Human Motion Capture
Based on Geometric Consistency
Using Visual and Inertial Sensors

February 2022

Tomoya Kaichi

A Thesis for the Degree of Ph.D. in Engineering

Human Motion Capture Based on Geometric Consistency Using Visual and Inertial Sensors

February 2022

Graduate School of Science and Technology,
Keio University

Tomoya Kaichi

Abstract

The automatic capture of human motion has attracted much attention in computer vision and biomechanics communities. The research in this area has aimed to obtain indicators that represent human movements, such as 3D human pose, joint torques, and the body center of mass (CoM). This thesis focuses on motion capture in real-world environments. An RGB camera is one of the most commonly used sensors for full-body motion capture since the visual sensor captures the kinematic and shape information of people in the image. Besides the RGB camera, inertial measurement units (IMUs) have become a prominent option for analyzing human motion in the last few years. Body-worn IMUs measure the movement of a person over a wide area without being disturbed by occlusions.

Exploiting and combining the benefits of these sensors, this thesis addresses motion capture under scenarios where a subject cannot wear the devices and can wear the sensors, respectively. For the former scene, this thesis proposes a multi-view image-based motion capture method. It reconstructs the human body by back-projecting 2D body keypoints (joints and face landmarks) and silhouettes into a 3D space. In the experiment, the trajectories of the CoM position during a baseball swing were estimated as an application to verify the effectiveness of the proposed method. For the latter scene, this thesis presents a method for 3D human pose estimation using a single camera and multiple body-worn IMUs. Unlike the existing visual-inertial motion capture approaches that require multi-view cameras to locate the body parts, the proposed method localizes the body by constraining the foot-ground contact positions with a single camera. The single-view setting expands the measurement range. The experiments on a standard benchmark dataset demonstrated that the proposed approach reduced mean joint position errors by 68.8% and 80.9% compared to the image-based and IMU-based methods, respectively. As with conventional IMU-based motion capture systems, the proposed visual-inertial motion capture requires complicated preprocesses to measure movements. To relax this limitation, this thesis lastly presents a framework to automate the preprocessing of IMU-based human motion measurements, which can be applied to a wide range of applications that attach multiple IMUs to the body segments. A deep neural network model is proposed to predict the segment to which each IMU is attached and the relative IMU-to-segment pose from the IMU signals during a few seconds of walking. The experiments on publicly available datasets showed that the proposed method improved the accuracy of identifying the segment on which each IMU is mounted by 8.5% compared with the state-of-the-art method.

Overall, the thesis proposed two motion capture methods and an IMU calibration framework for measuring human motion in real-world settings. The thesis includes the future potential applications and the impact on the corresponding research communities.

Acknowledgments

First and foremost, I would like to express the deepest appreciation to my supervisor, Prof. Hideo Saito, for his extraordinary mentoring. Prof. Saito is not only an outstanding researcher but also a person of integrity. I am deeply grateful to Prof. Saito for giving me great collaboration opportunities with world-leading researchers and broadening my perspective of the world. Thanks to his immense and kind supports, I was able to concentrate completely on my research.

I am also thankful for Dr. Dan Mikami, Dr. Mitsunori Tada, and Dr. Yuta Sugiura for being on my reading and defense committee. I greatly appreciate their insightful suggestions and comments on my thesis.

I am also grateful to my lab mates for sharing brilliant time with me by working together and by talking about things other than just our research.

I also give sincere thanks to my family, particularly my parents and grandparents, for providing an open space for my growth and always letting me follow my passions.

I would like to express my gratitude to my wife, Mahoko. Mahoko is always with me, both when I am thriving and when I am in difficulty. Without her tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

This work was supported by Grant-in-Aid for JSPS Fellows (19J22153).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The automatic capture of human motion has been actively studied for many years due both to the number of potential applications and its inherent complexity. The research in this area aims to locate the human body in three dimensions and obtain full-body motion indicators, which include 3D human pose, joint torques, and the body center of mass (CoM). This thesis focuses on motion capture in real-world environments where measurers construct a motion measurement system. Many researchers and practitioners in the biomechanics community collect motion data in this setting [80, 85, 120]. Since the real-world measurements allow subjects to move in their living space, the data of motion in context can be obtained [60, 66]. Further, real-world motion measurements provide various data because subjects interact with the environment in practice [92].

For human motion capture in real-world environments, an RGB camera is one of the most commonly used sensors due to its portability and ability to capture the entire body with a single unit. The visual sensor captures the kinematic and shape information of people in the image. Furthermore, image-based measurement systems are applicable for motion capture under environments where the subject cannot wear devices, such as during sports games. The recent developments of deep learning techniques have brought remarkable breakthroughs in this field [10, 17, 56, 94]. Many researchers have significantly improved 2D pose estimators using convolutional neural networks (CNNs) and extended their works to 3D human pose estimation methods [64, 72, 76, 108]. However, capturing a 3D human pose from a single view still remains as a challenging task due to the depth ambiguity. The literature reported that the developed 3D pose estimators are typically trained and evaluated on 3D datasets recorded in well-controlled environments, and their performance degrades when applied to in-the-wild data [4, 121].

Inertial measurement units (IMUs), which measure 3D acceleration, angular velocity, and magnetic field, have become a prominent option for analyzing real-world human motion. Body-worn IMUs provide 3D rotational and, sometimes, translational motion of the attached body segments. IMU-based measurement systems perform well in outdoor recordings and scenarios with occlusions. These benefits have led many researchers to extract motion-related features from IMUs attached to the body parts for estimating 3D human pose [41,65,103] and other motion indicators (e.g., human velocities [112] and gait phases [53]). However, IMUs have 3D position ambiguities as well as a monocular camera because 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.

