Marker-less 3D Feature Tracking for Mesh-based Human Motion Capture

Edilson de Aguiar [1], Christian Theobalt [2], Carsten Stoll [1]
and Hans-Peter Seidel [1]

[1] MPI Informatik, Germany
[2] Stanford University, USA
{edeaguia, stoll, hpseidel}@mpi-inf.mpg.de
theobalt@cs.stanford.edu

Abstract. We present a novel algorithm that robustly tracks 3D trajectories of features on a moving human who has been recorded with multiple video cameras. Our method does so without special markers in the scene and can be used to track subjects wearing everyday apparel. By using the paths of the 3D points as constraints in a fast mesh deformation approach, we can directly animate a static human body scan such that it performs the same motion as the captured subject. Our method can therefore be used to directly animate high quality geometry models from unaltered video data which opens the door to new applications in motion capture, 3D Video and computer animation. Since our method does not require a kinematic skeleton and only employs a handful of feature trajectories to generate lifelike animations with realistic surface deformations, it can also be used to track subjects wearing wide apparel, and even animals. We demonstrate the performance of our approach using several captured real-world sequences, and also validate its accuracy.

1 Introduction

Nowadays, generating realistic and lifelike animated characters from captured real-world motion sequences is still a hard and time-consuming task. Traditionally, marker-based optical motion capture systems [1] reconstruct the motion of a moving subject by measuring the 3D trajectories of optical beacons attached to her body. The optical markers are then mapped to a kinematic skeleton structure [2]. Marker-free methods also exist that are able to measure human motion in terms of a kinematic skeleton without any intrusion into the scene. Thereafter, the model geometry and the skeleton need to be connected such that the surface deforms realistically with the body motion by specifying the influence of each bone on both rigid and non-rigid surface deformation [3].

Stepping directly from a captured real-world sequence to the corresponding realistic moving character is still challenging. Several methods in the literature are able to partly solve this problem. Since marker-based and marker-free motion capture systems measure the motion in terms of a kinematic skeleton, they have to be combined with other scanning technologies to capture the time-varying

shape of the human body surface [4–6]. However, dealing with people wearing arbitrary clothing from only video streams is still not possible. Time-varying scene representations can also be reconstructed by means of shape-from-silhouette approaches [7], or with combined silhouette- and stereo-based methods [8]. Unfortunately, the measured models often lack detail if only a small number of input camera views is available and it is hard to preserve topological correspondences over time. Researchers have also used physics-based methods to track simple human motions if a kinematic skeleton is available [9]. However, the methods can not be directly applied to objects made of a variety of different materials, and they are not able to track arbitrarily dressed humans completely passively.

Instead, we present a robust skeleton-less approach to automatically capture the motion of a moving human subject and generate plausible and realistic surface deformations from multiple video streams without optical markers. Our algorithm is simple and versatile and enables us to directly animate a high quality static human scan from unaltered video footage which enables potential new applications in motion capture, computer animation and 3D Video.

The main contribution of this paper is a simple and robust method to automatically identify features on a moving human wearing everyday apparel, and track their 3D trajectories. It does not employ any a priori information about the subject, e.g. a kinematic skeleton, and can therefore be straightforwardly applied to other subjects, e.g. animals or mechanical objects. We also present a fast mesh deformation approach that uses only a handful of feature trajectories to directly and realistically animate a static human body scan making it performs the same motion as the captured subject. Our algorithm handles humans wearing arbitrary and sparsely textured clothing. As an additional benefit, it also preserves the mesh's connectivity over time.

The remainder of this paper is structured as follows: Sect. 2 reviews the most relevant related work and Sect. 3 briefly describes our overall framework. Thereafter, Sect. 4 details our automatic approach to identify features and track their 3D trajectories without optical markers. Sect. 5 describes our fast deformation scheme that is used to animate the static human model over the whole sequence according to the constraints derived from the estimated 3D point trajectories. Experiments and results with several captured real-world sequences are shown in Sect. 6, and the paper concludes in Sect.7.

2 Related Work

In our research we capitalize on ideas that have been published in the fields of object tracking, motion capture and scene reconstruction. For the sake of brevity, we refer the interested reader to overview articles on object tracking [10, 11]. The following, is by no means a complete list of references from the other two research topics, 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 [1] that parameterize the data in terms of kinematic skeletons. Unfortunately, these approaches can not 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 [5] for deformation capture, combine a motion capture system with a range scanner [4, 12] or a shape-from-silhouette approach [6], or jointly use a body and a cloth model to track the person [13]. Although achieving good results, most of these methods require active interference with the scene or require hand-crafted models for each individual.

Alternatively, shape-from-silhouette algorithms [7], multi-view stereo approaches [14], or methods combining silhouette and stereo constraints [8] can be used to reconstruct dynamic scene geometry. To obtain good quality results, however, several cameras are required and it is hard to generate connectivity-preserving dynamic mesh models.

