Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes

Luca Ballan Guido Maria Cortelazzo

Department of Information Engineering
University of Padova
35131 Padova, Italy
E-mail: ballanlu,corte@dei.unipd.it

Abstract

We explore a new approach to marker-less motion tracking of a priori known skinned meshes using both optical flow and silhouette information. We present a formulation which considers in a unified way both these two kinds of information and accounts for the non-rigid deformations of the object skin modeling them using the Skeletal Subspace Deformation (SSD). We then demonstrate the effectiveness of our technique showing its performance in a four camera set-up tracking a subject modeled by a skeleton with 46 degrees of freedom.

1. Introduction

Motion capture finds application in a variety of fields such as character animation in film and game industries, bio-mechanical analysis, ergonomics and surveillance. Current commercial motion capture products typically use optical, magnetic, inertial or mechanical motion capture devices which are rather expensive and force users to wear, in the best case, markers all over their body. Marker-less motion capture systems are a very attractive non-invasive alternative since they are not restricted to motion information associated with markers and weave users from the inconvenience of wearing special garments or devices. Therefore it doesn't come as a surprise that, in the last decade, marker-less motion capture has been a highly active research area. Two recent surveys [12], [13] list over 350 published works on this topic from 2000 to 2006.

Marker-less motion capture methods assume that a subject is observed by a single or by multiple video cameras and that the acquired images are processed in order to estimate the subject's pose at every observation time.

According to [13] the class of approaches which received most attention in the literature is the *Multiple views 3D Pose estimation with Direct Model Use*. This class of methods reconstructs the pose at time t from the pose at time $t-1$ based on an explicit representation of the kinematic structure of the human subjects. The analysis-by-synthesis approach is typically adopted in order to optimize a functional representing the similarity between observed and estimated data. The optimization is generally performed by gradient descent techniques. Other approaches use stochastic tracking techniques like particle filtering [9] essentially for handling abrupt pose changes with the drawback of a considerably increased computational complexity. Recent approaches use stochastic search methods [5] in order to avoid local minima and outliers problems with the added benefit of a considerable computational performance improvement.

The works in this class can be sub-classified according to the type and the domain of the used motion cues. The most used motion cue is silhouette information [3], [15], [5], [11], [14] but also stereo [16], [17] and/or optical flow [18], [19] are exploited.

The motion cue domains found in the literature can either be 3D or 2D. Methods belonging to the first class use the input images in order to get a course estimate of the 3D shape of the viewed human, i.e., the 3D cues. Then they try to fit their own subject model within such a reconstruction. The model's pose which best fits the reconstruction will be the solution to the pose problem. Typically, these methods use visual hull techniques [15], [11], [14]. In this case, the reconstruction quality critically depends on the number of used cameras, since the visual hull is an upper bound of the real surface and the fewer views are available, the higher are likely to be the discrepancies between the visual hull and the model to be estimated. With very few cameras the number of pose ambiguities becomes high and the pose represented by the visual hull may not be recognizable even by an human operator. Visual-hull methods may contribute false boundary information in the presence of concavities of the subject in action. In order to avoid the visual hull

limitations, some methods like [16], [17], [5] add stereo information to deal with concavities and to increase the 3D reconstruction quality. Other methods like [19] use 3D motion field as additional information.

The methods based on 2D motion cues minimize an objective function without directly inferring the 3D of the scene. For instance, [3] defines a silhouette similarity metric based on the XOR operator and minimizes it by Powell's method in order to obtain the pose estimate. The above mentioned reasons behind an imperfect match between the observed data and the estimated model typical of 3D cues do not hold in the case of 2D motion cues.

This paper presents a new approach to the tracking problem belonging to the class of *Multiple views 3D Pose estimation with Direct Model Use*. It is based on a priori known skinned mesh model of the human body, in contrast with other approaches representing the human body as a collection of rigid elements (such as sticks, ellipsoids, segmented body parts, etc...). The major advantage of the skinned mesh representation is that it allows to effectively account for non-rigid body skin deformations. The skinned mesh at a base pose is acquired by a body scanner.

