




Letter

# Resolving Position Ambiguity of IMU-Based Human Pose with a Single RGB Camera

Tomoya Kaichi <sup>1,\*</sup> , Tsubasa Maruyama <sup>2</sup>, Mitsunori Tada <sup>2</sup>  and Hideo Saito <sup>1</sup> 

<sup>1</sup> Graduate School of Science and Technology, Keio University, Yokohama 223-8522, Japan; hs@keio.jp

<sup>2</sup> Human Informatics Research Institute, National Institute of Advanced Industrial Science and Technology, Koto-ku 135-0064, Japan; tbs-maruyama@aist.go.jp (T.M.); m.tada@aist.go.jp (M.T.)

\* Correspondence: kaichi@keio.jp

Received: 1 September 2020; Accepted: 20 September 2020; Published: 23 September 2020



**Abstract:** Human motion capture (MoCap) plays a key role in healthcare and human–robot collaboration. Some researchers have combined orientation measurements from inertial measurement units (IMUs) and positional inference from cameras to reconstruct the 3D human motion. Their works utilize multiple cameras or depth sensors to localize the human in three dimensions. Such multiple cameras are not always available in our daily life, but just a single camera attached in a smart IP devices has recently been popular. Therefore, we present a 3D pose estimation approach from IMUs and a single camera. In order to resolve the depth ambiguity of the single camera configuration and localize the global position of the subject, we present a constraint which optimizes the foot-ground contact points. The timing and 3D positions of the ground contact are calculated from the acceleration of IMUs on foot and geometric transformation of foot position detected on image, respectively. Since the results of pose estimation is greatly affected by the failure of the detection, we design the image-based constraints to handle the outliers of positional estimates. We evaluated the performance of our approach on public 3D human pose dataset. The experiments demonstrated that the proposed constraints contributed to improve the accuracy of pose estimation in single and multiple camera setting.

**Keywords:** human pose estimation; inertial measurement units; single view; sensor fusion

---

## 1. Introduction

Inertial measurement units (IMUs) and RGB cameras are utilized for online human pose estimation in real-world settings. IMUs comprise accelerometers and gyroscopes providing measurements of 3D acceleration and calculated 3D orientation. The acceleration and orientation of the IMU attached to each body segment helps infer human motion [1–3]. RGB cameras are the most commonly used optical sensors and offer two-dimensional (2D) visual information of the environment. Recent image-based human pose estimation methods detect joints of the human body on the image that offer a robust 2D human pose [4–8]. Both devices are widely used in various motion analysis applications; however, they have physical limitations. IMUs suffer from measuring translational motion due to the integration-drift problem. The position error accumulates in time to reach a remarkable value if it is not reset or compensated, so IMUs cannot provide accurate 3D joint positions in the global coordinates. For RGB cameras, it remains difficult to obtain 3D human pose in the wild using a single view due to depth ambiguity, i.e., the 3D position of the points projected onto the 2D image are indefinite in the optical axis direction.

To compensate for these limitations, researchers have developed full-body motion capture (MoCap) systems that incorporate information from IMUs and RGB cameras. 3D human posture and position are simultaneously optimized to be consistent with the orientation of the IMUs and the

silhouettes or joints obtained through convolutional neural networks (CNN) on the images. They have achieved accurate and stable performance in MoCap, but images from multiple viewpoints are required to localize the 3D human position.

Most measurement environments in the real world consist of a single camera rather than multi-viewpoint cameras. In everyday life, cameras (e.g., cameras for surveillance, care systems in the homes of the elderly, and worker-safety systems in factories) are placed to fully cover the space to be monitored. The optimal camera arrangement is to place a minimal number of cameras so that the area where fields of view overlap is small [9]. Assuming these cameras are utilized to capture human posture for the purposes of health care [10,11] or human–robot collaboration [12], a technique for online MoCap in a single-camera environment is desirable. Moreover, inertial sensors have become affordable, and many studies have analyzed human motion using IMUs [13–15]. Recently, IMUs have been embedded in many cellphones and smartwatches, and further spread of IMUs is expected.

In this paper, we present an optimization-based method for online 3D human pose estimation that resolves the positional ambiguity of IMU-based poser with a single camera. Single-camera settings impose two challenges on pose reconstruction: (1) A single-view image cannot constrain the position of the human body in three dimensions due to depth ambiguity, and (2) the results of pose estimation are greatly affected by the failure of image-based constraints, such as outlier detection of the joints. For the first problem, we present 3D positional constraints of ground contact. The timing of the contact is determined from acceleration of IMUs, and the contact position is calculated by back-projecting the 2D foot joints on the image into the floor plane. The joints on the image are detected by a CNN-based method [6]. The proposed objective function is designed to handle the outlier detection of the joint detector, which resolves the second problem.

We experimentally evaluated our method using the public 3D dataset TotalCapture [16], which includes all-synchronized videos, IMU data, and ground-truth human pose. The experiments demonstrated that the cost terms incorporated into our objective function contributed to the accuracy and stability of pose estimation.

## 2. Related Work

### 2.1. IMU-Based Motion Capture

Many approaches for IMU-based MoCap have been proposed over the last decade. Huang et al. regressed the pose parameter of the human model from a small set of IMUs and achieved semi-realtime human pose estimation [1]. However, their method does not provide the global position of the solved human model. Although IMU provides accurate orientation in a high frame rate, it is susceptible to drift in global position. A survey reported that a commercial marker-less motion capture suit composed of 17 IMUs suffers from large positional error [17].

