Int J Comput Vis (2010) 87: 156–169
DOI 10.1007/s11263-009-0284-3

# Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation

**Stefano Corazza · Lars Mündermann · Emiliano Gambaretto · Giancarlo Ferrigno · Thomas P. Andriacchi**

Received: 8 May 2008 / Accepted: 24 July 2009 / Published online: 2 September 2009


**Abstract** An approach for accurately measuring human motion through Markerless Motion Capture (MMC) is presented. The method uses multiple color cameras and combines an accurate and anatomically consistent tracking algorithm with a method for automatically generating subject specific models. The tracking approach employed a Levenberg-Marquardt minimization scheme over an iterative closest point algorithm with six degrees of freedom for each body segment. Anatomical consistency was maintained by enforcing rotational and translational joint range of motion constraints for each specific joint. A subject specific model of the subjects was obtained through an automatic model generation algorithm (Corazza et al. in IEEE Trans. Biomed. Eng., 2009) which combines a space of human shapes (Anguelov et al. in Proceedings SIGGRAPH, 2005) with biomechanically consistent kinematic models and a pose-shape matching algorithm. There were 15 anatomical body segments and 14 joints, each with six degrees of freedom (13 and 12, respectively for the HumanEva II dataset). The overall method is an improvement over (Mündermann et al. in Proceedings of CVPR, 2007) in terms of both accuracy and robustness. Since the method was originally developed for ≥8 cameras, the method performance was tested both (i) on the HumanEva II dataset (Sigal and Black, Technical Report CS-06-08, 2006) in a 4 camera configuration, (ii) on a series of motions including walking trials, a very challenging gymnastic motion and a dataset with motions similar to HumanEva II but with variable number of cameras.

**Keywords** Markerless motion capture · Tracking · 3D reconstruction · Human body model · Shape from silhouette

## 1 Introduction

Currently, the capture and analysis of human motion includes many areas spanning from the entertainment industry to clinical and sports applications. For entertainment purposes, motion capture data is used to animate 3D characters in movies and games, while clinical applications include diagnosis and treatment of conditions associated with movement pathologies. In sports, motion capture and analysis focuses on injury prevention and improving performances. A more widespread use of motion capture and analysis has been limited because the available methods for accurate capture of three-dimensional human movement require a laboratory environment and the attachment of markers, fixtures or sensors to the skin's surface of the subject's body. Testing in a laboratory environment and attaching markers to the skin is time consuming, expensive, can introduce experimental skin-motion artifacts, and cannot be used in natural setting like sports fields. In addition, marker-based motion capture is very sensitive to skin movement relative to the underlying bone with errors up to of 7° (Leardini et al. 2005) in the estimation of the joint angles between bone segments.

**Electronic supplementary material** The online version of this article (http://dx.doi.org/10.1007/s11263-009-0284-3) contains supplementary material, which is available to authorized users.

S. Corazza (✉) · L. Mündermann · T.P. Andriacchi
Stanford University, Stanford, CA, USA
e-mail: stefanoc@ccrma.stanford.edu

E. Gambaretto · G. Ferrigno
Politecnico di Milano, Milano, Italy

T.P. Andriacchi
Bone and Joint RR&D, VA Palo Alto Hospital, Palo Alto, CA, USA

A markerless system that can address the limitations of marker based methods described above would represent a major breakthrough in the analysis of human motion and greatly expand the application of human motion capture. While developing an accurate markerless system is technically challenging, recent publications (Cheung et al. 2005; Mündermann et al. 2007; Corazza et al. 2006; Rosenhahn et al. 2006; Balan et al. 2007) have demonstrated a computational framework that enables markerless human motion capture with sufficient accuracy, even for the most demanding applications such as biomechanical and clinical analysis.

This paper presents an innovative formulation of the joint constraints that enhances previous approaches for tracking articulated models. For the first time it has the capacity to account for the complete six degrees of freedom (dof) movement of individual joints. The formulation uses an automated method to generate a subject specific model from a continuous database of human body shapes and implemented optimal joint center identification to track the visual hull sequences with a free form mesh subject specific model.

## 2 Previous Work

A large variety of vision-based systems have been proposed for tracking human motion in the past years. An extensive review of these methods can be found in Moeslund et al. (2006). These systems vary in the number of cameras (and camera configuration), the representation of captured data, types of tracking algorithms, and the application to specific or whole body models. Configurations ranged from a single camera (Hogg 1983; Lee and Chen 1985; Wagg and Nixon 2004) to multiple cameras (Gavrila and Davis 1996; Narayanan et al. 1995; Kakadiaris and Metaxas 1998; Kanade et al. 1998). Many different algorithms have been proposed for estimating human motion including constraint propagation (O'Rourke and Badler 1980), optical flow (Yamamoto and Koshikawa 1991), medial axis transformation (Bharatkumar et al. 1994), stochastic propagation (Isard and Blake 1996), search space decomposition based on cues (Gavrila and Davis 1996), statistical models of background and foreground (Wren et al. 1997), silhouette contours (Legrand et al. 1998), annealed particle filtering (Deutscher et al. 2000), silhouette based techniques (Bottino and Laurentini 2001), shape-encoded particle propagation (Moon et al. 2001), and fuzzy clustering process (Marzani et al. 2001). These algorithms typically derive features directly from the 2D image plane(s) (Isard and Blake 1996; Bregler and Malik 1997). When using multiple cameras a 3D representation could be also utilized for estimating human body kinematics (Gavrila and Davis 1996). Most recent works include (Kohli et al. 2008) addressing pose estimation and segmentation simultaneously, using graphic cuts, and (Knossow et al. 2008) using a kinematic parameterization of extremal contours.

The majority of approaches have been model-based in which an a priori model with relevant anatomic and kinematic information was tracked or matched to 2D image planes or 3D representations. Several recent surveys of computer vision approaches have provided additional detail about the state of the art and classified existing methods into different categories (Moeslund et al. 2006; Mündermann et al. 2006; Aggarwal and Cai 1999; Gavrila 1999; Cedras and Shah 1995). The method presented in this paper is categorized as a model based hierarchical tracking approach. The method differs from Bregler and Malik (1997) because the position of the model limb surfaces (3D point positions instead of 2D point motions) is linearized as a function of the models parameters and then a Levenberg Marquard minimization is performed. 3D surfaces are registered (Visual Hulls) with 6 dof joints while in Bregler and Malik (1997) an articulated moving object typically with 1 dof joints from one or more 2D views was tracked. Previous studies used articulated ICP (Iterative Closest Point) type algorithms (Delamarre and Faugeras 1999; Demirdjian 2004) and silhouettes or stereo, while in Niskanen et al. (2005) the authors proposed to register 1 dof joints in a 22 dof 3D model to a quadratic patch in which the 3D surface is reconstructed by combining images coming from multiples views. With respect to Niskanen et al. (2005) the proposed approach uses a subject specific free form mesh to describe the model. Moreover a more complete definition of the joint motion (6 dof) and constraints is provided.

In Rosenhahn et al. (2006) the authors compared two common algorithms for shape registration: the ICP and a level set method enabled through optical flow. The latter describes that the two images need to be registered through two functions whose zero level indicate the object contour. The registration finds the transformation which brings these two functions the closest through a minimization algorithm. The advantage of the zero level formulation is that it has fewer local minima than the ICP algorithm and therefore is more robust in the case of bad (far from the optimum) initializations. However, the ICP algorithm is more robust with noisy data. It shows better accuracy than the level set methods and it can be easily generalized to the case of three-dimensional articulated objects.