Our method uses only 2D motion cues, namely optical flow and silhouette information. Furthermore, it minimizes an objective function which considers both information and simultaneously accounts for the non-rigid deformations of the body skin. This allows also to estimate small body parts, such as clavicles and spinal bones whose deformations are hard to approximate by rigid sub-elements. The availability of an accurate body model and the use of only 2D motion cues allow a very good fit between estimated and observed data.

Experimental evidence of the effectiveness of the proposed technique in a four camera set-up is provided using a subject modeled by a skeleton with 46 degrees of freedom.

This paper is organized in three parts. Section 2 describes the basic concepts and notation used in this work. Section 3 describes the proposed algorithm, Section 4 shows the experimental results and Section 5 draws the conclusions.

2. Human body model: basic concepts and notation

In this work we assume that the 3D mesh of the human to be tracked is available at a known fixed position. This 3D model is obtained using a body scanner. The human moving over time is modeled as a time-varying 3D mesh supported by an inner skeleton, according to the Skeletal Subspace Deformation (SSD) [4]. This section recalls the concepts of kinematic tree and linear skinning of the SSD and introduces the notation used in this paper.

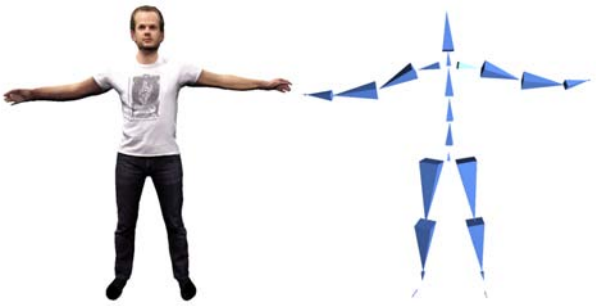

Figure 1. Textured articulated deformable model and its underlying skeleton.

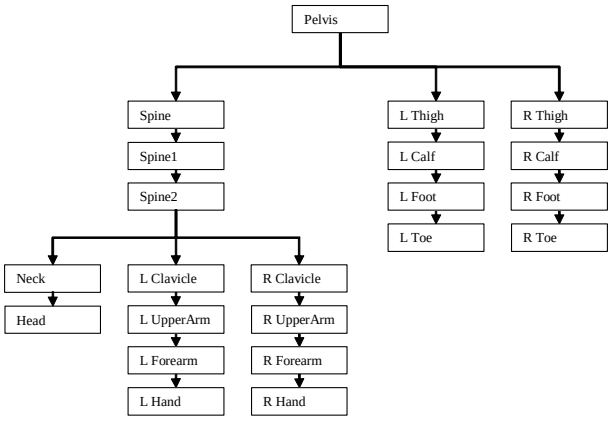

Figure 2. Hierarchical organization of the bones.

A kinematic tree is a set of m reference systems (called *bones* or *links*) organized in a tree structure (see Fig. 1 and Fig. 2). Let $AM_1, \ldots, AM_m$ be the homogeneous matrices associated to every reference system of the kinematic tree and let $\Phi = (\{1, \ldots, m\}, \phi)$ be the graph which relates the reference systems in a tree structure, where ϕ is a tree relationship such that, for a pair of bones h_1 and h_2, $(h_1, h_2) \in \phi$ if and only if h_1 is the father of h_2. The kinematic tree can be represented by the ordered pair $(\{AM_1, \ldots, AM_m\}, \Phi)$. The origins of all the reference systems belonging to a kinematic tree, excluding the one of the root, are called *joints*. A kinematic chain $\Lambda = \{h_1, \ldots, h_l\}$ of Φ is a subset of bones of Φ where $(h_i, h_{i+1}) \in \phi$ for $i = 1, 2, \ldots, l$, i.e., all the bones are father and son of each other.

The configuration (or *state*) of a kinematic tree can be represented by a vector θ of space $\Re^{6m}$ where each axis represents a degree of freedom of a bone in the tree. Note that each bone has six degree of freedom, three for rotation and three for translation. Rotations are parameterized by exponential maps.

