material. In consequence, even subtle aspects of shape, such as folds in attire, have to be measured at a sufficient level of detail.

Second, performance capture approaches reconstruct spatio-temporally coherent shape sequences. Here, coherence means that the correspondences between surface points over time are known. This is an important feature since it allows for simpler post-processing, editing and representation of the captured performances. Establishing these correspondences is one of the hardest problems in visual scene reconstruction. As we will see later, different strategies have been explored to achieve coherence. One class of methods uses spatio-temporal cross-parameterization techniques. Another class of approaches starts off with a detailed shape model of the performer, e.g. from a laser scan, that is then deformed to match the input multi-view video data.

Finally, performance capture approaches require no optical modification of the captured scene, e.g. in the form of intentionally placed visual markings, and they impose little restrictions on the type of apparel that a person can wear. As we will show, the majority of algorithms can even handle people in wide and wavy apparel, such as skirts or dresses. This puts them apart from the vast majority of marker-based and marker-less motion capture approaches that have been proposed up to now. In the following sections, we review a few representative examples of performance capture algorithms, and go into a slight bit more detail about a mesh-deformation-based approach that we have developed as part of our research.

3.1 Garment Capture

Capturing the motion of garment is a sub-problem of performance capture by the previously given definition. However, due to their complex deformation behavior, pieces of apparel are among the most difficult elements of dynamic scenes to be reconstructed from video. A recently presented algorithmic recipe to approach the problem shares many similarities to full performance capture, and it is therefore instructive to include it into our overview.

Most previously proposed methods for garment capture require active scene modification, e.g. in the form of color patterns printed on the captured attire (see also Sect. 2). Therefore, despite good results, they fall short to fulfill one of the main characteristics of what we call performance capture approaches in this chapter. In contrast, the recent approach by Bradley et al. captures spatio-temporally coherent geometry of moving pieces of apparel from multi-view video without any marker pattern [8]. In their method, a person wearing the piece of apparel to be reconstructed moves in front of a multi-view video camera setup. The method starts by reconstructing a 3-D mesh of the piece of garment at each time step of video by means of a multi-view stereo approach, that captures the detailed geometry of the fabric, including folds and creases, at each time step of video. Naturally, the meshes found at each time step may contain holes due to occlusions, and there is no spatio-temporal coherence in the mesh connectivity over time. To obtain a spatio-temporally coherent 3-D model representation and to fill in holes, Bradley et al. suggest a spatio-temporal cross parametrization approach that remaps the geometry from each time step to a template 3-D model. Figure 1 shows the acquisition setup, a test subject wearing a t-shirt to be reconstructed, and a 3-D mesh model of a reconstructed shirt illustrating nicely that both the overall shape as well as dynamic folds can be faithfully reconstructed.

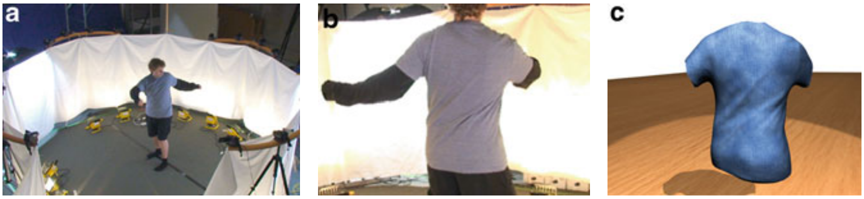

Fig. 1 Garment capture from multi-view video using the method of Bradley et al. [8]. (**a**, **b**) Input camera setup with test person. (**c**) Reconstructed 3-D mesh model of the t-shirt at the same time step. (Images courtesy of Derek Bradley, University of British Columbia, Vancouver)

3.2 Surface Capture

One of the first full performance capture approaches in the literature, by this we mean a method to capture entire humans, is the work by Starck and Hilton [48]. Input to their algorithm are eight HD video streams from a fully-calibrated camera setup. Via chroma-keying, the silhouette of the person in each frame is extracted.

In a first pass, their algorithm reconstructs an individual 3-D geometry model for each time step of multi-view video. To this end, a combination of visual hull and stereo reconstruction is used. The visual hull defines an outer boundary for the shape. By using a combination of sparse multi-view line feature matching and a graph-cut based stereo reconstruction, the very coarse visual hulls can be refined and concavities in the surfaces recovered, See Fig. 2 for an example. A surface texture for each captured pose can be created by projectively blending the input video frames on the 3-D surface.

Also here, one of the biggest challenges is to establish spatio-temporal correspondences. Similar to Bradley et al., Sect. 3.1, Starck and Hilton also use a spatio-temporal re-parameterization approach to remesh the individual triangle

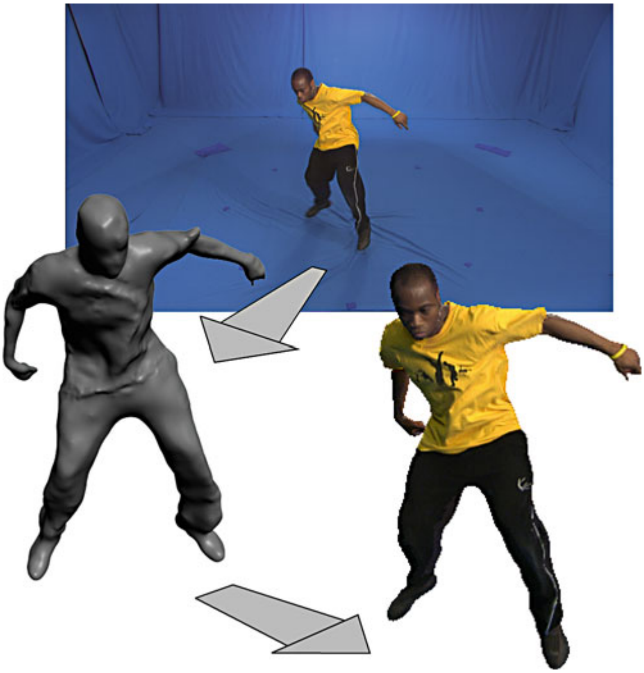

Fig. 2 Surface Capture method by Starck and Hilton [48]: 3-D models are reconstructed from multi-view video by means of a combination of shape-from-silhouette and stereo constraints. Spatio-temporal coherence in the meshes (at least for sub-sequences) is established during post-processing by means of spatio-temporal re-parametrization. (Images courtesy of Jonathan Starck, University of Surrey)

meshes from each time step to a temporally consistent triangulation [46]. In essence, they cut the surface open to achieve a genus zero surface that can then be parameterized over a sphere. On the sphere an adaptive subdivision and remeshing is performed such that eventually a mesh with the same graph structure is used to represent at least subsequences of an entire multi-view data set. Spatio-temporal reparameterization is a non-trivial problem and it is not guaranteed that under all circumstances the quality of the correspondences will be sufficient. Therefore, other researchers resorted to some form or prior model that is matched to each frame of video. This way, spatio-temporal correspondences are implicitly established.