To handle this potential hurdle, von Marcard et al. reconstructed human motion using global optimization [2]. As a result that their method optimizes the pose in all frames simultaneously, it is offline. Another approach focused on human–object contact, which constrains one or more positions the subject touches [3]. This method works well when the contact positions are predefined. However, it accumulates the positional error when the contact positions are determined online. Inspired by the contact constraints on pose reconstruction, our approach utilizes RGB images to compensate for the contact’s position ambiguity.

## 2.2. Image-Based Motion Capture

Improvements of deep neural networks have gained the attention of many researchers in human pose estimation. A recent data-driven method estimates 3D human configuration using only a single RGB camera [18–22]. The image-based 3D posers can be roughly divided into two approaches: estimating the 3D position of keypoints (joints and face landmarks), and inferring the pose parameters of a pre-defined human model. The former approaches do not provide the limbs orientation. The latter estimates the full-body posture including the limbs orientation; however, the literature noted that these framewise estimators are typically trained and evaluated on 3D datasets recorded in constrained and unrealistic environments [23]. On the other hand, the accuracy of 2D pose estimators, which detect human keypoints, has been improved by a number of studies over the last decade [4–8]. Due to its performance stability, we utilized one of the open-source 2D joint detectors [6].

## 2.3. Motion Capture Fusing IMUs and Other Sensors

A line of research on combining IMU and visual information has aimed to achieve full-body MoCap free from positional drift. Images from multi-view cameras are utilized to constrain the subject's position three-dimensionally [16,24–27]. The posture and the global position of the subject is optimized by minimizing the difference between the human silhouettes on the images and the solved human model projected onto the images [24]. Other studies have found that joint positions on 2D images obtained by a CNN-based keypoints detector improve the performance of 3D MoCap [16,25]. The above-mentioned IMUs and image fusion approaches optimize the pose parameter of the human model using the silhouettes and keypoints. Recent work estimate the 3D joint position by lifting 2D multi-view keypoints to the 3D space [27]. As a result that it directly infers the joint position, it does not provide the limb's orientation. Although these approaches are appealing because of their stability and accuracy, at least two viewpoints are required to resolve depth ambiguity and localize the subject.

Researchers have addressed pose estimation combining IMU and single view. Some studies have performed 3D human tracking with IMUs and a single depth sensor, such as Kinect [10,28]. However, the measurement accuracy of Kinect decreases outdoors. The only study that has dealt with 3D MoCap with IMUs and a single RGB camera simultaneously optimizes human pose for a certain period of frames, and the global optimization is processed offline [29]. An offline method uses all frames in a sequence to optimize the human pose of a certain frame in the sequence. Offline methods are used for motion analysis after the movement of the subject, especially in the sports and rehabilitation field. On the other hand, online methods that use current frame and/or previous frames to estimate the human pose can be applied to human–robot interactions and monitoring the subjects for healthcare. To the best of our knowledge, no study addressed online MoCap using IMUs and a single RGB camera.

## 3. Methods

### 3.1. Pose Parameterization and Calibration

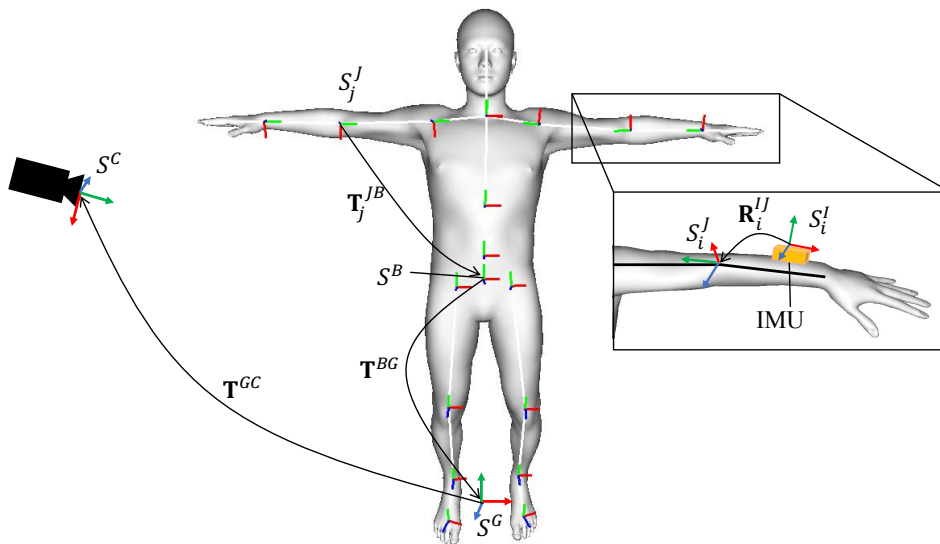
We parameterize the subject's pose using a Digital Human Model (DHM) [30] that consists of a 48 degrees of freedom (DoF) link configuration. The model provides kinematics and the body mesh when the pose including the global translation  $\theta$  ( $\in \mathbb{R}^{51}$ ) is determined. We extend the IMU-based MoCap method [3] for pose parameterization and optimization.

