Marker-less Deformable Mesh Tracking for Human Shape and Motion Capture

Edilson de Aguiar, Christian Theobalt, Carsten Stoll and Hans-Peter Seidel
MPI Informatik
Saarbrücken, Germany
{edeaguia,theobalt,stoll,seidel}@mpi-inf.mpg.de

Abstract

We present a novel algorithm to jointly capture the motion and the dynamic shape of humans from multiple video streams without using optical markers. Instead of relying on kinematic skeletons, as traditional motion capture methods, our approach uses a deformable high-quality mesh of a human as scene representation. It jointly uses an image-based 3D correspondence estimation algorithm and a fast Laplacian mesh deformation scheme to capture both motion and surface deformation of the actor from the input video footage. As opposed to many related methods, our algorithm can track people wearing wide apparel, it can straightforwardly be applied to any type of subject, e.g. animals, and it preserves the connectivity of the mesh over time. We demonstrate the performance of our approach using synthetic and captured real-world video sequences and validate its accuracy by comparison to the ground truth.

1. Introduction

3D video processing is a young and challenging field that aims at reconstructing time-varying models of real-world scenes from multi-view video in order to display them from synthetic viewpoints. The most important and most difficult to reconstruct part of these models, in particular if human actors are in the center of the scene, is the representation of the scene's geometry and its motion. A variety of approaches have been proposed in the literature that are able to partly solve this problem.

On one end of the spectrum, there are marker-based and marker-less motion capture systems that measure human motion in terms of a kinematic skeleton [18]. Since a kinematic skeleton only enables tracking of rigid motions, they have to be combined with other scanning technologies to capture the time-varying shape of the human body surface [1, 19, 26]. Unfortunately, none of these methods can perform joint dynamic shape and motion estimation of people wearing arbitrary clothing from unmodified raw video material.

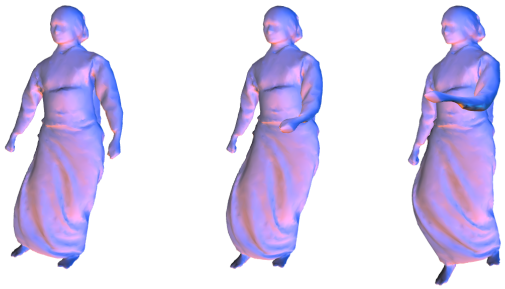

Figure 1. Our method realistically captures the motion and the dynamic shape of a woman wearing a Japanese kimono from only eight video streams.

Time-varying scene geometry can also be reconstructed by means of shape-from-silhouette approaches [9], or combined silhouette- and stereo-based methods [8]. However, the measured models often lack detail if only a handful of input camera views is available and it is hard to preserve topological correspondence between subsequent time steps.

On the other end of the spectrum, there are methods to track deformable models. Physics-based models can be applied to track garment [25], tissue in medical scanner data [16], or simple human motion if a kinematic skeleton is also available [17]. However, none of these approaches can trivially be applied to objects made of a variety of different materials, and none of them has yet tracked arbitrarily dressed humans using passive methods.

The main contribution of this paper is a method to fully-automatically track the motion and the time-varying non-rigid surface deformation of people from a handful of multi-view video streams. The algorithm can handle humans wearing arbitrary clothing, including wide t-shirts, skirts and even kimonos. It employs a high-quality laser-scan of the tracked subject as underlying scene representation. By means of an optical flow-based 3D correspondence estimation, the laser scan is deformed over time such that it follows the motion of the actor in the input streams. Deformations are computed via a new fast Laplacian tracking scheme that is robust against errors in the 3D flow. Our method does

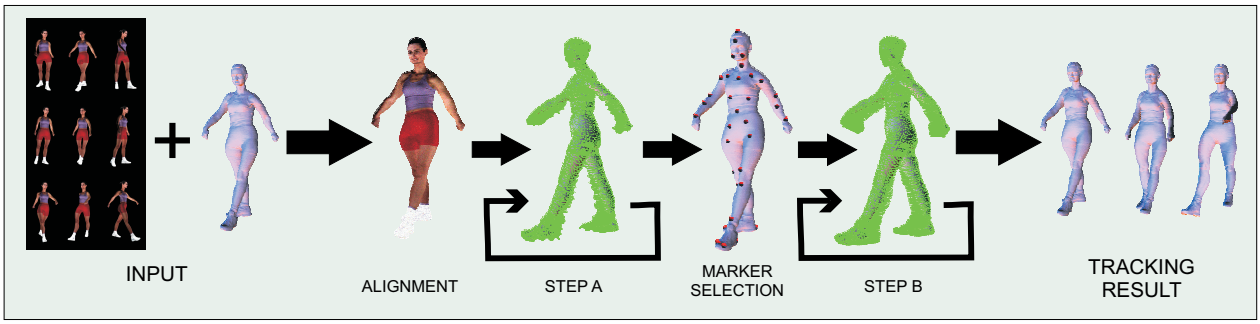

Figure 2. Overview of our marker-less deformable mesh tracking framework: Given a laser-scan of a person and a multi-view video sequence showing her motion, the method deforms the scan in the same way as its real-world counterpart in the video streams.

not employ a kinematic skeleton and it can directly be applied to track other subjects, e.g. animals. As an additional benefit, the connectivity of the model over time is trivially preserved.