A method similar to the one by Starck and Hilton has been proposed by Nobuhara et al. [33]. They also reconstruct shape-from-silhouette volumes and employ a deformation-based correspondence finding approach to establish spatio-temporal coherence. Their results show that they are able to successfully handle some cases of topology change.

3.3 Simultaneous Surface and Skeleton Capture

One approach that uses such a prior model is the work by Vlasic et al. [53]. Their approach also uses synchronized multi-view video sequences of human performers as input. The main conceptual difference to the previous two approaches lies in the fact that it employs a form of template model whose motion is tracked. This model comprises a surface triangle mesh and an underlying kinematic skeleton that is coupled to the surface via linear-blend skinning. The surface mesh is either reconstructed by means of a shape-from-silhouette approach, or obtained from a full-body laser scan of the person.

The algorithm commences by reconstructing a shape-from-silhouette 3-D model for each time step of video. The actual performance capture pipeline comprises two stages. In the first stage, only the skeleton part of the model is fitted into each visual hull, in order to capture the general body pose of the actor. The tracker minimizes an energy functional that drives the skeleton close to the medial axis of each visual hull, enforces temporal coherence, and ensures that the extremities of the skeleton are correctly positioned into the respective parts of the visual hulls. An additional term in the energy function takes into account user-defined position constraints that are required in difficult postures where automatic pose determination is likely to fail. To improve tracking accuracy, the authors suggest to use both a forward and a backward tracking pass.

The second stage of the pipeline deforms the surface of the template model such that the silhouettes in all camera views correctly match the outline of the reprojected model. This surface adaptation comprises of several sub-steps itself. First, the template surface is deformed into a new pose via skinning only. In general, this will not bring the mesh into agreement with the silhouettes, since, for instance, nonrigid deformations are only coarsely approximated. Further, skinning deformation artifacts may have deteriorated the surface. The authors therefore suggest an iterative deformation scheme which starts off the skinning pose of the mesh, but purposefully reduces its geometric complexity and iteratively deforms the reduced complexity meshes to match the silhouette boundaries. While iterating, high-frequency geometric surface detail of the template mesh is gradually re-introduced. To serve this purpose, a variant of Laplacian surface deformation [7] is used which allows for such gradual control of surface detail.

The final output is a sequence of skeleton poses together with the template mesh deformed in such a way that the multi-view silhouettes are matched. Overall, the visual quality of the results is very high. The algorithm has been shown to also handle sequences with more complex clothing, such as woman wearing a skirt.

The method has several advantages over the previous two approaches. It captures a rigged skeleton-based character which directly matches the animation pipeline used in most applications. It is also comparably fast, requiring only several tens of seconds of computation time per frame. This is a significant performance benefit over algorithms involving multi-view stereo. One of the disadvantages is that the tracking process of most sequences will require supervision by the user, since in difficult poses manual correction may be required. Further on, even though the quality of the recovered geometry is very convincing in general, it can naturally only capture the deformation of the surface as it is visible in the silhouette boundaries. True waving of cloth (including true folds and creases), as it is mostly observed in the interior of silhouettes, is not actually captured but in a sense pretended by the employed surface adaptation approach. In other words, high-frequency shape detail stays fixed and deforms with the underlying base surface. Here, multi-view stereo approaches are able to capture more true shape detail. Another potential problem is that the skeleton model, although it facilitates tracking, also imposes a prior on motion which is incorrect for wide pieces of apparel whose motion is not explained by skinning. Some of these problems were attacked by another template-based performance capture approach which is detailed in the following sections.

4 Mesh-Based Performance Capture

Template-based performance capture approaches bear several advantages over algorithms doing without strong a priori model assumptions. The template imposes a prior on geometry and motion which can be exploited to make scene reconstruction more robust and correspondence finding easier. The price to be paid is often measured in loss of flexibility since only scenes for which a template is easy to obtain can be reconstructed. Nonetheless, template-based approaches prove very successful for reconstructing performances of humans, as it was shown in the previous section where a kinemtic body model was used. However, a kinematic skeleton with surface skinning is obviously not the right prior model for representing wavy cloth. Although the previous method has shown that on a coarse scale cloth tracking is feasible with a kinematic prior and surface deformation, cloth tracking artifacts are likely to occur.

The method in this chapter intentionally abandons the skeleton-component of the template and uses a deformable surface model as scene representation [11]. This idea has been motivated by the fact that recently many new animation design [7], animation editing [58], deformation transfer [51] and animation capture methods [5] have been proposed that employ shape deformation approaches with great success. The explicit abandonment of kinematic parameterizations makes performance capture a much harder problem, but bears the striking advantage that it enables more reliable capturing of both rigidly and non-rigidly deforming surfaces with the same underlying technology.

First approaches that implemented this idea in the context of full body performance capture were suggested by de Aguiar et al. [12, 13]. Both approaches reconstruct a deformable human template model from a laser scan of the subject to be tracked. The mathematical deformation approach used in either case is a variant of Laplacian surface editing. Performances are retrieved by extracting features from the multi-view image streams and using their 3-D trajectories as deformation handles to change the model pose. Although these approaches can track performances of people wearing complex apparel at high reliability, the algorithms are subject to a few important limitations. Surface-based deformation represents a relatively "weak" prior on 3-D motion that may lead to erroneous deformation if the measured features are starkly noise contaminated, or if there are large regions with no deformation handle at all. This frequently happens when the motion is very fast and thus the image displacement of features is big. Therefore, rapid movements are hard to track with both these approaches. Further on, both methods share the limitation of the skeleton-based approach from the previous section that true high-frequency shape detail cannot be reconstructed.

This chapter describes a new deformation-based performance capture method that exceeds the abilities of the aforementioned algorithms in several ways. First, a new analysis-through-synthesis tracking framework enables capturing of motion that shows a very high complexity and speed. Secondly, we propose a volumetric deformation technique that greatly increases robustness of pose recovery. Finally, in contrast to previous related methods, our algorithm explicitly recovers small-scale dynamic surface detail by applying model-guided multi-view stereo.

Related to our approach are also recent animation reconstruction methods that jointly perform model generation and deformation capture from scanner data [54]. However, their problem setting is different and computationally very challenging which makes it hard for them to generate the visual quality that we achieve by resorting to an explicit prior model. The approaches proposed in [50] and [44] are able to deform mesh-models into active scanner data or visual hulls, respectively. Unfortunately, neither of these methods has shown to match our method's robustness, or the quality and detail of shape and motion data which our approach produces from video only.