The transformation matrices among global coordinates  $S^G$ , camera coordinates  $S^C$ , body coordinates  $S^B$ ,  $j$ -th joint coordinates  $S_j^J$ , and  $i$ -th IMU local coordinates  $S_i^I$  are required for fusing the sensors on motion tracking. Figure 1 shows relations between the coordinates and transformation matrices. The transformations between the global coordinates and the camera coordinates  $T^{GC}$  is determined using a checkerboard [31]. In our configuration, the checkerboard is placed on the floor. The Z-axis of the global coordinates ( $X_w, Y_w, Z_w$ ), defined by the checkerboard, points in the opposite direction of gravity, and the  $Z_w = 0$  plane coincides with the floor. Note that the checkerboard can be removed after the camera is calibrated and fixed. After the camera setup, the subject wearing

IMUs takes a calibration pose (e.g., T-pose: standing upright and keeping both arms horizontal). The rotational transformation from each IMU to the joint coordinate is obtained from

$$\mathbf{R}_i^J = \mathbf{R}_i^J(\theta_0) \cdot (\mathbf{R}_i^I(t_0))^{-1}, \quad (1)$$

where  $\mathbf{R}_i^I(t_0)$  represents the  $i$ -th IMU sensor orientation in the global coordinates when the subject takes the calibration pose, and  $\mathbf{R}_i^J(\theta_0)$  denotes the rotation matrix of the model joint belonging to the bone to which the IMU is attached in the global coordinates.  $t_0$  and  $\theta_0$  represent the frame and pose parameter of the calibration pose, respectively. As illustrated in Figure 1,  $\mathbf{R}_i^J(\theta_0)$  can be represented by the conversion of the coordinates from the global coordinates  $S^G$  to the local coordinates of each joint  $S_j^J$  of the human model. It can be calculated by transformation matrix  $\mathbf{T}_j^{IB}(\theta_0)$  and  $\mathbf{T}^{BG}(\theta_0)$ .  $\mathbf{T}_j^{IB}(\theta_0)$  denotes the transformation from  $S_j^J$  to the body coordinates  $S^B$ . In our method,  $S^B$  is defined to correspond with the local coordinates of the pelvis joint of the human model. The transformation  $\mathbf{T}_j^{IB}(\theta_0)$  can be obtained from the forward kinematics of predefined link configuration of the model.  $\mathbf{T}^{BG}(\theta_0)$ , transformation from the body coordinates to the global coordinates, is determined by the position and orientation of the subject taking the calibration pose.



**Figure 1.** Relations among the local coordinate systems.

For synchronizing the data from IMUs and a camera, a physical cue that can be detected from both the camera and IMUs can be used when it is difficult to synchronize a camera and multiple IMUs with a signal synchronizing apparatus. For example, a footstamp is applicable because, for the camera, the timing of the cue is obtained from the motion of ankle joint detected on the image, and for the IMUs, the timing can be calculated from the acceleration measurements of the IMU attached to foot. The synchronization should be performed after the calibration pose.

### 3.2. Full-Body Pose Optimization

We follow the paradigm of constraint-based motion tracking. More specifically, we minimize the following total cost function composed of multiple cost terms on a per-frame basis.

$$E(\theta) = E_O(\theta) + \lambda_{RoM} E_{RoM}(\theta) + \lambda_P E_P(\theta) + \lambda_G E_G(\theta), \quad (2)$$

where  $E_O(\theta)$  and  $E_{RoM}(\theta)$  constrain the orientation and the range of motion of the model joints, respectively.  $E_P(\theta)$  and  $E_G(\theta)$  represent the positional error of the joints and the ground contact points, respectively. We design these positional error terms so as to stably estimate the human

pose in an under-constrained environment. Every term is weighted by a corresponding weight  $\lambda$ . The quasi-Newton algorithm [32] is applied to solve the optimization problem.

### 3.2.1. IMU-Based Constraints

The orientation of the kinematic links is estimated from the measured orientation of IMU sensors. The cost term is represented as the sum of the orientation differences between IMU measured and estimated bone orientation. Here, the  $i$ -th IMU offers its orientation in each local coordinates. Using the transformation matrix from the sensor coordinates to the joint coordinates  $\mathbf{R}_i^{\text{IJ}}$  (Equation (1)), the cost  $E_O(\theta)$  can be expressed as

$$E_O(\theta) = \sum_{i=1}^{N_I} \|\mathbf{R}_i^{\text{IJ}} \cdot \mathbf{R}_i^{\text{I}} - \mathbf{R}_i^{\text{I}}(\theta)\|_{\text{F}}^2, \quad (3)$$

where  $\mathbf{R}_i^{\text{I}}$ , and  $\mathbf{R}_i^{\text{I}}(\theta)$  denote the sensor measurement and solved value of bone orientation in the current frame, respectively.  $N_I$  describes the number of IMUs.

The other IMU-based constraint,  $E_{\text{RoM}}(\theta)$ , adds cost when the estimated joint angle exceeds or falls short of the RoM  $\psi$ .  $\psi$  defines the minimum and maximum joint angles, i.e.,  $\psi \in \{(\psi_r^{\min}, \psi_p^{\min}, \psi_y^{\min}), (\psi_r^{\max}, \psi_p^{\max}, \psi_y^{\max})\}$ , where  $r$ ,  $p$ , and  $y$  represent the three principal axes in the joint coordinates. The cost for each joint is calculated according to

$$e_{\text{RoM}}(\phi(\theta), \psi) = \sum_{k \in \{r, p, y\}} \begin{cases} \rho((\phi_k(\theta) - \psi_k^{\min})^2) & (\phi_k(\theta) < \psi_k^{\min}) \\ \rho((\phi_k(\theta) - \psi_k^{\max})^2) & (\phi_k(\theta) > \psi_k^{\max}) \\ 0 & (\text{otherwise}) \end{cases}, \quad (4)$$

where  $\phi_k(\theta)$  represents the estimated rotation around the  $k$ -axis of the joint.  $\rho(\cdot)$  is a loss function detailed in Section 3.2.2. Then, we can compute the RoM cost for the entire body by

$$E_{\text{RoM}}(\theta) = \sum_{j=1}^{N_J} e_{\text{RoM}}(\phi^{(j)}(\theta), \psi^{(j)}), \quad (5)$$

where  $N_J$ ,  $\phi^{(j)}(\theta)$ , and  $\psi^{(j)}$  denote the number of joints whose rotation is estimated, the  $j$ -th joint angles, and the  $j$ -th joint RoM, respectively. We adopt the RoM defined in the commercial Digital Human Model [30].