The remainder of this paper is structured as follows: Sect. 2 reviews the most relevant related work. Sect. 3 details our automatic marker-less deformable mesh tracking technique. Validation experiments and results with both synthetic and captured real-world data are described in Sect. 4, and the paper concludes in Sect. 5.

2. Related Work

In our work we capitalize on ideas from research on model tracking and scene reconstruction. The following is by no means a complete list of references but merely a summary of the most related categories of approaches.

Human motion is normally measured by marker-based or marker-less optical motion capture systems [11, 18] that parameterize the data in terms of kinematic skeletons. Unfortunately, these approaches cannot directly measure time-varying body shape and they even fail to track people wearing loose apparel. To overcome this limitation, some methods use hundreds of optical markings [19] for deformation capture, combine a motion capture system with a range scanner [1] or a shape-from-silhouette approach [26], or jointly use a body and a cloth model to track the person [24]. Although achieving good results, most of these method require active interference with the scene or require hand-crafted models for each individual.

Alternatively, shape-from-silhouette algorithms [9], multi-view stereo approaches [34], or methods combining silhouette and stereo constraints [7, 8] can be used to reconstruct dynamic scene geometry. To obtain good quality results, however, many cameras are needed and it is hard for these algorithms to generate connectivity-preserving dynamic mesh models [28], such as our method produces them.

Related to our approach are also previous methods for deformable model tracking. Some passive methods extract 3D correspondences from images to track simple deformable objects [6] or cloth [23]. Similar algorithms can be used to track tissue in medical data [16]. Passive methods can also be employed to jointly capture kinematic motion parameters and surface deformations of tightly dressed humans [5, 22]. None of these algorithms, though, can track arbitrarily dressed people at a level of accuracy comparable to ours. Statistical models have also shown their potential to track confined deformable surface patches [30] and moving hands [10]. Researchers have also used physics-based shape models to track garment [25] or simple articulated humans [17, 21]. Unfortunately, none of these methods is able to track arbitrarily dressed people completely passively. It may also be difficult to apply them for tracking a human wearing different garments, since there the specification of material parameters is non-trivial.

In contrast, we propose a new method that captures high-quality deformable human geometry completely passively from only a handful of input video streams. It combines 3D flow estimation and Laplacian mesh editing [27, 29] to track the deformation of a high-quality a priori shape model which makes it robust against errors in 3D correspondence estimation. By relying on differential coordinates, shape deformations for large models can be computed at almost interactive frame rates without having to specify explicit material parameters. Our algorithm is highly flexible, easy to implement and captures both rigid motion and non-rigid surface deformation of the tracked subjects. In addition, it delivers triangle mesh geometry that maintains its connectivity over time. To the best of our knowledge, this is the first system of its kind that can capture the motion and non-rigid surface deformations of arbitrary subjects from only a handful of cameras.

3. Video-based Tracking of Scanned Humans

An overview of our technique is shown in Fig. 2. The input comprises of a static laser-scanned triangle mesh $M = (V, E)$ (V=set of vertices, E=set of edges) of the moving subject, and a multi-view video (MVV) sequence that shows the person moving arbitrarily. After data acquisition, we first align the laser scan to the pose of the person in the