4.1 Overview

Input data to our method are a full-body laser scan of the subject in its current apparel and a multi-view video stream of the subject recorded with eight synchronized geometrically and photometrically calibrated video cameras. We perform a color-based background subtraction to all video footage to yield silhouette images of the captured performers.

We convert the raw 3-D scan into a high-quality surface mesh $\mathbf{T}_{tri}$ using a robust surface reconstruction algorithm, which yields a water-tight high quality mesh. We also create a coarser tetrahedral version of the surface scan $\mathbf{T}_{tet}$ by applying a quadric error decimation and a subsequent constrained Delaunay tetrahedralization (see Fig. 3 (r)). Typically $\mathbf{T}_{tri}$ contains between 30,000 and 40,000 triangles, and the corresponding tet-version between 5,000 and 6,000 tetrahedrons. We register both models to the first pose of the actor in the input footage by means of a procedure based on iterative closest points (ICP). Since we asked the actor to strike in the first frame of video a pose similar to the one that she/he was scanned in, pose initialization is greatly simplified, as the model is already close to the target pose.

Since our capture method explicitly abandons a skeletal motion parametrization and resorts to a deformable model as scene representation, we are facing a much harder tracking problem. On the other hand we gain an intriguing advantage: we are now able to track non-rigidly deforming surfaces (like wide clothing) in the same way as rigidly deforming models and do not require prior assumptions about material distributions or the segmentation of a model.

We capture performances in a multi-resolution way to increase reliability. In a first step we employ an analysis-through-synthesis method to estimate the global pose of an actor at each frame on the basis of the tetrahedral input model, Sect. 4.3.

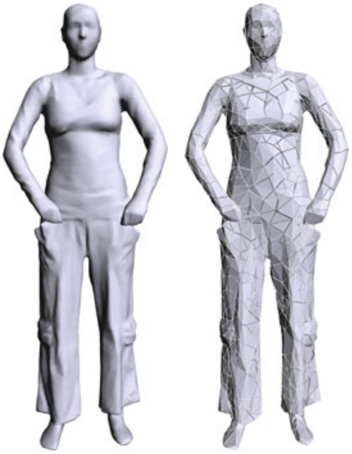

Fig. 3 A surface scan $\mathbf{T}_{tri}$ of an actress (l) and the corresponding tetrahedral mesh $\mathbf{T}_{tet}$ in an exploded view (r)

Afterwards we capture the high-frequency aspects of the performances, Sect. 4.3.4. This is achieved by transferring the pose to the high-detail surface and refining the mesh to fit closely to the input video. The output is a dense representation of the performance in both space and time. One important ingredient to achieve this is a fast and reliable shape deformation framework which we will detail in the following section.

4.2 A Deformation Toolbox

We use two variants of Laplacian shape editing in our performance capture technique. For low-frequency tracking, we use an iterative volumetric Laplacian deformation algorithm which is based on our tetrahedral mesh $\mathbf{T}_{tet}$. For recovery of high-frequency surface details, we transfer the captured pose of $\mathbf{T}_{tet}$ to the high-resolution surface scan. Being already roughly in the correct pose, we can resort to a simpler variant of surface-based Laplacian deformation to infer shape detail from silhouette and stereo constraints.

4.2.1 Volumetric Deformation

We want to deform the tetrahedral mesh $\mathbf{T}_{tet}$ as naturally as possible under the influence of a set of position constraints. To this end, we iterate a linear Laplacian deformation step and a subsequent update step, which compensates the (mainly rotational) errors introduced by the nature of the linear deformation. This procedure minimizes the amount of non-rigid deformation each tetrahedron undergoes, and thus exhibits qualities of an elastic deformation. Our technique implicitly preserves certain shape properties, such as cross-sectional areas, after deformation. This greatly increases tracking robustness since non-plausible model poses (e.g. due to local flattening) are far less likely.

Our algorithm is related to [45] and it is based on the following steps:

- Solve the linear tetrahedral Laplacian system given the current constraints
- Extract the transformation of each tetrahedral element and split it into rotational and non-rotational components
- Update the right hand side of the linear system using the extracted rotations
- Iterate the procedure

This procedure minimizes the amount of non-rigid deformation E_D remaining in each tetrahedron with each iteration. While our subsequent tracking steps would work with any physically plausible deformation or simulation method, our technique has the advantages of being extremely fast, of being very easy to implement, and of producing plausible results even if material properties are unknown. Further details on the deformation technique can be found in [49].

4.2.2 Deformation Transfer

To transfer a pose from the tetrahedral mesh $\mathbf{T}_{tet}$ to the high-resolution mesh $\mathbf{T}_{tri}$, we express the position of each vertex of $\mathbf{T}_{tri}$ as a linear combination of the vertices of the tetrahedral mesh. The coefficients for this are calculated in the rest pose and can be used afterwards to update the pose of the triangle mesh.

The coefficients are calculated as a weighted sum of the barycentric coordinates of nearby tetrahedra for each vertex. Using more than a single set of barycentric coordinates ensures that we get smooth deformations over the whole mesh. The weights for each tetrahedron are based on the respective distance from the initial vertex using a radial basis function.

4.2.3 Surface-Based Deformation

Our surface-based deformation relies on a simple least-squares Laplacian system as it has been widely used in recent years (see [7] for an overview). The linear system is calculated using cotangent weights and is used to deform the surface under the influence of a set of weighted position constraints. This simple surface based Laplacian deformation allows for a much wider and more detailed range of deformations than the tetrahedral deformation presented above.

4.3 Capturing the Global Model Pose

The first step in global model pose capture recovers for each time step of video a global pose of the tetrahedral input model that matches the pose of the real actor. In summary, our framework first computes deformation constraints from each pair of subsequent multi-view input video frames at times t and $t+1$, and then it applies the volumetric shape deformation procedure to modify the pose of $\mathbf{T}_{tet}$ at time t until it aligns with the input recorded data at time $t + 1$.

Our pose recovery process is divided into three steps and it begins with the extraction of 3-D vertex displacements from reliable image features which brings our model close to its final pose even if scene motion is rapid or complex. Subsequently, two additional steps are performed that exploit silhouette data to fully recover the global pose. The first step refines the shape of the outer model contours until they match the multi-view input silhouette boundaries and the second one optimizes 3-D displacements of key vertex handles until optimal multi-view silhouette overlap is reached. The additional steps are important since 3-D features on the model surface are dependent on scene structure, e.g. texture, and can, in general, be non-uniform or sparse.