### 3.2.2. Image-Based Constraints

$E_P(\theta)$  constrains positional differences between keypoints on an image  $\mathbf{p}^{\text{C}}$  detected by a CNN-based 2D pose estimator [6] and corresponding 3D joint positions projected onto the image  $\hat{\mathbf{p}}^{\text{C}}$ . The 3D point of the solved model in the body coordinates  $\hat{\mathbf{P}}^{\text{B}}$  can be projected to the camera coordinates by

$$\hat{\mathbf{p}}^{\text{C}}(\theta) = \mathbf{T}^{\text{GC}} \mathbf{T}^{\text{BG}}(\theta_0) \hat{\mathbf{P}}^{\text{B}}(\theta), \quad (6)$$

where  $\mathbf{P}_j$  denotes the 4D column vector, which represents the 3D joint position in a homogeneous coordinate system.  $\mathbf{T}^{\text{GC}}$  and  $\mathbf{T}^{\text{BG}}(\theta_0)$  are the  $4 \times 3$  translation matrices described in Section 3.1.

As a result that the global position of the estimated model is constrained by visual information from only one RGB camera, the failure of the 2D joint detector seriously compromises motion tracking accuracy. To improve the robustness to such outlier detection of keypoints, we extend Tukey's biweight. Specifically, the cost term of a joint is less weighted when the joint-position estimate is far from the model joint in the previous frame. The weight is calculated by

$$w_p = \begin{cases} \exp\left(-\frac{d_p^2}{2s^2k_p^2}\right) & (d_p \leq \beta_d s k_p) \\ 0 & (\text{otherwise}) \end{cases}, \quad (7)$$

where  $p$  ( $1 \leq p \leq N_p$ ),  $\beta_d$ , and  $s$  are the index of detected joints, a hyperparameter that controls the range of nonzero weight, and the scale of distribution, respectively. Here,  $N_p = 18$ ,  $\beta_d = 2$ , and  $s = 140$  in our experiments.  $d_p$  represents the Euclidean distance between the detector estimate and the projected point of the corresponding joint in the previous frame, and  $k_p$  denotes the standard deviation of the weight distribution. The distribution of keypoints detected by the data-driven 2D pose estimator depends on the keypoint type. For example, the distribution of an eye must be smaller than that of hips. The value of  $k_p$  is defined by object keypoint similarity (OKS) [33], which is used to evaluate the performance of the 2D keypoint detectors; that is, keypoint detectors ensure accuracy in this distribution. The positional cost weighted with  $w_p$  is expressed as

$$E_p(\theta) = \sum_{p=1}^{N_p} \rho(w_p c_p^{\text{im}} \|\mathbf{p}_p^{\text{C}} - \hat{\mathbf{p}}_p^{\text{C}}(\theta)\|_{\mathbb{F}}^2), \quad (8)$$

where  $c_p^{\text{im}}$  represents the confidence score from the keypoint detector.

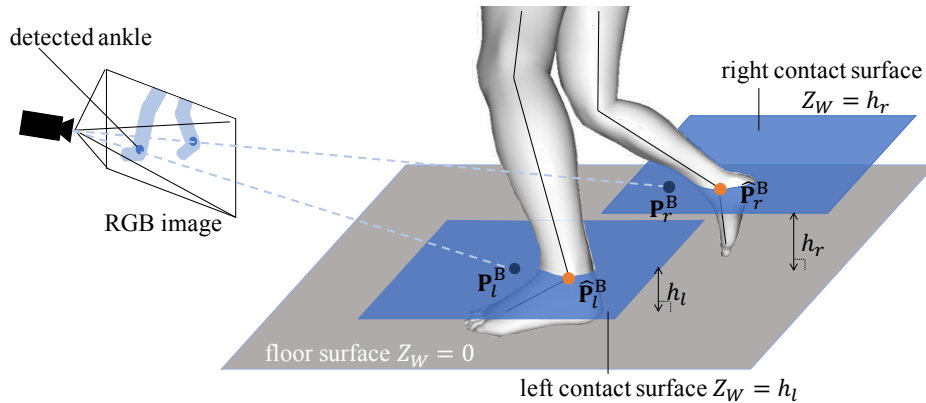
In our single-camera setting,  $E_p(\theta)$  alone cannot localize the global position of the model due to the camera's depth ambiguity. To optimize the model position three dimensionally, we present the ground contact cost term  $E_G(\theta)$ . Fusing IMU acceleration and positional measurement from the camera,  $E_G$  minimizes the distance between foot position and ground contact point.

We define the cost as depicted in Figure 2. Let  $\hat{\mathbf{p}}_g^{\text{B}}(\theta)$ , where  $g \in \{\text{left\_foot}, \text{right\_foot}\}$  is the left or right ankle position of the estimated model, and let  $\mathbf{p}_g^{\text{B}}$  be the intersection between the contact surface and the line where the 2D ankle keypoint is back-projected into three dimensions. The contact surfaces are the planes parallel to the floor plane, and each contact surface passes through each ankle of the solved model. The floor plane can be determined by camera calibration as described in Section 3.1. The confidence score  $c_g^{\text{G}}$  that the foot is on the ground is determined from the acceleration of the foot-attached IMU and the height of the foot. The resulting ground contact cost is calculated according to

$$E_G(\theta) = \sum_g \rho(c_g^{\text{G}} w_g c_g^{\text{im}} \|\mathbf{p}_g^{\text{B}} - \hat{\mathbf{p}}_g^{\text{B}}(\theta)\|_{\mathbb{F}}^2), \quad (9)$$