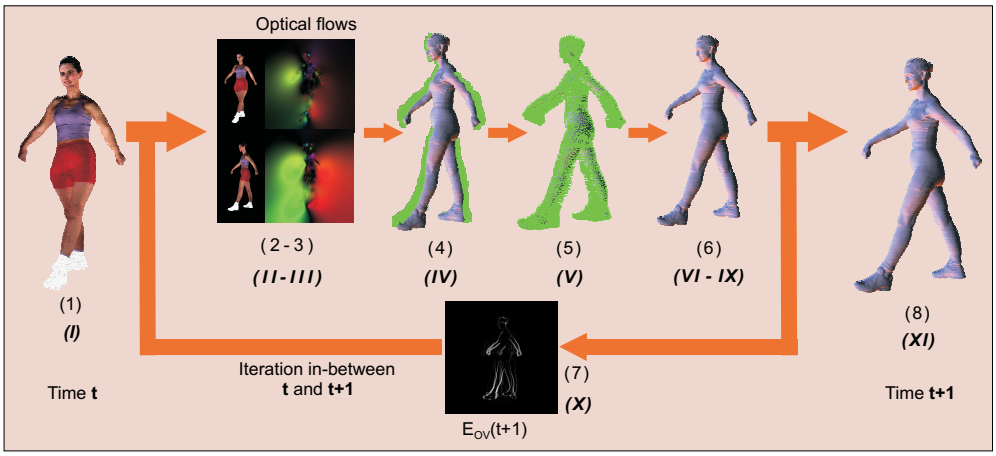

Figure 3. The workflows of tracking steps A and B are very similar: Arabic numerals indicate the workflow specific to step A as it is described in Sect. 3.2, whereas Roman numerals denote step B which is detailed in Sect. 3.4.

first time step of video, Sect. 3.1. Our framework comprises of two different tracking procedures, step A and step B, that are subsequently applied. In step A we apply an iterative 3D flow-based deformation scheme to extract the motion information of each vertex over time from the images, Sect. 3.2. The results of step A quickly deteriorate due to accumulation of correspondence estimation errors. Nonetheless, they give us the possibility to automatically identify N marker vertices that can be tracked reliably, Sect. 3.3. Tracking step B, Sect. 3.4, is more robust against flow errors since it implicitly enforces structural integrity of the underlying mesh. It uses the moving marker vertices as deformation constraints to drive a Laplacian deformation framework that makes all vertices correctly follow the motion of the actor in all video frames.

3.1. Acquisition and Initial Alignment

For each test person, we acquire a model and several MVV sequences in our studio. The triangle mesh is captured with our Vitus Smart^{TM} full body laser scanner. After scanning, the subject immediately moves to the nearby area where she is recorded with eight synchronized video cameras that run at 25 fps and provide 1004x1004 pixels frame resolution. The calibrated cameras are placed in an approximately circular arrangement around the center of the scene. After acquiring the sequence, silhouette images are calculated via color-based background subtraction.

In an initial alignment we register the scanned mesh with the pose of the person in the first time step of video. To this end, she initially strikes the same pose that she was scanned in. By means of an ICP-like registration the mesh is first coarsely aligned to a shape-from-silhouette reconstruction of the person. Thereafter, we run our flow-based Laplacian deformation scheme to correct for subtle pose differences.

3.2. Step A: Purely Flow-driven Mesh Tracking

After initial alignment, we iteratively deform each individual vertex of the mesh based on 3D optical flow fields that have been reconstructed from the multi-view images. Although this simple approach is not robust against errors in the 3D flow field, it allows us to deduce valuable motion information about certain vertices on the surface which we can capitalize on in step B. Using subsequent time steps t and $t+1$ our purely flow-driven mesh tracking approach consists of the following steps (see Fig. 3):

1. Projectively texture the model using the images $I_t^0 \cdots I_t^{K-1}$ recorded with the K cameras at time step t and blend them according to the weights described in [4]. From now on, for all deformation iterations between t and $t+1$ the texture coordinates are fixed.

2. Generate K temporary images $T_t^0 \cdots T_t^{K-1}$ by projecting the textured model back into all K camera views.

3. Calculate K 2D optical flow fields $\vec{o^k}(T_t^k, I_{t+1}^k)$ between image T_t^k and I_{t+1}^k with $k = \{0 \cdots K-1\}$.

4. Given the model, calibrated cameras and the optical flow fields for all camera views, we can compute the 3D motion field, also known as the scene flow, by solving a linear system for each vertex v_i that is visible from at least two camera views [31]. The generated 3D flow field $\vec{f}(v_i) = (x_i, y_i, z_i)$ is parameterized over the mesh's surface and it describes the displacement by which v_i should move from its current position.

5. Filter the 3D motion field $\vec{f}(\cdot)$ to remove noise and outliers. During the filtering process, the 3D flow vectors for all vertices are first classified as valid or invalid according to a silhouette-consistency criterion. $\vec{f}(\cdot)$ is valid if the position of v_i after displacement projects

inside the silhouette images for all camera views and it is invalid otherwise. Thereafter, a Gaussian low-pass kernel is applied over the entire flow field. All invalid displacements $\vec{f}(\cdot)$ are set to zero.

6. Using the filtered version of $\vec{f}(\cdot)$, update the model by moving its vertices according to the computed displacements. Add the displacements $\vec{f}(\cdot)$ to the accumulated displacement field $\vec{d}_{\text{ACCUM}}(\cdot)$ according to the rule: $\vec{d}_{\text{ACCUM}}(v_i) = \vec{d}_{\text{ACCUM}}(v_i) + \vec{f}(v_i)$. $\vec{d}_{\text{ACCUM}}(\cdot)$ describes the complete displacement of all vertices from captured time step t to the current intermediate position.

