2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

A Visual-Inertial Approach to Human Gait Estimation

Ahmed Ahmed and Stergios Roumeliotis[1]

Abstract— **This paper addresses the problem of gait estimation using visual and inertial data, as well as human motion models. Specifically, a batch least-squares (BLS) algorithm is presented that fuses data from a minimal set of sensors [two inertial measurement units (IMUs), one on each foot, and a head-mounted IMU-camera pair] along with motion constraints corresponding to the different walking states, to estimate the person's head and feet poses. Subsequently, gait models are employed to solve for the lower-body's posture and generate its animation. Experimental results against the VICON motion capture system demonstrate the accuracy of the proposed minimal sensors-based system for determining a person's motion.**

I. Introduction and Related Work

Human-motion modeling and estimation is of critical importance to physical therapy and rehabilitation (for assessing, diagnosing, and planning treatment [1]), the movie and gaming industries (for motion capturing and character animation [2]), and robotics (e.g., for modeling bipedal walk [3] and indoor localization [4]). Existing motion capturing systems can be classified into the following categories:

(i) *External camera-based systems*, or outside-in systems (e.g., the VICON system [5]), estimate the body-posture by tracking markers attached to the user. They provide high-accuracy measurements in real time, but their high cost, complex infrastructure, limited coverage area, and the required cumbersome markers' suit restrict their use.

(ii) *Body-mounted camera systems* attach the cameras, instead, to the person's body and observe the surrounding environment, as an inside-out system. For example, [6] used 16 cameras to capture general motion, while [7] used only 2 head-mounted cameras in a way that observes both the moving person and the surroundings. These systems provide high accuracy under sufficient motion and environment conditions, they do not impose the infrastructure and coverage area constraints, and have less cost compared to the previous ones. Their high computational demands, however, limit their operation to only offline, and still they have inconvenient setup due to the numerous body-mounted cameras.

(iii) *Body-mounted IMU systems* combine angular velocity and linear acceleration data from IMUs attached to different body segments. They follow a *sensor-based* approach and employ motion constraints (e.g., maintain the body dimensions, and ensure zero velocity of the feet during stance periods) to improve estimation accuracy. These systems impose few limitations on the area of operation, and achieve real-time performance. They are costly, however, and require cumbersome sensor suit setup (e.g., 17 IMUs for the Xsens MVN motion-capture suit [8]).

(iv) *Peripheral IMU-based systems* use prior motion models to reduce the number of body-attached sensors, hence, to overcome the previous shortcomings. For example, [9] used 4 IMUs (attached to hands and feet) to estimate the full body posture, while [10] used one foot-mounted IMU to estimate gait parameters and represent the motion with a simple 2D model. Lastly, [11] also used one IMU but only to estimate the foot trajectory. These attempts addressed the system's usability and cost constraints, but reducing the number of IMUs comes at the expense of lower estimation accuracy.

Our objective in this work is to combine the body-mounted camera system's high accuracy with the peripheral IMU-based system's low cost and usability within a minimal sensor-based framework. In particular, we employed two foot-mounted IMUs and a head-mounted camera-IMU pair (Google Glass) to estimate the person's trajectory and generate a 3D animation of their corresponding motion. This setup can be used to improve the quality of pervasive healthcare (by providing personalized monitoring and incidence detection [12]), and virtual reality (VR) applications (by enabling natural interaction and bringing immersive experiences [13]). To achieve our objectives, we need to address two key challenges:

- Since the camera-IMU pair is attached to the person's head while the two IMUs are on their feet, the relative transformations between these sensors are unknown and vary during motion. To fuse information from the three sensing modalities, these transformations have to be estimated.
- The lower-body posture corresponding to the person's motion has to be computed, given *only* the input from three body-attached sensors,

To this end, we employ the gait model, which describes the body-posture's time evolution during walking, and thus allows us to relate the sensors' poses and compute the lower body-posture. In particular, the gait model consists of gait-events and nominal joint-angle profiles. The gait-events define transition states for a complete walking step (e.g., foot ground contact and foot swing). Therefore, they impose pose-related motion constraints at the times of their occurrence timings [1]. The joint-angle profiles determine the nominal body-posture during walking, which is specified by 27 joint angles of an articulated human model. Our human model is defined as a set of joints connecting body-segments, with

[1]A. M. Ahmed and S. I. Roumeliotis are with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, 55455, USA, medhat|stergios@cs.umn.edu. This work was supported by the University of Minnesota and the National Science Foundation (IIS-1328722).

lengths computed as a percentage of the body-height input parameter [14].

In the epicenter of our approach is a batch least squares (BLS) estimator, that combines all visual and inertial data along with the motion constraints to estimate the kinematic state (pose, linear and rotational velocity, and linear acceleration) of the person's head and feet. These estimates are then combined with the gait model to determine the body's posture. To summarize, our main contributions are the following:

- To the best of our knowledge, we present the first visual-inertial-based motion-capture system, which takes advantage of the high accuracy of body-mounted camera systems and the usability, low price and reduced processing requirements of peripheral IMU-based systems.
- We introduce a twofold approach that (i) Fuses motion constraints with visual and inertial measurements to estimate the trajectories of the three time-varying head/feet coordinate frames, (ii) Generates a motion animation by using the estimated trajectories and the prior gait model, and solving a non-linear optimization problem, in closed forms to compute the body pose.

In the next section, we provide an overview of the proposed system. Details of each component are presented in subsequent sections as follows: Motion analysis and constraint generation in Sec. III; Trajectory estimation along with inertial, visual, and motion constraints are described in Sec. IV; Motion synthesis in Sec. V. Lastly, we present our experimental results in Sec. VI, and our concluding remarks in Sec. VII.

II. Approach Overview

Our objective is to estimate the trajectory of a person walking (along with the body posture) using few only sensors (head-mounted camera-IMU and two foot-mounted IMUs). The proposed system comprises (see Fig. 1):

- *Input* corresponds to the visual/inertial sensor data, the human model, and the gait model.
- *Gait-Event Detector (GED)* tracks the walking transition states (e.g., foot ground contact and foot swing). Specifically, it detects gait-event occurrences by applying thresholds on the accelerometer measurements, and thus it generates *motion constraints* corresponding to these events.
- *Trajectory estimator* computes the 3D pose of the 3 body-mounted IMUs (Google Glass-IMU and two foot-mounted IMUs), as well as, their time-varying relative transformations by fusing visual and inertial data with the proposed motion constraints in a BLS formulation.
- *Motion Synthesizer* generates the body animation. In particular, it applies nominal joint angles to the human model using the timings computed from the GED. Then, it places the model along the estimated trajectories by solving a non-linear optimization problem.