We gain further tracking robustness by subdividing the surface of $\mathbf{T}_{tet}$ into approximately 100–200 regions of similar size during pre-processing [59]. Rather than inferring displacements for each vertex, each individual step is applied to a representative vertex handle for each region, as explained in more details in the following sections.

4.3.1 Pose Initialization from Image Features

Given two sets of multi-view video frames from subsequent time steps, we first extract image features. SIFT features are chosen since they are largely invariant under illumination and out-of-plane rotation and enable reliable correspondence finding even if the scene motion is fast or complex [27].

In order to transform the feature data into deformation constraints, we first associate image features from time t with vertices in the model. After creating the spatial feature associations across camera views, we establish temporal correspondences between the features from time t and $t + 1$, Fig. 4a. Outliers are reduced by using a robust spectral matching [25] technique.

The positions of the 3-D deformation constraints are found by calculating the pseudo-intersection point of the reprojected rays passing through the image feature locations at $t + 1$. The 3-D constraints are applied to deform $\mathbf{T}_{tet}$ using a step-wise procedure which, in practice, is unlikely to converge to implausible model configurations. We resort to the set of regions on the surface of the tet-mesh and find for each one the best handle from all candidate handles that lie in that region. If no handle is found for a region, we constrain the center of that region to its original 3-D position to prevent unconstrained surface areas from arbitrary drifting.

For each region handle, new intermediate target positions are calculated such that the corresponding vertices in $\mathbf{T}_{tet}(t)$ move in directions as similar as possible to their original normal directions. This step-wise deformation is repeated until the multi-view silhouette overlap error $SIL(\mathbf{T}_{tet}, t + 1)$ (computed as pixel-wise XOR) cannot be improved further. At the end of this step, a feature-based pose estimate $\mathbf{T}^{F}_{tet}(t + 1)$ has been obtained.

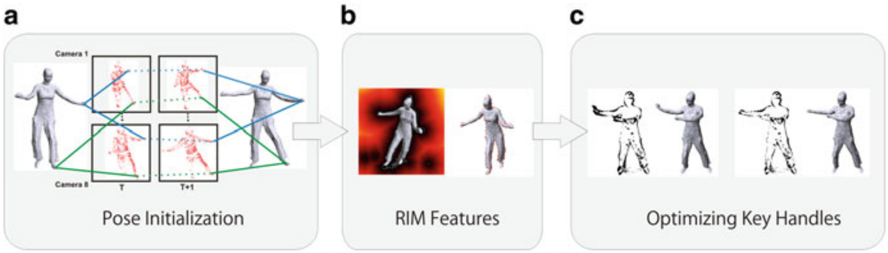

Fig. 4 (**a**) 3-D correspondences from corresponding SIFT features are used to deform the model into a first pose estimate for $t + 1$. (**b**) Color-coded distance field and rim vertices with respect to one camera view marked in red on the 3-D model. (**c**) Model and silhouette overlap after the rim step. Slight pose inaccuracies in the leg and the arms are removed and the model strikes a correct pose after key vertex optimization

4.3.2 Refining the Pose Using Silhouette Rims

In image regions with sparse or low-frequency textures, the pose of $\mathbf{T}_{tet}^{F}(t+1)$ may not be entirely correct as only few SIFT features could potentially be found. We therefore resort to another constraint that is independent of image texture and has the potential to correct for such misalignments.

We derive additional deformation constraints for a subset of vertices on $\mathbf{T}_{tet}^{F}(t+1)$ that lie on the silhouette contour. By displacing the constraints along their normals until alignment with the respective silhouette boundaries in 2D is reached, we are able to improve the pose accuracy for the model at $t+1$. The result is a new model configuration $\mathbf{T}_{tet}^{R}(t+1)$ in which the projections of the outer model contours more closely match the input silhouette boundaries, Fig. 4b.

4.3.3 Optimizing Key Handle Positions

In the majority of cases, the pose of the model in $\mathbf{T}_{tet}^{R}(t+1)$ is already close to a good match. However, in particular if the scene motion was fast or the initial pose estimate from the first step was not entirely correct, residual pose errors remain. We therefore perform an additional optimization step that corrects such residual errors by globally optimizing the positions of a subset of deformation handles until good silhouette overlap is reached, Fig. 4c.

We only optimize the position of typically 15–25 key vertices, previously selected by the user in a pre-processing step, until the tetrahedral deformation produces optimal silhouette overlap. Tracking robustness is increased by designing our energy function such that surface distances between key handles are preserved, and pose configurations with low distortion energy E_D (see Sect. 4.2.1) are preferred.

Tracking robustness is increased by preserving the distances between key handles, and by generating pose configurations with low distortion energies.

The output of this step is a new configuration of the tetrahedral model $\mathbf{T}_{tet}^{O}(t+1)$ that captures the overall stance of the model and serves as a starting point for the subsequent surface detail capture.

The above sequence of steps (Sect. 4.3.1–4.3.3) is performed for each pair of subsequent time instants. Typically the second step (silhouette rims) is performed once more after the last silhouette optimization step which, in difficult poses, leads to a better model alignment. Surface detail capture commences after the global poses for all frames were found.

4.3.4 Capturing Surface Detail

After recovering the global pose for each frame we transfer the poses of the tetrahedral mesh $\mathbf{T}_{tet}$ to the triangle mesh $\mathbf{T}_{tri}$ using the algorithm from Sect. 4.2.2. This sequence of high resolution triangle meshes will now be further refined in order to capture small-scale surface detail. We again match the reprojected model to the