Exploiting and combining the benefits of these visual and inertial sensors, this thesis addresses motion capture in the real world under scenarios where a subject cannot wear the devices (e.g., during sports games) and can wear the sensors (e.g., daily-life movement), respectively. The two proposed methods for human motion capture resolve the 3D position ambiguity based on geometric consistency using multiple visual and inertial sensors.

This thesis first presents a non-invasive motion capture method using multi-view cameras that localizes the subject by making full use of the cross-view geometric constraints. Since many motion analysis applications for sports and medicine require not only human pose but also purpose-oriented motion indicators, this work focuses on the body CoM. The trajectories of the CoM represent the global movement of the subject and can be a key to analyzing the subject's motion (e.g., risk assessment of injuries and optimization of athletes' performance, such as the swing of a baseball bat and a golf club [32, 79, 105]). The presented method exploits the benefit that body kinematic and shape information can be extracted from the images, which enables CoM estimation based on the intrinsic weight of each body segment.

The secondly proposed method captures human motion fusing a single camera and body-worn IMUs. This method can be used in a measurement system constructed in a living field where people perform their daily activities. The center panel of Figure 1.1 illustrates this sensor configuration, and the left panel depicts the multi-view setting. The single-view setup expands the measurement area. The proposed method reconstructs the 3D human pose, localizing the global position of the subject by constraining and optimizing the foot-ground contact points.

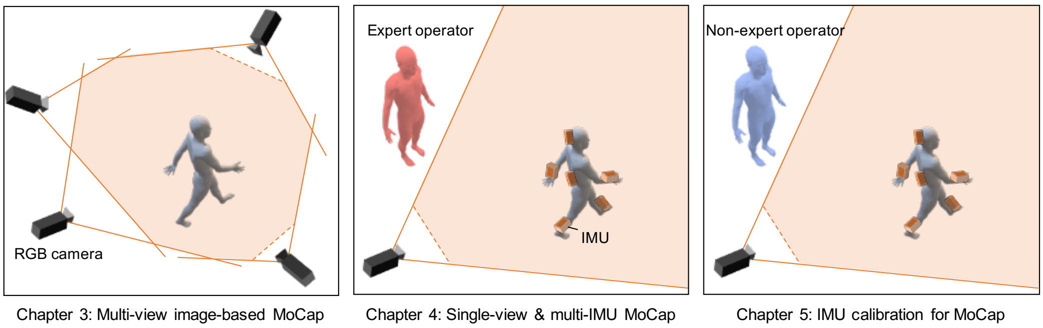

RGB camera
Expert operator
IMU
Non-expert operator
Chapter 3: Multi-view image-based MoCap
Chapter 4: Single-view & multi-IMU MoCap
Chapter 5: IMU calibration for MoCap

Figure 1.1: Motion capture environments in this thesis.

As with the conventional IMU-based motion measurements, the secondly proposed method requires complicated preprocessing to measure movement using IMUs. The preprocessing includes assigning each IMU to the body segment and calculating the relative orientation of the IMUs to the attached segments. These two preprocesses are collectively referred to as IMU calibration in this thesis. To improve the usability of the IMUs to motion capture, this thesis lastly presents a framework for automatic IMU calibration. Whereas the vision-based methods have well-established camera calibration techniques for multi-view configurations [117], the motion measurements based on multiple body-worn IMUs require special skills to build an appropriate setup. The proposed framework simplifies the operation of the IMU calibration and allows a non-expert to operate the calibration, as illustrated in the right panel of Figure 1.1. Specifically, the proposed framework allows the measurement operator to mount IMUs on arbitrary body segments, and to start measurements without making a subject take a calibration pose. The proposed calibration framework is applicable to general motion capture systems using multiple body-mounted IMUs. To extract the features for the sensor calibration, a network architecture that learns the global body motion and

Table 1.1: Comparison of the proposed approaches and related studies.

Method	Camera	IMU	Reconstruction	Localization
Monocular [64, 76, 108]	Single	-	Kinematics/Model	Depth inference
Multi-view [43, 83, 119]	Multiple	-	Kinematics/Model	Cross-view consistency
Chapter 3	Multiple	-	Weighted volume	Cross-view consistency
DIP [41]	-	6	Model	-
SIP [103]	-	6	Model	Offline optimization
Zhang *et al.* [118]	Multiple	6-15	Model	Cross-view consistency
von Marcard *et al.* [101]	Single	13	Model	Offline optimization
Chapter 4	Single	6-13	Model	Foot-ground contact

the interdependencies of the sensors is proposed.

As shown in Figure 1.1, this thesis proposes two 3D motion capture methods and an IMU calibration framework for the real-world motion measurements. Table 1.1 summarizes the two proposed motion capture methods and the related representative approaches using visual and inertial sensors in terms of sensor configuration, reconstruction form, and localization technique. To the best of our knowledge, the work introduced in Chapter 3 is the first attempt for multi-view body CoM estimation. Further, the 3D pose estimation proposed in Chapter 4, is also the world-first online approach for motion capture that combines a single camera and multiple IMUs. This thesis also presents a framework for automatic IMU calibration, which can be used for extensive IMU-based human body measurements including the proposed visual-inertial human pose estimation. These works contribute to both researchers and practitioners since the problem settings in this thesis are important in practical view.

1.2 Thesis Outline

This thesis investigates 3D human motion capture in settings where previous research has not addressed but that are desired in real-world scenarios. Specifically, motion capture in terms of 3D CoM estimation with multi-view cameras is presented in Chapter 3, and a method for 3D human pose estimation from a monocular camera and multiple IMUs is proposed in Chapter 4. Subsequently, Chapter 5 provides a calibration framework of inertial sensors that can be used in extensive IMU-based human motion measurements. The remainder of this thesis is organized as follows:

Chapter 2 reviews related works in human motion capturing. It starts with 2D human pose estimation, followed by image-based, IMU-based, and sensor-fusion approaches for motion capture.