We first describe the gait model and how to detect the events in Sec. III. Then, we present the trajectory estimation along with the generated motion constraints in Sec. IV. Finally, we describe the motion synthesizer in Sec. V.

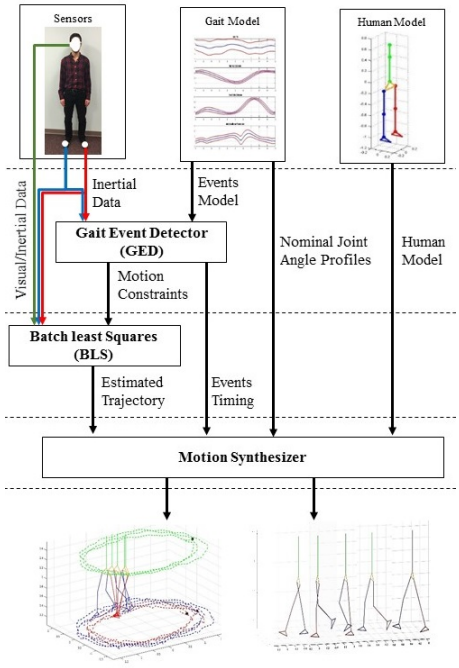

Fig. 1. Proposed system's block diagram.

III. Motion Analysis and Constraint Generation

The gait-event detector (GED) tracks the body-posture's transition states (e.g., foot ground contact and foot swing) according to the gait-events' model, and thus introduces motion constraints that relate the three head/feet trajectories. Moreover, in order to compute the joint-angles necessary for generating the 3D animation, we need detect the body posture corresponding to the person's motion at each time step. To this end, the GED provides information for relating the sensors' poses and to computing the lower-body's posture. The GED detects the gait events by applying simple thresholds to the magnitude of the linear acceleration measurements of the foot-mounted IMUs. As a result, it computes gait-event timings and generates *motion constraints*. The next subsections present the gait-events' model and the process of detecting them.

A. Gait Model

Our gait model consists of gait-events and nominal joint-angle profiles. The gait-events define the transition states of a complete walking cycle, i.e., the pose-related events during the time interval between two successive same-foot ground contacts. Detecting gait-events allows us to impose motion constraints at the time their occurrence. Therefore, it is important to understand these events and the associated body postures. Specifically, the gait-cycle (GC) is divided into the following states based on foot-ground-contact events (see Fig. 2):

1) Loading Response ($0 - 10\%$ of the GC): In this transition state, both feet are in contact with the ground

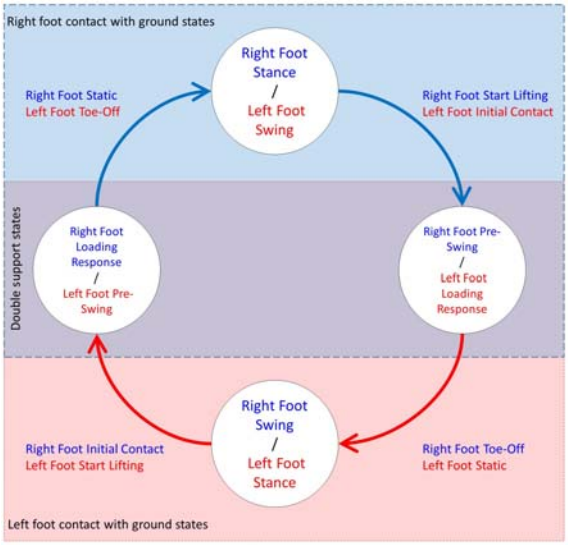

Fig. 2. Gait Cycle State Diagram.

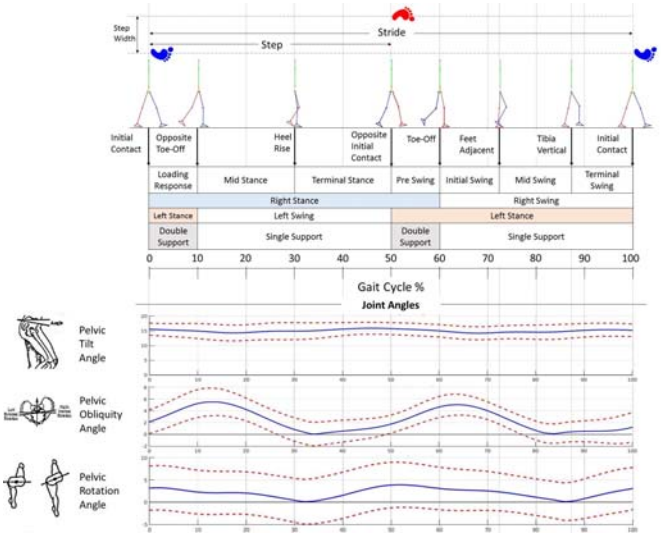

Fig. 3. Gait Model: Gait states and events during the gait cycle along with the joint-angle profiles.

(double support period). It is triggered by one foot's initial floor contact event and continues until the other foot is lifted off the ground for swing (opposite foot toe-off event).

2) Stance ($10-50\%$ of the GC): After the opposite toe-off, the foot becomes completely static and in full contact with the ground during a single support period (i.e., single-foot contact with ground), while the other foot is swinging. During this state motion constraints (e.g., static foot constraint) are applied [see Sec. IV-C].
3) Pre-Swing ($50-60\%$ of the GC): This state mirrors the Loading Response on the other foot. It starts when the opposite swinging foot hits the ground (opposite initial contact event), while the stance foot is lifted for the swing. The state ends with the foot toe-off event.
4) Swing ($60-100\%$ of the GC): After toe-off, the foot progresses forward until it hits the ground at the next initial contact event. During this state the motion constraints (e.g., static foot constraint) are applied to the opposite foot [see Sec. IV-C].

Fig. 2 illustrates the four gait cycle states and the transition events between them.[1] The joint-angle profiles, on the other hand, specify the time evolution of the 27 joint angles (constituting our human model) during the gait-cycle. Hence, they define the body-posture during a walking step. Fig. 3 illustrates the body posture at the corresponding detected gait-events, along with a subset of the nominal joint angle profiles during the gait-cycle. [2] A detailed analysis of the gait process and states are presented in [1].

