Analyzing and Evaluating Markerless Motion Tracking Using Inertial Sensors

Andreas Baak[1], Thomas Helten[1], Meinard Müller[1],
Gerard Pons-Moll[2], Bodo Rosenhahn[2], Hans-Peter Seidel[1]

[1]Saarland University & MPI Informatik, Germany. [2]Leibniz Universität Hannover, Germany.

Abstract. In this paper, we introduce a novel framework for automatically evaluating the quality of 3D tracking results obtained from markerless motion capturing. In our approach, we use additional inertial sensors to generate suitable reference information. In contrast to previously used marker-based evaluation schemes, inertial sensors are inexpensive, easy to operate, and impose comparatively weak additional constraints on the overall recording setup with regard to location, recording volume, and illumination. On the downside, acceleration and rate of turn data as obtained from such inertial systems turn out to be unsuitable representations for tracking evaluation. As our main contribution, we show how tracking results can be analyzed and evaluated on the basis of suitable limb orientations, which can be derived from 3D tracking results as well as from enhanced inertial sensors fixed on these limbs. Our experiments on various motion sequences of different complexity demonstrate that such limb orientations constitute a suitable mid-level representation for robustly detecting most of the tracking errors. In particular, our evaluation approach reveals also misconfigurations and twists of the limbs that can hardly be detected from traditional evaluation metrics.

1 Introduction

In the field of computer vision, markerless motion capturing (mocap) with the objective to estimate 3D pose information of a human actor from image data is a traditional field of research in computer vision [2, 4, 21, 26, 34]. Even though motion capturing has been an active research field for more than two decades [14], recent tracking procedures still tend to produce many tracking errors. In particular, when dealing with more involved settings like only few cameras, difficult lighting conditions, or challenging motion sequences, tracking errors are likely to occur.

In the process of developing and improving tracking algorithms, the analysis and evaluation of tracking results play a crucial role. In practice, the tracking results are often evaluated by manually inspecting the reconstructed 3D motion sequences or by looking at the differences between the 2D projections of these sequences and the original image data. To automate and objectify the evaluation process, one requires independent ground truth 3D information used for evaluation in addition to the image sequences. So far, only few benchmark data sets with non-synthetic data such as [27] are publicly available that make a fully automatic evaluation possible. Such benchmark data sets are generated by running a marker-based optical mocap system as a reference in addition to a multiview video camera system. However, marker-based mocap

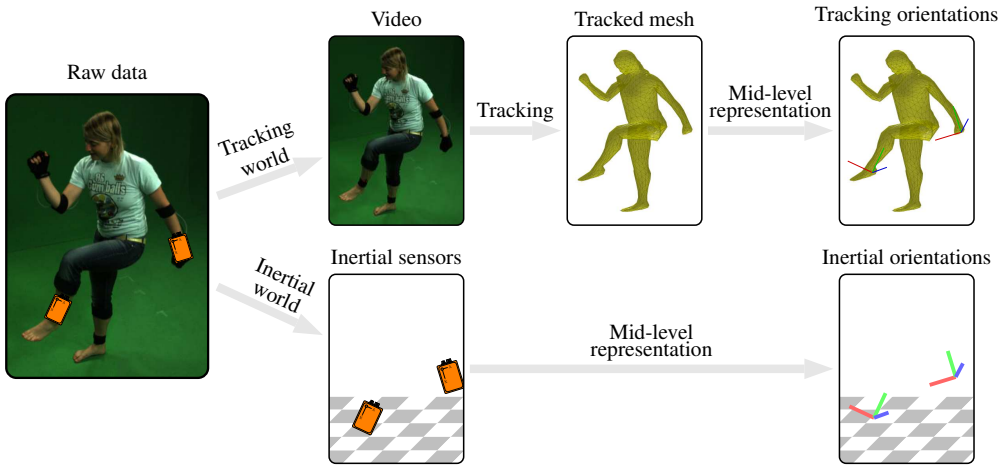

Fig. 1. To compare data of the inertial and the tracking world, orientation data turns out to be a suitable common mid-level representation.

systems are costly and inconvenient to set up, and typically pose additional constraints on the recording environment (e. g., illumination, volume, indoor). As an alternative to recording human motions with real cameras, rendering software can be used to generate synthetic semi-realistic images, yielding a ground truth representation in a natural way [1]. However, these images do not represent real recording scenarios well.

In this paper, we present an approach for automatically analyzing and evaluating 3D tracking results using an inertial-based sensor system to generate suitable reference information. In the following, to clearly distinguish between these two types of data, we speak of the *tracking world* to refer to data derived from markerless motion tracking, and we speak of the *inertial world* to refer to data derived from an inertial system, see also Fig. 1. In contrast to marker-based reference systems, inertial sensors impose comparatively weak additional constraints on the overall recording setup. Furthermore, inertial systems are relatively inexpensive as well as easy to operate and maintain. On the downside, the acceleration and rate of turn data obtained from such inertial systems cannot be directly compared with the tracking result which is given in form of 3D positions or joint angles. To make such data comparable, one could integrate the inertial data to obtain 3D positional data. Integration, however, is not practical since inertial data is prone to noise. Even small potions of noise accumulate during numerical integration, leading to diverging positional data [31]. On the other hand, one could differentiate the 3D positional data of the tracking result to obtain velocities and accelerations. Such data, however, is very local in nature with respect to the temporal dimension, making comparisons on this level susceptible to short-time artifacts and unwanted outliers.