Some passive methods extract 3D correspondences from images to track simple deformable objects [15] or cloth [16]. They can also be employed to jointly capture kinematic motion parameters and surface deformations of tightly dressed humans [17, 18]. Researchers have also used physics-based shape models to track textiles [19, 20] or simple articulated humans [9]. Unfortunately, none of these methods is able to track people dressed in arbitrary everyday apparel completely passively.

In contrast, we propose a skeleton-less method to directly capture the poses of a moving human subject and generate plausible surface deformations from only a handful of input video streams. This is achieved by first robustly identifying and tracking image features in 3D space. Thereafter, using the 3D trajectories of the features as constraints in a Laplacian mesh editing setting [21], the human model is realistically animated over time. By relying on differential coordinates, plausible shape deformations for the human scan are computed without having to specify explicit material parameters. Our algorithm is simple, robust, easy to implement and works even for moving subjects wearing wide and loose apparel.

3 Overview

An overview of our approach is shown in Fig. 1. Our system expects as input a multi-view video (MVV) sequence that shows the person moving arbitrarily. After acquiring the sequence, silhouette images are calculated via color-based background subtraction and we use the synchronized video streams to extract and track features in 3D space over time.

Our hybrid 3D point tracking framework jointly uses two techniques to estimate the 3D trajectories of the features from unmodified multi-view video streams. First, features in the images are identified using the Scale Invariant Feature Transform (SIFT). Furthermore, SIFT is able to match a feature to its corresponding one from a different camera viewpoint. This allows us to generate a set of pairwise pixel correspondences between different camera views for each time step of input video. Unfortunately, tracking the features over time using only local descriptors is not robust if the human subject is wearing sparsely textured clothing. Therefore, we use a robust dense optical flow method as an

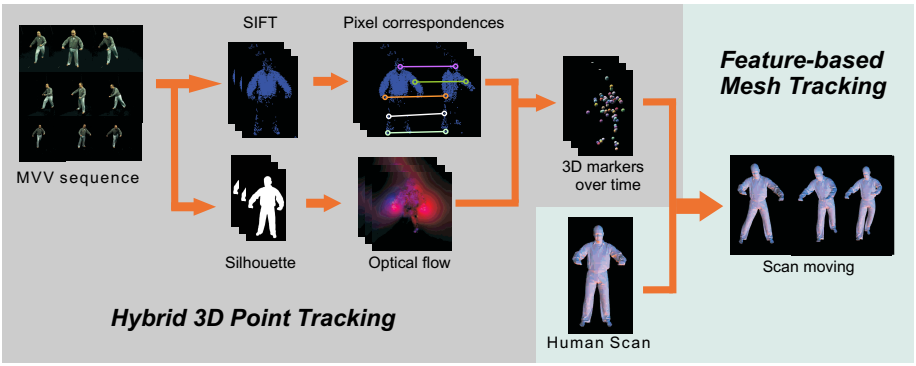

Fig. 1. Overview of our framework: given a multi-view video sequence showing a human performing, our method automatically identifies features and tracks their 3D trajectories. By applying the captured trajectories to a static laser-scan of the subject we are able to realistically animate a human model making it move the same as its real-world counterpart in the video streams.

additional step to track the features for each camera view separately to fill the gaps in the SIFT tracking. By merging both source of information we are able to reconstruct the 3D trajectories for many features over the whole sequence.

Our hybrid technique is able to correctly identify and track many 3D points. In addition to the estimation of 3D point positions, our approach also calculates a confidence value for each estimation. Using confidence-weighted feature trajectories as deformation constraints, our system robustly brings a static laser-scanned triangle mesh M of the subject into life by making it follow the motion of the actor recorded in the video frames.

4 Hybrid 3D Point Tracking

Our hybrid framework jointly employs local descriptors and dense optical flow to identify features and estimate their 3D positions over time from multiple calibrated camera views. In contrast to many other approaches [22–24], we developed an automatic tracking algorithm that works directly on the images without any a priori knowledge about the moving subject. It is our goal to create a simple and generic algorithm that can be used to track features on rigid bodies, articulated objects and non-rigidly deforming subjects in the same way.

The input to our algorithm comprises of synchronized video streams recorded from K cameras, each containing N video frames (Fig. 2a). In the first step, we automatically identify L important features, also called keypoints, for each camera view k and time step t and generate a set of local descriptors $F_{k,t} = \{f^0_{k,t}, \ldots, f^L_{k,t}\}$ using SIFT [25], Fig. 2b. We extract these features using the interest point detector proposed by Lowe [26] that is based on local 3D extrema in the scale-space pyramid built with difference-of-Gaussian filters. The local

descriptors are built as a distinctive representation of the feature in an image from a patch of pixels in its neighborhood.