In Demirdjian (2004) an articulated ICP method was introduced which operated by running several standard rigid-body ICP on each limb of an articulated model. The motions of individual limbs were then constrained using ball and socket joints to connect adjacent limbs; a linearization of this constraint was performed around the current model configuration. The motion was further constrained into a subspace of the articulated body configurations using support vector machines trained with reference motion capture

sequences. In the present paper we propose to embed the articulated body constraints in the mathematical formulation of the model, in order to avoid the need of projecting an intermediate solution into an articulation-feasible space. In this solution the optimization is run at the same time on all degrees of freedom.

With respect to Mündermann et al. (2007) the presented algorithm represents a significant improvement. The mathematical formulation of Mündermann et al. (2007) is not robust for applications where three degrees of freedom of rotation are needed at the joints. In fact, it solves the minimization problem considering the motion of every body segment in a global coordinate system. Thus, this solution does not allow the definition of local joints motion (referred to the parent body segment), which is the only way to enforce joint angles bounds. Without joint angle range bounds, silhouette-based ICP methods are prone to large errors along the long axis of the body segment. This results in misleading interpretation of the rotational degrees of freedom of the joint along the anatomical axis. Opposed to Mündermann et al. (2007), the presented algorithm provides a method which is both automatic and subject specific for the generation of the model used for tracking.

A robust method for model generation is another critical step in developing an accurate and efficient framework for markerless motion capture. Different from Balan et al. (2007), the model in this work is estimated prior to the tracking since changes in the subject's body shape during motion are negligible with respect to the system resolution. Several approaches for the generation of the model used for tracking have used geometric representation such as ellipsoidal metaballs to approximate body segments (Plankers and Fua 2003; Mikic et al. 2003). However, reconstructing the subject's body shape with the best fidelity requires a free form surface. Lee (Lee et al. 2000) reported a seamless, robust and efficient method to generate a human model although that study did not provide automated identification of the joint centers location. Rosenhahn (Rosenhahn and Klette 2005) used a free form surface to model the upper body that required multiple subject poses. Baran (Baran and Popovic 2007) published a relevant method for automatic rigging of characters even though no quantitative assessment of the methods accuracy has been provided. A seminal work for very accurate model generation used a space of human shapes database (Anguelov et al. 2004). However, the algorithm needed manual initialization of data and model mesh correspondences and has been demonstrated to work only on laser scan quality data. The method presented in this paper addresses both the need for automated joint center locations, initialization of data and mesh correspondence. With respect to Balan et al. (2007), in the proposed method the shape of the subject is estimated prior to tracking. This allows tracking of the motion using a more robust articulated model for the subject and limits the degrees of freedom of the search space. In the presented approach the optimal location of joint centers has been learned from a new set of subjects, which enriches the human body model generated by the SCAPE database.

Very few tracking algorithms have been quantitatively evaluated in the past. Instead qualitative tests and visual inspections have been most frequently used for assessing approaches introduced in the field of computer vision and machine learning. Evaluating existing approaches within a common quantitative framework is essential to establish development guidelines and compare different approaches. This work takes advantage of the common framework provided by the HumanEva II dataset to provide qualitative and quantitative evaluation of a markerless method. The study is also extended to sequences captured in the authors' lab with various camera configurations and a marker based gold standard.

## 3 Methods

The proposed approach tracks a 3D subject-specific articulated model over a series of visual hulls (Laurentini 1994) constructed from multiple camera views.

### 3.1 From Video to 3D Representation

#### *3.1.1 Data Acquisition: Testing Dataset 1*

Forthe the testing dataset 1 full body movements during gymnastics performances and walking trials were captured using eight AVT Pike VGA color cameras ($640 \times 480$ pixels) synchronized at 120 fps (200 fps for the gymnastic trial). Internal and external camera parameters and a common global reference frame were obtained through offline calibration. The captured and tracked motions were a gymnast flip (subject A) and three walking trials (subject B). The accuracy of the markerless method presented was quantified for subject B by comparing the joint centers time histories obtained with both the markerless and marker based systems. The marker based system employed was an 8 camera Qualysis system, operating at 120 Hz. The protocol adopted for the determination of the joint center locations over a whole walking cycle of a subject was the Point Cluster Technique (Andriacchi et al. 1998), a state of the art protocol for marker based analysis. This was done only for the walking trials since marker placement and tracking is very challenging for gymnastics movements.

#### *3.1.2 Data Acquisition: HumanEva II dataset*

The HumanEva II dataset (http://vision.cs.brown.edu/humaneva (Sigal and Black 2006)) was used for a quantitative evaluation that can be compared among different research groups and different approaches.

This data was obtained using 4 VGA color cameras (640 × 480 pixels) synchronized at 60 fps. Internal and external camera parameters as well as a common global reference frame were obtained through offline calibration. The analyzed motions of subjects S2 and S4 ranged from walking to running to balancing. A quantitative assessment of the presented markerless method's accuracy was achieved by comparing the tracked joint centers with the results from a marker based system. The marker based system employed was a 12 camera Vicon system, operating at 120 Hz with a marker protocol proposed by the manufacturer.

#### *3.1.3 Data Acquisition: Testing Dataset 2*

Dataset 2 was used to compare the effect of different multi-camera configurations (12, 10, 8, 6 and 4 cameras) and frame rate on the tracking error. A dataset mimicking the HumanEva II motions (walking, running, balancing) was captured using 12 NAC 1.3 MPixel cameras and capturing at 120 frames per second. The results obtained with the best configuration (12 cameras at 120 fps) were considered as reference to analyze the results obtained with a lower camera number (10, 8, 6 and 4) and a lower frame rate (60 and 30 fps).

#### *3.1.4 3D Representation*

In testing dataset 1 and 2 the subject was separated from the background in each color camera's image sequence using an intensity and color threshold.

In the HumanEva II dataset background subtraction was achieved using the mixture of Gaussian distributions model and using software provided with the dataset. The 3D representation was achieved through visual hull construction from multiple 2D camera views, as described in Laurentini (1994), Mündermann et al. (2005). Visual hulls were generated with 1 cm voxel resolution.

### 3.2 Human Body Modeling

#### *3.2.1 Subject Specific Articulated Model*

An articulated, subject-specific model was created from direct measurement of the subject's outer surface using either a laser scan (subject A, subject B, subject S4) or visual hull (subject S2) frame (Fig. 1). A recently published method (Corazza et al. 2009) that builds on a database of human body shapes (Anguelov et al. 2005) (SCAPE) was integrated in the markerless system to perform and iterate pose-shape registration of the model mesh, and to automatically segment anatomical regions (limbs, torso, head etc.) from the seamless surface mesh of the subject (Fig. 1).

In the SCAPE database, a continuum space of human shapes was learned from a series of full body laser scans. In (1) $D$ accounts for changes in body shape between different individuals. The method (Anguelov et al. 2005) used, deforming triangle edges $v_{k,j} = X_{k,j} - X_{k,1}$, (triangle $k$ is defined by points $(X_{k,1}, X_{k,2}, X_{k,3})$) of a template mesh using up to three transformations. The deformed edge is given by:

$$v'_{k,j} = T_k D_k Q_k v_{k,j}, \quad j = 2, 3, \tag{1}$$

where $T$ specifies a rigid pose transformation and $Q$ a non-rigid transformation specifying changes in pose due conditions such as changes in muscle contraction or tone in different positions. Components in the transformation matrix $D$ express model deformation and take into account changes in body shapes. These components enabled a better fit of the experimental data (lasers scans). PCA was used to reduce the dimensionality of the space of variation of matrix $D$ in order to represent very reasonable variation in weight and height, gender, abdominal fat and chest muscles, and bulkiness with as few as 10–20 principal components. Considering a standard pose, each point $X_i$ on the surface of the model can be represented through an affine transformation function of the $\beta$ principal component vector $X_i = U_i \times \beta + \mu_i$ (where $U_i$ is a 3-by-$N$ matrix and $\mu_i$ is the position of the $i$-th vertex in the standard model). The space of human body shapes was used to develop an iterative pose-shape registration to identify the optimal model comprising anatomical segmentation information. In the shape identification step shape parameters $\beta_j$ are identified, by iteratively minimizing the functional

$$\arg\min_{\beta} \sum_{i=0\ldots N} \|U_i\beta + \mu_i - CP_i\|^2, \tag{2}$$

where, $CP_i$ is the closest point in the input mesh to model point $i$. For the pose identification step (10) was used.