7. Iterate from step 2 until the overlap error $E_{ov}(t+1)$ between the rendered model silhouettes (see Fig. 3) and the video-image silhouettes at time $t+1$ in all camera views is below TR_{OV}. $E_{ov}(t+1)$ is efficiently implemented on the GPU as a pixel-wise XOR [4].

8. Update the complete motion field $\vec{d}(t, v_i)$, which describes the displacement of each vertex v_i from time step 0 to t, according to $\vec{d}(t, v_i) = \vec{d}(t-1, v_i) + \vec{d}_{\text{ACCUM}}(v_i)$.

The mesh is tracked over the whole sequence by applying the previously described steps to all pairs of consequent time steps. As a result, a complete motion field $\vec{d}(t, v_i)$ is generated for each vertex v_i that describes its displacement over time.

Since our scheme calculates $3D$ displacements without taking into account a priori information about the shape of the mesh, deformation errors accumulate over time. Step B solves this problem by explicitly enforcing structural properties of the mesh during tracking. To this end, the model is deformed based on constraints derived from reliably tracked marker vertices. These vertices are automatically selected from the results of step A based on the method described in the following section.

3.3. Automatic Marker Selection

Based on the deformation results of step A our approach selects N marker vertices of the model that were accurately tracked over time. To this end we first choose L candidate vertices for markers that are regularly distributed over the model's surface (Fig. 4A). To find these candidates, we segment the surface of the mesh by means of a curvature-based segmentation approach [32]. This algorithm creates surface patches with similar numbers of vertices whose boundaries do not cross important shape features. In each region the vertex located nearest to the center of gravity is selected as a candidate.

A candidate v_i is considered a marker vertex if it has a low error according to the two spatio-temporal selection criteria $tsc(\cdot)$ and $mov(\cdot)$. $tsc(\cdot)$ penalizes marker candidates that do not project into the silhouettes in all camera views and at all time steps. $mov(\cdot)$ penalizes candidates whose motions are not consistent with the average motion of all vertices in the model. This way, we can prevent the placement of constraints in surface areas for which the flow estimates might be inaccurate. The two functions are defined as follows:

$$tsc(v_i) = \frac{1}{N_F * K} \sum_{t=0}^{N_F} \sum_{k=0}^{K} (1 - PROJ_{sil}^{k}(p_i + \vec{d}(t, v_i), t)) \tag{1}$$

$$mov(v_i) = \frac{1}{N_F} \sum_{t=0}^{N_F} (\|\vec{d}(t, v_i) - \frac{1}{N_V} \sum_{j=0}^{N_V} \vec{d}(t, v_j)\|) \tag{2}$$

N_F is the number of frames in the sequence, N_V is the number of vertices in the model, and p_i is the position of v_i at the first time step. $PROJ_{sil}^{k}(x, t)$ is a function that evaluates to 1 if a 3D point x projects inside the silhouette image of camera view k at time step t, and it is 0 otherwise. A candidate v_i is accepted as a marker vertex if $tsc(v_i) < \text{TR}_{\text{TSC}}$ and $mov(v_i) < \text{TR}_{\text{MOV}}$. Appropriate thresholds TR_{TSC} and TR_{MOV} are found through experiments. The index i of each marker v_i is then stored in the set $\mathcal{Q}$.

3.4. Step B: Flow-driven Laplacian Mesh Tracking

In step B we extract rotation and translation constraints from the motion of the N marker vertices to drive a Laplacian mesh deformation approach. By this means we can extract novel motion fields $\vec{d}(t, v_i)$ for each vertex that make the model correctly move and deform like the recorded individual. Our Laplacian scheme encodes the knowledge about structural details of the model M in the mesh's differential coordinates d. They are computed by solving a linear system of the form $d = Lv$, where L is the discrete Laplace operator based on the cotangent-weights [14] and v is the vector of vertex coordinates. The individual steps of the Laplacian tracking scheme are very similar to step A (Fig. 3), but differ in the details of the deformation method. For two subsequent time steps t and $t+1$ tracking works as follows:

I-V are identical to steps 1-5 in Sect. 3.2.

VI From the motion of each marker vertex m_i=v_i, with $i \in \mathcal{Q}$, relative to the default position, a set of rotation and translation constraints is computed. We propose a new approach to automatically determine rotation constraints. Local coordinate frames for each m_i are derived from a graph connecting the markers whose structure is computed once during preprocessing. Each marker corresponds to a node in the graph.