It is customary to impose that θ belongs to a set of valid configurations Ω usually shaped as an m-dimensional box, i.e., to impose $\{\theta_{j,min} \leqslant \theta_j \leqslant \theta_{j,max}\}$ with $j = 1, 2, \ldots, m$.

The concept of kinematic tree describes the body skeleton. We model the body exterior surface (the skin) as a *deformable model* Ψ which is a 3D triangular mesh where each vertex v_i, $i = 1, 2, \ldots, n$, can freely move in space over time maintaining smoothness. In particular, the skin is modeled by a (linearly skinned) *articulated deformable model* which is an ordered triplet $(\Psi(\theta), K(\theta), \Gamma)$, where $\Psi(\theta)$ is the deformable model with n vertices described above, $K(\theta)$ is a kinematic tree with m bones and $\Gamma = (\alpha_{i,k})_{i,k}$ is a $n \times m$ matrix where each row has sum equal to one. Both the kinematic tree and the deformable model depend on configuration θ, hence both the absolute matrices of K and the vertices of Ψ depend on θ. The dependence of K from θ is given above, while the dependence of Ψ from θ is given by the following relationship

$$v_i(\theta) = LSK_i(\theta) \cdot v_i(0) \tag{1}$$

where $v_i(\theta)$ are the homogeneous coordinates of the i-th vertex of Ψ at configuration θ and $LSK_i(\theta)$ is the *linear skinning operator* defined as

$$LSK_i(\theta) = \sum_{k=1}^{m} \alpha_{i,k} AM_k(\theta) * (AM_k(0))^{-1} \tag{2}$$

Coefficients $\Gamma = (\alpha_{i,k})_{i,k}$ are called the skinning parameters of Ψ and describe how the mesh deforms itself according to the underlying skeleton configuration. Configuration 0 is called *base pose*. Matrices $\{AM_k(0)\}_k$ are the rigging parameters which describes the initial pose of the kinematic tree. Matrices $\{AM_k(\theta)\}_k$ are the motion parameters since they include motion information affecting both skeleton and mesh. In computer graphics, the process of searching for the $(\alpha_{i,k})_{i,k}$ values is called skinning and that of searching for the $\{AM_k(0)\}_k$ values is called rigging.

3. The Proposed Tracking Algorithm

The proposed motion tracking algorithm addresses the estimate of the motion state $\theta(t)$ of the articulated deformable model, describing the human body in the scene, given the previous motion state $\theta(t-1)$ and the set of the images of the body's action taken at time t and at time $t-1$ from a set of q cameras.

Our algorithm first pre-processes these images in order to extract motion cues, then uses such information in order to define an objective function $g(\theta)$ having its minimum at the current motion state $\theta(t)$. In the end, it minimizes $g(\theta)$ using as starting point for the minimization $\theta(t-1)$.
Objective function $g(\theta)$ is first examined and next the algorithm is described.

3.1. The Objective function

Given a set of z correspondences between the vertices of the articulated deformable model and their projections on a set of q views $\{V_1, \ldots, V_q\}$, let's denote one of such correspondences as (i_s, c_s, p_s), where i_s is the 3D mesh vertex index, c_s is the camera view index[1] and $p_s \in \Re^2$ is the actual projection of v_{i_s} on the view V_{c_s}. Let's denote with $Pr_V(\cdot)$ the projection of a point in the 3D space to a 2D point in the image space of view V, i.e.,

$$Pr_V(\cdot) = \frac{1}{P_V^z(\cdot)} \begin{bmatrix} P_V^x(\cdot) \\ P_V^y(\cdot) \end{bmatrix} \tag{3}$$

where $P_V(\cdot) = (P_V^x(\cdot), P_V^y(\cdot), P_V^z(\cdot))$ is the transformation from world space coordinates to camera space coordinates of V defined as

$$P_V(x) = R_V \cdot x + T_V \tag{4}$$

where $\begin{bmatrix} R_V & T_V \end{bmatrix} = K_V E_V$ and K_V, E_V are respectively the intrinsic and the extrinsic matrix of the view V.