The subject specific model was completed by locating accurate joint center positions relative to the surface mesh. Optimal joint center locations were learned using a 10 subject population to identify the optimal linear combination of mesh vertex coordinates to better express the $(x, y, z)$ position in space of the joint centers. Learning was achieved by solving the linear system to minimize the generalization error, i.e. the joint location error for subjects outside the training set ($k$-off cross validation) (Corazza et al. 2009). The automated model generation framework provides a model with biomechanical fidelity (Fig. 1) for tracking of the visual hull sequence through a multi-segment registration algorithm. Despite the poor quality of the visual hull mesh used to generate the S2 model, the model obtained is both anatomically consistent and in accord with the acquired 3D data (Fig. 1). The overestimation of subject's S2 volume is related to the natural visual hull behavior of overestimating volumes. As a general rule, a more complex anatomical description of the body will provide information with a higher

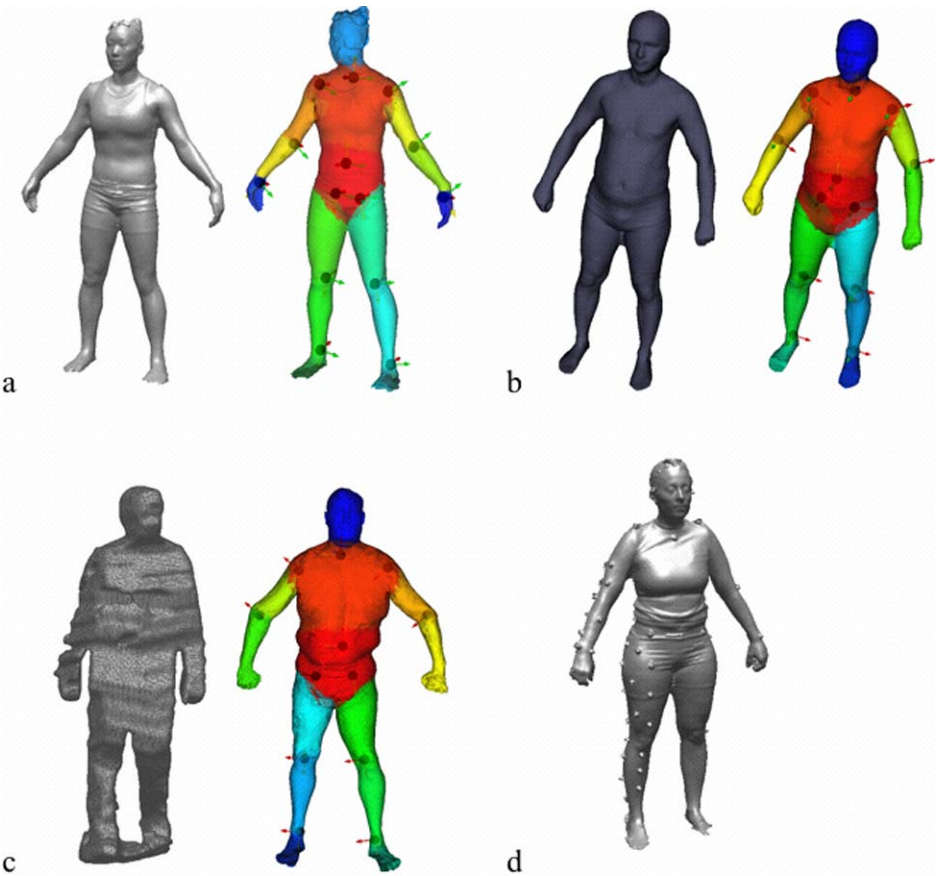


**Fig. 1** The generated Models of subject A (**a**) and subject S4 (**b**) using the laser scans and of subject S2 (**c**) using visual hull. Figure 1d shows the laser scan of subject B in the reference pose with markers, used to register the joint centers locations of marker based and markerless methods, for validation purposes

fidelity to the natural 3D movement, but requires solving for more degrees of freedom in the multi-body registration problem. In this paper the automatically generated models (Fig. 1) included 15 body segments and 14 joints, for a total of 90 degrees of freedom including the rigid body motion of the root segment (13 body segments and 12 joints leading to 78 degrees of freedom for the HumanEva II subjects S2 and S4). For subject B a laser scan of the reference pose (Fig. 1d) with the markers attached was used for the validation process described in the methods section. The method was used to generate the model of S4 made available by the authors for the HumanEva II dataset.

### 3.3 Multi-segment Registration

The implemented articulated ICP algorithm was a generalization of the standard ICP algorithm (Besl and McKay 1992) to articulated models with the introduction of a new mathematical formalism to handle articulated models and joints constraints.

#### *3.3.1 Articulated ICP*

In ICP, at each iteration points from one set are aligned to points on the other one by applying a rigid transformation to one set of points to minimize the squared sum of the distances between the point sets. Convergence to a local minimum is assured and attained after a few iterations. If the sum of the squared distances after applying the rigid transformation is small enough, the ICP should converge to the global minimum.

The algorithm presented here extended the ICP algorithm to articulated objects. Segments were connected with 6 dof joints. This hierarchical organization allowed each joint to propagate motion along the kinematic chain, —unlike other methods (Mündermann et al. 2007) that treated each rigid segment independently—allowing the extraction from the tracker of anatomically meaningful and rigorous data. Each point on the model surface was expressed as a function of joint parameters of all the articulations that connected that point to the root segment of the articulated model. This function was linearized, as described in the following section, in order to solve the registration problem by minimizing the following ICP cost function (expressed over $N$ matched points) through the solution of a series of linear least squares minimization problems:

$$H = \sum_{i=0\ldots N} \| P_i - CP_i \|^2, \tag{3}$$

where, $CP_i$ is a point on the data which corresponds to point $P_i$ on the model. At each iteration $CP$ points were estimated through a closest point criterion with respect to points $P$. The normal on the visual hull surface in proximity of $CP_i$ was compared to the model surface normal in proximity of $P_i$. If these norms differed excessively (typically more than 90 degrees) this corresponding point couple was excluded from the minimization problem.

With such parameterization the absolute position of a point on the model mesh ${}_A P$ was a function of the position of that point with respect to its local frame of reference ${}_L P$, and of the configuration of each joint connecting its correspondent body element to the world reference frame:

$$ {}_A P = \prod_{d=0}^{L} \begin{bmatrix} R_d & t_d \\ \bar{0} & 1 \end{bmatrix} \cdot \begin{bmatrix} {}_L P \\ 1 \end{bmatrix}. \tag{4} $$

Index $d$ goes from the root model segment "0" to segment "$L$", $R_d$ and $t_d$ are the rotational and translational components that describe the motion of the coordinate frame of segment "$d$" with respect to its parent. Translational dof were constrained using a set of inequalities defining a bounding box on a "3×number of body segments" Euclidean space.

The $Z$ axis of each local coordinate frame was defined as parallel to the segment connecting the joint centres of the parent and child segments. For segments at the end of the kinematic chain the orientation of the $Z$ axis of the parent segment was used. Rotations of body segments could then be decomposed in two rotations: first, "swing", moving the $Z$ axis and second, "twist", rotating around the local $Z$ axis. All movements were defined relative to a reference configuration.