Contributions. As the main contribution of this paper, we introduce a novel inertial-based evaluation framework, where we use orientation data as a kind of common mid-level representation. On the one hand, we derive *tracking orientations* of certain limbs from the estimated 3D pose parameters given by a tracking result. On the other hand,

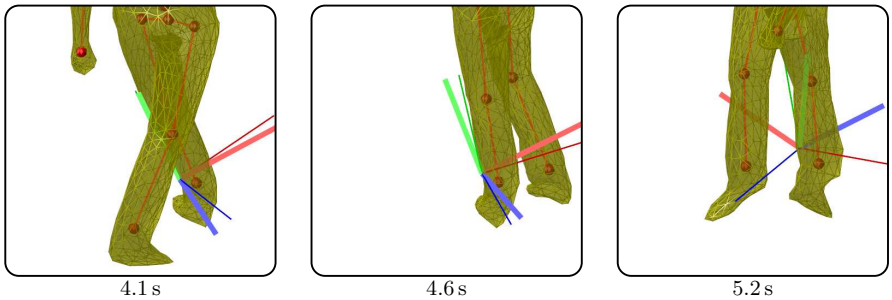

Fig. 2. Snapshots of a tracking result at the given timestamps of the tracked sequence. Basis axes of the limb coordinate systems of the left lower leg are drawn, once extracted from the tracking result (thin axes, dark colors), and once from an enhanced inertial sensor (bold axes, light colors).

we use enhanced inertial sensors rigidly attached to suitable limbs to derive *inertial orientations*. Introducing a robust calibration scheme, we show how these two types of orientations can be used to reliably detect tracking errors in markerless motion tracking. In contrast to using velocities and accelerations, our orientation-based approach particularly suits this purpose since typical tracking errors stem from misconfigurations of certain limbs that effect the tracking result over an entire period of time rather than occurring at certain instances of time.

Standard error metrics are based on Euclidean distances between positions of joints or markers which reflect positional errors fairly well. However, orientation errors, in particular misestimated rotations of cylindrical limbs, can lead to small deviations in the Euclidean distance metric. Moreover, these tracking errors are difficult to spot from visual cues. By contrast, our evaluation approach reveals twists of rotationally symmetric body parts by an orientation-based distance metric.

The remainder of this paper is organized as follows. After discussing related work in Sect. 2, we discuss in Sect. 3 the two types of orientation data. In particular, we introduce a robust calibration method for making the tracking orientations and inertial orientations comparable. In Sect. 4, we present our evaluation framework. Furthermore, we report on extensive experiments conducted on the basis of 24 different motion sequences exemplarily using a state-of-the-art markerless tracking system. Finally, conclusions and prospects on future work are given in Sect. 5.

2 Related Work

To the best of the authors' knowledge, this is the first approach for evaluating markerless tracking using inertial sensors. However, there are several papers that deal with the combination of inertial sensors and cameras. Works in this field have in common that the relative offset between both systems has to be obtained as a subtask. Starting with works in robotics [6, 17, 24, 29], this task has also been approached in the vision community, e. g., [22]. Also, [11] identifies the task with the gray-box problem in the area of system identification. Application scenarios include the estimation of an offset

between a robot's end effector and a visual sensor attached to it [24, 29], or between an inertial sensor and a camera [11, 22]. Analytically, both scenarios can be described by the *hand-eye calibration* equation $AX = XB$, to which we relate our work in Sect. 3.

For motion tracking, [18] uses orientation data obtained from a small set of inertial sensors attached to the outer extremities to stabilize a markerless motion tracking approach. The authors, however, do not discuss the option of using inertial data to evaluate a purely markerless tracking approach. Moreover, they do not discuss the essential step of spatial alignment of both worlds, to which we present a solution in this paper.

For activity recognition, [13] evaluates the influence of sensor displacement on a certain body limb for recognition performance and proposes a heuristic for improving detection results when the exact sensor position on the body limb is not known. Recently, CMU made a multi-modal activity database publicly available, also containing inertial data [5]. In biomedics, the authors of [7] use inertial sensors fixed on a lower leg to reconstruct the one-dimensional knee angle in the sagittal plane. To study biomechanical properties of outdoor activities, GPS information can be combined with inertial sensors [3]. Using a combination of inertial, magnetometer and GPS information, [8] shows that accurate position estimation for pedestrians is possible. In [28], the motion of an arm model is reconstructed using inertial sensors [30]. [28] retrieves motions from a database using few inertial sensor signals to obtain a full body motion. Using only inertial and magnetic sensors, [19] shows that a full body motion can be reconstructed. Having many sensors in a custom motion capture suit [33], a plausible motion model in everyday surroundings can be reconstructed. For home entertainment, inertial sensors are used actively in the recent years. User interfaces based on such sensors have been studied, e. g., in [23].

3 Orientation Data

After introducing the basics of orientation data, we describe how to obtain inertial as well as tracking orientation data (Sect. 3.2). In Sect. 3.3, we present a robust and efficient solution for the calibration problem that one needs to solve to make the two types of orientation data comparable.

3.1 Basics

Suppose a fixed global coordinate system F^{G} that is represented by a right-handed orthonormal basis (like all coordinate systems in this paper). Furthermore, suppose a local coordinate system F^{L} that is moving for a static observer in F^{G}. The relative orientation of F^{L} with respect to F^{G} can be modeled as a rotation. Given the basis vectors X^L, Y^L, and $\mathrm{Z}^L \in \mathbb{R}^{3\times 1}$ of F^{L} in coordinates of F^{G}, the rotation is defined by a rotation matrix with column vectors X^L, Y^L, and Z^L.