We say that the actual model configuration θ is the one which minimizes the following functional

$$g(\theta) = \sum_{s=1}^{z} \left\| Pr_{V_{c_s}}(v_{i_s}(\theta)) - p_s \right\|^2 \tag{5}$$

[1] $s \in \{1, \ldots, z\}$ where z is the number of founded correspondences. $i_s \in \{1, \ldots, n\}$ and $c_s \in \{1, \ldots, q\}$.

and belongs to Ω, i.e., the set of all the allowed kinematic configurations. Note that, $v_{i_s}(\theta)$ is the same function defined above in Eq. (1).

Constrained minimization of Eq. (5) can be solved by classical gradient descent approaches with hard constraints since the gradient of $g(\theta)$ can be easily derived in closed form.

$$\frac{\partial g(\theta)}{\partial \theta_j} = 2\sum_{s=1}^{z}\left(Pr_{V_{c_s}}(v_{i_s}(\theta)) - p_s\right)\cdot\frac{\partial\left(Pr_{V_{c_s}} \circ v_{i_s}\right)}{\partial\theta_j}(\theta) \tag{6}$$

The Levenberg-Marquardt method [6], [10] was found rather stable and reliable with respect to other methods especially when configuration θ approaches the optimal solution. In this case only the closed form of the jacobian $\partial\left(Pr_{V_{c_s}} \circ v_{i_s}\right)/\partial\theta$ is needed.

3.2. Motion Tracking

The use of the objective functional described in Eq. (5) requires to extract a set of valid correspondences (i_s, c_s, p_s) from the given images. To this aim we used two types of motion cues, namely optical flow and silhouette information.

Optical flow information is extracted by the KLT [8] operator independently applied to each video stream. The result of this process is a large set of 2D correspondences on consecutive frames. Let's call (y_{t-1}, y_t) a pair of such correspondences[2] viewed by the camera c_s. Since $\theta(t-1)$ is known, we can easily find which vertex of the deformable model is projected onto image point y_{t-1}. This can be done by computing the z-buffer and finding the visible mesh vertex with the nearest projection to y_{t-1}. If such a vertex has index i_s, then a valid correspondence for our algorithm is (i_s, c_s, y_t).

Silhouette information is extracted and used similarly to the ICP approach [1] with the difference that our method does not operate in the 3D space but in the image space. Differently than optical flow information, silhouette information is updated also during the minimization procedure by the following method. Assuming that we are in the process of estimating $\theta(t)$ with our tracking algorithm, call $\widetilde{\theta}$ the estimate of configuration $\theta(t)$ at the current algorithm iteration and call $I_{c_s}(t)$ the segmented image of the video stream of camera c_s at time t. For each camera c_s, find all the vertices v_{i_s} of the deformable model with configuration $\widetilde{\theta}$ which belongs to the silhouette viewed from c_s. Project such vertices on the view c_s. For each projected vertex find the closest point of the border of $I_{c_s}(t)$ with similar local characteristics and call it y. Define (i_s, c_s, y) to be a valid correspondence. The concept of border characteristic similarity is determined upon local gradient values.

[2] $y_{t-1}, y_t \in \Re^2$

	frames	mean [%]	st. dev. [%]
walk	390	8.84	0.93
pirouette & jump	490	11.74	2.24
somersault	170	11.05	4.8
hand stand	200	12.03	3.9
press up	280	11.34	1.9

Table 1. Pixel discrepancy error statistics of the tested movements.

Outliers are pruned from the correspondences set by a simple technique which nevertheless performs rather satisfactorily. At each frame the average number of optical flow correspondences is far less than the number of silhouette ones. Typically there are about one hundred of the former versus one thousand of the latter. Therefore, in order to account for such proportions, the algorithm weights by a factor of 10 the contribution of optical flow versus that of silhouettes.

In the end, once obtained a valid set of correspondences, the algorithm defines the function $g(\theta)$ according to Eq. (5) and minimizes it under the constraint $\theta \in \Omega$ using the Levenberg-Marquardt algorithm [7].