The "swing" movement of the $Z$ axis has two degrees of freedom and is parameterized as a point on the surface of a sphere $S^2$. Constraints on this movement are applied by forcing the $Z$ axis to lie inside a sector of the $S^2$ sphere. Instead of defining this sector using a spherical polygon as suggested in Liu and Prakash (2003), for simplicity, it was defined by inequalities applied to the two projected angles parametrizing the $Z$ axis.

The "twist" residual rotation around the $Z$ axis, parametrized through said projected angles is constrained through one inequality. Joint ranges constraints can also be defined in an asymmetric way to account for joints in which the reference configuration does not represent the centre of the range of motion (e.g. knee flexion).

The joint angle ranges are defined according to established anatomical and biomechanical knowledge.

### 3.3.2 Differential of Surface Point Position

As described in the previous paragraph the absolute position of a point in the hierarchical model is a nonlinear function of joint parameters. In order to implement a gradient based optimization method, the differential of each point position was computed. Then the derivative of position point $P$ as a function of rotational and translational parameters of the coordinate frame associated with the generic joint "$\circ$" were computed.

The partial development of (4) leads to:

$$ {}_A P = {}^A_B R({}^B_\circ R({}^\circ_L R_L P + {}^\circ_L t) + {}^B_\circ t) + {}^A_B t, \tag{5} $$

where ${}^A_B R$ is the rotation which expresses frame $B$ (that is the parent of frame $\circ$) with respect to the world reference frame $A$, while ${}^\circ_L R$ expresses the local frame of reference of the segment associated with point $P$ with respect to frame $\circ$. The same considerations are valid for translational parameters ${}^A_B t$ and ${}^\circ_L t$.

In order to compute the differential of point $P$ with respect to rotational parameters ${}^B_\circ R$ of frame $\circ$, a rotational perturbation was applied to this term using the exponential map ${}^B_\circ \delta\omega : \exp({}^B_\circ \delta\hat{\omega}) \cdot {}^B_\circ R$, the exponential map was linearized with $\exp({}^B_\circ \delta\hat{\omega}) = I + {}^B_\circ \delta\hat{\omega}$.

The operator $\hat{}$ denotes the $3 \times 3$ skew symmetric matrix associate with a three-dimensional vector:

$$ \hat{\omega} = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}. $$

The differential of $P$ is then:

$$ \partial_A P' = {}_A P' - {}_A P = {}^A_B R \cdot {}^B_\circ \delta\hat{\omega} \cdot {}^B_\circ R({}^\circ_L R \cdot {}_L P + {}^\circ_L t). \tag{6} $$

In linear regime (small perturbations) the adjoin representation on Lie group SO(3) allows the following substitution ${}^A_B R^{-1} \cdot {}^A_\circ \delta\hat{\omega} \cdot {}^A_B R = {}^B_\circ \delta\hat{\omega}$, rearranging (6) yields to:

$$ \begin{aligned} \partial_A P' &= {}^A_\circ \delta\hat{\omega} \cdot ({}^A_\circ R \cdot {}^\circ_L R \cdot {}_L P + {}^A_\circ t - {}^A_\circ t + {}^A_\circ R \cdot {}^\circ_L t) \\ &= {}^A_\circ \delta\hat{\omega} \cdot ({}_A P - {}^A_\circ t) = {}^A_\circ \delta\hat{\omega} \cdot {}^A_\circ t\overline{P} = {}^A_\circ \delta\omega \times {}^A_\circ t\overline{P} \\ &= -{}^A_\circ t\overline{P} \times {}^A_\circ \delta\omega = -{}^A_\circ t\hat{\tilde{P}} \cdot {}^A_\circ \delta\omega. \end{aligned} \tag{7} $$

Where we noted ${}^A_\circ \overline{tP} = ({}_A P - {}^A_\circ t)$ the vector going from the origin of segment $\circ$ frame to point $P$ (in absolute coordinate) and $\times$ the cross product.

Then, parametrizing the rotation perturbation of joint $\circ$ as an absolute exponential map the differential of point position with respect to the three exponential map parameters leads to the skew symmetric matrix $-{}^A_\circ t\hat{\tilde{P}}$.

The differential of the point position with respect to translational degrees of freedom is immediately found:

$$ \partial_A P = {}^A_B R \cdot {}^B_\circ \delta t = {}^A_\circ \delta t. \tag{8} $$

The total differential of $P$ with respect to all parameters of the articulated model was obtained by summing up the

partial differentials of point $P$ with respect to the joint parameters of all frames connecting the local frame (attached to point $P$) to the world reference frame:

$$d_A P = \sum_{d \in G_P} -{}^A_d t \hat{\tilde{P}} \cdot {}^A_d \delta\omega + {}^A_d \delta t, \tag{9}$$

where, $G_P$ is the ensemble of frames which connects the local frame of $P$ to the world reference.

The cost function in (3) was linearized using the total differential of point position with respect to the model parameters $\delta\omega$ and $\delta t$ (see (9)) leading to the following linear least square minimization problem solved at each iteration of the ICP algorithm:

$$\arg\min_{\delta\omega,\delta t} \left\{ \sum_i \left\| P_i - C P_i + \sum_{d \in G_P} -{}^A_d t \hat{\tilde{P}} \cdot {}^A_d \delta\omega + {}^A_d \delta t \right\|^2 + \mu \cdot \lambda \cdot \|\delta\omega\|^2 + \mu \cdot \|\delta t\|^2 \right\}, \tag{10}$$

where the last two terms are norms of all the rotational and translational dof and are used to force small steps in accord to the Levenberg-Marquardt method. The joint constraints previously described are enforced by projecting every step vector within the feasible region. Typically 20 to 40 iterations of (10) are required for one frame registration. The damping parameter $\mu$ update rule was chosen as specified in Nielsen (1999). Parameter $\lambda$ distributes the damping on rotational and translational dof and it was chosen to be constant $\lambda = 20$ based on experimental evidence.

The results achieved with the presented formulation, which represents a substantial improvement with respect to Mündermann et al. (2007), are shown in the following section where a qualitative and quantitative comparison is provided.

#### *3.3.3 Validation Technique*

To rigorously compare the marker based and markerless methods of estimating joint centre positions and limb rotations during human motion, the initial joint center locations and rigid body segment coordinate systems were registered in a reference pose (Fig. 1d). In the marker based protocol (Andriacchi et al. 1998), a cluster of technical markers (as opposed to anatomical markers, technical markers are typically not positioned in correspondence of specific anatomical landmarks) was associated with each body segment of interest and used to define a segment coordinate frame. In a similar way, for the markerless method, the coordinate frames describe the segment rigid body motion. Joint centres were defined at once for both methods, in a reference pose frame captured using a laser scanner (Cyberware, Monterey, CA) (Fig. 1d). The 3D mesh obtained from the laser scanner was used to generate a subject specific model, comprising joint centre locations as described in Corazza et al. (2009). At the same time the 3D position of every marker was digitized by fitting spheres to the mesh. The joint centre positions were then registered to the coordinate frames described by the markers on each body segment. This procedure eliminated bias related to different joint centre locations definition for markerless and marker based methods, allowing for an appropriate comparison.

Concerning Testing Dataset 2, the mean absolute error considered in the evaluation of the different camera configurations (number of cameras and frame rate) was calculated as the mean of the distances between the joint centres tracked with 12 cameras at 120 fps (used as gold standard) and the joint centres tracked with every other cameras configuration.