Since the SIFT descriptors are invariant to image scale, rotation, change in viewpoints, and change in illumination, they can be used to find corresponding features across different camera views. Given an image $I_{k,t}$, from camera view k and time step t, and the respective set of SIFT descriptors $F_{k,t}$, we try to match each element of $F_{k,t}$ with the set of keypoints from all other camera views. We use a matching function similar to [25], which assigns a match between $f^i_{k,t}$ and a keypoint in $F_{j,t}$ if the Euclidean distance between their invariant descriptor vectors is minimum. In order to discard false correspondences, nearest neighbor distance ratio matching is used with a threshold T_{MATCH} [27] .

After matching the keypoints across all K camera views at individual time steps, we gather all R correct pairwise matches into a list of pixel correspondences $C_t = \{c^0_t, \ldots, c^R_t\}$ by using all reliable matches found for each time step t (Fig. 2c). Each element $c^r_t = ((cam_u, P^i_t), (cam_v, P^j_t))$ stores the information about a correspondence between two different camera views, i.e. that pixel P^i_t in camera cam_u corresponds to pixel P^j_t in camera view cam_v at time t.

Unfortunately, tracking the features over time using only the list of correspondences C and connecting their elements at different time steps is not robust, because it is very unlikely that the same feature will be found at all time instants. This is specially true if the captured images show subjects performing fast movements, where features can be occluded for a long period of time, or when the subject wears everyday apparel with sparse texture. In the latter case, SIFT only detects a small number of keypoints per time step, which is usually not enough for tracking articulated objects. Therefore, in order to robustly reconstruct the 3D trajectories for the features we decided to use optical flow to track both elements of all c^r_t for each camera view separately, i.e. the pixel P^i_t is tracked using camera view cam_u and the pixel P^j_t using camera view cam_v.

The 2D flow-based tracking method works as follows: for each camera view k, we track all pixels over time using the warping-based method for dense optical flow proposed by Brox et al. [28]. After calculating the optical flow $\boldsymbol{o^t_k}(I_{k,t}, I_{k,t+1})$ between time step t and $t+1$ for camera k, we use $\boldsymbol{o^t_k}$ to warp the image $I_{k,t}$ and we verify for each pixel in the warped image if it matches the corresponding pixel in $I_{k,t+1}$. We eliminate the pixels that do not have a partner in $t+1$ and the pixels that belong to the background by comparing the warped pixels with the pre-computed silhouette $SIL_{k,t+1}$. This process is repeated for all consecutive time steps and for all camera views. As a result, we construct a tracking list $D_k = \{E^0, \ldots, E^g\}$ with G pixel trajectories for each camera view k (Fig. 2d). Each element $E^i = \{P^i_0, \ldots, P^i_N\}$ contains the positions of the pixel P^i_t for all time steps t.

The last step of our hybrid tracking scheme merges the optical flow tracking information with the list of correspondences to reconstruct the 3D trajectories for all features. We take pixel correspondences from all time steps into account. For instance, if a matching c^r_t is detected by SIFT only at the end of the sequence we

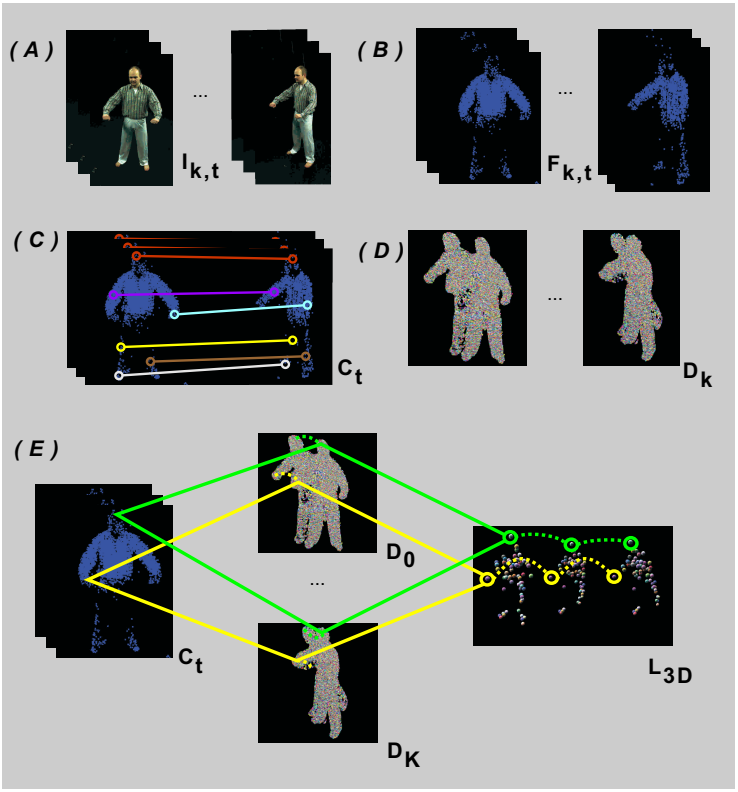

Fig. 2. Using the synchronized video streams as input (A), our hybrid approach first identifies features in the images using SIFT (B) and then matches these features between different pairs of camera views based on their descriptors (C). In addition, we track these features for each camera view separately using optical flow (D). At the end, reliable 3D trajectories for the features are reconstructed by merging both information (E).

are still able to recover the anterior positions of the feature by using the optical flow information.