In the following, we represent a rotation (or orientation) by a unit length quaternion $\mathsf{q} \in \mathbb{R}^4, \|\mathsf{q}\|_2 = 1$, which is a more compact representation than rotation matrices, see [9, 25]. The composition of two rotations represented by q_1 and q_2 is then given as the composition $\mathsf{q}_2 \circ \mathsf{q}_1$. Furthermore, the inverse rotation of q is given by the quaternion conjugate $\overline{\mathsf{q}}$. Further, we define a distance function $d_{\mathrm{quat}} : \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R}$,

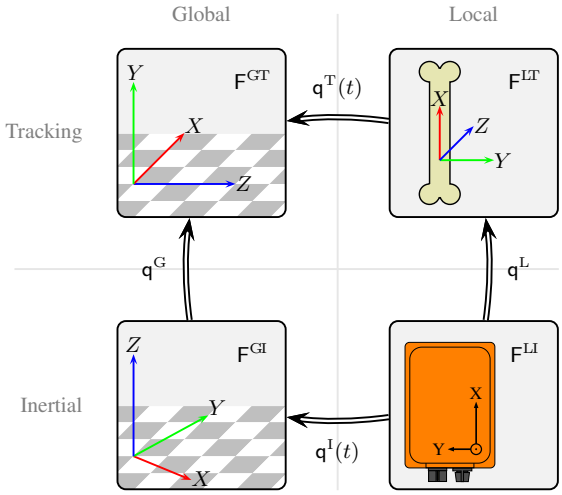

Fig. 3. Relation between the orientation of coordinate systems.

$d_{\mathrm{quat}}(\mathsf{q}_1, \mathsf{q}_2) = {}^{360}/_{\pi} \cdot \arccos \|\langle \mathsf{q}_1, \mathsf{q}_2 \rangle\|$, which denotes the angle in degrees between the rotations defined by q_1 and q_2, see [12] for a proof. We use the notation

$$\mathsf{F}^{\mathrm{A}} \overset{\mathsf{q}}{\Longrightarrow} \mathsf{F}^{\mathrm{B}} \tag{1}$$

to describe a transformation of coordinate systems F^{A} to F^{B} using the rotation defined by q. For time dependent quantities we append a discrete frame index (t) and assume that co-occurring quantities are subject to the same sampling rate.

3.2 Obtaining Two Types of Orientation Data

We now describe how to obtain orientation data in the inertial as well as in the tracking world. In the *inertial world*, as described in [10], an orientation estimation device can be used to measure its orientation in a static global coordinate system $\mathsf{F}^{\mathrm{GI}} = (\mathrm{X}^{GI}, \mathrm{Y}^{GI}, \mathrm{Z}^{GI})$. In this coordinate system, the Z^{GI} axis points to the negative gravity direction, the X^{GI} direction is the orthogonalized direction of the magnetic North, and Y^{GI} is chosen to form an orthonormal right-handed basis. Measurements of accelerometers, gyroscopes, and a magnetic field sensor are fused in a Kalman filter method, which provides drift free estimates of the sensor's orientation $\mathsf{q}^{\mathrm{I}}(t)$. This orientation maps from the sensor's local coordinate system F^{LI} to F^{GI}, see Fig. 3. We refer to $\mathsf{q}^{\mathrm{I}}(t)$ with the term *inertial orientation*. In our experiments, we use an orientation estimation device MTx provided by Xsens [35].

In the *tracking world*, a global coordinate system F^{GT} is defined by camera calibration. Tracking results are typically given by a mesh-based surface representation for every frame in coordinates of F^{GT}. To obtain the orientation of a certain limb in the mesh, one needs to define a local coordinate system F^{LT} that is rigidly attached to the limb. By selecting three non-collinear vertices of the limb, an orthonormal basis of F^{LT}

can be build. To ensure that the coordinate system is well defined, one has to claim one-to-one vertex correspondence throughout the entire motion sequence. In many cases, tracking results are given as joint angles of a skeletal kinematic chain, which drives the animation of the mesh surface. In this case, without having to resort to the vertices of the mesh surface, a local coordinate system for every limb can be defined by forward kinematics [15]. This way, a *tracking orientation* $\mathsf{q}^{\mathrm{T}}(t)$ can be obtained, see Fig. 3.

In order to make $\mathsf{q}^{\mathrm{I}}(t)$ and $\mathsf{q}^{\mathrm{T}}(t)$ comparable, one needs a correspondence between the global coordinate systems F^{GI} and F^{GT} as well as between the two local coordinate systems F^{LI} and F^{LT}. These correspondences, however, are generally not known. The global inertial coordinate system F^{GI} is defined by physical quantities, whereas F^{GT} is defined by an arbitrary placement of a calibration cube in the recording volume. Let q^{G} denote the resulting offset. Furthermore, the local coordinate system F^{LI} is defined by the placement of the sensor on the human actor, whereas F^{LT} is defined either by mesh vertices or by means of a kinematic chain. Let q^{L} denote the resulting offset. The determination of q^{G} and q^{L} is referred to as calibration, which is a tedious and error-prone task when done manually. Therefore, automated calibration is an important concern that we deal with in Section 3.3.

3.3 Calibration and Error Measure

In this section we describe a robust method how q^{L} and q^{G} can be obtained. We show that the described problem is closely related to the prominent *hand-eye calibration* task in robotics [32]. The orientation $\mathsf{q}^{\mathrm{I}}(t)$ can be described by two distinct compositions of rotations in the diagram of Fig. 3, once with tracking and once with inertial orientations:

$$\mathsf{F}^{\mathrm{LI}} \xrightarrow{\mathsf{q}^{\mathrm{L}}} \mathsf{F}^{\mathrm{LT}} \xrightarrow{\mathsf{q}^{\mathrm{T}}(t)} \mathsf{F}^{\mathrm{GT}} \xrightarrow{\overline{\mathsf{q}^{\mathrm{G}}}} \mathsf{F}^{\mathrm{GI}} \qquad \mathsf{F}^{\mathrm{LI}} \xrightarrow[\mathsf{q}^{\mathrm{I}}(t)]{} \mathsf{F}^{\mathrm{GI}} \tag{2}$$

With quaternion algebra, this equality can be expressed as

$$\mathsf{q}^{\mathrm{I}}(t) = \overline{\mathsf{q}^{\mathrm{G}}} \circ \mathsf{q}^{\mathrm{T}}(t) \circ \mathsf{q}^{\mathrm{L}} \; . \tag{3}$$