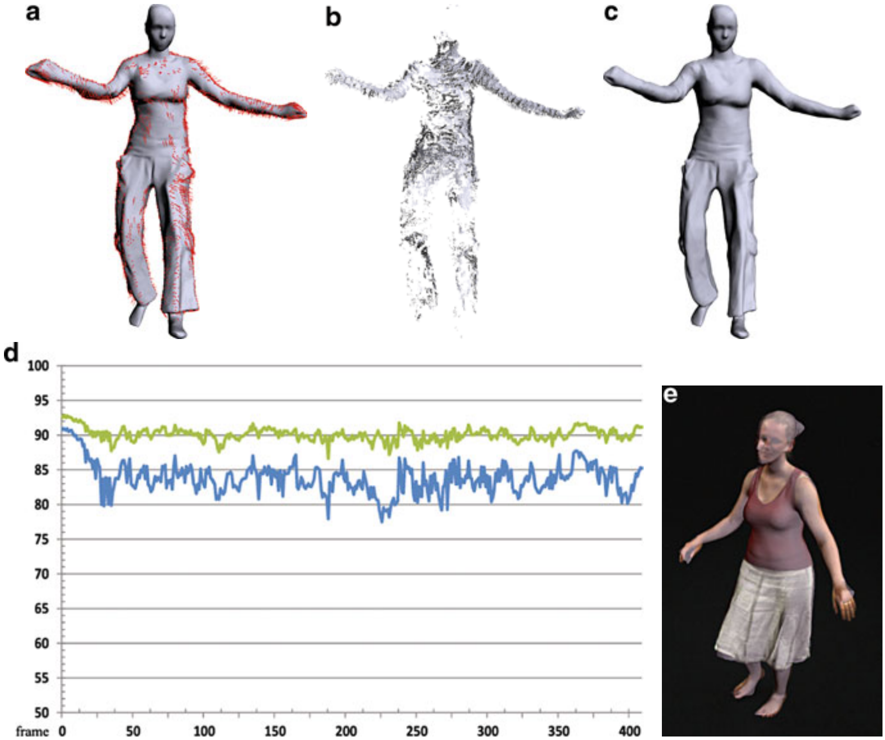

Fig. 5 **Capturing small-scale surface detail**: (**a**) First, deformation constraints from silhouette contours, shown as red arrows, are estimated. (**b**) Additional deformation handles are extracted from a 3-D point cloud that was computed via model-guided multi-view stereo. (**c**) Together, both sets of constraints deform the surface scan to a highly accurate pose. – **Evaluation**: (**d**) per-frame silhouette overlap in per cent after global pose estimation (*blue*) and after surface detail reconstruction (*green*). (**e**) Blended overlay between an input image and the reconstructed model showing the almost perfect alignment of our result

silhouette rims to better fit the input data and recover deformation detail in the interior of the silhouette with help of a multi-view stereo reconstruction algorithm. The details of the employed stereo reconstruction approach can be found in [11]. We extract position constraints from both of these cues and deform the triangle mesh using our surface Laplacian scheme from Sect. 4.2.3 to match the constraints as closely as possible. A typical set of found position constraints and the result of surface refinement are illustrated in Fig. 5a–c). After a temporal smoothing pass this yields our final output, a dense representation of the performance in both space and time matching the input video as closely as possible.

4.4 Results

The multi-view video data used in our tests comprise of 12 sequences that show four different actors and that feature between 200 and 600 frames each. To show the large application range of our algorithm, the performers wore a wide range of different apparel, ranging from tight to loose, and made of fabrics with prominent texture as well as plain colors only. Also, the recovered set of motions ranges from simple walks, over different dance styles, to fast capoeira sequences. The images in Figs. 6 and 7, as well as the results in the video that can be obtained from

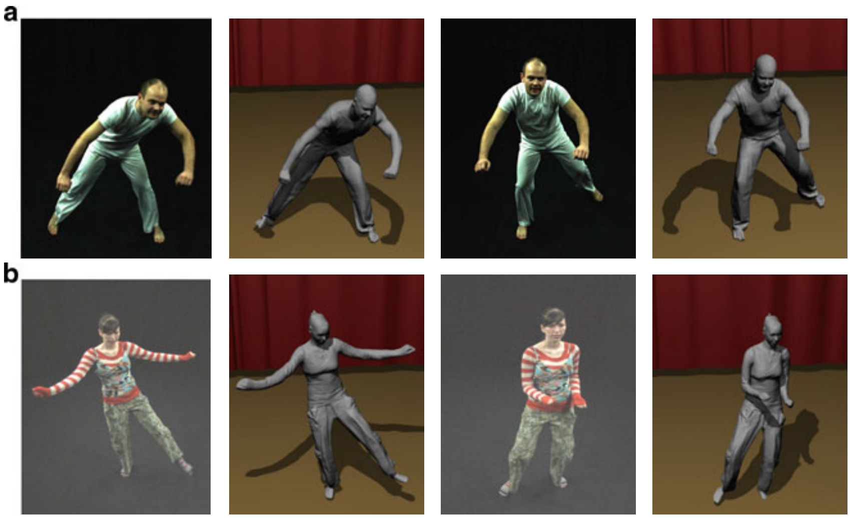

Fig. 6 (**a**) Poses from a fast capoeira performance. (**b**) Jazz dance posture with reliably captured inter-twisted arm motion (Input camera viewpoint and virtual camera viewpoint differ minimally)

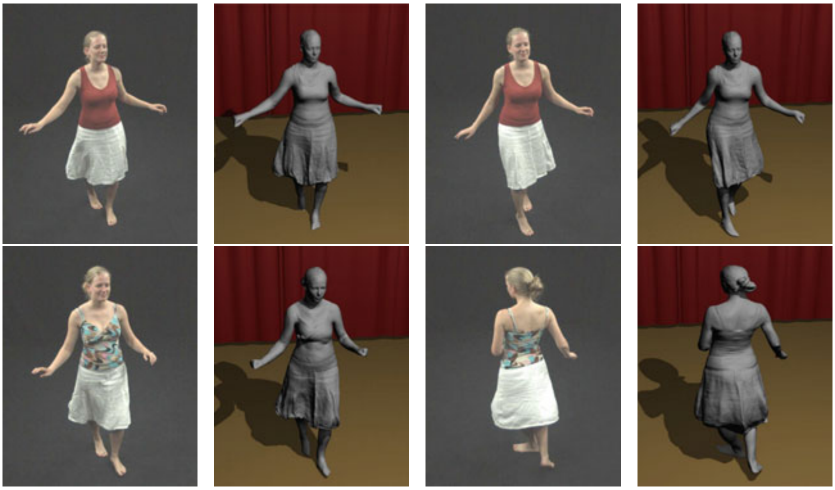

Fig. 7 Side-by-side comparison of input and reconstruction of a dancing girl wearing a skirt (input and virtual viewpoints differ minimally). Body pose and detailed geometry of the waving skirt, including lifelike folds and creases visible in the input, have been recovered

http://www.mpi-inf.mpg.de/resources/perfcap/ show that our algorithm faithfully reconstructs this wide spectrum of scenes.

Figure 6a shows two captured poses of a very rapid capoeira sequence in which the actor performs a series of turn kicks. Despite the fact that in our 24 fps recordings the actor rotates by more than 25 degrees in-between some subsequent frames, both shape and motion are reconstructed at high fidelity. The resulting animation even shows deformation details such as the waving of the trouser legs (see video). Furthermore, even with the plain white clothing that the actor wears in the input and which exhibits only few traceable SIFT features, our method performs reliably as it can capitalize on rims and silhouettes as additional sources of information.