[1]For illustration purposes, we associate red and blue colors with left and right feet events, respectively, in all figures of the paper.

[2]Fig. 3 is created using the experimental data provided by Clinical Gait Analysis (CGA) Normative Gait Database [15] and modified illustrative images acquired from various medical websites.

B. Gait-Event Detector

The gait event detector (GED) applies thresholds to the magnitude of the accelerometer measurements from the foot-mounted IMUs (they are expected to be equal to gravity during a stance period) to detect the foot-ground-contact events. Fig. 4 illustrates the gait-event detection process. Specifically, the two lower plots show the left-right foot accelerometer measurements' magnitude during the gait-cycle, along with the GC states and events. As evident, the linear acceleration's magnitude equals that of gravity during the foot stance periods (i.e., the foot-mounted IMU is static). The feet elevation trajectories (computed from the VICON tracking system and shown in Fig. 4) confirm the accuracy of the detected gait-event timings, and thus the validity of this method. Lastly, we note that the proposed system maintains the state-machine illustrated in Fig. 2 to track the four gait states.

IV. Trajectory Estimation

Our BLS estimator computes the 3D trajectories of the three body-mounted IMUs (Google Glass-IMU and the two foot-mounted IMUs) by fusing visual and inertial data. As mentioned before, the main challenge in this step is that the three IMUs are not rigidly attached to each other, i.e., their relative transformations vary as the person is walking. To address this issue, we estimate their time varying extrinsics and impose motion constraints on their trajectories. Specifically, our BLS estimates the following state vector:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_F^T & \mathbf{x}_G^T & \mathbf{x}_L^T & \mathbf{x}_R^T & \mathbf{x}_E^T \end{bmatrix}^T \tag{1}$$

where $\mathbf{x}_F$ is the Euclidean coordinates of the observed visual features, $\mathbf{x}_G$, $\mathbf{x}_L$, and $\mathbf{x}_R$ are the state vectors of the Glass, left-foot, and righ-foot IMUs, respectively, along the entire trajectory, i.e., $\mathbf{x}_G^T = \begin{bmatrix} \mathbf{x}_{G_1}^T & \dots \mathbf{x}_{G_k}^T & \dots & \mathbf{x}_{G_N}^T \end{bmatrix}^T$. Each $\mathbf{x}_{G_i}$, $i = 1, \dots, N$, comprises the position, attitude, linear velocity, and biases of the corresponding IMU [see (2)].

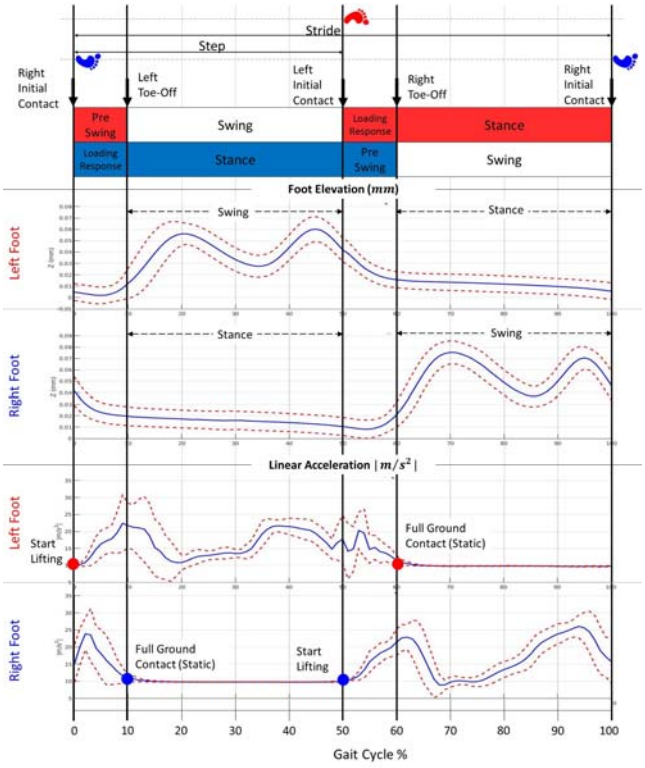

Fig. 4. Linear acceleration and feet elevation profiles during the gait cycle.

Finally, $\mathbf{x}_E$ denotes the extrinsic parameters (relative pose of the 3 IMUs and the pose of the Glass-camera w.r.t. the Glass-IMU). The BLS estimator computes $\mathbf{x}$ by iteratively minimizing a function comprising cost terms from the IMU measurements, the visual observations, and the motion constraints. Each of these are described in detail hereafter.

A. Inertial measurements cost term $\mathcal{C}_u(\widetilde{\mathbf{x}})$

Each IMU measures the sensor's rotational velocity and linear acceleration contaminated by white Gaussian noise and time-varying biases. The IMU state at time k is defined as

$$\mathbf{x}_{I_k} = \begin{bmatrix} {}^{I_k}\mathbf{q}_G^T & \mathbf{b}_{g_k}^T & {}^{G}\mathbf{v}_{I_k}^T & \mathbf{b}_{a_k}^T & {}^{G}\mathbf{p}_{I_k}^T \end{bmatrix}^T \tag{2}$$

where ${}^{I_k}\mathbf{q}_G$ is the quaternion representing the orientation of the global frame, $\{G\}$, in the IMU's frame of reference, $\{I_k\}$, ${}^{G}\mathbf{p}_{I_k}$ and ${}^{G}\mathbf{v}_{I_k}$ are the position and velocity of $\{I_k\}$ in $\{G\}$, and $\mathbf{b}_g$ and $\mathbf{b}_a$ are the gyroscope and accelerometer biases, respectively. The sensor's state evolution is described by the propagation model

$$\mathbf{x}_{I_{k+1}} = \mathbf{f}(\mathbf{x}_{I_k},\ \mathbf{u}_k) + \mathbf{w}_k \tag{3}$$

where $\mathbf{f}(\mathbf{x}_{I_k},\ \mathbf{u}_k)$ is a nonlinear function for integrating the inertial measurements $\mathbf{u}_k$ over the time interval $[t_k,\ t_{k+1}]$ (see [16] for details), and $\mathbf{w}_k$ is the discrete-time zero-mean white Gaussian measurement noise with covariance $\mathbf{Q}_k$, computed through IMU characterization [17]. Thus, every IMU measurement imposes a stochastic constraint between the consecutive IMU states $\mathbf{x}_{I_k}$ and $\mathbf{x}_{I_{k+1}}$. Linearizing (3), around the state estimates $\hat{\mathbf{x}}_{I_k}$ and $\hat{\mathbf{x}}_{I_{k+1}}$, results in the following error equation