Chapter 3 proposes a method to capture 3D human motion using a multi-view set of RGB cameras. This chapter focuses on the body CoM, considering that the position or trajectory of the CoM is often a parameter of interest when studying posture or movement. Different from conventional approaches that require large-scale measuring systems or attaching sensors to the subjects, the present study takes a multi-view vision-based approach, assuming the use in a situation where the sensors cannot be attached to the body. The proposed method first reconstructs subjects' body with voxels by back-projecting the body silhouettes and obtaining the intersection of all back-projection cones of the multi-view frames, also known as visual hull. Then, the method weights each voxel with body part-dependent weights to calculate a CoM. The content of this chapter is based primarily on [48].

Chapter 4 presents a novel 3D human pose estimation approach using a single RGB camera and a set of body-worn IMUs. In order to resolve the depth ambiguity of the single-camera configuration and localize the global position of the subject, this work presents an objective function that optimizes the foot-ground contact points. The timing and 3D positions of the ground contact are calculated from the acceleration of IMUs on the feet and geometric transformation of foot position detected on the image, respectively. Given that the results of the 3D pose estimation are greatly affected by the failure of the 2D joint detection, the image-based constraint is designed to handle outliers of the positional estimates of the 2D joints. The content of this chapter is based primarily on [46].

Chapter 5 provides a framework to automate the preprocessing of the IMU-based human motion measurements that attach multiple sensors to body segments. In the preprocessing, each IMU has to be attached to a predefined body segment, and the subject has to take a predetermined pose named calibration pose (e.g., T-pose: standing upright with hands open at the sides) to calculate the relative orientation from the sensor to the attached joint. The presented framework enables the user to attach sensors to arbitrary segments and start measuring motion without making the subject take the calibration pose. A novel end-to-end learning model that identifies the body segment on which each IMU is mounted is proposed. The model incorporates a global feature generation module and an attention-based mechanism. The former extracts the feature representing the motion of all attached IMUs, and the latter enables the model to learn the dependency relationships between the IMUs. The proposed model thus identifies the IMU placement based on the features from global motion and relevant IMUs. The experimental results have shown that this IMU-to-segment assignment model can be extended to an IMU-to-segment orientation alignment model that predicts the relative orientation from the IMU to the attached joint. The content of this chapter is based primarily on [47].

Finally, Chapter 6 summarizes this thesis and provides some suggestions for future work.

Chapter 2

Related Work

This section reviews previous works related to the focus of this study. This section first reviews prior attempts for 2D human pose estimation, including concurrent and recent efforts, since the proposed methods are built on those recent techniques. Then, single- and multi-view image-based approaches aiming to obtain 3D human motion, including body pose and other motion indicators, are introduced. Subsequently, IMU-based motion capture approaches are reviewed, followed by a survey of the IMU calibration. Lastly, visual and inertial sensor-fusion approaches for full-body pose estimation are introduced.

2.1 2D Pose Estimation

As one of the fundamental computer vision tasks, 2D human pose estimation has been actively studied for many years. Deep learning has shown great performance on many tasks, such as image classification [37], object detection [57], and semantic segmentation [86]. 2D human pose estimation also achieves rapid progress by employing deep learning technology, particularly convolutional neural networks (CNNs) [10, 17, 59, 94]. The progress directly contributes to improving the accuracy of a number of applications that employ off-the-shelf 2D pose detectors (e.g., action recognition [69], human tracking [28], and video surveillance [71]).

Deep learning-based 2D pose estimation methods can be divided into two categories: top-down and the bottom-up approaches. Top-down methods first detect humans in the image and then apply single-person pose estimators to each person box [27, 94]. Bottom-up methods first locate all the body keypoints (joints and face landmarks) in the input image and then group them to the corresponding subjects [10, 45]. Bottom-up methods usually have constant computation time, since they do not need to predict the pose for each person separately. OpenPose [10], AlphaPose [27], Cascade Pyramid Network [17], High-Resolution Net [94], and Multi-Stage Pose Net [56] have been extensively used as 2D pose detectors in a wide range of tasks, given that they have been maintained and updated regularly. In this thesis, the proposed methods are built on OpenPose [10], the pioneering work of the open-source 2D pose detector constructed with bottom-up framework, due to its usability and stable computation time.

2.2 Image-based Motion Capture

2.2.1 Monocular 3D Human Pose Estimation

Improvements of deep neural networks have gained the attention of many researchers in human motion capture using a single RGB camera [64,72,76,108].

A recent data-driven method that estimates 3D human configuration from an image can be roughly classified into keypoint regression and model-based approaches. The former estimates the 3D position of the body keypoints, and the latter infers the pose parameters of a pre-defined human model.

In the keypoint regression methods (e.g., [64, 72, 76]), 2D-to-3D lifting approaches that infer the 3D human pose from the intermediate representations of the estimated 2D human pose generally outperform approaches that directly regress the 3D pose because of the great performance of the 2D keypoint detectors [121]. Most 2D-to-3D approaches employ state-of-the-art 2D pose estimation networks, such as those exemplified in Section 2.1, and the second networks take the estimated 2D pose representation as an input, which finally regress the 3D pose. The better performance of the 2D-to-3D approaches than direct regressions suggests the effectiveness of incorporating off-the-shelf 2D joint detectors for 3D motion estimation. In this thesis, the proposed motion capture approaches reconstruct the human body using the 2D pose on the images estimated by the off-the-shelf 2D joint detector.