For each entry $c_t^r = ((cam_u, P_t^i), (cam_v, P_t^j))$, we verify if the pixel P_t^i is found in D_{cam_u} and if the pixel P_t^j is found in D_{cam_v}. In case both elements are found, we estimate the position of the respective 3D point, $mm_r(t)$, for the whole sequence (Fig. 2e), otherwise c_t^r is discarded. The 3D positions are estimated by triangulating the viewing rays that start at the camera views cam_u and cam_v and pass through the respective image plane pixel at P_t^i and P_t^j. However, due to inaccuracies, these rays will not intersect exactly at a single point. However, we can compute a pseudo-intersection point $pos_t^r = \{x, y, z\}$ that minimizes the sum of squared distance to each pointing ray. We also use the inverse of this distance, cv_r, as a confidence measure indicating how reliable a particular feature has been

located. If cv_r is below a threshold T_{CONF} we discard it, since it indicates that c_t^r assigns a wrong pixel correspondence between two different camera views.

We also discard a trajectory mm_r if it does not project into the silhouettes in all camera views and at all time steps. This way, we can prevent the use of 3D points whose trajectories degenerate over time as deformation constraints. We assess silhouette-consistency using the following measure:

$$TSIL(mm_r) = \sum_{t=0}^{N} \sum_{k=0}^{K} PROJ_{sil}^{k}(pos_t^r, t) \tag{1}$$

where $PROJ_{sil}^{k}(pos_t^r, t)$ is a function that evaluates to 1 if $mm_r(t)$ projects inside the silhouette image of camera view k at time step t, and it is 0 otherwise. We only consider mm_r a reliable 3D trajectory if $TSIL(mm_r) > TR_{SIL}$. Appropriate values for the thresholds are found through experiments.

After processing all elements of C for all time steps, we generate a list with reliable 3D trajectories for the features. The list $L_{3D} = \{mm_0, \ldots, mm_h\} = \{(LP_0, LE_0), \ldots, (LP_H, LE_H)\}$ assigns to each trajectory mm_i, a tuple (LP_i, LE_i) containing the 3D point positions, $LP_i = \{pos_0^i, \ldots, pos_N^i\}$, and the respective list of confidence values for each estimated 3D position, $LE_i = \{cv_0^i, \ldots, cv_N^i\}$. As shown in Sect. 6, our hybrid approach is able to identify and accurately track many 3D points for sequences where the human subject is performing fast motion, even when he is dressed in everyday apparel.

5 Feature-based Laplacian Mesh Tracking

It is our goal to animate the human scan making it move the same way as its real-world counterpart in the video streams by using the reconstructed 3D point trajectories as motion constraints. For this purpose, we first roughly align the human model with the 3D point positions at the first time step of video (our reference), Fig. 3(a). This is automatically done by applying a PCA-based alignment scheme to a reconstructed volumetric shape-from-silhouette model of the moving subject. Thereafter, we select H target vertices $V_T = \{v_{Th} | h \in \{0 \cdots H\}\}$ in the human model M by choosing vertices that are closest to the 3D point positions at the reference time step, Fig. 3(b). These target vertices V_T are used to guide the mesh deformation method.

We deform the human scan by employing a Laplacian mesh deformation scheme that jointly uses rotational and positional constraints on the target vertices V_T in a similar way as [29]. The details of the human model M are encoded in its differential coordinates. The differential coordinates d of M are computed once at the beginning of the sequence by solving a matrix multiplication of the form $d = Lv$, where L is the discrete Laplace operator based on the cotangent-weights, and v is the vector of M's vertex coordinates [30]. Thereafter we perform the following three processing steps for each time step t:

Since the differential coordinates d are rotation-dependent [31], we need to first calculate the local rotations that should be applied to d. We derive these