Now, we can express the rotation that is needed to transform F^{LI} at frame s to F^{LI} at frame t. In Fig. 3, there are two distinct compositions of rotations, starting at t in F^{LI} and ending at s in F^{LI}:

$$\mathsf{F}^{\mathrm{LI}} \xrightarrow{\mathsf{q}^{\mathrm{L}}} \mathsf{F}^{\mathrm{LT}} \xrightarrow{\mathsf{q}^{\mathrm{T}}(t)} \mathsf{F}^{\mathrm{GT}} \xrightarrow{\overline{\mathsf{q}^{\mathrm{G}}}} \mathsf{F}^{\mathrm{GI}} \xrightarrow{\mathsf{q}^{\mathrm{G}}} \mathsf{F}^{\mathrm{GT}} \xrightarrow{\overline{\mathsf{q}^{\mathrm{T}}(s)}} \mathsf{F}^{\mathrm{LT}} \xrightarrow{\overline{\mathsf{q}^{\mathrm{L}}}} \mathsf{F}^{\mathrm{LI}} \qquad \mathsf{F}^{\mathrm{GI}} \xrightarrow[\mathsf{q}^{\mathrm{I}}(t)]{} \qquad \xrightarrow[\overline{\mathsf{q}^{\mathrm{I}}(s)}]{} \mathsf{F}^{\mathrm{GI}} \tag{4}$$

Here, tracking orientations in the upper path and inertial orientations in the lower path are used. The equality of the paths can be expressed with quaternion algebra, where the offset q^{G} cancels out:

$$\overline{\mathsf{q}^{\mathrm{I}}(s)} \circ \mathsf{q}^{\mathrm{I}}(t) = \overline{\mathsf{q}^{\mathrm{L}}} \circ \overline{\mathsf{q}^{\mathrm{T}}(s)} \circ \mathsf{q}^{\mathrm{T}}(t) \circ \mathsf{q}^{\mathrm{L}} \; . \tag{5}$$

Substituting $\mathsf{q}^{\mathrm{A}} := \overline{\mathsf{q}^{\mathrm{I}}(s)} \circ \mathsf{q}^{\mathrm{I}}(t)$, $\mathsf{q}^{\mathrm{B}} := \overline{\mathsf{q}^{\mathrm{T}}(s)} \circ \mathsf{q}^{\mathrm{T}}(t)$, and $\mathsf{q}^{\mathrm{X}} := \overline{\mathsf{q}^{\mathrm{L}}}$, we get

$$\mathsf{q}^{\mathrm{A}} \circ \mathsf{q}^{\mathrm{X}} = \mathsf{q}^{\mathrm{X}} \circ \mathsf{q}^{\mathrm{B}} \; . \tag{6}$$

In robotics, a more general equation of the same form, in which homogeneous transformations are used instead of sole rotations, describes the hand-eye calibration problem. Manifold solutions to this problem have been published, see, e. g. [29] and references therein. Unique solutions can be found as soon as two measurements of q^{A} and q^{B} are available. However, in the presence of noise, an approximate solution using many measurements is preferable to diminish the influence of measurement errors. Therefore, we suggest to use $N \gg 2$ measurements based on a calibration tracking result. The solution of

$$\underset{\mathsf{q}^{\mathrm{X}}}{\operatorname{argmin}} \sum_{n \in [1:N]} \| \mathsf{q}_n^{\mathrm{A}} \circ \mathsf{q}^{\mathrm{X}} - \mathsf{q}^{\mathrm{X}} \circ \mathsf{q}_n^{\mathrm{B}} \| \tag{7}$$

yields a best approximate solution under the Euclidean norm. In [17], Park and Martin present an efficient and easy to implement solution for this subproblem of the hand-eye calibration using exponential coordinates, which we adapt for our needs. Denoting the real part of a quaternion q with q_{w} and the imaginary part with $\mathsf{q}_{\mathrm{xyz}}$, the quaternion logarithm is defined as

$$\log(\mathsf{q}) := 2 \arccos(\mathsf{q}_{\mathrm{w}}) \frac{\mathsf{q}_{\mathrm{xyz}}}{\|\mathsf{q}_{\mathrm{xyz}}\|} \in \mathbb{R}^{3 \times 1} \ . \tag{8}$$

Intuitively, $\log(\mathsf{q})$ extracts a representation for rotations in which the direction of $\log(\mathsf{q})$ denotes the axis and the length denotes the angle of the rotation. Then, we define the matrix M as

$$\alpha_n := \log(\mathsf{q}_n^{\mathrm{A}}), \quad \beta_n := \log(\mathsf{q}_n^{\mathrm{B}}) \tag{9}$$

$$M := \sum_{n \in [1:N]} \beta_n \cdot \mathrm{trans}(\alpha_n), \quad M \in \mathbb{R}^{3 \times 3}, \tag{10}$$

where $\mathrm{trans}(\alpha)$ is the transpose of α. The solution to Eq. (7) as a rotation matrix is given by

$$\mathrm{M}^X := (\mathrm{trans}(\mathrm{M}) \cdot \mathrm{M})^{-1/2} \cdot \mathrm{trans}(\mathrm{M}) \ . \tag{11}$$