The model-based approach estimates the full-body posture by inferring the parameters that represent the pose and shape of the parametric human model [54, 77, 108]. Many researchers have been appreciably interested in incorporating the human model into the motion estimation because the model-based methods allow poses to be explicitly constrained based on prior knowledge about the kinematic model, such as range of joint motion, fixed bone length, and skeletal joint connectivity information. In this thesis, a 3D pose estimation method presented in Chapter 4 adopts this model-based approach.

3D human pose datasets are usually established using optical motion capture systems (e.g., Vicon) in controlled environments to obtain 3D pose annotations [42, 93, 97]. The conventional monocular 3D pose estimation methods train and evaluate their models on these datasets; thus, the literature reports that their performance degrades when applied to in-the-wild data [4, 121].

2.2.2 Multi-view Motion Capture

Many existing studies for 3D human pose estimation using multi-view images take a simple two-step approach similar to 2D-to-3D lifting methods of the single-view motion capture: the backbone network extracts pose features from each image, merges them, and reconstructs the human pose through another network. Iskakov *et al.* merges 2D joint positions and their confidence scores to regress 3D joint positions [43]. Some studies fuse heatmaps from a 2D pose estimator and predict 3D poses [83, 119]. These methods designed their objective function based on cross-view consistency constraints, which impose the reprojection of the reconstructed model joints to be close to the detected keypoints on the 2D images.

Besides full-body pose estimation, conventional image-based studies for analyzing human motion address head pose estimation [5], facial feature detection [90], and hand pose estimation [70], which predict the movement of a part of the body. Other works for capturing full-body movement include the investigation of trajectories of movements [58], human-object interaction [124], and gait analysis [14]. To our knowledge, no study has estimated the CoM of a human body using RGB cameras, although the body

CoM can be a key to analyzing the human motion, especially in the sports scene [32, 79, 105].

2.3 IMU-based Motion Capture

2.3.1 Human Pose Estimation Methods

Many approaches for IMU-based 3D pose estimation 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 [41]. 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 the global position. A survey reported that a commercial marker-less motion capture suit composed of 17 IMUs suffers from large positional error [33].

To handle this potential hurdle, von Marcard *et al.* reconstructed human motion using global optimization [103]. Since 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 [65]. This method achieves good performance when the contact positions are predefined (e.g., contact between the hip and the plane of a chair). However, it accumulates the positional error when the contact positions are determined online. Inspired by the contact constraints on the body localization, the method proposed in Chapter 4 utilizes RGB images to compensate for the contact's position ambiguity.

2.3.2 IMU Calibration for Motion Capture

Human motion capture using IMUs requires two complicated procedures before starting measurements: IMU-to-segment (I2S) assignment and orientation alignment [123]. I2S assignment represents a process to map the segments on which each IMU is mounted. I2S orientation alignment denotes a procedure for calculating the relative orientation of the IMUs to the corresponding (attached) segments. A line of research on the I2S assignment has aimed to define effective feature representations based on signals from IMUs. The early work applied hand-crafted feature descriptors, such as root mean square and amplitudes of accelerations and classical classification algorithms, including support vector machines and decision trees [3,55,104]. The feature descriptors of these approaches are designed based on the intuition and experience of the researchers, with no agreement regarding the most suitable features for I2S assignments. A recent study proposed an approach that combines CNNs and recurrent networks [123] that were trained in an end-to-end manner without the need to manually design features. Their network provides both I2S assignment and orientation alignment predictions. However, this approach assumes that IMUs are attached to the lower limbs and assigns IMUs one by one, ignoring the signals from other IMUs. In this thesis, the models for I2S assignment and orientation alignment presented in Chapter 5 extract discriminative features by learning the interdependencies of the IMUs.

2.4 Visual and Inertial Sensor Fusion for 3D Human Pose Estimation

A line of research on combining visual and inertial information has aimed to achieve full-body motion capture free from positional drift. Images from multi-view cameras are utilized to constrain the subject's position three-dimensionally [61, 81, 97, 102, 118]. The posture and the global position of the subject are optimized by minimizing the difference between the human silhouettes on the images and the solved human model projected onto the images [102]. Other studies have shown that joint positions on 2D images obtained by a CNN-based keypoints detector improve the performance of 3D human pose estimation [39, 61, 97, 118]. Although these approaches are appealing because of their stability and accuracy, at least two viewpoints are required to resolve the depth ambiguity and localize the subject.

Researchers have addressed pose estimation by combining IMUs and a single view. Some studies have performed 3D human tracking with IMUs and a single depth sensor, such as Kinect [49, 122]. However, the measurement accuracy of Kinect decreases outdoors. The only study that has addressed 3D motion capture with IMUs and a single RGB camera simultaneously optimizes human pose for a certain period of frames, and the global optimization is processed offline [101]. An offline method uses all frames in a sequence to optimize the human pose of a certain frame. To the best of our knowledge, the proposed method introduced in Chapter 4 is the first attempt for online 3D pose estimation using IMUs and a single RGB camera.

Chapter 3

Motion Capture with Multi-view Cameras

3.1 Introduction

This chapter presents an approach for capturing human motion using multi-view images. This work focuses on estimating the position of the human CoM because the trajectory of the CoM indicates the global motion of the subject, and it plays a key role in many healthcare and sports applications [32,79,105]. While many approaches for 3D human pose estimation have been developed, this work is the first attempt for image-based human CoM estimation.