Figure 3. The capture environment consists of a set of four calibrated cameras arranged in a blue fabric covered room of about 24 square meters.

4. Experimental results

The acquisition set-up consists of a set of four calibrated cameras (Basler scA1000) arranged in a room of about 24 square meters (see Fig. 3). The cameras are synchronized by an hardware trigger and acquire video streams with a resolution of 1034x778 pixels at 21fps. Used optics consist in three $4.5mm$ lenses and one $3.5mm$ lens.

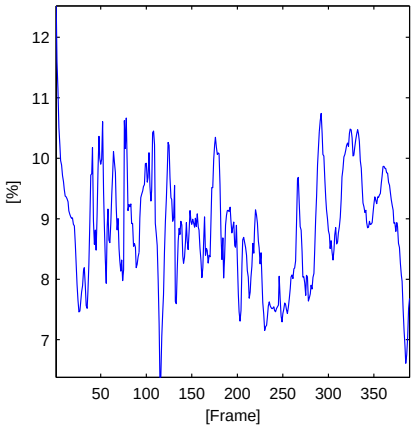

Figure 4. The pixel discrepancy error for each frame of the walking sequence.

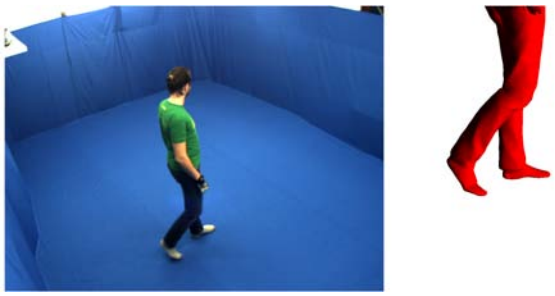

Figure 5. (Left) A frame of the walking sequence. (Right) Detail of the reconstructed model.

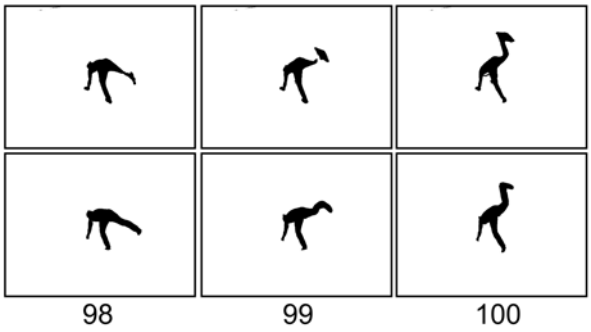

Figure 6. Hand stand sequence. Upper row: extracted silhouettes. Lower row: silhouettes of the estimated 3D model.

Consequently, the average size of a human seen by each camera is about the 5% of the entire frame resolution. The room is covered with blue fabric in order to simplify the silhouette extraction process. To this purpose, we used an HLS chroma-keying technique [21]. Our human models have been acquired by a passive 3D body scanner and their skeleton structure has 46 degrees of freedom arranged in 22 bones as follows: head(3), clavicles(2+2), upper arms(3+3), forearms(1+1), hands(2+2), three spine bones(3+3+1), pelvis(6), thighs(3+3), calves(1+1), foots(2+2) and toes(1+1). Bones are arranged as in Fig. 2 and are constrained to be inside a box shaped set of valid configurations $\Omega \subset \Re^{6\times 22}$.

We evaluated our algorithm on several video sequences with different types of motion such as walking, jumping, break-dancing, pirouettes, somersaults, hand stands and more.