The video also shows the captured capoeira sequence with a static checkerboard texture. This result demonstrates that temporal aliasing, such as tangential surface drift of vertex positions, is negligible, and that the overall quality of the meshes remains highly stable.

In Fig. 6b we show two poses from a captured jazz dance performance. As the comparison to the input in image and video shows, we are able to capture this fast and fluent motion. In addition, we can also reconstruct the many poses with complicated self-occlusions, such as the inter-twisted arm-motion in front of the torso.

Figure 7 shows that our algorithm is able to capture the full time-varying shape of a dancing girl wearing a skirt. Even though the skirt is of largely uniform color, our results capture the natural waving and lifelike dynamics of the fabric. In all frames, the overall body posture, and also the folds of the skirt were recovered nicely without the user specifying a segmentation of the model beforehand. We would also like to note that in all skirt sequences the benefits of the stereo step in recovering concavities are most apparent. In the other test scenes, the effects are less pronounced and we therefore deactivated the stereo step (Sect. 4.3.4) there to reduce computation time.

Apart from the scenes shown in the result images, the video contains three more capoeira sequences, two more dance sequences and two more walking sequences.

4.4.1 Validation and Discussion

Table 1 gives detailed average timings for each individual step in our algorithm obtained after code optimization of the version from [11]. These timings were obtained with a single-threaded code running on a Quad Core Intel Xeon Processor E5410 workstation with 2.33 GHz. We still see plenty of room for implementation improvement, and anticipate that parallelization can lead to significant further run time reduction.

To formally validate the accuracy of our method, we have compared the silhouette overlap of our tracked output models with the segmented input frames. We use this criterion since, to our knowledge, there is no gold-standard alternative capturing approach that would provide us with accurate time-varying 3D data. The re-projections of our final results typically overlap with over 85% of the input

Table 1 Average run times per frame for individual steps

Step	Time
SIFT step (Sect. 4.3.1)	~5 s
Global rim step (Sect. 4.3.2)	~4 s
Key handle optimization (Sect. 4.3.3)	~40 s
Capturing Surface Detail (Sect. 4.3.4)	~34 s

silhouette pixels, already after global pose capture only (blue curve in Fig. 5d). Surface detail capture further improves this overlap to more than 90% as shown by the green curve. Please note that this measure is slightly negatively biased by errors in foreground segmentation in some frames that appear as erroneous silhouette pixels. Visual inspection reveals almost perfect overlap, Fig. 5e.

All 12 input sequences were reconstructed fully-automatically after only minimal initial user input. As part of pre-processing, the user marks the head and foot regions of each model to exclude them from surface detail capture. Even slightest silhouette errors in these regions (in particular due to shadows on the floor and black hair color) would otherwise cause unnatural deformations. Furthermore, for each model the user once marks at most 25 deformation handles needed for the key handle optimization step, Sect. 4.3.3.

In individual frames of two out of three capoeira turn kick sequences (11 out of around 1,000 frames), as well as in one frame of each of the skirt sequences (2 frames from 850 frames), the output of global pose recovery showed slight misalignments in one of the limbs. Please note that, despite these isolated pose errors, the method always recovers immediately and tracks the whole sequence without drifting – this means the algorithm can run without supervision and the results can be checked afterwards. All observed pose misalignments were exclusively due to oversized silhouette areas because of either motion blur or strong shadows on the floor. Both of this could have been prevented by better adjustment of lighting and shutter speed, and more advanced segmentation schemes. In either case of global pose misalignment, at most two deformation handle positions had to be slightly adjusted by the user. In none of the over 3,500 input frames we processed it was necessary to manually correct the output of surface detail capture (Sect. 4.3.4).

Our method is subject to a few further limitations. The current silhouette rim matching may produce erroneous deformations in case the topological structure of the input silhouette is too different from the reprojected model silhouette. However, in none of our test scenes did this turn out to be an issue. In the future, we plan to investigate more sophisticated image registration approaches to solve this problem entirely. Currently, we are recording in a controlled studio environment to obtain good segmentations, but are confident that a more advanced background segmentation will enable us to handle outdoor scenes.

Moreover, there is a resolution limit to our deformation capture scheme. Some of the high-frequency detail in our final result, such as fine wrinkles in clothing or details of the face, has been part of the laser-scan in the first place. The deformation on this level of detail is not actually captured, but this fine detail is "baked in"

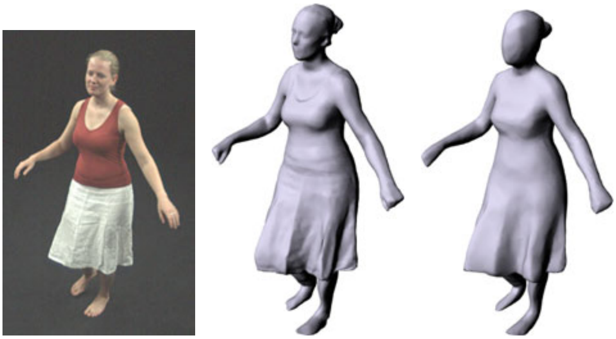

Fig. 8 Input frame (l) and reconstructions using a detailed (m) and a coarse model (r). Although the fine details on the skirt are due to the input laser scan (m), even with a coarse template, our method captures the folds and the overall lifelike motion of the cloth (r)

to the deforming surface. To illustrate the level of detail that we are actually able to reconstruct, we generated a result with a coarse scan only that lacks any fine surface detail. Figure 8 shows an input frame (l), as well as the reconstructions using the detailed scan (m) and the coarse scan (r). While, as noted before, finest detail in Fig. 8(m) is due to the high-resolution laser scan, even with a coarse scan, our method still captures the important lifelike motion and deformation of all surfaces at sufficient detail, Fig. 8(r), in particular cloth motion not visible in the silhouettes alone.

Also, since we rely on a laser scan with fixed topology, our system can currently not track sequences with arbitrarily changing apparent topology (e.g. the movement of hair or deep folds with self-collisions).

Our volume-based deformation technique essentially mimics elastic deformation, thus the geometry generated by the low-frequency tracking may in some cases have a rubbery look. For instance, an arm may not only bend at the elbow, but rather bend along its entire length. Surface detail capture eliminates such artifacts in general, and a more sophisticated yet slower finite element deformation could reduce this problem already at the global pose capture stage.

Despite these limitations, our skeleton-less method can robustly capture a large range of performances at very high detail.

5 Conclusion and Further Reading

Performance capture algorithms enable reconstruction of detailed spatio-temporally coherent scene geometry of scenes from video without having to rely on optical markers. This puts them apart from many previous approaches in the literature and opens up the perspective for many new applications. In this chapter, we presented several recent methods for video-based performance capture, and exemplified the different strategies used to represent geometry and to establish spatio-temporal coherence. The core of the chapter describes a mesh-deformation-based approach and analyzes its benefits and drawbacks in comparison to the other approaches.