Many researchers have attempted to develop methods to estimate 3D CoM, but most of them have relied on the hardware in the laboratory environment. For example, classical methods have used optical motion capture systems and force plates to measure the human body's CoM [13, 62]. These systems are designed only for special environments, such as research labs and studios, and markers are attached to subjects. Although such systems are reliable and accurate, measuring CoM under limited hardware resources has been of great interest. González *et al.* proposed the use of Kinect and Wii balance board together [30], and Najafi *et al.* adopted wearable inertial sensors to track the CoM [74]. However, many studies for casual CoM estimation require prior measurements for personalizing the weight of the body segments [13, 62].

This work addresses the 3D CoM estimation that only relies on a set of multi-view cameras. The multi-camera setting is motivated by many existing image-based approaches that analyze motion in real-world scenarios and in situations where mounting devices on the subjects' body is difficult [9,91,114]. The proposed method satisfies the following three conditions: it works in (1) outside scenarios, (2) with no wearable devices attached, and (3) with no prior personalization.

The proposed method first reconstructs the subject's body with voxels using multi-view RGB images. The 3D voxel model is divided into nine body parts, and weights, which depend on the body parts, are assigned to each part. Then, the weighted average of the parts is used to calculate the whole body's CoM. Since the proposed method uses only RGB images, outdoor CoM estimation is achieved (condition 1), and wearable devices are not necessary (condition 2). Moreover, the 3D shape reconstruction of the subject's body handles the differences in individuals' figures when calculating the CoM (condition 3). The proposed approach is the first attempt toward an end-to-end automated process for 3D CoM estimation using image inputs only, taking the volumetric properties into account.

The method was quantitatively and qualitatively evaluated through extensive experiments. The accuracy of the CoM estimation in static poses was

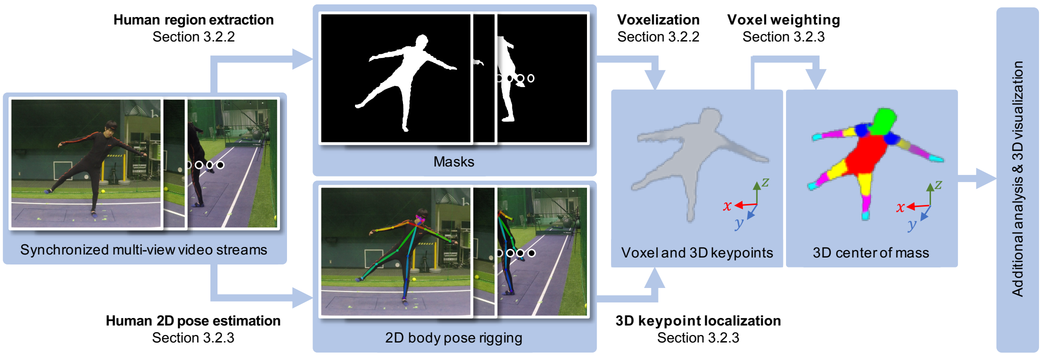


Figure 3.1: Overview of the proposed CoM estimation.

evaluated, and the effects of the number of cameras on the CoM estimation were discussed. The qualitative evaluation demonstrated the applicability of the proposed method utilizing the cooperation of professional and amateur baseball batters assuming a setup of an actual game.

3.2 Method

3.2.1 Overview

The proposed method estimates the CoM of a person in a process using N calibrated cameras ($N \geq 2$). A global summary of the proposed process is shown in Figure 3.1. The input consists of only RGB images taken from multiple viewpoints. Those images are used for 3D reconstruction of the body shape and a 3D kinematic structure estimation of the human body. Based on the joint positions obtained via the estimation of the body structure, the human body model is segmented into nine parts. Then, the CoM is obtained by assigning a weight to each part of the human body, as reported by [22].

3.2.2 3D Reconstruction of the Human Body

The 3D human body is reconstructed using Martin's method [63], which extracts the subject's 2D silhouette from the input images (e.g., using [98]) and reprojects the silhouettes into a 3D world. The common parts of the reprojected silhouette are the 3D shape of the body $\mathbf{V}(\ni \mathbf{v}_j)$. $\mathbf{V}$ denotes a set of voxels, where each voxel element $\mathbf{v}_j$ contains 3D positional information. In the case that the subject holds tools, users have a choice of whether to include or exclude the tools from the following CoM calculation. To retrieve a precise 3D model, cameras need to be arranged to observe the voxel space from various angles [73]. By reconstructing the 3D shape of the subject's body, an individual's unique figure can be reflected.

3.2.3 Human Kinematic Structure Estimation

If the frame of reference is at the body CoM, the CoM is the unique position at which the weighted position vectors of all the parts of a system add up to zero. Because each body part has a different density [7], assigning the appropriate weight to each part will lead to a more accurate CoM estimation. As shown in Figure 3.2, the 3D model reconstructed in Section 3.2.2 is divided into nine parts: head, body, shoulder, back arm, forearm, hand, thigh, calf, and foot. To this end, 2D keypoints, which represent the joints and the face of an individual, are obtained from the input images by applying the method of Cao *et al.* [10]. By applying the direct linear transform to each 2D keypoint $\mathbf{q}$ to triangulate them, the 3D position $\mathbf{p}$ of each $\mathbf{q}$ is obtained.

As shown in Figure 3.2, the 3D model $\mathbf{V}$ is segmented into $\mathbf{V}_i$ $(0 \leq$

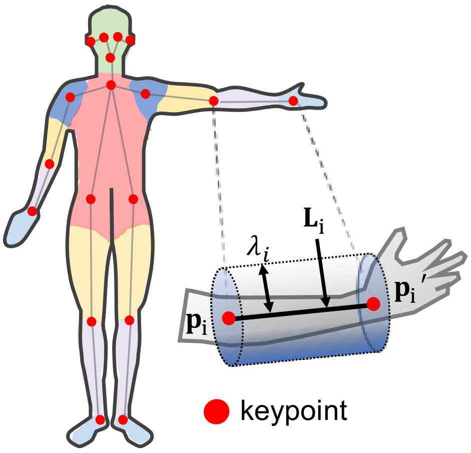


Figure 3.2: Variables in a segmented part.