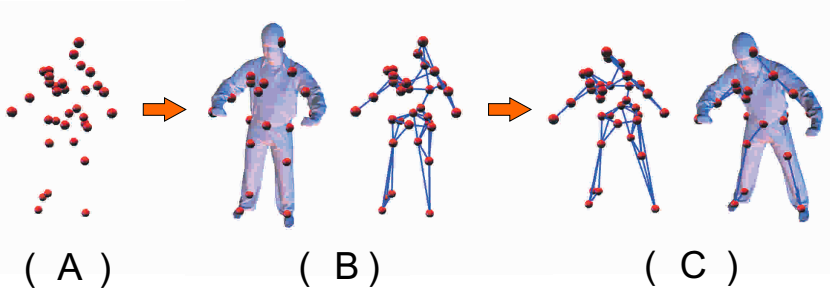

Fig. 3. After aligning the 3D point positions (A) with the human model at the reference time step (B), our method reconstructs a novel pose jointly using rotational and positional constraints on the target vertices which we derived from the 3D feature trajectories (C).

rotational constraints from the 3D trajectories. The local rotation for each target vertex v_{Th} of M is calculated from the rotation of the corresponding 3D point $mm_h(t)$ between reference time and time t by means of a graph-based method, Fig. 3. To this end, 3D points are considered as nodes in a graph, and edges between them are determined by constructing the minimal spanning tree [32] using the approximated geodesic distance as edge weights. For each 3D point $mm_h(t)$, we find the minimal rotation that makes its outgoing edges at the reference time match its outgoing edges at time t (i.e. using the Jacobian). This local rotation is then converted into a quaternion $q_{mm_h(t)}$. Since we want the target vertices V_T to perform the same rotations as the 3D points, we set $q_{v_{Th}} = q_{mm_h(t)}$ for all H 3D points.

Using the estimated rotations for the target vertices, we interpolate them over M using the idea proposed in [33] in order to estimate rotations for each vertex of the model. Each component of a quaternion $q = [w, q_1, q_2, q_3]$ is regarded as a scalar field defined over the entire mesh. A smooth interpolation is guaranteed by regarding these scalar fields as harmonic fields. The interpolation is performed by solving the Laplacian equation $Lq = 0$ over the whole mesh using the constrained target vertices as Dirichlet boundary conditions and normalizing the resulting quaternions.

At the end, we reconstruct the vertex positions v of M such that the mesh best approximates the rotated differential coordinates, as well as the positional constraints. This can be formulated as a least-squares problem of the form

$$\underset{v}{argmin}\{\|Lx - (q \cdot d \cdot \overline{q})\|^2 + \|Av - p\|^2\}. \tag{2}$$

which can be transformed into a linear system

$$(L^T L + A^T A)v = L^T(q \cdot d \cdot \overline{q}) + A^T p. \tag{3}$$

In Eq. 3, p is the vector of positional constraints of the form $v_j = pos_t^j$, $j \in \{1, \ldots, H\}$ specified for the H target vertices and derived from the position of

the 3D points at time t. The matrix A is a diagonal matrix containing non-zero weights $A_{ij} = c * cv_t^j$, c being a constant, only for constrained vertices j. We weight the target vertex position pos_t^j for v_{Tj} at time t proportional w.r.t its corresponding confidence value since small values for cv_t^j indicate inaccuracies in the estimated 3D position. As demonstrated in Sect. 6, this weighting scheme leads to a better visual animation quality for the animated human scans.

After applying this algorithm to the whole sequence, our mesh deformation approach is able to animate the human scan making it correctly follow the motion of the actor recorded in all video frames. As shown in Sect. 6, our approach preserves the details and features of the mesh and is able to generate plausible and realistic surface deformation for subjects wearing even loose everyday apparel.

6 Results and Discussion

We tested our method on several real-world sequences with different male and female test subjects recorded in our studio. Our acquisition procedure works as follows: we first acquire the scanned model with a Vitus Smart$^{\text{TM}}$ 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 and color-consistency across cameras is ensured by applying a color-space transformation to each camera stream. The captured video sequences are between 150 and 400 frames long and show a variety of different clothing styles. We captured different motions ranging from simple walking to yoga and capoeira moves.

As shown in the second and third columns of Table 1, our hybrid 3D point tracking approach is able to identify and track many features in 3D space accurately. The average confidence value (CV) for the 3D point positions are large, which corresponds to position errors of around $1.0-2.2cm$. Three different frames for the yoga (YOGA) and walking (WALK) sequences with selected features shown as dots can be seen in the upper row of Fig. 4. Looking at the temporal evolution one can see that the features are reliably tracked over time. The left images in the middle row of Fig. 4 also show two closeups on the legs in the walking sequence. Features were accurately tracked despite the appearance ambiguities caused by the trousers with homogeneous color. If we had used only SIFT descriptors to track these features, it would have been impossible to track them in these homogeneous areas.

High tracking accuracy and reliability even in such difficult situations is upheld by additionally taking into account optical flow information. Even if a correspondence was only found for one time step, we can reconstruct the complete trajectory for this feature by looking at the optical flow information. This second source of information also enables us to apply a very high threshold T_{MATCH} which eliminates unreliable 3D feature matches already at an early stage.