$$\widetilde{\mathbf{x}}_{I_{k+1}} = (\mathbf{f}(\hat{\mathbf{x}}_{I_k}, \mathbf{u}_k) - \hat{\mathbf{x}}_{I_{k+1}}) + \mathbf{\Phi}_k \widetilde{\mathbf{x}}_{I_k} + \mathbf{w}_k \tag{4}$$

where $\mathbf{\Phi}_k$ is the corresponding Jacobians w.r.t the state $\mathbf{x}_{I_k}$. The error state $\widetilde{\mathbf{x}}$ is defined as the difference between the true state $\mathbf{x}$ and the state estimate $\hat{\mathbf{x}}$ employed for linearization (i.e., $\widetilde{\mathbf{x}} \triangleq \mathbf{x} - \hat{\mathbf{x}}$), while for the quaternion $\mathbf{q}$ a multiplicative error model is used $\widetilde{\mathbf{q}} \triangleq \mathbf{q} \otimes \hat{\mathbf{q}}^{-1} \simeq \begin{bmatrix} \frac{1}{2}\delta\boldsymbol{\theta}^T & 1 \end{bmatrix}^T$, where $\delta\boldsymbol{\theta}$ is a minimal representation of the attitude error.

Thus, each inertial measurement $\mathbf{u}_k$ contributes to the BLS a linearized cost term of the form

$$\mathcal{C}_{u_k}(\widetilde{\mathbf{x}}_{I_k}, \widetilde{\mathbf{x}}_{I_{k+1}}) = || \begin{bmatrix} \mathbf{\Phi}_k & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{x}}_{I_k} \\ \widetilde{\mathbf{x}}_{I_{k+1}} \end{bmatrix} - (\hat{\mathbf{x}}_{I_{k+1}} - \mathbf{f}(\hat{\mathbf{x}}_{I_k}, \mathbf{u}_k))||^2_{\mathbf{Q}_k} \tag{5}$$

while multiple IMU measurements across time yield a cost term $\mathcal{C}_{u_I}(\widetilde{\mathbf{x}})$ resulting from the summation of the individual terms (5), i.e.,

$$\mathcal{C}_{u_I}(\widetilde{\mathbf{x}}) = \sum_{k=1}^{n} \mathcal{C}_{u_k}(\widetilde{\mathbf{x}}_{I_k}, \widetilde{\mathbf{x}}_{I_{k+1}}) \tag{6}$$

In our case, we have three IMU sensors, and thus the overall cost function $\mathcal{C}_u(\widetilde{\mathbf{x}})$ comprises the cost terms from the head-mounted IMU $\mathcal{C}_{u_G}(\widetilde{\mathbf{x}})$, the right-foot IMU $\mathcal{C}_{u_R}(\widetilde{\mathbf{x}})$, and the left-foot IMU $\mathcal{C}_{u_L}(\widetilde{\mathbf{x}})$, i.e.,

$$\mathcal{C}_u(\widetilde{\mathbf{x}}) = \mathcal{C}_{u_G}(\widetilde{\mathbf{x}}) + \mathcal{C}_{u_R}(\widetilde{\mathbf{x}}) + \mathcal{C}_{u_L}(\widetilde{\mathbf{x}}) \tag{7}$$

B. Visual observations cost term $\mathcal{C}_z(\widetilde{\mathbf{x}})$

As the person walks, the head-mounted camera observes the surrounding environment. To provide information about camera's motion, we extract and track static visual point features. Specifically, we extract Harris corners [18] (i.e., points of local maximum or minimum intensity) as 2D observations of the 3D features and track them through consecutive images using the Kanade-Lucas-Tomasi (KLT) feature tracker [19]. The resulting observed feature tracks impose visual constraints on the camera's (and hence the IMU's) motion relative to the observed scene. Note that this visual information is of critical importance fro reducing the IMU drift.

In particular, each visual measurement relates the camera pose at a specific time step k with the observed feature f through the following projection model

$$\mathbf{z}_{f_k} = \pi({}^{C_k}\mathbf{p}_f) + \mathbf{n}_{f_k} \tag{8}$$

with

$$\pi(\begin{bmatrix} x & y & z \end{bmatrix}^T) \triangleq \begin{bmatrix} \frac{x}{z} & \frac{y}{z} \end{bmatrix}^T,$$

where ${}^{C_k}\mathbf{p}_f$ is the feature position expressed in the camera frame of reference, $\mathbf{n}_{f_k}$ is zero-mean, white Gaussian noise with covariance $\sigma_f^2\mathbf{I}_2$, and $\mathbf{I}_2$ is the 2×2 identity matrix. Note that we express the feature measurement (8) in the normalized pixel coordinates, after performing intrinsic camera calibration offline [20]. Note that ${}^{C_k}\mathbf{p}_f$ is expressed w.r.t. the state vector elements (head IMU pose

${}^{G}\mathbf{p}_{I_k}$ and ${}^{I_k}\mathbf{q}_G$, feature position w.r.t. global ${}^{G}\mathbf{p}_f$, and camera-IMU extrinsic ${}^{C}\mathbf{p}_I$ and ${}^{C}\mathbf{q}_I$); i.e., ${}^{C_k}\mathbf{p}_f = {}^{C}\mathbf{p}_I + \mathbf{R}({}^{C}\mathbf{q}_I)\mathbf{R}({}^{I_k}\mathbf{q}_G)\left({}^{G}\mathbf{p}_f - {}^{G}\mathbf{p}_{I_k}\right)$, where $\mathbf{R}(\mathbf{q})$ is the rotation matrix corresponding to the quaternion $\mathbf{q}$. Linearizing (8), yields the error equation

$$\widetilde{\mathbf{z}}_{f_k} = \mathbf{H}_{f_k}\,\widetilde{\mathbf{x}} + \mathbf{n}_{f_k} \tag{9}$$

where $\mathbf{H}_{f_k}$ is the corresponding Jacobian evaluated at the state estimate $\hat{\mathbf{x}}$, which contributes a linearized cost term of the form

$$\mathcal{C}_{z_{(f,k)}}(\widetilde{\mathbf{x}}) = ||\mathbf{H}_{f_k}\widetilde{\mathbf{x}} - \widetilde{\mathbf{z}}_{f_k}||^2_{\sigma_f^2\mathbf{I}_2} \tag{10}$$