## 4 Results

### 4.1 Visual Hulls

Visual hulls for the different sequences analyzed are shown in Fig. 2. Camera number and placement highly influenced the quality of the visual hulls, as previously described in a virtual environment in Mündermann et al. (2005) and experimentally in the present paper. The HumanEva II dataset was

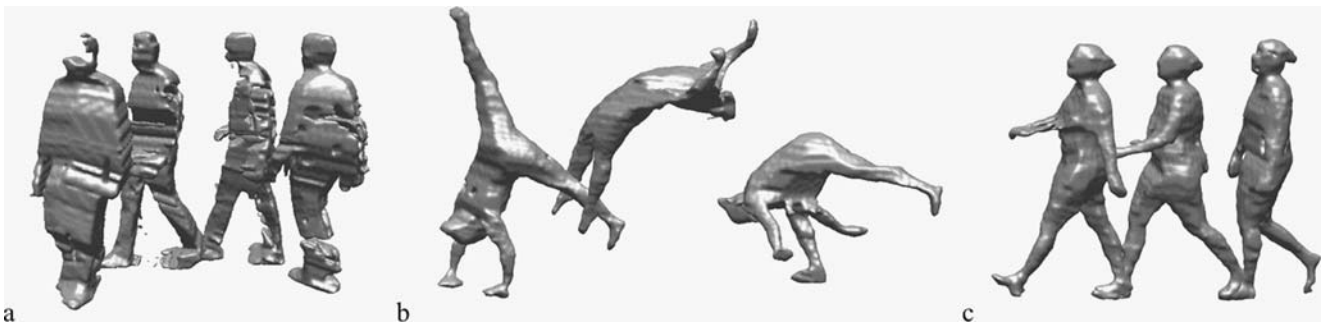


**Fig. 2** Visual hulls obtained from the S2 sequence of HumanEva II dataset (**a**), visual hulls from the test dataset 1 for subject A (**b**) and B (**c**)

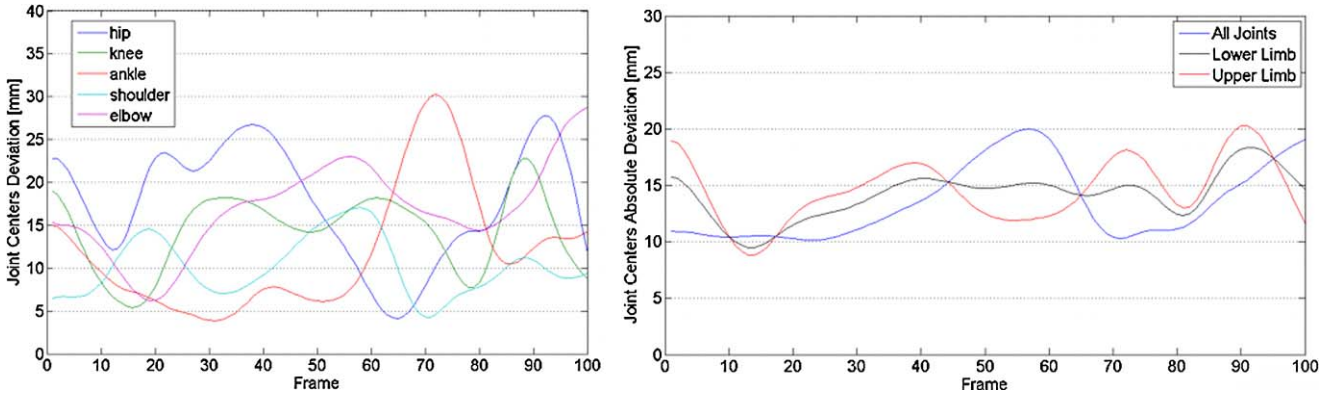


**Fig. 3a** The 3D error results for the test dataset 1. *Left*: The deviation in joint centers locations between marker based and markerless method. *Right*: The average deviations for lower limb, upper limb and all joints

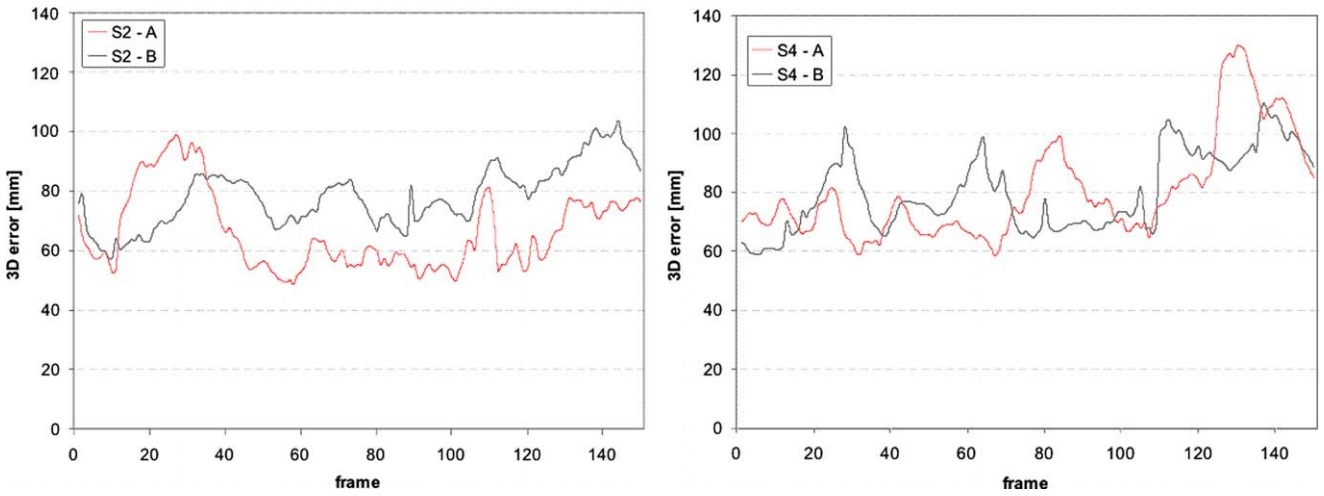


**Fig. 3b** The 3D error results for the HumanEva II dataset. The results (A) were obtained with the method described in Mündermann et al. (2007) while (B) represents results obtained with the presented method

captured with only 4 cameras positioned in a cross configuration, which is the worst placement for visual hull generation. These visual hulls were of much worse quality compared to the ones generated with 8 cameras (Fig. 2b and 2c) placed in a circular pattern. Visual hull generation is the intermediate step for tracking subject motion.

### 4.2 Tracking: Testing Set 1, Walking and Gymnast Flip Sequence

Figure 5 shows the agreement between video data and the calculated model pose for the gymnast flip.

A quantitative comparison was performed over 3 walking trials showing average deviations (Euclidean distance) between joint centers calculated with marker based and markerless systems of 15 mm mean absolute error and 10 mm standard deviation. Individual joint deviations are reported in Table 1 and Fig. 3a.

The results were obtained by comparing marker based state of the art protocol Point Cluster Technique (Andriacchi et al. 1998) with the joint centers location provided by the markerless system. The results also demonstrated excellent robustness for activities involving rapid movements such as the gymnastic flip (Fig. 5), since no re-initialization of the model pose was ever necessary for any of the tracked sequences.

### 4.3 Tracking: HumanEva II Results

In Table 2 the results on the HumanEva II dataset are reported for subject S2 and S4, using both the articulated ICP algorithm described in Mündermann et al. (2007) and the approach described in this paper, which employed rotational bound and described motion with 6 degrees of freedom for every joint of the body. The comparison was done for the first 150 frames of the walking (Fig. 3b) since after that tracking became less robust and both methods start failing

due to the reduced number of cameras employed and the sub-optimal camera configuration.

The results for the two sequences were calculated for all the joints of the body as specified in the HumanEva II dataset, apart from the wrist joints and the joint between the pelvis and the torso. The reason for neglecting the latter was related to its position as determined by the marker based system (ground truth) clearly not representative of the real joint.

In Fig. 4a qualitative comparison between the results obtained with the presented method and with Mündermann et al. (2007) shows how the rotational bounds prevent the tracker to get stuck in configurations that are not anatomically possible.