$$\text{where } c_g^{\text{G}} = \delta + \begin{cases} \beta_G / \|\mathbf{a}_g\| & (\beta_G / \|\mathbf{a}_g\| \leq 1) \\ 1 & (\text{otherwise}) \end{cases},$$

where  $\mathbf{a}_g$  and  $\beta_G$  represent the acceleration measured by the IMU attached to the foot  $g$  and a constant value to determine the gradient, respectively. For all experiments,  $\beta_G = 5$  and  $\beta_G / \|\mathbf{a}\|$  was calculated using  $\beta_G / (\|\mathbf{a}\| + \epsilon)$ ,  $\epsilon = 1.0 \times 10^{-6}$  to avoid zero division.  $\delta$  takes 1 when the lowest mesh of  $g$  is lower than that of the other foot, and 0 otherwise.  $w_g$  is also multiplied for handling outlier detection of foot keypoints. In our method, the Cauchy loss function,  $\rho(x) = \log(1 + x)$ , is used as a loss function  $\rho(\cdot)$  in the range of motion cost term  $E_{RoM}$ , image-based positional cost term  $E_p$ , and ground contact cost term  $E_G$ . The Cauchy loss function suppresses extremely large values so that the effect of the error of one joint on the total loss does not become too large in the process of the optimization calculation. An example of extremely large error is that when the distance from the camera to the subject is large and camera position is relatively low, the small 2D position error of detected joints on the image causes huge error in the 3D space.



**Figure 2.** Visualization of the ground contact constraint.

## 4. Evaluation

### 4.1. Dataset

We quantitatively evaluate the performance of our approach on 3D human pose dataset TotalCapture [16]. TotalCapture provides 60 fps of all-synchronized IMU data, HD videos from fixed cameras, and ground-truth human pose measured by optical MoCap. A total of 13 IMUs are attached on the head, sternum, pelvis, upper and lower limbs, and feet. Our method uses acceleration and orientation of IMUs, and an image sequence from a single camera. Note that optical MoCap data are not used for our approach. The original ground-truth of the joint position and orientation is obtained by fitting the marker position measured by optical motion capture system to the surface of the human model. The human model of the optical motion capture has a different definition of the link structure from that of DHM we used for pose estimation. For example, the pelvis joint to neck joint is divided into 5 segments in the original ground-truth, but it is divided into 3 segments in DHM. Therefore, it is not possible to make a strict comparison of the joint position and orientation between the estimated pose of DHM and the original ground truth. Hence, we determined the joint position and orientation of DHM so that the Vicon 57-point markers defined in advance on the DHM surface matches the marker position measured by optical motion capture [30], and used it as the ground-truth in this experiment.

We quantitatively evaluated our method following the standard evaluation protocol defined in [16]. In the protocol, the test set consists of 15 scenes in total including the scenes Walking 2 (W2), Acting 3 (A3), and Freestyle 3 (F3) of Subjects S1, S2, S3, S4, and S5. However, there are several sequences in which both feet are off the ground for several frames in a row, such as jumping, in S2-F3, S3-F3, and S5-A3. These scenes are excluded from our dataset and we used S2-ROM3 (S2-R3), S3-F1, and S5-F1 instead. The limitations on the scenes where our method is effective will be mentioned in Section 5.

### 4.2. Implementation Details

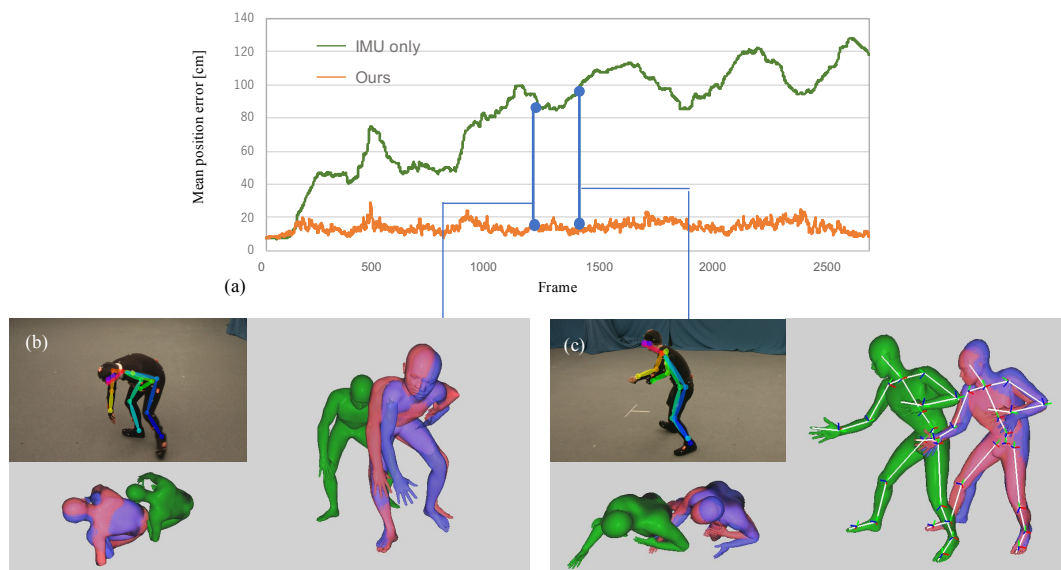
We utilized a human model generated statistically from the height and weight of the subject, which is offered by DHM software [30]. Before starting the pose estimation, the subject took T-pose as a calibration pose. During the calibration pose, the global coordinates  $(X_W, Y_W, Z_W)$  is defined so that the subject stands on the plane at  $Z_W = 0$ . For the model of the 2D joint detector used in image-based constraints, we utilized the weights of public pretrained model [6]. No additional training or finetuning is conducted.