Since the camera observes multiple features from different poses, we form the visual observations cost function $\mathcal{C}_z(\widetilde{\mathbf{x}})$ by accumulating the contributing cost terms (10) for all features and camera poses:

$$\mathcal{C}_z(\widetilde{\mathbf{x}}) = \sum_k \sum_f \mathcal{C}_{z_{(f,k)}}(\widetilde{\mathbf{x}}) \tag{11}$$

C. Motion constraints

As mentioned earlier, the main difference between our BLS formulation and existing visual-inertial navigation systems is that we are fusing measurements from a camera and 3 IMUs whose relative transformations are unknown and time varying. In order to estimate them, we introduce additional information in the form of motion constraints that relate the three IMUs' trajectories. Specifically, we take advantage of the fact that the sensors are mounted on a person walking on a planar surface, and seek to infer the kinematic state of each IMU during gait events. In particular, every time a foot hits the ground, we require it to be at the same height (i.e., we restrain both feet to intermittently be on the same ground plane). Additionally, detecting each foot's stance state imposes its velocity to be zero during this period. Lastly, we constrain the head's projection on the ground plane to be in between consecutive footsteps. These simple, yet efficient, constraints capture the person's motion and prevent the estimated IMU trajectories from diverging from each other. Note that, all the motion constraints stem from stance events detected by GED (see Sec. III). In what follows, we assume that these events are already detected and provided to the BLS estimator. Next, we present the formulation of each of these constraints within the BLS framework:

1) Zero-velocity constraint cost term $\mathcal{C}_v(\widetilde{\mathbf{x}})$: This constraint, applied during the foot stance states of the gait cycle (see Gait Model in Sec. III), sets the linear velocity of the foot to zero; this implies that the foot is not moving and is in full contact with ground. This stochastic constraint and its corresponding error equation at footstep k are expressed as:

$$\begin{aligned}\mathbf{z}_{v_k} &= {}^{G}\mathbf{v}_{I_k} + \mathbf{n}_{v_k} \\ \widetilde{\mathbf{z}}_{v_k} &= \mathbf{H}_{v_k}\widetilde{\mathbf{x}} + \mathbf{n}_{v_k}\end{aligned} \tag{12}$$

where $\mathbf{H}_{v_k}$ is the corresponding Jacobian, and $\mathbf{n}_{v_k}$ is zero-mean white Gaussian noise with covariance $\sigma_v^2\mathbf{I}_3$. The constraint contributes to the BLS a cost term of the form

$$\mathcal{C}_{v_k}(\widetilde{\mathbf{x}}) = ||\mathbf{H}_{v_k}\widetilde{\mathbf{x}} - \widetilde{\mathbf{z}}_{v_k}||^2_{\sigma_v^2\mathbf{I}_3} \tag{13}$$

For multiple footsteps n, we sum the contributing cost terms (13) corresponding to the left $\mathcal{C}_{v_{Lk}}$ and right $\mathcal{C}_{v_{Rk}}$ feet to form the cost function:

$$\mathcal{C}_v(\widetilde{\mathbf{x}}) = \sum_{k=1}^{n}\mathcal{C}_{v_{Lk}} + \sum_{k=1}^{n}\mathcal{C}_{v_{Rk}} \tag{14}$$

2) Constant-height constraint cost term $\mathcal{C}_h(\widetilde{\mathbf{x}})$: This constraint sets the foot elevation to be at the same height during the stance states of the gait cycle, i.e., it sets the z position of the IMU to zero. In other words, this constraint restrains the feet to move on a plane, following a planar walking motion model see Sec. III). This stochastic constraint and its corresponding error equation at footstep k can be written as:

$$\begin{aligned}\mathbf{z}_{h_k} &= \begin{bmatrix}0 & 0 & 1\end{bmatrix}{}^{G}\mathbf{p}_{I_k} + \mathbf{n}_{h_k} \\ \widetilde{\mathbf{z}}_{h_k} &= \mathbf{H}_{h_k}\widetilde{\mathbf{x}} + \mathbf{n}_{h_k}\end{aligned} \tag{15}$$

where $\mathbf{H}_{h_k}$ is the corresponding Jacobian, and $\mathbf{n}_{h_k}$ is zero-mean white Gaussian noise with variance σ_h^2. This constraint contributes to the BLS a cost term of the form

$$\mathcal{C}_{h_k}(\widetilde{\mathbf{x}}) = ||\mathbf{H}_{h_k}\widetilde{\mathbf{x}} - \widetilde{\mathbf{z}}_{h_k}||^2_{\sigma_h^2} \tag{16}$$

For multiple footsteps n, we sum the contributing cost terms (16) corresponding to the left $\mathcal{C}_{h_{Lk}}$ and right $\mathcal{C}_{h_{Rk}}$ feet to form the cost function:

$$\mathcal{C}_h(\widetilde{\mathbf{x}}) = \sum_{k=1}^{n}\mathcal{C}_{h_{Lk}} + \sum_{k=1}^{n}\mathcal{C}_{h_{Rk}} \tag{17}$$

3) Head/Feet relative position constraint cost term $\mathcal{C}_p(\widetilde{\mathbf{x}})$: This constraint requires the projection of the head's position on the plane to be in the middle of consecutive foot steps. In other words, it does not allow the feet to move far away from each other on the planar surface, and maintains the left/right feet positions relative to the head. The constraint can be represented with the following equation

$$\mathbf{S}\,{}^{G}\mathbf{p}_{h_k} = \frac{1}{2}\mathbf{S}\left({}^{G}\mathbf{p}_{l_k} + {}^{G}\mathbf{p}_{r_k}\right) \tag{18}$$

where $\mathbf{S} \triangleq \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$, ${}^{G}\mathbf{p}_{h_k}$ is the head position in the Glass global frame,[3] ${}^{G}\mathbf{p}_{l_k}$ and ${}^{G}\mathbf{p}_{r_k}$ are the positions of the left and right foot, respectively, during consecutive steps (i.e., both at stance phase) w.r.t. the Glass-IMU global frame $\{G\}$. Summarizing, this stochastic constraint and its corresponding error equation are expressed as:

$$\begin{aligned}\mathbf{z}_{p_k} &= \mathbf{S}\,{}^{G}\mathbf{p}_{h_k} + \mathbf{n}_{p_k} \\ \widetilde{\mathbf{z}}_{p_k} &= \mathbf{H}_{p_k}\widetilde{\mathbf{x}} + \mathbf{n}_{p_k}\end{aligned} \tag{19}$$

where where $\mathbf{H}_{h_k}$ is the corresponding Jacobian and $\mathbf{n}_{p_k}$ is zero-mean, white Gaussian noise with covariance $\sigma_p^2\mathbf{I}_2$.