In Fig. 5 the qualitative results on a challenging gymnastic sequence are shown. In Fig. 6 qualitative results of the tracking of S4 subject are shown. While in the first case (and generally) the algorithm tracked in a completely automatic way, for the HumanEva II Dataset after frame 150 reinitializations were necessary in order to complete the tracking of the sequence, due to poor visual hull quality.

Finally, Fig. 7a and 7b show the effect of the number of cameras and frame rate on the results quality for testing dataset 2, which mimics the HumanEva II dataset motions for a total of 2160 frames. In Fig. 7a the mean absolute errors reported for the configuration with 12 (considered gold standard), 10, 8, 6 and 4 cameras for different frame rates. Figure 7b shows the mean absolute error in the joint centers estimated over the whole sequence for the different camera configurations at 120 fps. For the configuration with 4 cameras the tracking fails around frame 400.

**Table 1** Results on the testing dataset 1, reported as distances between joint centers obtained using marker based and markerless method

| Walking | Mean abs error [mm] | Standard Deviation [mm] |
|---|---|---|
| Average | 15 | 10 |
| Upper | 14 | 9 |
| Lower | 16 | 11 |
| Hip | 16 | 7 |
| Knee | 14 | 7 |
| Ankle | 18 | 6 |
| Shoulder | 9 | 3 |
| Elbow | 19 | 5 |
| Wrist | 15 | 3 |

**Table 2** Results on the HumanEva II dataset in terms of absolute error for subject S2 and S4

| | Walking—AICP with Rot. Bounds | | Walking—AICP no Rot Bounds (Mündermann et al. 2007) | |
|---|---|---|---|---|
| | Mean abs error [mm] | Standard Deviation [mm] | Mean abs error [mm] | Standard Deviation [mm] |
| S2 | 78 | 10 | 73 | 11 |
| S4 | 80 | 13 | 87 | 17 |
| mean | 79 | 11.5 | 80 | 14 |

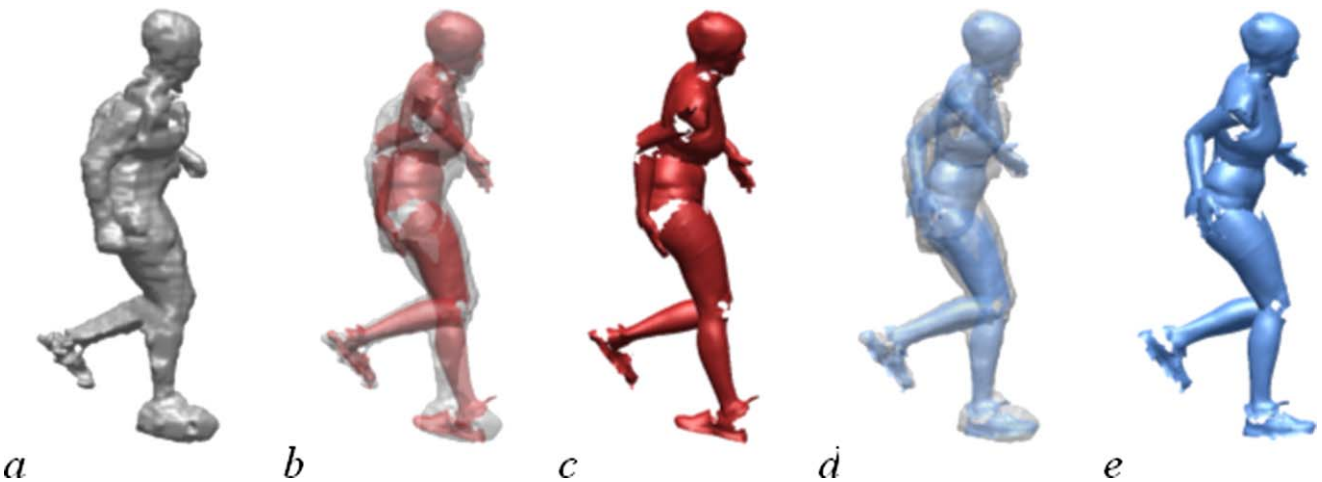


**Fig. 4** An illustration of the Visual hull (**a**). A comparison of the model matched to the visual using the approach in Mündermann et al. (2007) (**b**, **c**), A comparison of the presented approach matched to the visual hull which includes rotational constraints and a LM optimization scheme (**d**, **e**)

**Fig. 5** The video sequence of a gymnast flip (*top*), and calculated model pose for the corresponding frames (*below*)

## 5 Discussion

This study presented the development, application and testing of a novel approach for accurate markerless motion capture (MMC) that maintains anatomic fidelity when capturing human movement under natural conditions. The method represents a complete and automatic solution for markerless motion capture since it combines an accurate and anatomically consistent tracking algorithm with an automatic model generation method. The automatic generation of the model eliminated the most time consuming task in model based markerless motion capture, making the method usable for studies with high number of subjects.

In general the quality of the results depends mainly on the quality of the visual hulls and the quality of the articulated model of the subject. The quality of visual hulls depends on numerous aspects including camera calibration,