Fig. 4. (upper row) Selected features tracked in three different frames for the yoga and walking sequences; (middle and lower row) two frames of the walking sequence in detail and side-by-side comparisons between input video frames and reconstructed poses for the human scan. Our framework correctly tracks 3D trajectories of features even in the presence of occlusions or appearance ambiguities. By combining the 3D point trajectories with our mesh deformation method, our algorithm is able to directly animate a human body scan.

Before using the 3D point trajectories to guide the deformation of the human scan we first choose a subset of N_M points from the initial set of 3D trajectories L_{3D} at the reference time step. This subset of points should be distributed evenly on the model surface. This is done by ramdonly choosing a element in L_{3D} and all adjacent points next to it at the reference time step by using a distance threshold T_{DIST}. We compare the confidence values for this group of elements and choose the point with the maximum confidence value. We continue the same procedure choosing another point in L_{3D} until all selected points are separated by a distance T_{DIST}, and consequently distributed over the model's surface. We conducted several experiments with different values for T_{DIST} and found out that in general, values between $10cm$ and $20cm$ produce best results. For a typical sequence, this leads to around $20 - 50$ selected points. Note that

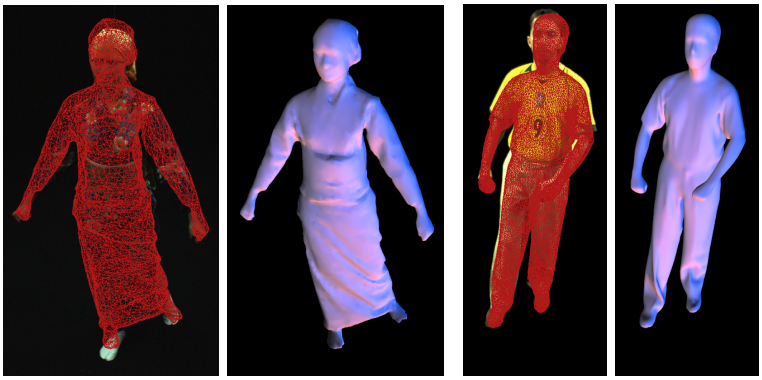

Fig. 5. Overlap between the reprojected model (red) and the input image for the female and male subjects. Our framework is able to correctly reconstruct their pose even when they are wearing wide and loose apparel.

although our hybrid 3D tracking approach is able to correctly track many more points over time, even a subset of points is sufficient to track body poses reliably. Our selection criteria also enable us to eliminate multiple trajectories of the same feature (stemming for different camera pairs) which bears no useful information

The middle (right) and lower row of Fig. 4 shows several side-by-side comparisons between input video frames and tracked poses of the human scan. Our algorithm reliably recovers the poses and creates plausible and realistic surface deformations for the male actor performing a capoeira move and even for the female subject wearing a long dress. Due to the occlusion of the limbs or the wide and loose apparel, tracking the motion of these subjects over time would have been hard with a normal motion capture system. More captured real-world results are shown in the accompanying video.

Due to the lack of ground truth for our experiments, we evaluate our results by overlapping the reprojected model with the input images as shown in Fig. 5. We also calculate a multi-view overlap measure by counting the average number of pixels that do not match between the reprojected model and the input image silhouettes for all camera views and all time steps. As shown in the plot in Fig. 6, our system automatically animates the human scan making it follow the motion of the real-world actor with a consistent silhouette-accuracy of more than 94%.

We also performed experiments to evaluate the performance of our framework in animating the human scan. Table 1 summarizes the results we have obtained employing quality and accuracy measures for several sequences. The column *VOL* shows the average volume change in the animated scan over the whole sequence. This measure is a numerical indicator for implausible deformations. The preservation of mesh quality is analyzed by looking at the average distortion of the triangles, *QLT*. It is computed by averaging the per-triangle Frobenius norm over the mesh and over time [34]. This norm is 0 for an equilateral triangle and approaches infinity with increasing degeneracy. Finally, the column labeled *OVLP* contains the average multi-view overlap between the reprojected model and the input image silhouettes over time.