The weighting parameter controls the contribution of each cost term to the overall cost Equation (2). The algorithm based on Tree-structured Parzen Estimator is used to seek the parameter values. Several scenes other than the test set are used for parameter tuning and the value found are  $\lambda_{RoM} = 0.01$ ,  $\lambda_P = 5.0 \times 10^{-4}$ , and  $\lambda_G = 5.0 \times 10^{-3}$ . The parameters are fixed through all experiments.

### 4.3. Contribution of the Proposed Cost Terms

We evaluated how the proposed cost term  $E_G(\theta)$  and the adaptive biweight  $w_p$  work in the constraint-based pose optimization. In this experiment, a full set of 13 IMUs and a single camera that captures entire movement in the field of view were used. The position error in this section represents the mean 3D Euclidean distance between the estimated model and the ground truth over the 16 joints.

The graph of Figure 3a represents per-frame mean Euclidean distance between the solved pose and ground-truth. Figure 3b,c visualize the output of the 2D joint detector [6], and the human models colored in green, red, and blue represent the 3D human pose solved by the IMU only method [3], the proposed method, and optical MoCap (ground-truth), respectively. The estimated 2D joints and 3D models in (b) and (c), respectively capture the same frame in the same scene.



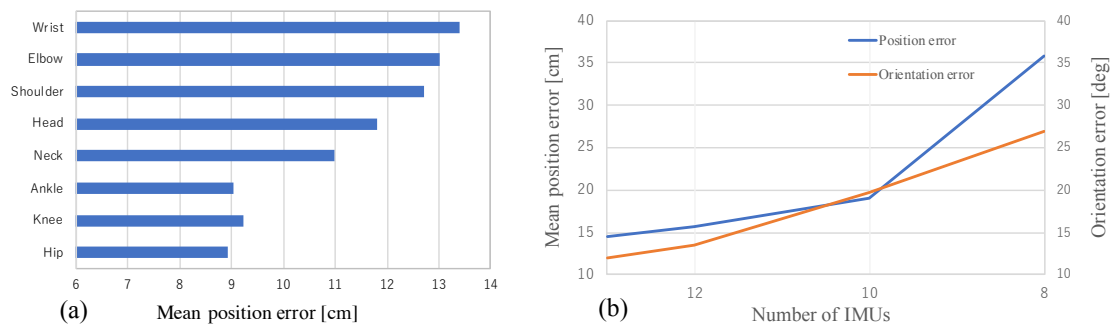
**Figure 3.** The top graph (a) represents per joint mean position error for each frame. The bottom figures (b) and (c) illustrate the the view of the used single camera and the detected joints by the 2D joint detector, OpenPose [6]. The human models colored in green, red, and blue represent the inference by IMU only, the proposed approach, and ground-truth from optical MoCap, respectively. It is observed that the position of the foot touching the ground is estimated correctly.

Figure 3a and the human model visualized from above revealed that our approach using a single camera prevented the accumulation of position error. The right foot in (c) is self-occluded and the misdetection occurred; however, our approach robustly optimized the 3D full-body pose. Focusing on the feet in (b) and (c), the foot touching the ground and fixed (right foot in (b) and left foot in (c)) are estimated with higher accuracy in these frames. It would be due to the proposed ground contact cost term.

Table 1 summarizes the quantitative results for pose estimation using the position error metric. RGB only [34] is the state-of-the-art of 3D human pose estimation using only a single RGB camera.  $F(E_O, E_{RoM}, E_P)$  estimates the human pose by minimizing the cost function composed of  $E_O(\theta)$ ,  $E_{RoM}(\theta)$ , and  $E_P(\theta)$ . The results revealed that the ground contact cost term  $E_G(\theta)$  improves the positional error.  $F(full, w_p = 1)$  optimizes the pose by Equation (2), but adaptive weight  $w_p$  is fixed to 1. Meanwhile, the proposed cost function  $F(full)$  calculates  $w_p$  according to Equation (7). Although the mean error of  $F(full)$  in the 15 scenes was smallest,  $F(full, w_p = 1)$  estimated the human pose with the highest accuracy in more than half of the test scenes. Especially in Walking 2 (W2),  $F(full, w_p = 1)$  outperformed  $F(full)$  in 4 out of 5 trials. The results indicate that in the scene where the 2D joint detector estimates the 2D pose of the subject with high accuracy, the 3D pose reconstruction accuracy is slightly lowered by the adaptive biweight  $w_p$ ; however,  $w_p$  stabilizes the 3D pose estimation when there are misdetections of the joints on a image due to the self-occlusion or unusual posture of the



subject (included in Freestyle 3 and Acting 3). The effect of the ground contact cost term is validated from Figure 4a. It represents per-joint position error of human model estimated by the proposed method with a single view and 13 IMUs. Although the estimation error of the hands and feet tends to be large because the limbs move a lot, the positional error of the ankle is relatively small due to the 3D positional constraints of the ground contact.



**Figure 4.** (a) Mean per-joint positional error of the human motion capture (MoCap) by the proposed method on all the scenes in the test set. The error values of wrist, elbow, shoulder, ankle, knee, and hip represent the average error of the both side of the segments, i.e., the error of the wrist denotes the average error of left wrist and right wrist. (b) Mean 3D position and orientation errors on subjects S3-F1 and S4-F3 with 8 to 13 IMUs.

The mean orientation error of joints is shown in the bottom of Table 1. The error of IMU only and the proposed method ( $F(full)$ ) were 8.75 degrees and 8.83 degrees, respectively, and no significant differences were observed.