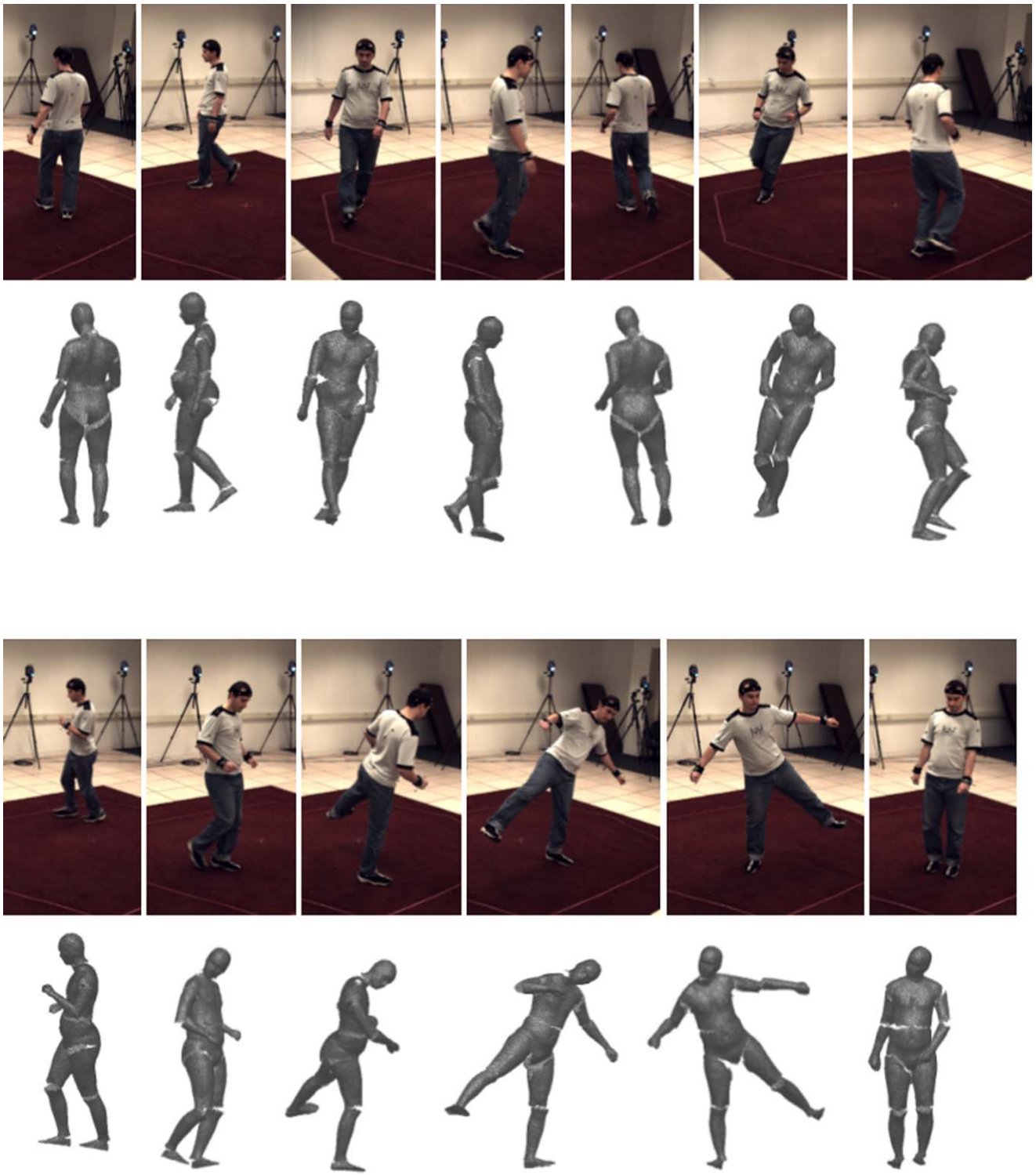

**Fig. 6** The sequence for subject S4, with one frame every 100 frames: video from Camera C1 and corresponding calculated model pose

number of cameras, camera configuration, image resolution and the accurate fore/background segmentation in the image sequences.

In testing dataset 1 provided by the authors, using 8 VGA cameras, qualitative comparison through visual inspection showed very robust and accurate tracking for the gymnastic flip sequence, while the quantitative comparison with the marker based method in a walking trial showed small deviations in the identification of the joint centers with respect to marker based results (average 15 mm deviation). The abil-

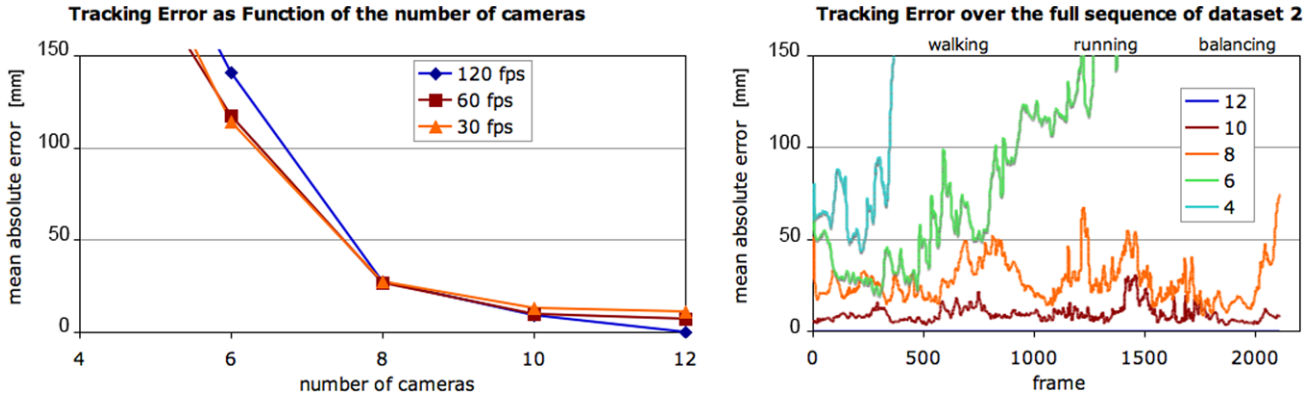


**Fig. 7** (*Left*): The mean error in the estimation of joint centers with different number of cameras and frame rate. (*Right*): The mean absolute error throughout the whole sequence (2160 frames) for different camera number

ity to allow six degrees of freedom at the joints was crucial to provide good tracking. In particular, for joints like the shoulder joint, the usual ball and socket joint model would be inappropriate and would not guarantee the high fidelity in the reconstruction of the human movement shown in Fig. 5. Another advantage of this formulation is that it is possible to define a custom kinematic chain and a specific set of bounds, both translational and rotational, for every single joint of the human body.

Considering the HumanEva II dataset, from a quantitative point of view the errors are sensibly higher for a series of reasons: (i) the cameras configuration (number and position) is not very suitable for methods based on shape from silhouette (visual hull in this case) representation, which require a minimum number of 8 cameras to work optimally (Fig. 7); moreover, having cameras in opposing configuration provides a couple of very similar silhouettes; (ii) the ground truth provided by the HumanEva II dataset is affected itself by large errors due to a number of factors such as (a) the non optimal placement of the markers to identify some of the joint centers, (b) the relative motion both of the clothes with respect to the underlying skin and of the skin with respect to the underlying bone, which is the part ultimately rigidly connected with the joint centers (the variable upon which the validation is performed). Although the absolute error is far from representing the real error generated by the markerless system, the standard deviation of such error can give us an idea about the stability and robustness of the tracking algorithm, showing rather low values if compared to the absolute error. Compared with Mündermann et al. (2007) the method provided slightly better results on the HumanEva II dataset, mainly due to poor quality of the visual hulls which was a prevailing factor over the quality of the tracker. A qualitative example reported in Fig. 4 shows evidence of better performances of the presented method since it prevents anatomically non allowed body poses. This is a very important factor especially to limit the rotations along the long axis of the segment and reduce the number of local minima in the configuration space. Moreover, the computational efficiency of the proposed method allows much faster tracking, with a 20X speed up with respect to Mündermann et al. (2007).

The results on testing dataset 2 showed the importance of the number of cameras in the robustness and accuracy of the presented method. Being based on the visual hull approach the method is not suitable for a configuration with less than 8 cameras and tracking fails with 4 cameras, consistent with what happens with the HumanEva II dataset. Analogous problems were found in the HumanEva II dataset and in the testing dataset 2, where with 4 cameras the tracking does not provide useful results. On the other hand the system is very accurate and robust when it operates with the proper number of cameras. Frame rate seems to be a less important factor in the quality of the tracking, emerging only when dealing with very high accuracies.

The last contribution of the paper is to trace the way for a rigorous comparison between markerless and marker based methods, in terms of (i) definition of optimal marker based protocols to increase the quality of the "gold standard", (ii) registration of the coordinate systems and joint center locations for the two methods in order to highlight the deviations purely due to different tracking results between the two methods.

## 6 Conclusions

This study presented a Markerless Motion Capture (MMC) approach for accurately measuring human motion. The method used multiple color cameras and combined an accurate and anatomically consistent tracking algorithm with a method for the automatic generation of subject specific models. Qualitative and quantitative comparison with different dataset was performed, including the HumanEva II dataset that represents the first large scale worldwide attempt to compare the work by different research groups on

markerless motion capture on a common basis. The quantitative results coming out of the 8 cameras dataset shows very small deviations with respect to marker based methods, demonstrating the effectiveness of the presented approach. With much fewer cameras, like in the HumanEva II conditions, the visual hull method generates much higher errors since it is not particularly suitable for this type of experimental setup, as confirmed by the sensitivity analysis performed on cameras number on testing dataset 2.

**Acknowledgement** Authors want to thank Kat Steele, M.Sc. for her work in the experimental comparison of marker based and markerless methods in testing dataset I. Funding provided by NSF#0325715 and VA#ADR0001129.

## References

Aggarwal, J., & Cai, Q. (1999). Human motion analysis: a review. *Computer Vision and Image Understanding*, *73*(3), 295–304.

Andriacchi, T. P., Alexander, E. J., Toney, M. K., Dyrby, C. O., & Sum, J. A. (1998). A point cluster method for in vivo motion analysis: applied to a study of knee kinematics. *Journal of Biomechanical Engineering*, *120*, 743–749.