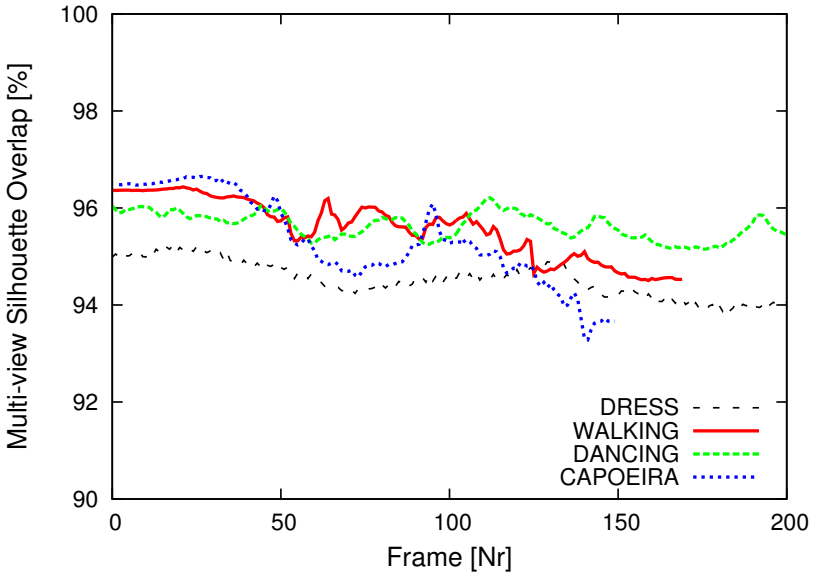

Fig. 6. Multi-view silhouette overlap for several captured sequences. Our system automatically makes the static human scan follow the motion of the captured real-world actor with high precision.

Table 1 shows that the volume change in the animated human scan is in the range of normal non-rigid body deformations, and that triangles remain in nice shape. It also shows that our mesh deformation approach reconstructs the poses of the scan with high accuracy, even if the subjects wear wide and loose everyday apparel.

We performed experiments to demonstrate the importance of the confidence value as a weight in Eq. 3 (Sect. 5) as well. Our experiments show that when using the confidence value in our mesh deformation approach, surface deformations are generated in a more reasonable and lifelike way, which leads to a better visual animation quality.

SEQ	FEAT	CV $[m^{-1}]$	OVLP	VOL	QLT
CAPO	1207	65.18	95.4%	3.2%	0.03
DANC	1232	58.30	95.6%	1.8%	0.01
YOGA	1457	112.23	93.7%	3.6%	0.10
WALK	2920	71.78	95.5%	1.5%	0.01
DRSS	3132	45.72	94.4%	2.0%	0.01

Table 1. For each captured real-world sequence, the number of identified features (FEAT) and the average confidence value (CV) are shown. We also employ accuracy and quality measures for the animated scan, i.e. changes in volume (VOL), distortion of triangles (QLT) and multi-view silhouette overlap (OVL), to demonstrate the performance of our framework.

Our results show that our purely passive tracking method can automatically identify and track the 3D trajectories of features on a moving subject without the need of any a priori information or optical markers. By combining it with our fast deformation technique it also enables us to directly and realistically animate a static human scan making it follow the same motion as its real-world counterpart even if he wears casual everyday apparel.

Nonetheless, our algorithm is subject to a few limitations. Currently, if the subject moves very quickly, the optical flow method may fail to track the 2D features. However, in such situations one might use one of the many high-speed camera models available today for capturing fast scenes. Another limitation is the run time of our tracking system. Currently, we need around 3-5 minutes per multi-view frame on a Pentium IV with 3GHz, with more than 90% of the time spent for the SIFT and optical flow calculations. We are planning to investigate the use of lower image resolutions for tracking without compromising the overall tracking accuracy. Our fast mesh deformation approach, on the other hand, can generate animations at 5 fps for models comprising of $20k$ to $50k$ triangles.

Another problem is that our mesh deformation method does not handle volume constraints [35]. In starkly under-constrained settings such a constraint would prevent inaccurate deformations, however at the cost of slower runtimes. Also, in some situations, e.g. very wide apparel, a volume constraint might even prevent correct deformations. Finally, although our algorithm correctly captures the body deformations at a coarse scale, the deformations of subtle details, such as small wrinkles, are not captured. We are planning to extend our method in the future to also capture these details by means of a multi-view stereo algorithm.

Despite these limitations our automatic method is a simple, flexible, easy to implement and reliable purely passive method to robustly track 3D trajectories of features on a moving human and even other subjects. These features can then be used to animate a static human body scan making it perform the same motion as the captured subject recorded from only a handful of cameras.

7 Conclusion

We have presented a new skeleton-less approach to automatically identify features and track them on a moving subject who has been recorded with only eight video cameras. Our algorithm does not require optical markings, does not need a priori information about the tracked subject, and behaves robustly even for humans wearing sparsely textured and wide apparel. By applying the captured feature trajectories as constraints in a fast mesh deformation approach, we can make a high-quality human scan move and deform in the same way as its real-world counterpart in the input video footage. We expect that our new mesh-based paradigm will pave the trail for many new applications in motion capture in general, 3D Video and character animation.

Acknowledgments 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. We would like to thank Thomas Brox and David Lowe for letting us use their code.