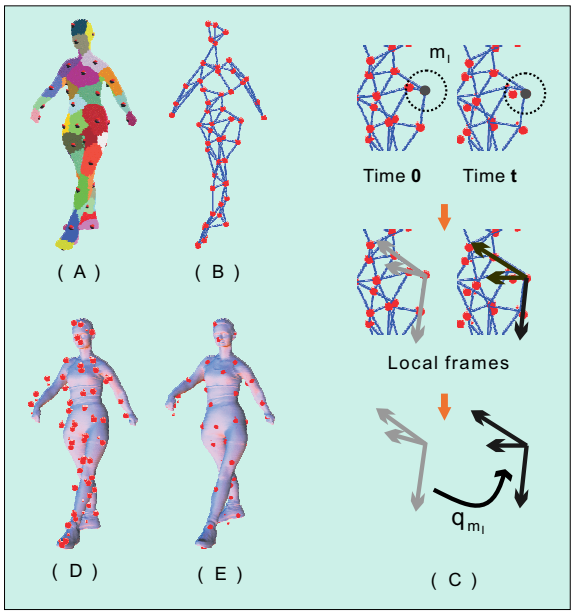

Figure 4. (A) Segmented mesh and marker candidates (Sect. 3.3); (B) Graph connecting marker vertices; (C) Rotation for each m_i calculated according to the change in its local frame from time 0 to t; (D-E) Model is reconstructed subject to constraints derived from the motion of the markers.

Edges are constructed by building a minimal spanning tree between them using geodesics as distance measure [13], Fig. 4B. For each marker m_i a local rotation is estimated from the change of its local frame between its reference orientation at $t = 0$ and its current orientation. This rotation is parameterized as a quaternion q_{m_i}, Fig. 4C.

VII The rotations for all markers m_i are interpolated over the model [33]. If each component of a quaternion $q_{m_i} = [w, q_1, q_2, q_3]$ is considered to be a scalar field defined over the entire mesh, a smooth interpolation is guaranteed by regarding these scalar fields as harmonic fields. The interpolation can be efficiently calculated by solving the Laplace equation $Lq = 0$ over the whole mesh with constraints q_{m_i} for all markers.

VIII The model M in its new target pose is reconstructed by solving the Laplace equation, subject to the constraints derived from the motion of the m_i (Fig. 4D-E). This can be formulated as a least-squares problem of the form [29]:

$$\underset{p^{REC}}{argmin}\{\|Lp^{REC} - (q \cdot d \cdot \overline{q})\|^2 + \|Ap^{REC} - b\|^2\}. \quad (3)$$

which can be transformed into a linear system

$$(L^T L + A^T A)p^{REC} = L^T(q \cdot d \cdot \overline{q}) + A^T b. \quad (4)$$

Here, $q \cdot d \cdot \overline{q}$ are the differential coordinates that have been rotated according to the interpolated rotation field q. b is the vector of positional constraints of the form $b_i = p_i + \vec{d}(t-1, m_i) + \vec{d}_{\text{ACCUM}}(m_i)$, and p^{REC} is the vector of reconstructed vertex positions. The field $\vec{d}_{\text{ACCUM}}(\cdot)$ stores the displacements for each vertex relative to time t that have been accumulated up to now during iterations of steps II to IX (see also point IX). Matrix A is a diagonal matrix containing non-zero weights $A_{i,i} = w_i$ only for markers m_i.

IX Update the accumulated displacement field for all vertices $\vec{d}_{\text{ACCUM}}(\cdot)$ according to the rule: $\vec{d}_{\text{ACCUM}}(v_i) = p_i^{REC} - p_i - \vec{d}(t-1, m_i)$, where p_i^{REC} is the reconstructed vertex position for v_i.

X Iterate from step II until the overlap error $E_{ov}(t+1)$ between rendered model silhouettes and video-image silhouettes in all cameras at $t+1$ is below a threshold TR_{OV}.

XI Update the complete motion field $\vec{d}(t, v_i)$ by $\vec{d}(t, v_i) = \vec{d}(t-1, v_i) + \vec{d}_{\text{ACCUM}}(v_i)$.

By applying this algorithm to all subsequent time steps we can track the mesh over the whole video sequence. Our Laplacian scheme reconstructs the mesh in its new pose in a way that preserves the differential surface properties of the original scan. Due to this implicit shape regularization, our tracking approach in step B is robust against inaccurate flow estimates and deforms the mesh in accordance to its real-world counterpart in the video streams.

4. Results and Discussion

We have tested our method on several synthetic and captured real-world data sets (see Sect. 3.1).

Synthetic sequences enable us to compare our results against the ground truth. They were generated by animating a textured scan of a woman ($26K\triangle$) provided by Cyberware™ (Fig. 2) with publically available motion capture files showing soccer moves and a simple walk. Output streams were rendered into eight virtual cameras (1004x1004 pixels, 25 fps) that were placed in a circular arrangement like in our real studio. Image noise was purposefully added to mimic the characteristics of our real cameras. We ran a series of experiments to evaluate the performance of different algorithmic alternatives and to decide on the best optical flow estimation scheme for our purpose.