To convert M^X to the quaternion q^{X}, we refer to [25]. Using this formulation, the offset q^{L} can be found efficiently from Eq. (5). Analogously, one can also regard the dual equation

$$\mathsf{q}^{\mathrm{I}}(s) \circ \overline{\mathsf{q}^{\mathrm{I}}(t)} = \overline{\mathsf{q}^{\mathrm{G}}} \circ \mathsf{q}^{\mathrm{T}}(s) \circ \overline{\mathsf{q}^{\mathrm{T}}(t)} \circ \mathsf{q}^{\mathrm{G}} \tag{12}$$

to find a solution for the global offset q^{G}. After alignment, we use a thresholding strategy based on Eq. (3) to detect whether a tracking error in frame t occurs by evaluating

$$d_{\mathrm{quat}} \left(\mathsf{q}^{\mathrm{I}}(t), \overline{\mathsf{q}^{\mathrm{G}}} \circ \mathsf{q}^{\mathrm{T}}(t) \circ \mathsf{q}^{\mathrm{L}} \right) > \tau \ . \tag{13}$$

4 Experiments

Data Acquisition. For our experiments, we recorded a comprehensive set of image data using eight synchronized cameras as well as inertial data for five different body parts

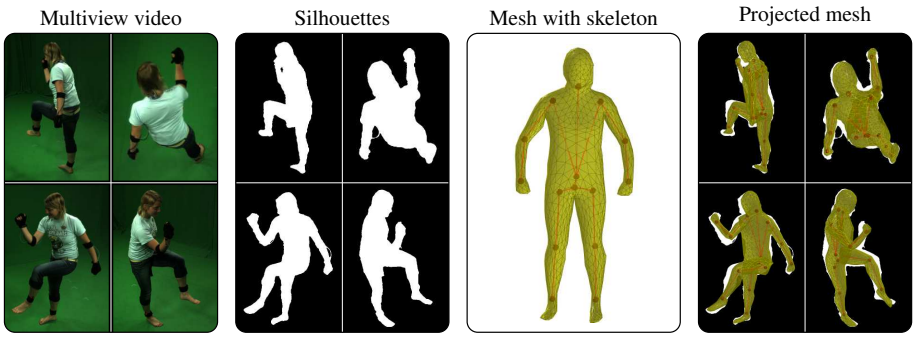

Fig. 4. Starting from multiview video, silhouettes are extracted by chroma keying. A generated skeleton-enhanced 3D model of the actor is then fit to the silhouettes based on optimization of joint angle parameters as well as the root orientation and translation.

using MTx devices [35]. By systematically recording two human actors performing various actions including motion classes such as walk, sit down and stand up, hop and jump, cartwheel, rotate arms, and throw, we obtained 24 takes with a total length of 14131 frames or 353 seconds of data.

We selected body points at different kinematic levels for fixing the sensors. Firstly, to represent body limbs that are influenced by a small number of degrees of freedom, we selected the lower legs as mounting position. Secondly, to represent body limbs that are influenced by a larger number of degrees of freedom, we selected the hands as mounting positions. Thirdly, the fifth sensor was fixed on the upper torso.

Finally, to temporally align the inertial data and the video data, we used a simple cross-correlation method applied to inertial absolute accelerations from both worlds. Here, since the offsets do not change with time, temporally local tracking errors do not play a crucial role in this step. All data streams were sampled at $40\,\mathrm{Hz}$.

Tracking. Our framework is thought for evaluating tracking results independent of the specific tracking method. In our experiments, we exemplarily used a tracking algorithm similar to [20], see Fig. 4. First, we extract silhouettes from captured images by chroma keying. We generate a surface mesh of the actor using a 3D body scanner and fit a skeletal kinematic chain to it. Then, the surface deformation of the mesh is defined by joint angle parameters as well as root orientation and translation of the kinematic chain. Using a local optimization approach, pose configuration parameters are determined to minimize the distance between the transformed 3D mesh projected back onto the 2D images and the silhouettes. This way, we generated tracking results for all 24 takes, which we then evaluated in our experiments.

Calibration. To compare orientation data from different worlds as explained in Section 3.2, the global coordinate system offset q^{G} and local offsets $\mathsf{q}_{\mathrm{s}}^{\mathrm{L}}$ for each of the sensors $s \in [1:5]$ have to be estimated. For this purpose, we propose a solution using a calibration take. There are only two small requirements for the calibration take that are easy to meet in practice. Firstly, the orientations of the limbs should be represented reasonably well by the tracking result. Secondly, to obtain unambiguous offsets, the take

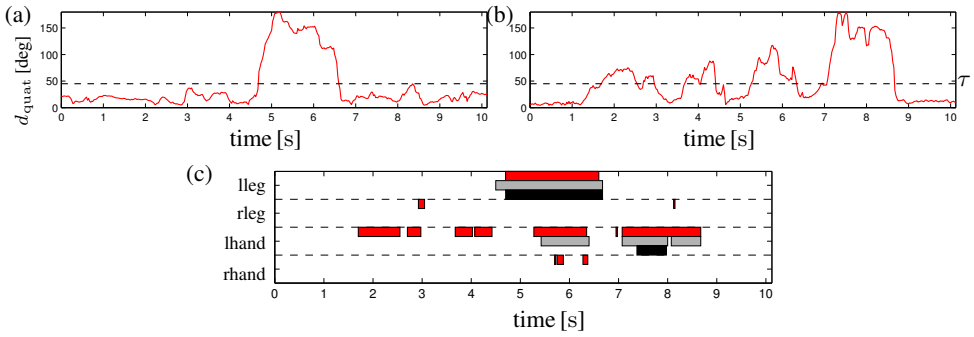

Fig. 5. Distance measure d_{quat} and threshold τ used to reveal tracking errors in an example tracking sequence for **(a)** the left leg and **(b)** the left hand. Using these curves, automatically detected tracking errors are marked by red boxes, see **(c)**. Manual annotations conducted by two subjects are marked with gray and black boxes, respectively.

should contain poses in different orientations. To this end, we selected a take containing relatively slow motions which are rather easy to track. Since the offset for the local and global orientations are constant for each actor, local tracking errors do not have a significant impact on the final estimations.