**Table 1.** 3D position error (cm) on TotalCapture dataset.

	S1			S2			S3			S4			S5			Mean
	W2	A3	F3	W2	A3	R3	W2	A3	F1	W2	A3	F3	W2	F1	F3	
Mean position error (cm)																
<i>RGB only</i> [34]	52.4	90.1	22.5	33.3	22.6	27.4	51.4	26.9	24.6	50.4	53.3	56.1	57.7	37.1	43.1	43.3
<i>IMU only</i> [3]	45.0	42.7	44.2	144	63.9	8.91	34.8	72.3	62.4	42.3	221	39.4	124	32.9	81.0	70.6
$F(E_O, E_{RoM}, E_P)$	54.4	41.7	29.4	142	63.3	12.2	33.0	68.8	68.5	42.8	224	39.2	124	28.2	78.1	70.0
$F(full, w_p = 1)$	<b>19.6</b>	<b>14.8</b>	<b>11.9</b>	<b>11.5</b>	<b>9.22</b>	<b>7.37</b>	<b>15.3</b>	<b>10.1</b>	<b>14.3</b>	<b>15.7</b>	<b>13.8</b>	<b>14.6</b>	<b>14.9</b>	<b>46.7</b>	<b>17.5</b>	<b>15.8</b>
$F(full)$	20.2	15.6	12.2	12.2	10.2	<b>7.32</b>	<b>15.2</b>	12.5	<b>11.1</b>	16.3	<b>12.3</b>	14.7	16.0	<b>10.0</b>	<b>16.9</b>	<b>13.5</b>
Mean orientation error (degrees)																
<i>IMU only</i> [3]	9.32	8.25	9.43	8.59	8.27	12.5	6.50	6.55	10.6	7.10	8.14	9.51	6.59	8.37	11.6	8.75
$F(full)$	9.38	8.45	9.45	8.74	8.51	12.5	6.65	6.63	10.9	7.07	8.20	9.52	6.72	8.37	11.3	8.83

The minimum error values are shown in bold.

The proposed method can easily be extended to use multi-view cameras by adding the image-based cost function  $E_P(\theta)$  and  $E_G(\theta)$  for each camera and simultaneously minimize the total cost. We performed the experiments using 8 cameras and 13 IMUs. The state-of-the-art approach for 3D MoCap that infers both joint position and orientation from IMUs and multiple images [25] extracted several images from TotalCapture to test their approach. The performance of our approach was compared with [16,25] on the same scenes as the test set of [25], excluding the scenes where the subject jumped. As shown in Table 2, in several scenes, our method outperformed the conventional approach that optimizes the pose parameter to reconstruct human motion. In the scene where our approach was inferior in accuracy (S2-R3), the subject frequently crouched and bent forward. It appears that these motions caused self-occlusion of the ankle and the ground contact constraint did not work. The experiments demonstrate that the proposed ground contact constraint contributes to improve the accuracy of 3D human pose estimation in multi-view camera setting as well as single-camera setting when the floor plane is pre-defined and the foot can be detected from the camera.

**Table 2.** 3D orientation error (degrees) on TotalCapture dataset.

	S1-F3	S2-R3	S3-F1	S4-F3	S5-F1	Mean
Trumble et al. [16]	9.4	9.3	13.6	11.6	10.5	10.9
Malleson et al. [25]	7.4	<b>3.9</b>	6.7	6.4	7.0	6.3
$F_{multi}(full)$	<b>6.25</b>	5.66	6.70	<b>6.32</b>	<b>5.91</b>	<b>6.17</b>

The minimum error values are shown in bold.

#### 4.4. The Number of IMUs

Wearing many IMUs takes time and hampers the subject's range of motion. Towards the real-world use of our method, we investigated the relation between the accuracy of the pose estimation and the number of IMUs. The experiments were conducted with (1) 13 IMUs: full set as described in Section 4.1, (2) 12 IMUs: full set without head, (3) 10 IMUs: IMUs on upper arms removed from (2), and (4) 8 IMUs: IMUs on upper legs removed from (3). 3D position and orientation errors in different IMU configurations are shown in Figure 4b.

The decrease of the IMUs largely affects the accuracy of both position and orientation. It would be because our single-camera approach does not constrain joint positions other than the foot in three dimensions. In the experiments on IMU only and  $F(E_O, E_{RoM}, E_P)$ , the objective function diverged with 8 IMUs. The proposed ground contact cost term  $E_G(\theta)$  and  $w_p$  contributed to the convergence of pose estimation.

## 5. Conclusions and Future Work

We have presented the first online approach to estimate the 3D human pose fusing IMUs and a single camera. In order to constrain the position of the solved model in three dimensions, the proposed cost term detects the timing and position of foot grounding. We handle the outlier of visual information by extending the biweighting algorithm. The experimental results showed that the proposed objective function stably estimated the 3D human pose, including the global position.

To calculate the confidence of foot grounding, it is assumed in Equation (9) that one foot is grounded. Therefore, the accuracy of the proposed approach degrades in a sequence in which a subject lifts both feet off the ground for long time, such as by jumping. We confirmed that the short period of foot takeoff does not seriously affect the accuracy by the experiment on S5-F1, which included side-skip steps. This limitation will be overcome by inferring ground contact confidence from visual context and IMU data.

**Author Contributions:** Conceptualization, T.K., T.M., M.T., and H.S.; methodology, T.K., T.M., M.T., and H.S.; software, T.K., T.M., and M.T.; validation, T.K., T.M., and M.T.; formal analysis, T.K., T.M., and M.T.; investigation, T.K., T.M., and M.T.; resources, T.M. and M.T.; data curation, T.K., T.M., and M.T.; writing—original draft preparation, T.K.; writing—review and editing, T.M., M.T., and H.S.; visualization, T.K.; supervision, M.T. and H.S.; project administration, M.T. and H.S.; funding acquisition, T.K. and H.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by Grant-in-Aid for JSPS Fellows (19J22153) and JST AIP-PRISM Grant Number JPMJCR18Y2.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M.J.; Hilliges, O.; Pons-Moll, G. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Trans. Graph. TOG* **2018**, *37*, 1–15. [CrossRef]
2. Von Marcard, T.; Rosenhahn, B.; Black, M.J.; Pons-Moll, G. Sparse inertial poser: Automatic 3D human pose estimation from sparse IMUs. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2017; Volume 36, pp. 349–360.