The latter question was answered by our first experiment. To test a representative set of alternative flow algorithms, we compared the results obtained by using our complete tracking framework (steps A and B) in conjunction with the local Lukas Kanade method [15] (*LK*), the dense optical flow method by Black et al. [2] (*BA*), and the warping-based method for dense optical flow by Brox et al. [3] (*BR*). The plot in Fig. 5 shows the average position errors between

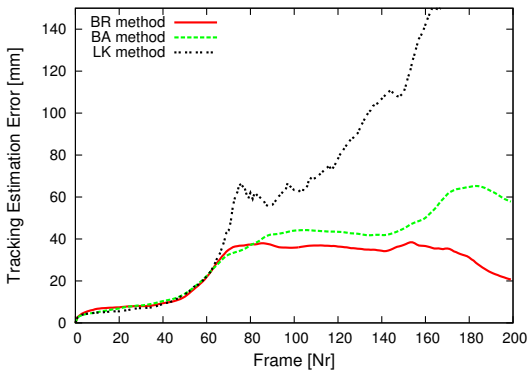

Figure 5. Three different optical flow methods have been tested with our framework. The average vertex position errors for each frame relative to the ground truth are plotted in this figure. The method by Brox et al. (red line) shows the best performance.

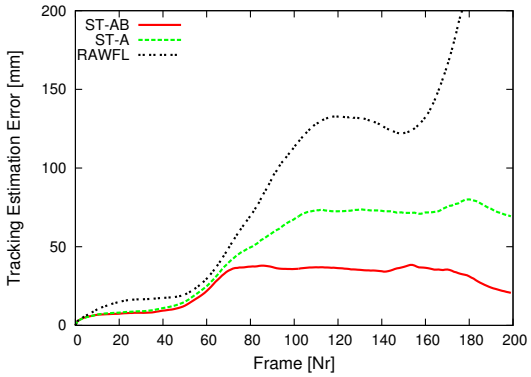

Figure 6. Average tracking error for all time steps of the synthetic walking sequence obtained with different mesh tracking alternatives. The pipeline we propose (ST-AB) clearly produces the best results.

ground truth and tracking results for each frame of a walking sequence. By using the local Lukas Kanade method, we are unable to track the mesh and the error constantly increases over time. The error plot for *BA* is much better, but it is clearly outperformed by *BR*. The positional inaccuracy obtained with *BR* never exceeds 4 cm and even decreases after a peak in the middle. Please note that the synthetic model (Fig. 2) is textured with very uniform colors which makes optical flow computation extremely hard. Even on such difficult data, *BR* tracks the mesh reliably, and thus the method by Brox et al. is our method of choice for flow estimation.

In a second experiment, we compared the different deformation alternatives, namely deformation along the unfiltered flow (*RAWFL*), deformation according to step A only (*ST-A*), and deformation with our complete pipeline (*ST-AB*). Fig. 6 plots the average vertex position error against the frame number. Using *RAWFL*, the measurement error grows almost exponentially. Tracking with a filtered flow field leads to significantly better results, but the absolute inaccuracy is still comparably high. In contrast, our complete

METHOD	TIME	VOLCHG	MQLT	ERROR
RAWFL	109s	17.65%	0.46	98.66mm
ST-A	111s	4.97%	0.30	49.39mm
BR / ST-AB	111s	2.79%	0.035	26.45mm
BA	426s	2.77%	0.029	35.28mm
LK	89s	10.73%	1.72	76.24mm

Table 1. Different algorithmic alternatives are compared in terms of run time, volume change (*VOLCHG*), mesh quality (*MQLT*), and position error (*ERROR*). Our proposed pipeline with the dense optical flow method by Brox et al. (*BR*/*ST-AB*) leads to the best results.

pipeline leads to a very low peak position error of 3.5 cm that even decreases over time.

Table 1 summarizes the results that we obtained by assessing different combinations of mesh tracking and flow computation methods according to additional criteria. The column *TIME* contains the time needed on a Pentium IV with 3GHz to compute the deformation from one video time step t to the next one $t + 1$. We also analyzed the average volume change over the whole sequence, *VOLCHG*, in order to get a numerical indicator for unreasonable deformations. The preservation of mesh quality is analyzed by looking at the average distortion of the triangles, *MQLT*. It is computed by averaging the per-triangle Frobenius norm over the mesh and over time [20]. This norm is 0 for an equilateral triangle and approaches infinity with increasing degeneracy. Finally, the column labeled *ERROR* contains the average of the position error over all vertices and time steps.