The algorithm performance was validated by both qualitative and quantitative evaluations. Qualitative evaluations are based upon visual comparisons of each reconstructed frame with the original one. Quantitative evaluations are based on the *pixel discrepancy error* (PDE) between the silhouettes extracted from the video stream and the silhouettes of the estimated model. We represent silhouettes (either obtained by segmenting the video stream or by rendering the reconstructed model) as binary images with the convention that background pixels are white pixels (1) and object pixels are black pixels (0). At every frame, a XNOR operator is applied between the extracted silhouette and the silhouette of the reconstructed model. The percentage of black pixels, with respect to the total number of pixels forming the extracted silhouette, is the pixels discrepancy error.
Table 1 reports the PDE statistics over the whole sequence for some of the tested movements. It is worth pointing out that, by definition, the PDE may be due to actual pose estimation errors but also to background subtraction error and to mismatches between the actual 3D shape of the human and the used 3D model. Therefore the values in Table 1 generally overestimate the actual pose estimation error especially in sequences like the somersault and the pirouette where fast movements and ground interaction make the silhouette extraction rather critical. Furthermore let's note that our actors wear casual clothes and have long hair. This increases the mismatches between the actual 3D shape and the used 3D model indeed, for instance, it is very difficult to tie their hair in the same way as during the 3D scanning process. All the above factors contribute to increase the PDE but this is not necessary related to an actual increase of the pose estimation error.

For instance, in order to exemplify the visual meaning of the PDE consider the case of frame 63 of the somersault sequence shown in Fig. 7 (row 2, column 2). It has a PDE of 10.2% which in spite of its considerable value has a visual