Automatic Evaluation and Discussion. In our experiments we resort to a studio setup for the multiview recordings. Going for outdoor recordings, one would require a more advanced tracking method than the one we currently use. However, our evaluation concepts transfer without modification to more advanced tracking scenarios. In particular, inertial sensors do not depend on a studio setup and are applicable for outdoor settings.

To automatically detect tracking errors, we evaluate Eq. (13) for every limb and frame. In Fig. 5, the quaternion distance functions for (a) the left leg and (b) the left hand are drawn. In Fig. 5 (c), the detected tracking errors for the body segments are marked with red boxes, which we also refer to as *automatic annotations*. In our experiments, we chose the quality threshold $\tau = 45\,^\circ$ (dashed line), which turned out to be a suitable trade-off between error detection capability and robustness. The threshold selection will be discussed later, see also Fig. 8.

Since we aim to assess the quality of the presented procedure, we asked two people (hereafter referred to as A1 and A2) of our working group to manually annotate each frame of the tracking results according to tracking errors in the limbs, see Fig. 5 (c). We refer to these annotations as *manual annotations*. For this task, the annotators were provided with the original multiview videos as well as with a tool to view the reconstructed 3D mesh from arbitrary viewpoints. As it turned out, both annotators did not notice any tracking errors in the torso. This is also reflected by our distance measure, which stays well within a small range of $14.4\,^\circ$ mean and $7.7\,^\circ$ standard deviation. Therefore, we only regard the other four sensors in our evaluation below.

In Fig. 5 (a), high distance values correspond to a tracking error in the left leg. The corresponding motion sequence is also indicated by Fig. 2. Here, both annotators as well as the automatic annotations agree. However, we found that the automatic annotation procedure generally marks more frames as erroneous than the annotators did. For an

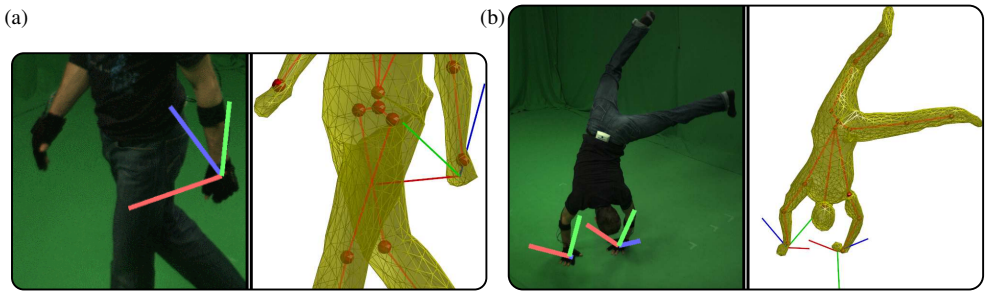

Fig. 6. (a) Left: Calibrated inertial orientation for point in time 4.2 s of the example tracking sequence. **(a) Right:** Tracking orientation. A tracking error can be detected means of orientation distances. **(b):** In a cartwheel sequence, both hands show tracking errors.

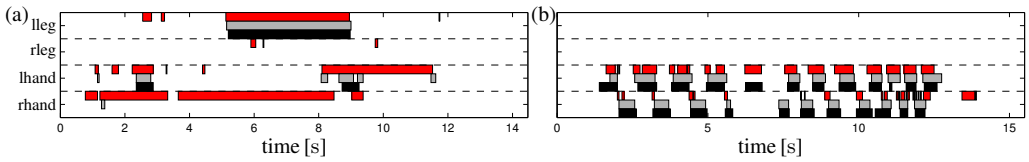

Fig. 7. Comparison of automatic (red) and manual annotations (gray, black) of **(a)** cartwheels and **(b)** locomotion.

annotator A, one needs to distinguish between false positives (automatic annotations, where A has not seen an error), and false negatives, where A has seen a tracking error, but the automatic annotation procedure did not detected it. In fact, by examining the false positives in more detail, we found that they often correspond to subtle tracking errors that are hardly visible when looking at the reconstructed mesh. For instance, in the example sequence at 4.2 s, the procedure has marked a tracking error in the left hand. Fig. 6 (a) (left) shows that the palm is facing to the actor's hip, represented by the blue axis of the calibrated inertial orientation. In the 3D reconstruction (right), however, the palm is facing backwards.

At this point we emphasize that such a tracking error might appear subtle and unimportant, because it is hardly noticeable in the visual appearance of an untextured 3D mesh. However, when using a textured mesh in a rendered scene, this kind of orientation error will lead to unwanted artifacts. Such an error is not well reflected by previous evaluation metrics like the one presented in [27]. In these metrics, ground truth marker trajectories are compared to trajectories extracted from the 3D mesh, where such an error results in only negligible differences on the positional level. With the proposed method based on orientation data, however, this error can be revealed.

Fig. 7 (a) shows the annotations of a take containing cartwheels. As an example for a false positive, consider the point in time 2.5 s. Both annotators agreed on a tracking error in the actor's left hand. In Fig. 6 (b), this error is visible even without the additionally drawn inertial orientations (left) and tracking orientations (right), since the left hand points into the wrong direction. By contrast, the tracking error in the right hand is much less visually apparent. In fact, the orientation of the whole arm is estimated incorrectly,

(a)

	A1		A2	
	P	R	P	R
Legs	0.65	0.91	0.64	0.96
Hands	0.36	0.78	0.45	0.69

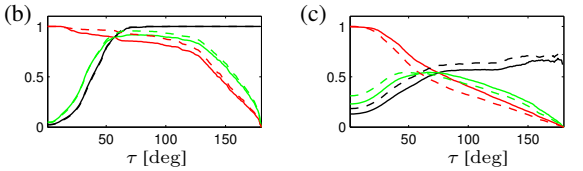

Fig. 8. (a): Precision and recall values for $\tau = 45\,^\circ$. Precision (black), recall (red) and F-measure (green) over variations of τ for **(b)** legs and **(c)** hands. Solid and dashed lines represent values belonging to A1 and A2, respectively.

coming from a misconfiguration in the shoulder joint. This error is revealed by the orientation error of the end-effector in the kinematic chain. Again, this error could not be captured well with traditional metrics.