$i < 9$) based on the distance between the line segments $\mathbf{L}_i$ connecting the adjacent keypoints $\mathbf{p}$ and each voxel $\mathbf{v}_j$. Algorithm 1 shows the segmentation procedure. A voxel $\mathbf{v}_j$, which exists within a distance λ_i from $\mathbf{L}_i$, is classified as $\mathbf{V}_i$. A voxel $\mathbf{v}_j$ located in the common area of two or more body parts is classified as the part with the smaller distance. All voxels $\mathbf{v}_j$ that are not classified as any body part are removed. The threshold λ_i is manually set to be large to not remove the body parts. The segmented model is weighted based on the weight of each part of the human body, as reported by de Leva [22]. The overall CoM of the human body $\mathbf{C}$ is computed via

$$\mathbf{C} = \frac{1}{M}\sum_{i=1}^{M} w_i \mathbf{v}_i \tag{3.1}$$

where M denotes the total number of voxels, and w_i represents the weight assigned to $\mathbf{V}_i$.

Algorithm 1: Proposed segmentation procedures

```
$\mathbf{V}_i$: A part of 3D human body model $\mathbf{V}$
$\mathbf{v}_j$ : A voxel costituting the 3D model $\mathbf{V}$
$\mathbf{p}_i$, $\mathbf{p}'_i$: Keypoints that divide $\mathbf{V}$ into $\mathbf{V}_i$
$\mathbf{L}_i(\mathbf{p}_i, \mathbf{p'}_i)$: Line segment between $\mathbf{p}_i$ and $\mathbf{p}'_i$
foreach $\mathbf{v}_j$ do
    foreach $\mathbf{L}_i(\mathbf{p}_i, \mathbf{p'}_i)$ do
        $D_i \leftarrow CalculateDistance(\mathbf{v}_j, \mathbf{L}_i)$
    end
    if $Min(D_i) < \lambda_i$ then
        stock $\mathbf{v}_j$ to $\mathbf{V}_i$
    else
        remove $\mathbf{v}_j$
    end
end
```

3.3 Experiments

This section provides two performance evaluations of the proposed method using real data. First, the proposed method and two baselines are compared concerning accuracy in terms of the center of pressure (CoP) error metric [113]. Second, the sequences of tracked CoM is visualized in 3D space, which is compared with CoMs measured using the wearable sensors. It demonstrates that the proposed method has the ability to provide meaningful 3D data for sports performance analysis.

3.3.1 CoM Estimation for Static Scenes

Setup. As shown in Figure 3.3, a force plate (TF-6090) and five cameras (GoPro, 30 frames per second (fps), 1920×1080 resolutions) were utilized in this evaluation. Internal and external camera parameters were obtained

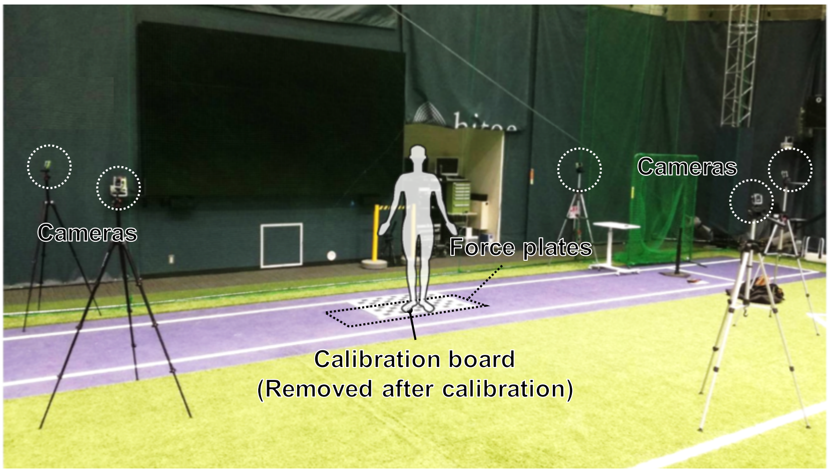


Figure 3.3: Experimental setup of a static scene.

through offline calibration with [117]. These cameras were placed to surround the force plate at $0°, 45°, 100°, 260°$, and $300°$ respectively, where $0°$ represents a face-on view of the subject, capturing the subject standing on the force plate. Three subjects (two males and one female) each stood on the force plate in four static postures: upright standing, single-leg standing, squatting, and bending forward. The human regions are extracted using a semi-automated manner implemented using GIMP2 [1] in this experiment, to confirm the pixel mask.

To validate the performance of the proposed approach, the following two baseline methods were developed:

Uniform: Voxels with a uniform weight

This method estimates the CoM as the center of a reconstructed 3D model, in which all parts are assigned a uniform weight. The CoM is computed using Equation 3.1 with all $w_i = 1$.

Articulated: Articulated joints model

This method estimates the CoM as the center of the weighted articulated joint model. The CoM is computed by

$$C = \frac{1}{M'} \sum_{i=1}^{M'} w_i \boldsymbol{j}_i, \tag{3.2}$$

where $\boldsymbol{j}_i$ represents the 3D positions of the mid-points between two connected joints (e.g., the CoM of a left lower arm is defined as a mid-point between a hand joint and an elbow joint), and M' denotes the number of the mid-points. The 3D joint positions are computed by triangulation with the 2D joints detected by [10].