The run times of the first three alternatives are almost identical since 109 s have to be spent on the calculation of the eight megapixel optical flow fields. Even in our complete pipeline the deformation itself runs at almost interactive frame rates since the involved linear systems can be solved quickly. As expected, the tracking error is highest if one deforms the mesh using the unfiltered flow, *RAWFL*. Furthermore, the mesh distortion is fairly high and the volume change rises to implausible values. The best overall performance is achieved when we use our full mesh motion capture pipeline *ST-AB*/*BR*. Here, the position error is lowest, the volume change is in the range of normal non-rigid body deformations, and the triangles remain in nice shape. Although the alternative *BA* produces a fairly low triangle distortion, its run time is four times slower than the best alternative and the resulting positional accuracy is almost 1 cm lower. *LK* is fastest, but leads to bad results according to all other criteria. Our tests thus confirm that the complete tracking pipeline in combination with a high-quality dense flow method can reliably track human motion from raw unmodified video streams. Admittedly, we cannot track people at a one millimeter accuracy. Nonetheless, given that a completely passive measurement setup was used and that

Figure 7. Side-by-side comparisons between an input video frame and the pose of the laser scan that our approach reconstructed. The poses of the persons and even the deformations of complex apparel, like the kimono, are faithfully reproduced.

the person stands several meters away from the cameras, an average tracking accuracy of roughly 2.6 cm is very good.

For our tests with real data we captured video footage and body models for different male and female test subjects using the setup described in Sect. 3.1. The captured video sequences are between 300 and 600 frames long and show a variety of different clothing styles, including normal everyday apparel and a traditional Japanese kimono. Many different motions have been captured ranging from simple walking to gymnastic moves.

Fig. 7 shows several side-by-side comparisons between input video frames and recovered mesh poses. The algorithm reliably recovers the pose and surface deformation for the male subject who wears comparably wide apparel. Our algorithm can even capture the motion and the cloth deformation for a woman wearing a kimono, Fig. 1 and Fig. 7. Since the limbs are completely obscured, this would not have been possible with a normal motion capture approach. Please note that spatial constraints in our studio limit the useable recording area to roughly 1.5x1.5 m. Thus, the motions in our test data are spatially confined. Our method could handle arbitrary large-scale motion in just the same way. More results with synthetic and real-world sequence are shown in the accompanying video.

The results show that our purely passive mesh-based tracking approach can automatically capture both pose and surface deformation of human actors. It illustrates that a skeleton-less algorithm is capable of tracking even complex deformations of different materials by means of the same framework. Our tracker neither requires any segmentation of the model into parts, e.g. clothing and body, nor does it expect the specification of explicit material parameters as they are often used in garment motion tracking. Both of this would be very difficult for a human wearing different kinds of fabrics. The combination of an a priori model, a fast Laplacian mesh deformation scheme, and a 3D flow-based correspondence estimation method enables us to capture complex shape deformations from only a few cameras. As an additional benefit, our method preserves the mesh's connectivity which is particularly important when it comes to our envisioned 3D video applications and dynamic shape encoding.

Nonetheless, our algorithm is subject to a few limitations. Currently, we cannot handle volume constraints [12]. In some situations such a constraint may prevent incorrect mesh deformations and thus compensate the effect of incorrect flow estimates. However, for some types of apparel, such as a long skirt or our kimono, a volume constraint may even prevent correct tracking. From this point of view our implementation is more flexible.

Another problem arises if the subject in the scene moves very quickly. In these situations, optical flow tracking may fail. To attack this problem, one might use one of the many high-speed camera models available today for capturing fast scenes.

Finally, our algorithm cannot capture the true shape variation of low-frequency surface details, such as wrinkles in clothing. While they globally deform with the model, they seem to be "baked in" to the surface. In typical 3D video applications, however, this inaccuracy will not play a major role. Nonetheless, we plan to extend our method in the future to capture these small details by means of a multi-view stereo algorithm.

Despite these limitations our method is a flexible, easy to implement and reliable purely passive method to capture the time-varying shape of humans from video. To our knowledge, this is the first system in the literature that can capture arbitrarily deforming meshes in a connectivity preserving way for such complex scenes.

5. Conclusion

We have presented a new algorithm for automatic marker-less tracking of deformable human models from a handful of video streams. The combination of a 3D scene flow-based correspondence estimation approach with a Laplacian mesh deformation scheme enables our method to make a laser scan of a subject move and deform in the same way as its real-world counterpart in video. Our algorithm is easy to implement and can handle a large range of human motions and clothing styles. Its robustness and reliability has been demonstrated on both real and synthetic input data.

As a direction of future work, we plan to integrate our

method into a larger approach to reconstruct high-quality 3D videos of humans wearing arbitrary apparel.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV and AIM@SHAPE, a Network of Excelence project (506766) within EU's Sixth Framework Programme.