[3]The translation between the Glass-IMU position and the head's middle point is calculated based on the human-body measurements.

Finally, the cost term contributed to the BLS in this case is:

$$\mathcal{C}_{p_k}(\widetilde{\mathbf{x}}) = ||\mathbf{H}_{p_k}\widetilde{\mathbf{x}} - \widetilde{\mathbf{z}}_{p_k}||^2_{\sigma_p^2\mathbf{I}_2} \tag{20}$$

For multiple footsteps n, we accumulate the contributing cost terms (20) to form the cost function:

$$\mathcal{C}_p(\widetilde{\mathbf{x}}) = \sum_{k=1}^{n} \mathcal{C}_{p_k} \tag{21}$$

D. Batch Least Squares

After adding all cost terms [see (7), (11), (14), (17), and (21)] together, we iteratively minimize the BLS function

$$\begin{aligned}\mathcal{C}(\widetilde{\mathbf{x}}) &= \mathcal{C}_u(\widetilde{\mathbf{x}}) + \mathcal{C}_z(\widetilde{\mathbf{x}}) + \mathcal{C}_v(\widetilde{\mathbf{x}}) + \mathcal{C}_h(\widetilde{\mathbf{x}}) + \mathcal{C}_p(\widetilde{\mathbf{x}}) \\ &= ||\mathbf{H}\widetilde{\mathbf{x}} - \widetilde{\mathbf{z}}||^2_{\mathbf{R}}\end{aligned} \tag{22}$$

using Gauss-Newton's method, The Jacobian $\mathbf{H}$ is evaluated at each iteration and the normal equations are formed by evaluating the residual $\mathbf{r}$ and the Hessian matrix $\mathcal{H}$, i.e.,

$$\begin{aligned}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\widetilde{\mathbf{x}} &= \mathbf{H}^T\mathbf{R}^{-1}\widetilde{\mathbf{z}} \\ \Rightarrow \mathcal{H}\widetilde{\mathbf{x}} &= \mathbf{r}\end{aligned} \tag{23}$$

The normal equations (23) are solved by employing Cholesky factorization of the (sparse) Hessian and consecutively solving two triangular systems using the SuitSparse numerical library [21]. The resulting estimated trajectories are then provided to the animator (see Sec. V) that generates the lower-body motion.

V. Motion Synthesis

Motion synthesizer generates the 3D animation of the articulated human model (head, trunk, and lower limps) during walking. In particular, it takes as input the estimated head/feet trajectories, the detected gait event timings, and the nominal joint angle profiles from a prior gait model (see Fig. 1), and applies the linear warping function of [2] to compute the model's joint angles across time, hence generating the corresponding human postures. Fig. 3 depicts a subset of the nominal joint-angle profiles during the gait cycle, which is used for generating the motion animation. Finally, the synthesizer computes the body's root (pelvis) pose to place it along the estimated trajectory.

These two main steps for generating the animation [(i) compute the joint angles (based on the gait-event timings and nominal joint-angle profiles), and (ii) evaluate the body pose (based on the estimated trajectories and body posture)] are described in detail in the following two sections.

A. Joint Angle Calculation

The synthesizer computes the joint angles at a given time step using the detected GC events and the nominal joint-angle profiles. In particular, a joint-angle θ at time step t is evaluated using a piecewise linear warping function [$\tau = w(t)$] that computes the equivalent GC percentage τ, and applies it to the nominal joint-angle profile $g(\tau)$, i.e.,

$$\theta(t) = g(\tau) = g(w(t)) \tag{24}$$

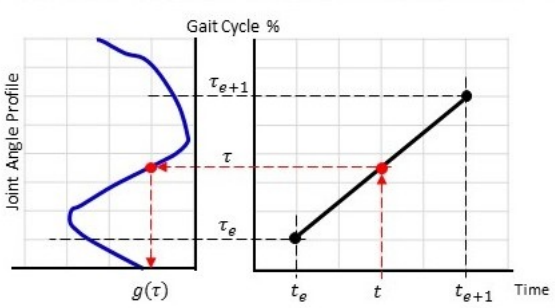

Fig. 5. Piecewise linear motion warping approach.

where, the warping function w is defined as a linear interpolation between the closest detected gait-event times (t_e and t_{e+1}) and their corresponding gait-cycle percentages (τ_e and τ_{e+1}). Fig. 5 illustrates the proposed motion warping approach and the computed variables of interest. Next, forward kinematics are employed to evaluate the head/feet positions w.r.t. the pelvis using the computed joint-angles.

B. Body Pose Evaluation

The final step, after applying the estimated joint-angles to the human model, is to place the model along the estimated head/feet trajectories. To do so, the position of the body root (pelvis) and its orientation (yaw angle, since the person is walking on a planar surface) need to be computed. Note that, it is *not* possible to use the estimated head orientation to place the model; as the head can independently point to different directions while walking. To address this issue, we formulate an optimization problem in the body position and yaw angle that seeks to minimize the distances between the articulated models' head/feet positions and the estimated 3D trajectories, i.e., it computes the body pose that best aligns with the trajectories.

In particular, we introduce a closed-form solution for determining the body position ${}^G\mathbf{p}_B$ w.r.t. the global frame and the yaw angle ${}^G\theta_B$, given the estimated trajectory positions (${}^G\mathbf{p}_i$, $i = 1, 2, 3$ representing the three head/feet points) and their corresponding human model's positions (${}^B\mathbf{p}_i$, computed from applying forward kinematics). The error ϵ_i is defined from the geometric constraint as the following

$$\epsilon_i = {}^G\mathbf{p}_i - ({}^G\mathbf{R}_z({}^G\theta_B){}^B\mathbf{p}_i + {}^G\mathbf{p}_B) \tag{25}$$

where $\mathbf{R}_z(\theta)$ is the rotation matrix around the z-axis with yaw angle θ. Therefore, we solve the minimization problem

$${}^G\mathbf{p}_B^*, {}^G\theta_B^* = \underset{{}^G\mathbf{p}_B, {}^G\theta_B}{\operatorname{argmin}} \left\{\mathcal{C}({}^G\mathbf{p}_B, {}^G\theta_B) \triangleq \frac{1}{2}\sum_{i=1}^{3} ||\epsilon_i||^2\right\} \tag{26}$$