Anguelov, D., Koller, D., Pang, H., Srinivasan, P., & Thrun, S. (2004). Recovering articulated object models from 3D range data. In *Proceedings UAI*.

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., & Davis, J. (2005). SCAPE: shape completion and animation of people. In *Proceedings SIGGRAPH*.

Balan, A. O., Sigal, L., Black, M. J., Davis, J. E., & Haussecker, H. W. (2007). Detailed human shape and pose from images. In *Proceedings CVPR*.

Baran, I., & Popovic, J. (2007). Automatic rigging and animation of 3D characters. In *Proceedings of SIGGRAPH*.

Besl, P., & McKay, N. (1992). A method for registration of 3D shapes. *Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 239–256.

Bharatkumar, A. G., Daigle, K. E., Pandy, M. G., Cai, Q., & Aggarwal, J. K. (1994). Lower limb kinematics of human walking with the medial axis transformation. In *IEEE Workshop on Non-Rigid Motion*, Austin, USA (pp. 70–76).

Bottino, A., & Laurentini, A. (2001). A silhouette based technique for the reconstruction of human movement. *Computer Vision and Image Understanding*, *83*, 79.

Bregler, C., & Malik, J. (1997). Tracking people with twists and exponential maps. In *Proceedings CVPR*.

Cedras, C., & Shah, M. (1995). Motion-based recognition: a survey. *Image and Vision Computing*, *13*(2), 129–155.

Cheung, K., Baker, S., & Kanade, T. (2005). Shape-from-silhouette across time part I: Theory and algorithm. *International Journal of Computer Vision*, *62*, 221–247.

Corazza, S., Mündermann, L., Chaudhari, A. M., Demattio, T., Cobelli, C., & Andriacchi, T. P. (2006). A markerless motion capture system to study musculoskeletal biomechanics: visual hull and simulated annealing approach. *Annals Biomedical Engineering*, *34*(6), 1019–1029.

Corazza, S., Gambaretto, E., Mündermann, L., & Andriacchi, T. (2009). Automatic generation of a subject specific model for accurate markerless motion capture and biomechanical applications. *IEEE Transactions on Biomedical Engineering*, in press.

Delamarre, Q., & Faugeras, O. (1999). 3D articulated models and multiview tracking with silhouettes. In *Proceedings ICCV*.

Demirdjian, D. (2004). Combining geometric- and view-based approaches for articulated pose. In *Proceedings ECCV04* (Vol. III, pp. 183–194).

Deutscher, J., Blake, A., & Reid, I. (2000). Articulated body motion capture by annealed particle filtering. In *Proceedings CVPR* (pp. 2126–2133).

Gavrila, D. (1999). The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, *73*(3), 82–98.

Gavrila, D., & Davis, L. (1996). 3-D model based tracking of humans in action:a multiview approach. In *Proceedings CVPR* (pp. 73–80).

Hogg, D. (1983). Model-based vision: a program to see a walking person. *Image and Vision Computing*, *1*, 5.

Isard, M., & Blake, A. (1996). Estimating 3D hand pose using hierarchical multi-label classification. In *Proceedings of 4th European Conference on Computer Vision*, Cambridge, UK.

Kakadiaris, I. A., & Metaxas, D. (1998). Three-dimensional human body model acquisition from multiple views. *International Journal of Computer Vision*, *30*, 191.

Kanade, T., Saito, H., & Vedula, S. (1998). *The 3D Room: Digitizing time-varying 3D events by synchronized multiple video streams* (Tech. report CMU-RI-TR-98-34). Robotics Institute, Carnegie Mellon University.

Knossow, D., Ronfard, R., & Horaud, R. P. (2008). Human motion tracking with a kinematic parameterization of extremal contours. *International Journal of Computer Vision*, *79*(2), 247–269.

Kohli, P., Rihan, J., Bray, M., & Torr, P. H. S. (2008). Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *International Journal of Computer Vision*, *79*(3), 285–298.

Laurentini, A. (1994). The Visual Hull concept for silhouette base image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*, 150–162.

Leardini, A., Chiari, L., Della Croce, U., & Cappozzo, A. (2005). Human movement analysis using stereophotogrammetry. Part 3: Soft tissue artifact assessment and compensation. *Gait and Posture*, *21*, 221–225.

Lee, H. J., & Chen, Z. (1985). Determination of 3D human body posture from a single view. *Computer Vision, Graphics, and Image Processing*, *30*, 148–168.

Lee, W., Gu, J., & Magnenat-Thalmann, N. (2000). Generating animatable 3D virtual humans from photographs. In *Proceedings Computer Graphics Forum—Eurographics* (pp. 1–10).

Legrand, L., Marzani, F., & Dusserre, L. (1998). A marker-free system for the analysis of. movement disabilities. *Medinfo*, *9*, 1066–1070.

Liu, Q., & Prakash, E. C. (2003). The parametrization of joint rotation with the unit quaternion. In *Proceedings of 7° Digital Image Computing*.

Marzani, F., Calais, E., & Legrand, L. (2001). A 3-D marker-free system for the analysis of movement disabilities—an application to the legs. *IEEE Transactions on Information Technology in Biomedicine*, *5*(1), 18–26.

Mikic, I., Trivedi, M., Hunter, E., & Cosman, P. (2003). Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, *53*, 199–223.

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

Moon, H., Chellappa, R., & Rosenfeld, A. (2001). 3D object tracking using shape-encoded particle propagation. In *Proceedings ICCV*.

Mündermann, L., Corazza, S., Chaudhari, A. M., Alexander, E. J., & Andriacchi, T. P. (2005). Most favorable camera configuration for a shape-from-silhouette markerless motion capture system for biomechanical analysis. *Proceedings of SPIE-IS&T Electronic Imaging*, *5665*, 278–287.

Mündermann, L., Corazza, S., & Andriacchi, T.P. (2006). The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of Neuroengineering and Rehabilitation*, *3*(1).

Mündermann, L., Corazza, S., & Andriacchi, T. (2007). Accurately measuring human movement using articulated ICP with soft-joint constraints and a repository of articulated models. In *Proceedings of CVPR*.

Narayanan, P. J., Rander, P., & Kanade, T. (1995). *Synchronous capture of image sequences from multiple cameras* (Technical Report CMU-RI-TR-95-25). Robotics Institute, Carnegie Mellon University.

Nielsen, H. B. (1999). *Damping parameter in Marquardt's method* (Technical Report IMM-REP-1999-05). Technical University of Denmark.

Niskanen, M., Boyer, E., & Horaud, R. (2005). Articulated motion capture from 3-D points and normals. In *Proceedings of BMVC'05*.

O'Rourke, J., & Badler, N. I. (1980). Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2*, 522–536.

Plankers, R., & Fua, P. (2003). Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 1182–1187.

Rosenhahn, B., & Klette, R. (2005). Automatic human model generation. *Computer Analysis of Images and Patterns*, 230–237.

Rosenhahn, B., Brox, T., Kersting, U. G., Smith, A. W., Gurney, J. K., & Klette, R. (2006). A system for marker-less motion capture. *Künstliche Intelligenz (KI)*, *1*, 45–51.

Sigal, L., & Black, M. J. (2006). *HumanEva: synchronized video and motion capture dataset for evaluation of articulated human motion* (Technical Report CS-06-08). Brown University.

Wagg, D. K., & Nixon, M. S. (2004). Automated markerless extraction of walking people using deformable contour models. *Computer Animation and Virtual Worlds*, *15*, 399–406.

Wren, C. R., Azarbayejani, A., Darrell, T., & Pentland, A. P. (1997). Pfinder—real-time tracking of the human body. *Transactions on Pattern Analysis and Machine Intelligence*, *19*, 780–785.

Yamamoto, M., & Koshikawa, K. (1991). Human motion analysis based on a robot arm model. In *Proceedings Computer Vision and Pattern Recognition*.