3. Maruyama, T.; Tada, M.; Toda, H. Riding Motion Capture System Using Inertial Measurement Units with Contact Constraints. *Int. J. Autom. Technol.* **2019**, *13*, 506–516, doi:10.20965/ijat.2019.p0506. [[CrossRef](#)]
4. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
5. Luvizon, D.C.; Tabia, H.; Picard, D. Human pose regression by combining indirect part detection and contextual information. *Comput. Graph.* **2019**, *85*, 15–22. [[CrossRef](#)]
6. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
7. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
8. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
9. Gonzalez-Barbosa, J.J.; García-Ramírez, T.; Salas, J.; Hurtado-Ramos, J.B. Optimal camera placement for total coverage. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 844–848.
10. Kalkbrenner, C.; Hacker, S.; Algorri, M.E.; Blechschmidt-Trapp, R. Motion Capturing with Inertial Measurement Units and Kinect. In Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies, Angers, France, 3–6 March 2014; Volume 1, pp. 120–126.
11. Haynes, S.; Williams, K. Impact of seating posture on user comfort and typing performance for people with chronic low back pain. *Int. J. Ind. Ergon.* **2008**, *38*, 35–46. [[CrossRef](#)]
12. Liu, H.; Wang, L. Gesture recognition for human-robot collaboration: A review. *Int. J. Ind. Ergon.* **2018**, *68*, 355–367. [[CrossRef](#)]
13. Bousdar Ahmed, D.; Munoz Diaz, E.; García Domínguez, J.J. Automatic Calibration of the Step Length Model of a Pocket INS by Means of a Foot Inertial Sensor. *Sensors* **2020**, *20*, 2083. [[CrossRef](#)]
14. Zihajehzadeh, S.; Yoon, P.K.; Kang, B.S.; Park, E.J. UWB-aided inertial motion capture for lower body 3-D dynamic activity and trajectory tracking. *IEEE Trans. Instrum. Meas.* **2015**, *64*, 3577–3587. [[CrossRef](#)]
15. Garofalo, G.; Argones Rúa, E.; Preuveneers, D.; Joosen, W. A Systematic Comparison of Age and Gender Prediction on IMU Sensor-Based Gait Traces. *Sensors* **2019**, *19*, 2945.
16. Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; Collomosse, J. *Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors*; BMVC: Guildford, UK, 2017; Volume 2, p. 3.
17. Guo, L.; Xiong, S. Accuracy of base of support using an inertial sensor based motion capture system. *Sensors* **2017**, *17*, 2091. [[CrossRef](#)] [[PubMed](#)]
18. Veges, M.; Lorincz, A. Multi-Person Absolute 3D Human Pose Estimation with Weak Depth Supervision. *arXiv* **2020**, arXiv:2004.03989.
19. Martinez, J.; Hossain, R.; Romero, J.; Little, J. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2659–2668.
20. Nibali, A.; He, Z.; Morgan, S.; Prendergast, L. 3d human pose estimation with 2d marginal heatmaps. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1477–1485.
21. Moreno-Noguer, F. 3D Human Pose Estimation From a Single Image via Distance Matrix Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1561–1570.
22. Xiang, D.; Joo, H.; Sheikh, Y. Monocular total capture: Posing face, body, and hands in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10965–10974.
23. Arnab, A.; Doersch, C.; Zisserman, A. Exploiting temporal context for 3D human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3395–3404.

24. Von Marcard, T.; Pons-Moll, G.; Rosenhahn, B. Human pose estimation from video and IMUs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1533–1547. [[CrossRef](#)] [[PubMed](#)]
25. Malleson, C.; Gilbert, A.; Trumble, M.; Collomosse, J.; Hilton, A.; Volino, M. Real-time full-body motion capture from video and IMUs. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 449–457.
26. Pons-Moll, G.; Baak, A.; Helten, T.; Müller, M.; Seidel, H.P.; Rosenhahn, B. Multisensor-fusion for 3d full-body human motion capture. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 12–18 June 2010; pp. 663–670.
27. Zhang, Z.; Wang, C.; Qin, W.; Zeng, W. Fusing Wearable IMUs With Multi-View Images for Human Pose Estimation: A Geometric Approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020; pp. 2197–2206.
28. Zheng, Z.; Yu, T.; Li, H.; Guo, K.; Dai, Q.; Fang, L.; Liu, Y. HybridFusion: real-time performance capture using a single depth sensor and sparse IMUs. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 384–400.
29. Von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 601–617.
30. Endo, Y.; Tada, M.; Mochimaru, M. Dhaiba: Development of virtual ergonomic assessment system with human models. In Proceedings of the 3rd International Digital Human Symposium, Tokyo, Japan, 20–22 May 2014; pp. 1–8.
31. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [[CrossRef](#)]
32. Dennis, J.E., Jr.; Moré, J.J. Quasi-Newton methods, motivation and theory. *SIAM Rev.* **1977**, *19*, 46–89. [[CrossRef](#)]
33. Ruggero Ronchi, M.; Perona, P. Benchmarking and error diagnosis in multi-instance pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 369–378.
34. Moon, G.; Chang, J.Y.; Lee, K.M. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 10133–10142.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).