A comparison with the Uniform method clarifies the effectiveness of the proposed method for considering the weight of each part. A comparison with the Articulated method reveals the influence of the volume of the human body on the CoM estimation accuracy.

In these evaluations, the CoM estimation error of each method was evaluated as the Euclidean distance of the 2D coordinates of the CoP $\boldsymbol{g}$, which represents the vertical projection of the estimated CoM as

$$E_{COP} = \|\boldsymbol{g} - \boldsymbol{g}_f\|_2, \tag{3.3}$$

where $\boldsymbol{g}_f$ denotes the CoP estimated from the force plate.

Results. Figure 3.4 shows the input images from one view (first row); the reconstructed 3D model, showing the joint positions (second row); and the labeled 3D model, which is based on the joint positions and estimated CoM (third row). The results in the second and third row show that the estimated

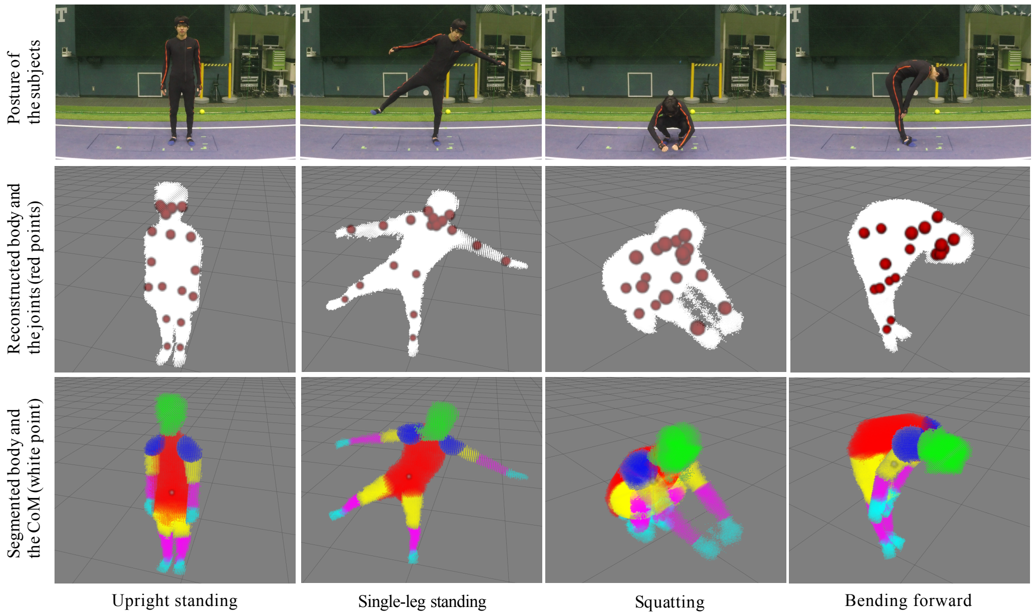

Posture of the subjects
Reconstructed body and the joints (red points)
Segmented body and the CoM (white point)
Upright standing
Single-leg standing
Squatting
Bending forward

Figure 3.4: Experimental results regarding static posture.

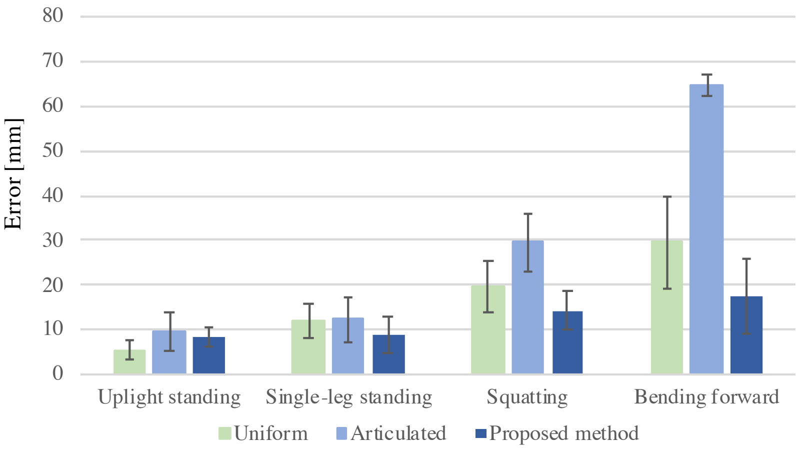


Figure 3.5: Distance between the CoP and the vertically projected CoM in four postures.

3D joint positions were sufficient to assign each voxel to the appropriate body parts. The average estimation errors of each method are shown in Figure 3.5.

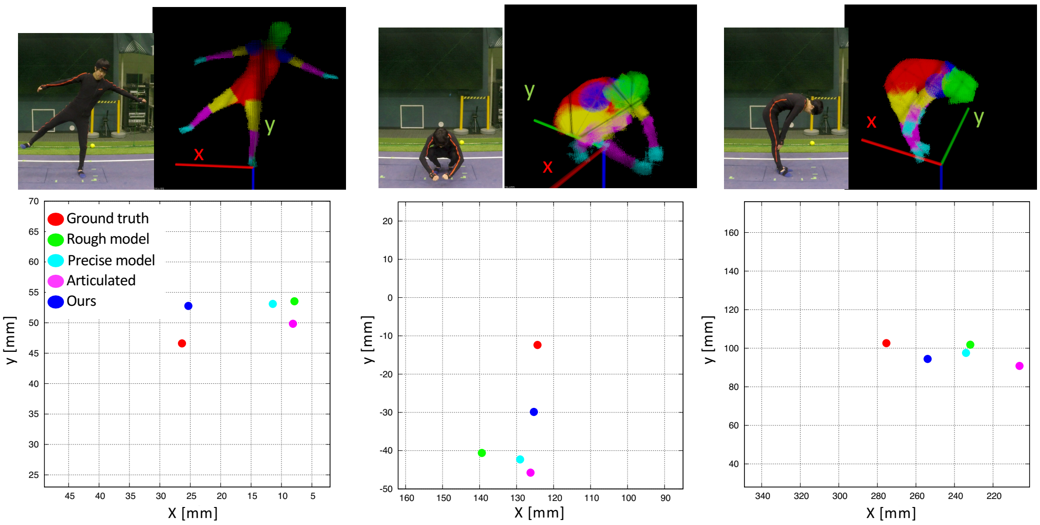


Figure 3.6: Visualization of the CoP and vertically projected CoMs in “Single-leg standing”, “Squatting”, and “Bending forward”.