References

1. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. CVIU **81**(3) (2001) 231–268
2. Silaghi, M.C., Plänkers, R., Boulic, R., Fua, P., Thalmann, D.: Local and global skeleton fitting techniques for optical motion capture. In: CAPTECH '98, London, UK, Springer-Verlag (1998) 26–40
3. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: ACM Trans. Graph. (2000) 165–172
4. Allen, B., Curless, B., Popović, Z.: Articulated body deformation from range scan data. ACM Trans. Graph. (SIGGRAPH '02) (2002) 612–619
5. Park, S.I., Hodgins, J.K.: Capturing and animating skin deformation in human motion. ACM Trans. Graph. (SIGGRAPH 2006) **25**(3) (August 2006)
6. Sand, P., McMillan, L., Popovic, J.: Continuous capture of skin deformation. ACM Trans. Graph. **22**(3) (2003) 578–586
7. Goldluecke, B., Magnor, M.: Space-time isosurface evolution for temporally coherent 3d reconstruction. In: CVPR 2004. Volume I. (2004) 350–355
8. Furukawa, Y., Ponce, J.: Carved visual hulls for image-based modeling. In: ECCV(1). (2006) 564–577
9. Metaxas, D., Terzopoulos, D.: Shape and nonrigid motion estimation through physics-based synthesis. IEEE Trans. Pattern Anal. Mach. Intell. **15**(6) (1993) 580–591
10. Lepetit, V., Fua, P.: Monocular model-based 3d tracking of rigid objects. Found. Trends. Comput. Graph. Vis. **1**(1) (2006) 1–89
11. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. **38**(4) (2006) 13
12. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. ACM Trans. Graph. **24**(3) (2005) 408–416
13. Rosenhahn, B., Kersting, U., Powel, K., Seidel, H.P.: Cloth x-ray: Mocap of people wearing textiles. In: Pattern Recognition 2006,DAGM. (2006) 495–504
14. 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) (2004) 600–608
15. Decarlo, D., Metaxas, D.: Optical flow constraints on deformable models with applications to face tracking. Int. J. Comput. Vision **38**(2) (2000) 99–127
16. Pritchard, D., Heidrich, W.: Cloth motion capture. In: Eurographics. (September 2003) 263–271
17. de Aguiar, E., Theobalt, C., Magnor, M., Seidel, H.P.: Reconstructing human shape and motion from multi-view video. In: CVMP'05. (2005) 42–49
18. Plänkers, R., Fua, P.: Articulated soft objects for multiview shape and motion capture. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9) (2003) 1182–1187
19. Hasler, N., Asbach, M., Rosenhahn, B., Ohm, J.R., Seidel, H.P.: Physically based tracking of cloth. In: Proc of VMV 2006, Aachen, Germany (2006) 49–56

20. Salzmann, M., Ilic, S., Fua, P.: Physically valid shape parameterization for monocular 3-d deformable surface tracking. In: British Machine Vision Conference. (2005)
21. Sorkine, O.: Differential representations for mesh processing. Computer Graphics Forum **25**(4) (2006)
22. Balan, A.O., Black, M.J.: An adaptive appearance model approach for model-based articulated object tracking. In: Proc. of CVPR '06, Washington, DC, USA, IEEE (2006) 758–765
23. Brox, T., Rosenhahn, B., Cremers, D., Seidel, H.P.: High accuracy optical flow serves 3-d pose tracking: Exploiting contour and flow based constraints. In: ECCV (2). (2006) 98–111
24. Kehl, R., Gool, L.V.: Markerless tracking of complex human motions from multiple views. Comput. Vis. Image Underst. **104**(2) (2006) 190–209
25. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. In: International Journal of Computer Vision. Volume 20. (2004) 91–110
26. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. of ICCV. (1999) 1150–1157
27. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. (2003)
28. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: ECCV. Volume 3024. (2004) 25–36
29. de Aguiar, E., Theobalt, C., Stoll, C., Seidel, H.P.: Rapid animation of laser-scanned humans. In: IEEE Virtual Reality 2007, Charlotte, USA, IEEE (2007) 223–226
30. Lipman, Y., Sorkine, O., Cohen-Or, D., Levin, D., Rössl, C., Seidel, H.P.: Differential coordinates for interactive mesh editing. In: SMI 2004. (2004) 181–190
31. Stoll, C., Karni, Z., Rössl, C., Yamauchi, H., Seidel, H.P.: Template deformation for point cloud fitting. In: Symposium on Point-Based Graphics. (2006) 27–35
32. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. of the American Mathematical Society **7** (1956) 48–50
33. Zayer, R., Rössl, C., Karni, Z., Seidel, H.P.: Harmonic guidance for surface deformation. In: Proc. of Eurographics 2005. Volume 24. (2005) 601–609
34. Pebay, P.P., Baker, T.J.: A comparison of triangle quality measures. In Proc. of the 10th International Meshing Roundtable (2001) 327–340
35. Huang, J., Shi, X., Liu, X., Zhou, K., Wei, L.Y., Teng, S.H., Bao, H., Guo, B., Shum, H.Y.: Subspace gradient domain mesh deformation. ACM Trans. Graph. **25**(3) (2006) 1126–1134