References

[1] B. Allen, B. Curless, and Z. Popović. Articulated body deformation from range scan data. *ACM Trans. Graph. (SIGGRAPH '02)*, pages 612–619, 2002. 1, 2

[2] M. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proc. of ICCV93*, pages 231–236, 1993. 5

[3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, volume 3024, pages 25–36, 2004. 5

[4] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph. (Proc. of SIGGRAPH'03)*, 22(3):569–577, 2003. 3, 4

[5] E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Reconstructing human shape and motion from multi-view video. In *CVMP'05*, pages 42–49, 2005. 2

[6] D. Decarlo and D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *Int. J. Comput. Vision*, 38(2):99–127, 2000. 2

[7] C. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *CVIU*, 96(3):367–392, 2004. 2

[8] Y. Furukawa and J. Ponce. Carved visual hulls for image-based modeling. In *ECCV(1)*, pages 564–577, 2006. 1, 2

[9] B. Goldluecke and M. Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. In *CVPR 2004*, volume I, pages 350–355, 2004. 1, 2

[10] T. Heap and D. Hogg. Towards 3d hand tracking using a deformable model. In *FG '96*, page 140, 1996. 2

[11] L. Herda, P. Fua, R. Plänkers, R. Boulic, and D. Thalmann. Skeleton-based motion capture for robust reconstruction of human motion. In *Computer Animation*, page 77ff, 2000. 2

[12] J. Huang, X. Shi, X. Liu, K. Zhou, L.-Y. Wei, S.-H. Teng, H. Bao, B. Guo, and H.-Y. Shum. Subspace gradient domain mesh deformation. *ACM Trans. Graph.*, 25(3):1126–1134, 2006. 7

[13] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50, 1956. 5

[14] Y. Lipman, O. Sorkine, D. Cohen-Or, D. Levin, C. Rössl, and H.-P. Seidel. Differential coordinates for interactive mesh editing. In *SMI 2004*, pages 181–190, 2004. 4

[15] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. DARPA IU Workshop*, pages 121–130, 1981. 5

[16] T. McInerney and D. Terzopoulos. Deformable models in medical images analysis: a survey, Medical Image Analysis, 1(2):91–108, 1996. 1, 2

[17] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(6):580–591, 1993. 1, 2

[18] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *CVIU*, 81(3):231–268, 2001. 1, 2

[19] S. I. Park and J. K. Hodgins. Capturing and animating skin deformation in human motion. *ACM Trans. Graph. (SIGGRAPH 2006)*, 25(3), Aug. 2006. 1, 2

[20] P. P. Pebay and T. J. Baker. A comparison of triangle quality measures. *In Proceedings to the 10th International Meshing Roundtable*, pages 327–340, 2001. 6

[21] A. Petland and B. Horowitz. Recovery of nonrigid motion and structure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(7):730–742, 1991. 2

[22] R. Plänkers and P. Fua. Articulated soft objects for multiview shape and motion capture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1182–1187, 2003. 2

[23] D. Pritchard and W. Heidrich. Cloth motion capture. In *Eurographics*, pages 263–271, Sept. 2003. 2

[24] B. Rosenhahn, U. Kersting, K. Powel, and H.-P. Seidel. Cloth x-ray: Mocap of people wearing textiles. In *Pattern Recognition 2006,DAGM*, pages 495–504, 2006. 2

[25] M. Salzmann, S. Ilic, and P. Fua. Physically valid shape parameterization for monocular 3-d deformable surface tracking. In *British Machine Vision Conference*, 2005. 1, 2

[26] P. Sand, L. McMillan, and J. Popovic. Continuous capture of skin deformation. *ACM Trans. Graph.*, 22(3):578–586, 2003. 1, 2

[27] O. Sorkine. Differential representations for mesh processing. *Computer Graphics Forum*, 25(4), 2006. 2

[28] J. Starck and A. Hilton. Spherical matching for temporal correspondence of non-rigid surfaces. *IEEE ICCV*, pages 1387–1394, 2005. 2

[29] C. Stoll, Z. Karni, C. Rössl, H. Yamauchi, and H.-P. Seidel. Template deformation for point cloud fitting. In *Symposium on Point-Based Graphics*, pages 27–35, 2006. 2, 5

[30] L. Torresani and A. Hertzmann. Automatic non-rigid 3d modeling from video. In *ECCV(2)*, pages 299–312, 2004. 2

[31] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *ICCV-99*, volume II, pages 722–729, 1999. 3

[32] H. Yamauchi, S. Gumhold, R. Zayer, and H.-P. Seidel. Mesh segmentation driven by gaussian curvature. *Visual Computer*, 21(8-10):649–658, 2005. 4

[33] R. Zayer, C. Rössl, Z. Karni, and H.-P. Seidel. Harmonic guidance for surface deformation. In *Proc. of Eurographics 2005*, volume 24, pages 601–609, 2005. 5

[34] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph. (Proc. SIGGRAPH'04)*, 23(3):600–608, 2004. 2