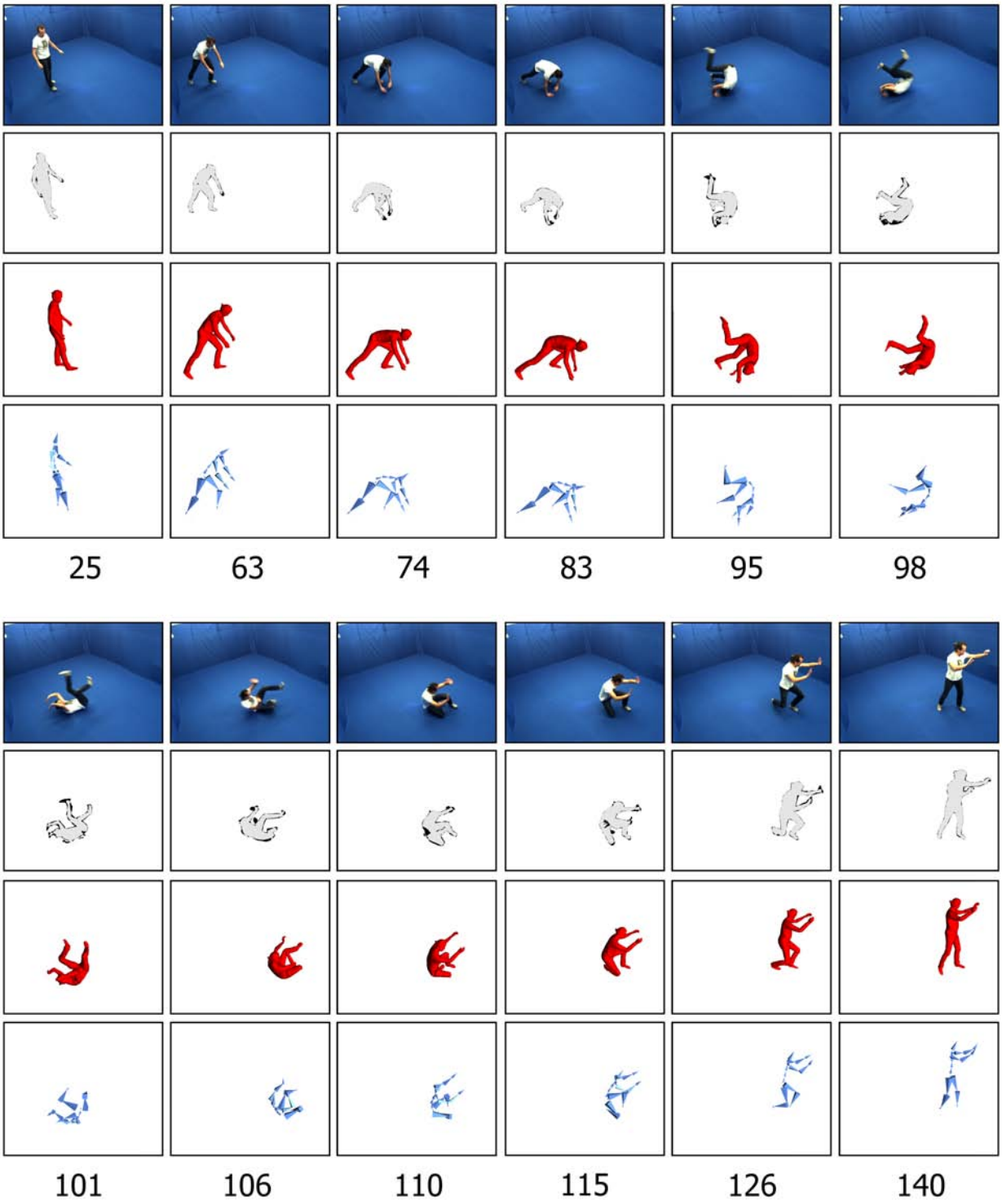

Figure 7. Some frames of the somersault sequence. From top to bottom: real images acquired from one of the cameras of the acquisition system, pixel discrepancy error, reconstructed model and the reconstructed skeleton.

impact rather contained.

In case of simple sequences like the walk (390 frames), where the actor performs a walk around the room and rises one of his legs rotating it, the algorithm can reconstruct it rather accurately both in terms of discrepancy error and in terms of visual quality. Indeed, as shown in Figure 4, the PDE is rather low around its average (8.84%). Figure 5 shows a reconstructed frame of the walk sequence where also small details, such as the right foot articulation, are accurately reproduced.

Fast movements pose a twofold problem. Indeed, in this case the implicit assumption of the tracking algorithm that $\theta(t-1)$ is a good starting point for the minimization does not hold anymore. Furthermore fast movements cause motion blur, because of the finite aperture time of the cameras, which deteriorates the quality of silhouette information typically altering the real object dimensions. These artifacts are clearly shown in the upper row of Figure 6 where the actor leg, during an hand stand, becomes smaller and some parts disappear. However, optical flow information in these situations effectively overcomes such problems since the visual quality of the reconstruction remains remarkably good, as shown in the lower row of Figure 6.

The reconstruction power of the proposed method can be appreciated by the somersault sequence which combines fast movements, spine bending and clavicles rotations and a lot of self occlusion given by the fact that actor bends it self on the ground.

Figure 7 shows some frames of such a sequence. The first rows represents the original video recorded by one of the cameras of the acquisition system. The third row shows the reconstructed actor while the second one shows the PDE between the reconstructed silhouette and the silhouette of the estimated 3D model. In particular, the black values represent the discrepancy errors while both white and light grey values represent agreement between reconstructed and observed data. In particular, a light grey value means that the the reconstructed and observed data are both black while a white value that they are both white. Finally, the forth row represents the moving skeleton.

It is worth noting that thanks to the SSD model the back of the actor is perfectly tracked. Indeed the reconstructed silhouette lies exactly on the real one. Furthermore, this sequence could not be successfully processed without optical flow information since, in force of the fact that the actor bends himself on the ground, a number of bones are not part of any silhouette in any view. optical flow indeed, overcomes the intrinsic limitation of silhouette cues that is, the fact that they supply information only about the body parts which belong to a silhouette.

In this sequence, hands are not correctly tracked, e.g. see frames 83, 106, 115 and 126. This is simply due to the fact that the used human model does not have an accurate description of the shape of its hands and its skeleton does not contain finger bones. Therefore it cannot model widely opened hands such as in frames 83, 106, 115 and 126.

The current implementation of the algorithm is not time optimized. Its running time for processing a single frame depends on the actor's movement speed, i.e., how far the actual solution is from its initial guess. Typical running time on a single core 3.4Ghz Intel P4 with 2Gb of RAM is up to 30 seconds per frame.

In light of the limited number of cameras and the unavoidable occlusions the obtained reconstruction is rather satisfactory.