Taking derivatives w.r.t. the body position yields:

$$\frac{\partial\mathcal{C}({}^G\mathbf{p}_B, {}^G\theta_B)}{\partial {}^G\mathbf{p}_B} = 0 \implies {}^G\mathbf{p}_B^* = \frac{1}{3}\sum_{i=1}^{3}\left[\mathbf{R}_z({}^G\theta_B){}^B\mathbf{p}_i - {}^G\mathbf{p}_i\right] \tag{27}$$

Substituting 27 in 26 results in a cost function w.r.t. the yaw angle ${}^G\theta_B$

$$\mathcal{C}'({}^G\theta_B) = \frac{1}{2}\sum_{i=1}^{3} ||\mathbf{R}_z({}^G\theta_B)\mathbf{v}_i - \mathbf{u}_i||^2 \tag{28}$$

where, $\mathbf{v}_i = [v_{xi}\ v_{yi}\ v_{zi}]^T \triangleq \frac{1}{3}\sum_{j=1}^{3} {}^B\mathbf{p}_j - {}^B\mathbf{p}_i$ and $\mathbf{u}_i = [u_{xi}\ u_{yi}\ u_{zi}]^T \triangleq \frac{1}{3}\sum_{j=1}^{3} {}^G\mathbf{p}_j - {}^G\mathbf{p}_i$. Since $\mathbf{R}_z({}^G\theta_B)$ is a rotation matrix around the z-axis, we reformulate the problem as a constraint optimization problem by introducing the variable $\mathbf{x} \triangleq [\cos {}^G\theta_B\ \ \sin {}^G\theta_B]^T$, i.e.,

$$\begin{aligned} \mathbf{x}^* \quad = \quad & \underset{x}{\text{argmin}} \quad \left\{\frac{1}{2}\sum_{i=1}^{3} ||\mathbf{A}_i\mathbf{x} - \mathbf{w}_i||^2\right\} \\ & \text{s.t.} \qquad\qquad \mathbf{x}^T\mathbf{x} = 1 \end{aligned} \tag{29}$$

where, $\mathbf{A}_i \triangleq \begin{bmatrix} v_{xi} & -v_{yi} \\ v_{yi} & v_{xi} \end{bmatrix}$ and $\mathbf{w}_i \triangleq [u_{xi}\ u_{yi}]^T$. Lastly, the body yaw angle is computed from the solution of this constraint optimization problem as $\mathbf{x}^* = \frac{\sum_{i=1}^{3}\mathbf{A}_i^T\mathbf{w}_i}{||\sum_{i=1}^{3}\mathbf{A}_i^T\mathbf{w}_i||}$.

VI. Experimental Results

Our system comprises a Google Glass and two 100Hz Navchip IMU each attached to a foot. For the Google Glass, 100Hz inertial data is received from the Invensense 9150 IMU and 15Hz 320×240 narrow field of view ($45°$) images from the camera. The system estimates head poses every 25 cm and feet poses at every step. The experiments took place within an area of 3×4 meters. We used VICON system for our ground truth comaprisons. Specifically, for position accuracy evaluation, we compared the head/feet trajectory estimates against the VICON ground truth. Similarly for assessing the motion synthesis' accuracy, we compared the computed head/feet positions (evaluated from applying forward kinematics to the articulated model) against the VICON ground truth.

Table I shows the computed root mean square error (RMSE) at each stage of the system (i.e., trajectory estimation and motion synthesis), along with the trajectory lengths. As evident, the error is less than 2% of the distance travelled. Note also that, the trajectory estimator's accuracy is typically better than the motion synthesizer's. This is due to the fact that the estimated joint angles need to be refined during motion. For example, we can use inverse kinematics to refine the estimated joint-angles given head/feet position estimates. Fig.s 6 and 7 illustrate the outputs of the trajectory estimator and motion synthesizer for Dataset 3, while Fig. 8 depicts the position errors (note that the z-position error is very small for both feet due to the incorporated motion constraints). Finally, a video of the corresponding animations of the walking experiments along with the estimated trajectories is shown on the project's webpage [22].

VII. Conclusions

In this paper, we presented, to the best of our knowledge, the first visual-inertial-based motion capture system. The proposed approach takes advantages of the high accuracy of body-mounted camera systems and the usability, low price and computational requirements of the peripheral IMU-based systems within a minimal sensor-based framework. To do so, it addresses two main challenges: (i) estimating the trajectories of different sensors with time-varying relative transformations, and (ii) determining the body posture and joint angles corresponding to the person's motion using only three body-attached IMUs. In particular, our system fuses visual and inertial measurements along with motion constraints in a batch least squares formulation, and incorporates a prior human gait model to generate motion animation. In our implementation, we use a Google Glass and two IMUs attached to each foot. The results of our experiments support the feasibility of the proposed method for decreasing the number of required sensors in motion capture systems. As part of our future work, we plan to improve the estimation accuracy by incorporating the joint-angle profile models within the trajectory estimation process. Finally, we aim to refine the generated animation by considering the effect of additional gait parameters, such as the walking speed and cadence.

TABLE I
RMSE Trajectory Estimation and Motion Synthesis Errors

	Estimator RMSE(m)	Synthesizer RMSE(m)
Dataset 1 - Trajectory Length 29.13m		
Feet	0.13	0.12
Head	0.05	0.12
Dataset 2 - Trajectory Length 3.18m		
Feet	0.07	0.09
Head	0.13	0.09
Dataset 3 - Trajectory Length 8.31m		
Feet	0.09	0.09
Head	0.06	0.14

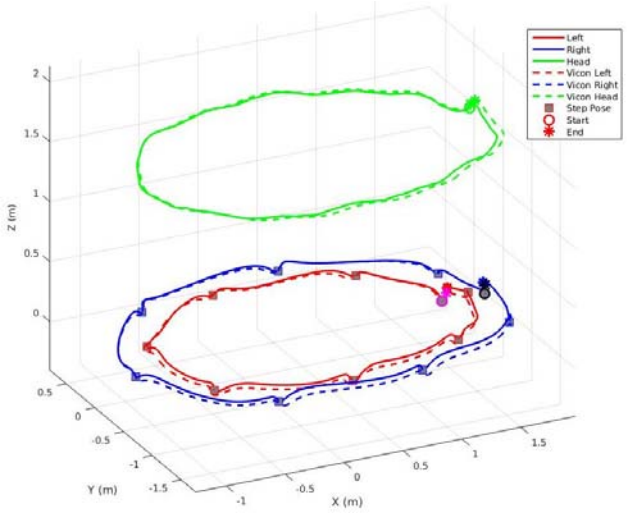

Fig. 6. Dataset 3: Estimated 3D trajectories vs. VICON ground truth.