The first methods reviewed did not use an a priori template model, but used either stereo, or a combination of stereo and visual hulls to reconstruct a base model to be used as scene representation. From a high-level perspective, the advantage of this strategy is that it makes a method more flexible, and many different scenes can be captured, even if a laser-scan is not available. The conceptual disadvantage is that robustness is much harder to achieve and spatio-temporal coherence is much harder to establish. The methods discussed use some clever yet often computationally expensive cross-parameterization algorithms to solve the latter problem. 3D correspondence finding is itself one of the most challenging problems in dynamic scene reconstruction. The following methods propose a few different strategies to approach this problem that we have not discussed in this chapter [1, 3, 47]. This is not a complete list of references but merely meant to give the reader a starting point.

The second class of algorithms discussed uses a stricter form of a priori shape model to capture performances of humans in general apparel. The first method from this category which we discussed employs a kinematic template model with a loosely deformable surface to retrieve human shape and motion from multi-view video. The kinematic prior greatly helps to make tracking fast and robust, and the silhouette-based surface deformation makes retrieval of cloth motion at a coarse scale feasible. Despite its benefits for tracking the human body itself, however, a skeleton (with surface skinning) generally introduces a wrong bias when tracking cloth regions of a model. The fourth approach discussed in this chapter tries to overcome some of these limitations by explicitly abandoning a skeleton model and using deformable shapes as scene representation. Additionally, the deformation-based approach described also uses a multi-view stereo method to recover true time-varying surface movement also in areas away from silhouette boundaries. This way, more time-varying surface detail than in the purely skeleton-based method can be recovered, lending the final results a more lifelike look. This is particularly visible in the tracking results of the dancer in a skirt shown previously in this chapter where true fold motion, at least at medium resolution, is apparent. The price to be paid, however, is a longer run-time compared to the skeleton-based method and the fact that a kinematic skeleton is not directly available. In our research we were able to show, however, that kinematic skeletons can automatically be learned from moving deforming surfaces which reduces the latter mentioned disadvantage [14].

Overall, this chapter as shown that performance capture techniques open up a new chapter in dynamic scene reconstruction and allow for retrieval of real world performances at such a high level of detail that new levels of quality can be expected in future video game, virtual environment and 3D video productions.

References

1. Ahmed, N., Theobalt, C., Rössl, C., Thrun, S., Seidel, H.P.: Dense correspondence finding for parametrization-free animation reconstruction from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), pp. 1–8 (2008)
2. Allen, B., Curless, B., Popović, Z.: Articulated body deformation from range scan data. ACM Trans. Graph. **21**(3), 612–619 (2002)
3. Anguelov, D., Koller, D., Srinivasan, P., Thrun, S., Pang, H.C., Davis, J.: The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces. In: Advances in Neural Information Processing Systems (NIPS 2004) (2004)
4. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: Proc. CVPR (2007)
5. Bickel, B., Botsch, M., Angst, R., Matusik, W., Otaduy, M., Pfister, H., Gross, M.: Multi-scale capture of facial geometry and motion. In: Proc. of SIGGRAPH, p. 33 (2007)
6. Botsch, M., Pauly, M., Wicke, M., Gross, M.: Adaptive space deformations based on rigid cells. Comput. Graph. Forum **26**(3), 339–347 (2007)
7. Botsch, M., Sorkine, O.: On linear variational surface deformation methods. IEEE Trans. Visual. Comput. Graph. **14**(1), 213–230 (2008)
8. Bradley, D., Popa, T., Sheffer, A., Heidrich, W., Boubekeur, T.: Markerless garment capture. In: SIGGRAPH '08: ACM SIGGRAPH 2008 Papers, pp. 1–9. ACM (2008)
9. Byrd, R., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comp. **16**(5), 1190–1208 (1995)
10. Carranza, J., Theobalt, C., Magnor, M., Seidel, H.P.: Free-viewpoint video of human actors. In: Proc. SIGGRAPH, pp. 569–577 (2003)
11. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: SIGGRAPH '08: ACM SIGGRAPH 2008 papers, pp. 1–10. ACM (2008)
12. de Aguiar, E., Theobalt, C., Stoll, C., Seidel, H.: Marker-less 3d feature tracking for mesh-based human motion capture. In: Proc. ICCV Workshop on Human Motion, pp. 1–15 (2007)
13. de Aguiar, E., Theobalt, C., Stoll, C., Seidel, H.P.: Marker-less deformable mesh tracking for human shape and motion capture. In: Proc. CVPR, pp. 1–8. IEEE (2007)
14. de Aguiar, E., Theobalt, C., Thrun, S., Seidel, H.P.: Automatic conversion of mesh animations into skeleton-based animations. Comput. Graph. Forum (Proc. Eurographics EG'08) **27**(2), 389–397 (2008)
15. Einarsson, P., Chabert, C.F., Jones, A., Ma, W.C., Lamond, B., Hawkins, T., Bolas, M., Sylwan, S., Debevec, P.: Relighting human locomotion with flowed reflectance fields. In: Proc. Eurographics Symposium on Rendering, pp. 183–194 (2006)
16. Exluna, Inc.: Entropy 3.1 Technical Reference (2002)
17. Fedkiw, R., Stam, J., Jensen, H.W.: Visual simulation of smoke. In: Fiume, E. (ed.) Proceedings of SIGGRAPH, pp. 15–22. ACM (2001)
18. Goesele, M., Curless, B., Seitz, S.M.: Multi-view stereo revisited. In: Proc. CVPR, pp. 2402–2409 (2006)
19. Gross, M., Würmlin, S., Näf, M., Lamboray, E., Spagno, C., Kunz, A., Koller-Meier, E., Svoboda, T., Gool, L.V., Lang, S., Strehlke, K., Moere, A.V., Staadt, O.: Blue-c: a spatially immersive display and 3D video portal for telepresence. ACM Trans. Graph. **22**(3), 819–827 (2003)
20. Jobson, D.J., Rahman, Z., Woodell, G.A.: Retinex image processing: improved fidelity to direct visual observation. In: Proceedings of the IS&T Fourth Color Imaging Conference: Color Science, Systems, and Applications, vol. 4, pp. 124–125 (1995)
21. Kanade, T., Rander, P., Narayanan, P.J.: Virtualized reality: constructing virtual worlds from real scenes. Proc. IEEE MultiMedia **4**(1), 34–47 (1997)
22. Kartch, D.: Efficient rendering and compression for full-parallax computer-generated holographic stereograms. Ph.D. thesis, Cornell University (2000)

23. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proc. SGP, pp. 61–70 (2006)
24. Landis, H.: Global illumination in production. In: ACM SIGGRAPH 2002 Course #16 Notes (2002)
25. Leordeanu, M., Hebert, M.: A spectral technique for correspondence problems using pairwise constraints. In: Proc. ICCV (2005)
26. Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., Shade, J., Fulk, D.: The digital michelangelo project. In: Akeley, K. (ed.) Proceedings of SIGGRAPH, pp. 131–144 (2000)
27. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. ICCV, vol. 2, p. 1150ff (1999)
28. Matusik, W., Buehler, C., Raskar, R., Gortler, S., McMillan, L.: Image-based visual hulls. In: Proc. SIGGRAPH, pp. 369–374 (2000)
29. Menache, A.: Understanding Motion Capture for Computer Animation and Video Games. Morgan Kaufmann, San Francisco (1999)
30. Mitra, N.J., Flory, S., Ovsjanikov, M., Gelfand, N., AS, L.G., Pottmann, H.: Dynamic geometry registration. In: Proc. Symposium on Geometry Processing, pp. 173–182 (2007)
31. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Understand. **104**(2), 90–126 (2006)
32. Müller, M., Dorsey, J., McMillan, L., Jagnow, R., Cutler, B.: Stable real-time deformations. In: Proc. of SCA, pp. 49–54. ACM (2002)
33. Nobuhara, S., Matsuyama, T.: Deformable mesh model for complex multi-object 3D motion estimation from multi-viewpoint video. In: 3DPVT06, pp. 264–271 (2006)
34. Paramount: Beowulf movie page. http://www.beowulfmovie.com/ (2007)
35. Park, S.I., Hodgins, J.K.: Capturing and animating skin deformation in human motion. ACM Trans. Graph. (SIGGRAPH 2006) **25**(3) (2006)
36. Park, S.W., Linsen, L., Kreylos, O., Owens, J.D., Hamann, B.: Discrete sibson interpolation. IEEE Trans. Visual. Comput. Graph. **12**(2), 243–253 (2006)
37. Parke, F.I., Waters, K.: Computer Facial Animation. A. K. Peters, Natick (1996)
38. Pellacini, F., Vidimče, K., Lefohn, A., Mohr, A., Leone, M., Warren, J.: Lpics: a hybrid hardware-accelerated relighting engine for computer cinematography. ACM Trans. Graph. **24**(3), 464–470 (2005)
39. Poppe, R.: Vision-based human motion analysis: an overview. Comput. Vis. Image Understand. **108**(1–2), 4–18 (2007)
40. Rosenhahn, B., Kersting, U., Powel, K., Seidel, H.P.: Cloth x-ray: Mocap of people wearing textiles. In: LNCS 4174: Proc. DAGM, pp. 495–504 (2006)
41. Sako, Y., Fujimura, K.: Shape similarity by homotropic deformation. Vis. Comput. **16**(1), 47–61 (2000)
42. Sand, P., McMillan, L., Popović, J.: Continuous capture of skin deformation. ACM Trans. Graph. **22**(3) (2003)
43. Scholz, V., Stich, T., Keckeisen, M., Wacker, M., Magnor, M.: Garment motion capture using color-coded patterns. Comput. Graph. Forum (Proc. Eurographics EG'05) **24**(3), 439–448 (2005)
44. Shinya, M.: Unifying measured point sequences of deforming objects. In: Proc. of 3DPVT, pp. 904–911 (2004)
45. Sorkine, O., Alexa, M.: As-rigid-as-possible surface modeling. In: Proc. SGP, pp. 109–116 (2007)
46. Starck, J., Hilton, A.: Spherical matching for temporal correspondence of non-rigid surfaces. In: IEEE Int. Conf. Computer Vision, pp. 1387–1394 (2005)
47. Starck, J., Hilton, A.: Correspondence labelling for wide-timeframe free-form surface matching. In: Proc. ICCV , pp. 1–8 (2007)
48. Starck, J., Hilton, A.: Surface capture for performance based animation. IEEE Comput. Graph. Appl. **27(3)**, 21–31 (2007)
49. Stoll, C., de Aguiar, E., Theobalt, C., Seidel, H.P.: A volumetric approach to interactive shape editing. Research Report MPI-I-2007-4-004, Max-Planck-Institut für Informatik (2007)

50. Stoll, C., Karni, Z., Rössl, C., Yamauchi, H., Seidel, H.P.: Template deformation for point cloud fitting. In: Proc. SGP, pp. 27–35 (2006)
51. Sumner, R.W., Popović, J.: Deformation transfer for triangle meshes. In: SIGGRAPH '04, pp. 399–405 (2004)
52. Vedula, S., Baker, S., Kanade, T.: Image-based spatio-temporal modeling and view interpolation of dynamic events. ACM Trans. Graph. **24**(2), 240–261 (2005)
53. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Trans. Graph. **27**(3), 1–9 (2008)
54. Wand, M., Jenke, P., Huang, Q., Bokeloh, M., Guibas, L., Schilling, A.: Reconstruction of deforming geometry from time-varying point clouds. In: Proc. SGP, pp. 49–58 (2007)
55. Waschbüsch, M., Würmlin, S., Cotting, D., Sadlo, F., Gross, M.: Scalable 3D video of dynamic scenes. In: Proc. Pacific Graphics, pp. 629–638 (2005)
56. White, R., Crane, K., Forsyth, D.: Capturing and animating occluded cloth. In: ACM TOG (Proc. SIGGRAPH) (2007)
57. Wilburn, B., Joshi, N., Vaish, V., Talvala, E., Antunez, E., Barth, A., Adams, A., Horowitz, M., Levoy, M.: High performance imaging using large camera arrays. ACM Trans. Graph. **24**(3), 765–776 (2005)
58. Xu, W., Zhou, K., Yu, Y., Tan, Q., Peng, Q., Guo, B.: Gradient domain editing of deforming mesh sequences. In: Proc. SIGGRAPH, p. 84ff. ACM (2007)
59. Yamauchi, H., Gumhold, S., Zayer, R., Seidel, H.P.: Mesh segmentation driven by gaussian curvature. Vis. Comput. **21**(8–10), 649–658 (2005)
60. Yee, Y.L.H.: Spatiotemporal sensistivity and visual attention for efficient rendering of dynamic environments. Master's thesis, Cornell University (2000)
61. Zitnick, C.L., Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High-quality video view interpolation using a layered representation. ACM Trans. Graph. **23**(3), 600–608 (2004)