5. Conclusions

This paper addresses the marker-less motion tracking of the whole human body by a new method which uses an a priori known skinned mesh model and deals with only 2D motion cues. An interesting contribution of this work is a new objective function which considers in a unified way both optical flow and silhouette information and it accounts for the non-rigid deformations of the body skin, within the SSD framework.

Experimental evidence shows that the proposed technique can satisfactorily track complex human movements with a large number of occlusions in a real four cameras environment with models of 46 degrees of freedom. Tests proved that the use of an SSD-based model in the minimization gives the capability of correctly estimating also small body parts, such as clavicles and spinal bones, whose deformations are hard to approximate by rigid sub-elements.

Tests concerning fast human movement, producing images with a lot of motion blur, proved the usefulness of the optical flow cues for a correct motion estimation.

Further work will be devoted to increase the robustness of the proposed method by integrating in it a prediction step based on extended Kalman filtering [20]. This will bring to a more accurate starting point for $\theta\,(t)$ avoiding possible problems related to movements faster than the current frame rate. A further important improvement could be given by the use of the stochastic meta descent algorithm (SMD) [2] in place of the Levenberg-Marquardt method in order to increase the the algorithmic robustness against outliers. Finally, we plan to verify the performance of the method with a lower (3 or 2) number of cameras.

References

[1] P. Besl and H. McKay. A method for registration of 3d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:239–256, 1992.

[2] M. Bray, E. Koller-Meier, N. N. Schraudolph, and L. V. Gool. Stochastic meta-descent for tracking articulated struc-

tures. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, Washington, DC, USA, 2004. IEEE Computer Society.

[3] J. Carranza, C. Theobalt, M.Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Transaction on Computer Graphics*, 22(3), July 2003.

[4] D. Jacka, A. Reid, B. Merry, and J. E. Gain. A comparison of linear skinning techniques for character animation. In *Afrigraph*, pages 177–186. ACM, 2007.

[5] R. Kehl and L. V. Gool. Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding*, 104(2):190–209, 2006.

[6] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathematics*, 2:164–168, 1944.

[7] M. Lourakis. levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++. [web page] `www.ics.forth.gr/~lourakis/levmar/`, Jul. 2004.

[8] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision (darpa). In *Proceedings of the 1981 DARPA Image Understanding Workshop*, pages 121–130, April 1981.

[9] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, London, UK, 2000. Springer-Verlag.

[10] D. Marquardt. An algorithm for least-squares estimation of non-linear parameters. *Journal of the Society of Industrial and Applied Mathematics*, 11:431–441, 1963.

[11] C. Ménier, E. Boyer, and B. Raffin. 3d skeleton-based body pose recovery. In *Proceedings of the 3rd International Symposium on 3D Data Processing, Visualization and Transmission, Chapel Hill*, June 2006.

[12] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[13] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006.

[14] L. Mundermann, S. Corazza, and T. Andriacchi. Accurately measuring human movement using articulated icp with soft-joint constraints and a repository of articulated models. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, June 2007.

[15] K. Ogawara, X. Li, and K. Ikeuchi. Marker-less human motion estimation using articulated deformable model. *IEEE International Conference on Robotics and Automation*, pages 46–51, April 2007.

[16] R. Plänkers and P. Fua. Articulated soft objects for multi-view shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1182–1187, 2003.

[17] F. Remondino, N. D'Apuzzo, G. Schrotter, and A. Roditakis. Markerless motion capture from single or multi-camera video sequence. In *Int. Workshop on Modelling and Motion Capture Techniques for Virtual Environments*, pages 8–12, December 2004.

[18] A. Sundaresan and R. Chellappa. Markerless motion capture using multiple cameras. In *Computer Vision for Interactive and Intelligent Environment*, pages 15–26, November 2005.

[19] C. Theobalt, J. Carranza, M. A. Magnor, and H.-P. Seidel. Combining 3d flow fields with silhouette-based human motion capture for immersive video. *Graphical Models*, 66(6):333–351, 2004.

[20] S. Wachter and H.-H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding*, 74(3):174–192, 10 June 1999.

[21] S. Wright. *Digital Compositing for Film and Video*. Focal Press, May 2006.