References

[1] D. Levine, J. Richards, and M. W. Whittle, *Whittle's Gait Analysis*, 5th ed. Churchill Livingstone, 2012.

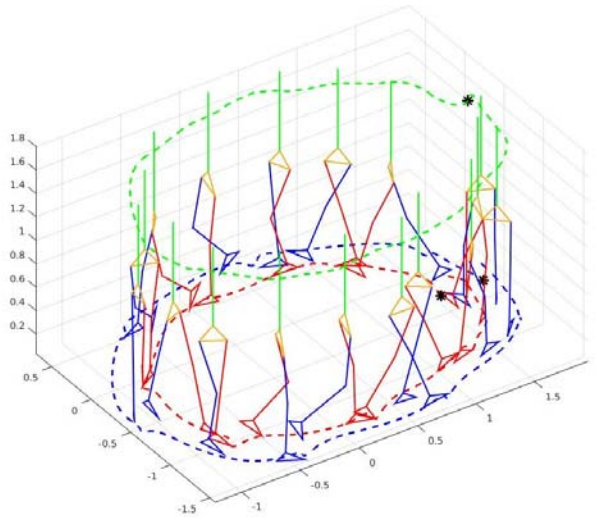

Fig. 7. Dataset 3: Estimated 3D trajectories and animated model.

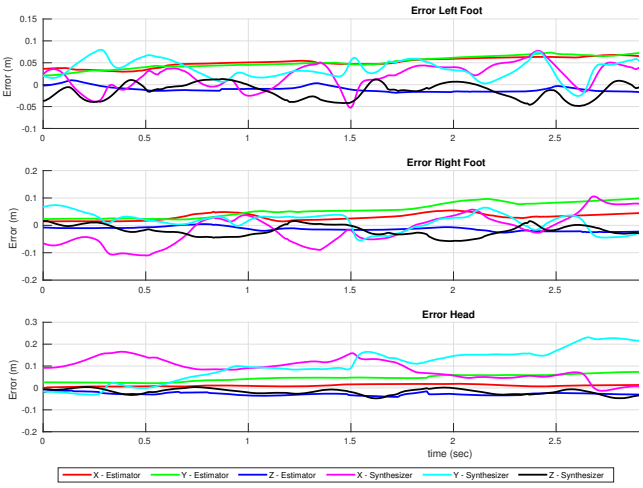

Fig. 8. Dataset 3: Trajectory and motion synthesizer position errors.

[2] A. Witkin and Z. Popovic, "Motion warping," in *Proc. of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, Los Angeles, CA, August 6 – 11 1995, pp. 105–108.

[3] J. Koenemann, F. Burget, and M. Bennewitz, "Real-time imitation of human whole-body motions by humanoids," in *Proc. of IEEE International Conference on Robotics and Automation*, Hong Kong, China, May 31 – June 7 2014, pp. 2806–2812.

[4] D. G. Kottas and S. I. Roumeliotis, "An iterative Kalman smoother for robust 3D localization on mobile and wearable devices," in *Proc. of IEEE International Conference on Robotics and Automation*, Seattle, Washington, May 26 – 30 2015, pp. 6336–6343.

[5] "VICON: Motion capture system," *Online*, https://www.vicon.com/.

[6] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins, "Motion capture from body-mounted cameras," *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 31:1–31:10, 2011.

[7] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt, "Egocap: Egocentric marker-less motion capture with two fisheye cameras," *ACM Transactions on Graphics*, vol. 35, no. 8, pp. 162:1–162:11, 2016.

[8] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors," Xsens Technologies, Tech. Rep., 2009.

[9] J. Tautges, A. Zinke, B. Krüger, J. Baumann, A. Weber, T. Helten, M. Müller, H.-P. Seidel, and B. Eberhardt, "Motion reconstruction using sparse accelerometer data," *ACM Transactions on Graphics*, vol. 30, no. 3, pp. 18:1–18:12, 2011.

[10] Y. Ketema, D. Gebre-Egziabher, M. Schwarts, C. Matthews, and R. Kriker, "Use of gait-kinematics in sensor-based gait monitoring: A feasibility study," *Journal of Applied Mechanics*, vol. 81, no. 4, 2013.

[11] Y. S. Suh, "Inertial sensor-based smoother for gait analysis," *Sensors*, vol. 14, no. 12, pp. 24 338–24 357, 2014.

[12] U. Varshney, *Pervasive Healthcare Computing: EMR/EHR, Wireless and Health Monitoring*, 1st ed. Springer Publishing Company, 2009.

[13] S. Tregillus and E. Folmer, "VR-STEP: Walking-in-place using inertial sensing for hands free navigation in mobile VR environments," in *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose, CA, May 07 – 12 2014, pp. 1250–1255.

[14] R. Drillis, R. Contini, and M. Bluestein, "Body segment parameters: A survey of measurement techniques," *Artificial Limbs*, vol. 8, no. 1, pp. 44–66, 1964.

[15] "Clinical gait analysis CGA normative gait database," available at http://www.clinicalgaitanalysis.com/data/.

[16] N. Trawny and S. I. Roumeliotis, "Indirect Kalman filter for 3D attitude estimation," University of Minnesota, Tech. Rep., March 2005.

[17] R. O. Allen and D. H. Change, "Performance testing of the systron donner quartz gyro," JPL Engineering Memorandum, Tech. Rep., January 1993.

[18] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. of the Alvey Vision Conference*, Manchester, UK, August 31 – September 2 1988, pp. 147–151.

[19] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of the International Joint Conference on Artificaial Intelligence*, Vancouver, British Columbia, August 24 – 28 1981, pp. 674–679.

[20] J.-Y. Bouguet, "Camera calibration toolbox for matlab," 2006, available at http://www.vision.caltech.edu/bouguetj/calibdoc/, version 1.6.0.

[21] T. A. Davis, "SUITESPARSE: A suite of sparse matrix software," available at http://faculty.cse.tamu.edu/davis/suitesparse.html.

[22] "The project webpage," *Online*, http://mars.cs.umn.edu/research/human_motion_project.php.