To evaluate the accuracy on all takes, we calculated precision and recall values, taking each of the manual annotations as baseline. We separately report on the values for the hands and the legs representing two kinematic levels, see Figure 8 (a). For both the legs and hands the automatic annotations show relatively small precision values of around 0.65 and 0.36, respectively. As discussed above, the low precision is coming from a large amount of automatically detected tracking errors that the annotators did not see. This shows that the manual evaluation of tracking results is not sufficient to find all tracking errors. By contrast, the recall values for the legs are quite high, showing that the automatic annotation procedure detected nearly all manually annotated errors. The hands, however, show a lower recall in comparison to the legs. Note that this is mainly due to the per-frame annotations we pursued. In case of short tracking errors that mainly occur in the tracking results of the hands, small misalignments in the results lead to low recall values, see Fig. 7 (b). Although most of the boxes coming from manual annotations have a certain overlap with an automatic annotation, the automatic annotations achieve a low recall. Here, segment-based rather than frame-based values may be a more suitable measure.

For quantitative evaluations a combined recording setup with a marker-based optical motion capture system would have been beneficial. In our setup we did not have a marker-based reference system at hand. Different sources of errors like sensor noise and bias, calibration errors, sensors getting out of place, or errors due to the approximation of the human body with a rigged surface mesh are thus difficult to quantify. However, our experiments show that the influence of all sources of noise are small. For example, the distance measure of the upper torso sensor over all 14131 frames of our evaluation data stays within a small error range with a mean of $14.4\,^\circ$ and a standard deviation of $7.7\,^\circ$, and the manual inspection shows that there are no noticeable tracking errors in the torso region. This observation suggests that the overall noise lies within this small order of magnitude. In particular, it follows that the accuracy of the obtained inertial orientations is high enough for a quantitative evaluation of tracking results. Moreover, our experiments show that the proposed distance metric is able to cover most of the manually observed tracking errors, which is supported by high recall values. Finally, a manual inspection showed that the false positive detections correspond to tracking errors that were difficult to perceive for the manual annotators. This supports

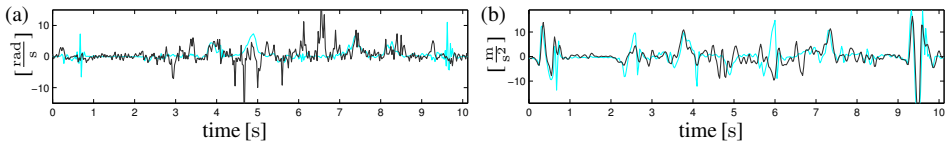

Fig. 9. Calibrated inertial data (cyan) and tracking data (black) for the left leg of the example sequence used in Fig. 5 (a). **(a)**: Y-component of the local rate of turn data. **(b)**: Z-component of the local acceleration data. The tracking errors are hardly detectable.

the statement that our orientation-based distance measure is well suited for detecting tracking errors.

To evaluate the influence of the threshold parameter τ, we computed precision, recall, and F-measure for variations of τ, see Fig. 8. Selecting a low τ yields in a high recall, since many parts of the evaluation takes are annotated. However, also many parts unrelated to tracking errors are annotated, yielding a low precision. Our final choice of $\tau = 45°$ is motivated by the request of having high recall values without having too many false detections.

As described in Sect. 3.2, orientation data from the inertial world is obtained by combining different sensors. These sensors naturally provide 3D acceleration and rate of turn data. Thus, a method comparing these types of data with corresponding data generated from the tracking world could also reveal tracking errors. In practice, however, this does not work well. In Fig. 9, we show a comparison of (a) the rate of turn data and (b) the acceleration data corresponding to the left leg for the example tracking sequence also used in Fig. 5 (a). Here, we only present the Y-component of the rate of turn and Z-component of the acceleration data, which show the most significant differences. One can see that the tracking error in the left leg, occurring from $4.6\,\mathrm{s}$ to $6.6\,\mathrm{s}$, is hardly revealed on the basis of such data. Firstly, these quantities are very local in nature with respect to the temporal dimension. This makes it hard to detect the duration as well as the starting and ending point in time of an error. Secondly, filtering techniques necessary to determine meaningful acceleration and rate of turn data may not only suppress the sensor noise but may also smooth out peaks coming from actual tracking errors. Thirdly, slowly moving limbs generate low amplitudes in these quantities, which makes it hard, if not infeasible, to detect errors for such motions. With orientation data, as shown in the paper, these considerations do not hold, thus yielding a robust procedure for tracking error detection.

5 Conclusions

As a main result of this paper, we showed that limb orientations are a suitable mid-level representation for detecting tracking errors in markerless motion capturing. In contrast to traditional evaluation techniques with marker-based optical systems, the usage of inertial sensors provides an unobtrusive and affordable way to generate ground truth data. Furthermore, inertial sensors impose comparatively weak additional constraints on the overall recording setup with regard to location, recording volume, and illumination. In particular, our procedure enables the detection of tracking errors that come from

rotationally symmetric body parts. Such errors can hardly be identified by traditional evaluation metrics which are based on visual cues or positional information.

Sensor fusion for motion tracking, recognition, and retrieval applications has become a vital strand of research. Apart from detecting tracking errors, the integration of inertial data into tracking algorithms as additional prior information constitutes a promising approach to stabilizing motion tracking in complex scenarios such as outdoor settings, fast motions, or presence of occlusions. Furthermore, we plan to apply our framework for orientation-based motion retrieval and reconstruction. Finally, we contribute to these fields by making the multimodal data set and the Matlab-implementation of the calibration method used in this paper publicly available at [16].

Acknowledgments. This work has been supported by the German Research Foundation (DFG CL 64/5-1 and DFG MU 2686/3-1). Meinard Müller is funded by the Cluster of Excellence on Multimodal Computing and Interaction.

References

1. A. Agarwal and B. Triggs. Learning to track 3D human motion from silhouettes. In *ICML '04: Proceedings of the 21st International Conference on Machine Learning*, pages 2–9, New York, NY, USA, 2004. ACM.
2. C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinetics. *International Journal of Computer Vision*, 56(3):179–194, 2004.
3. M. Brodie, A. Walmsley, and W. Page. Fusion motion capture: a prototype system using inertial measurement units and GPS for the biomechanical analysis of ski racing. *Sports Technology*, 1(1):17–28, 2008.
4. T. Brox, B. Rosenhahn, U. Kersting, and D. Cremers. Nonparametric density estimation for human pose tracking. In *Pattern Recognition*, volume 4174 of *LNCS*, pages 546–555, Berlin, Germany, Sept. 2006. Springer.
5. CMU. CMU multi-modal activity database. `http://kitchen.cs.cmu.edu`, 2010.
6. K. Daniilidis. Hand-eye calibration using dual quaternions. *The International Journal of Robotics Research*, 18(3):286–298, 1999.
7. H. Dejnabadi, B. M. Jolles, E. Casanova, P. Fua, and K. Aminian. Estimation and visualization of sagittal kinematics of lower limbs orientation using body-fixed sensors. *IEEE Transactions on Biomedical Engineering*, 53(7):1385–1393, 2006.
8. E. Foxlin. Pedestrian tracking with shoe-mounted inertial sensors. *IEEE Computer Graphics and Applications*, 25(6):38–46, 2005.
9. F. S. Grassia. Practical parameterization of rotations using the exponential map. *Journal of Graphics, GPU, and Game Tools*, 3(3):29–48, 1998.
10. T. Harada, T. Mori, and T. Sato. Development of a tiny orientation estimation device to operate under motion and magnetic disturbance. *The International Journal of Robotics Research*, 26(6):547–559, 2007.
11. J. D. Hol, T. B. Schön, and F. Gustafsson. Relative pose calibration of a spherical camera and an IMU. In *7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 21–24, Sept. 2008.
12. D. Q. Huynh. Metrics for 3D rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, 2009.
13. K. Kunze and P. Lukowicz. Dealing with sensor displacement in motion-based onbody activity recognition systems. In *UbiComp '08: Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 20–29, New York, NY, USA, 2008. ACM.

14. D. Lowe. Solving for the parameters of object models from image descriptions. In *Image Understanding Workshop*, pages 121–127, College Park, Apr 1980.
15. M. Müller. *Information Retrieval for Music and Motion*. Springer, 2007.
16. Multimodal human motion database MPI08. http://www.tnt.uni-hannover.de/project/MPI08_Database/.
17. F. C. Park and B. J. Martin. Robot sensor calibration: solving $AX = XB$ on the Euclidean group. *IEEE Transactions on Robotics and Automation*, 10(5):717–721, Oct. 1994.
18. G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), to appear*, June 2010.
19. D. Roetenberg. Inertial and magnetic sensing of human motion. *These de doctorat*, 2006.
20. B. Rosenhahn, T. Brox, and H.-P. Seidel. Scaled motion dynamics for markerless motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1203–1210, Minneapolis, Minnesota, 2007. IEEE.
21. C. Schmaltz, B. Rosenhahn, T. Brox, D. Cremers, J. Weickert, L. Wietzke, and G. Sommer. Region-based pose tracking. In *Pattern Recognition and Image Analysis*, volume 4478 of *LNCS*, pages 56–63, Girona, Spain, June 2007. Springer.
22. Y. Seo, Y.-J. Choi, and S. W. Lee. A branch-and-bound algorithm for globally optimal calibration of a camera-and-rotation-sensor system. In *IEEE 12th International Conference on Computer Vision (ICCV)*, Sept. 2009.
23. T. Shiratori and J. K. Hodgins. Accelerometer-based user interfaces for the control of a physically simulated character. In *ACM SIGGRAPH Asia*, pages 1–9, New York, NY, USA, 2008. ACM.
24. Y. C. Shiu and S. Ahmad. Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form $AX = XB$. *IEEE Transactions on Robotics and Automation*, 5(1):16–29, Feb. 1989.
25. K. Shoemake. Animating rotation with quaternion curves. *ACM SIGGRAPH Computer Graphics*, 19(3):245–254, July 1985.
26. H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *Proc. European Conference on Computer Vision*, volume 2353 of *LNCS*, pages 784–800. Springer, 2002.
27. L. Sigal and M. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, USA, 2006. Available at `http://vision.cs.brown.edu/humaneva/` .
28. R. Slyper and J. Hodgins. Action capture with accelerometers. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, July 2008.
29. K. Strobl and G. Hirzinger. Optimal hand-eye calibration. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4647–4653, Oct. 2006.
30. Y. Tao, H. Hu, and H. Zhou. Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. *International Journal of Robotics Research*, 26(6):607–624, 2007.
31. Y. K. Thong, M. S. Woolfson, J. A. Crowe, B. R. Hayes-Gill, and D. A. Jones. Numerical double integration of acceleration measurements in noise. *Measurement*, 36(1):73–92, 2004.
32. R. Tsai and R. Lenz. Real time versatile robotics hand/eye calibration using 3D machine vision. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 1, pages 554–561, Apr 1988.
33. D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik, and J. Popović. Practical motion capture in everyday surroundings. *ACM Transactions on Graphics*, 26(3):35, 2007.
34. D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):1–9, 2008.
35. Xsens Motion Technologies. http://www.xsens.com/, Accessed November 19th, 2009.