It shows that the proposed method outperformed the baseline methods and robustly estimated the CoM with errors of approximately 10 mm for the CoP in all postures.

In the case of standing upright, the precision of all methods was similar because of the symmetry of the posture. The precision of all methods was greater in the case of single-leg standing than in squatting and bending forward. This would be caused by self-occlusions affecting the precision of both the reconstructed 3D model and the estimation of the joint positions. For example, the chest portion of the bending forward 3D model appeared to be thicker than the subject’s chest. This would be because the five cameras were placed at the same height, as depicted in Figure 3.3, and no cameras were able to observe under the chest. Mundermann *et al.* [73] found that cameras positioned in a geodesic dome configuration produce the best results to build a visual hull model of the human body. Figure 3.6 visualizes the CoP measured with the force plate and vertically projected CoMs estimated by the proposed and baseline methods. The CoMs estimated by Articulated tend to be plotted in the direction of the subject's upper body including the head. It suggests that the CoM position of the upper body segment defined in the Articulated method reduces the accuracy.

3.3.2 CoM Estimation for Dynamic Motion

Compared with the 2D CoP estimation approaches, utilizing a force plate, the vision-based approach including the proposed method could estimate 3D positions of the CoM, which is an important advantage for analyzing a player's performance in a sports scene. In particular, CoP estimated with a force plate does not match the CoM projections when the subject is in motion [106]. Here, the results suggest that the proposed method could estimate the 3D positions of CoM in such a challenging situation.

Setups. As illustrated in Figure 3.7, the four cameras (three 640×360 resolution cameras and one 640×480 resolution camera) are placed outside the baseball field, assuming that the proposed method would be used to observe a baseball batter. The camera position and example images taken with each camera are shown in Figure 3.7. All cameras were fixed and pre-calibrated with [117]. The 48 frames are extracted during each swing from 30 fps videos. The two subjects, an expert baseball batter and an amateur batter, swung a bat twice without batting a ball, assuming (a) an inside pitch and (b) an outside pitch. the subjects' regions are extracted in the same manner as in Section 3.3.1 (i.e., the bat held by the subject was excluded by masking it out). The estimated time-sequential trajectories of the CoM were compared

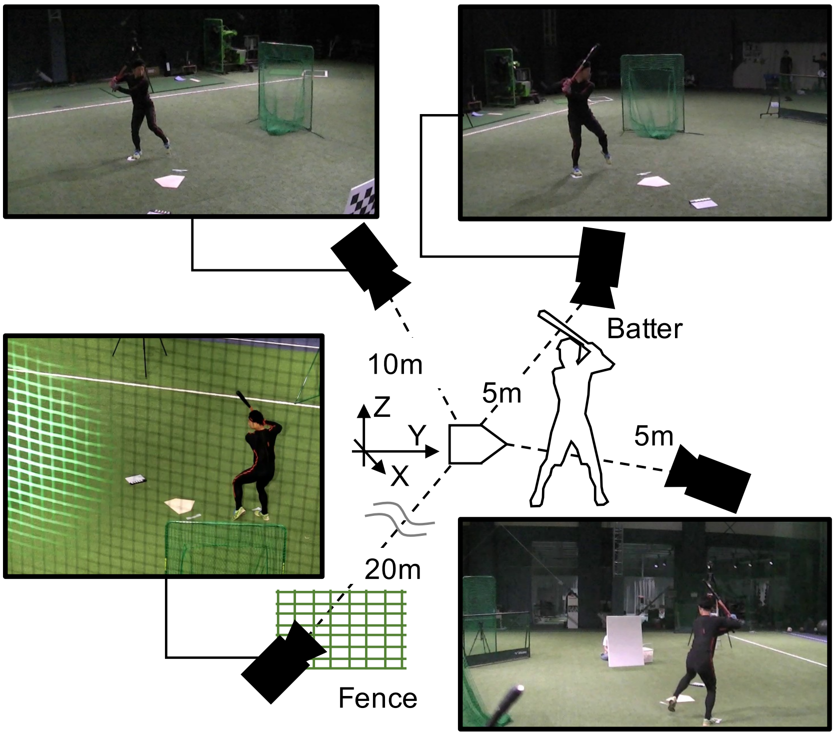


Figure 3.7: Experimental setup for the CoM estimation in a dynamic scene.

with those calculated by a set of wearable sensors.

Results. Figure 3.8 illustrates the 3D trajectories of the estimated CoM. The expert (two left graphs of Figure 3.8) and the amateur (two right graphs) swung a bat twice. The graphs on the upper left and lower left represent the same swing, but are visualized from different perspectives. The same applies to the right graphs. The red and blue trajectories correspond to the cases of (a) inside and (b) outside pitch, respectively. The CoM sequences are plotted on the graphs so that the sum of the distance of each CoM is minimized because calibrating the coordinates of the proposed method's CoM and the coordinates of the CoM measured with IMUs is difficult. The subjects were right-handed batters and assumed that a ball was coming from a negative to

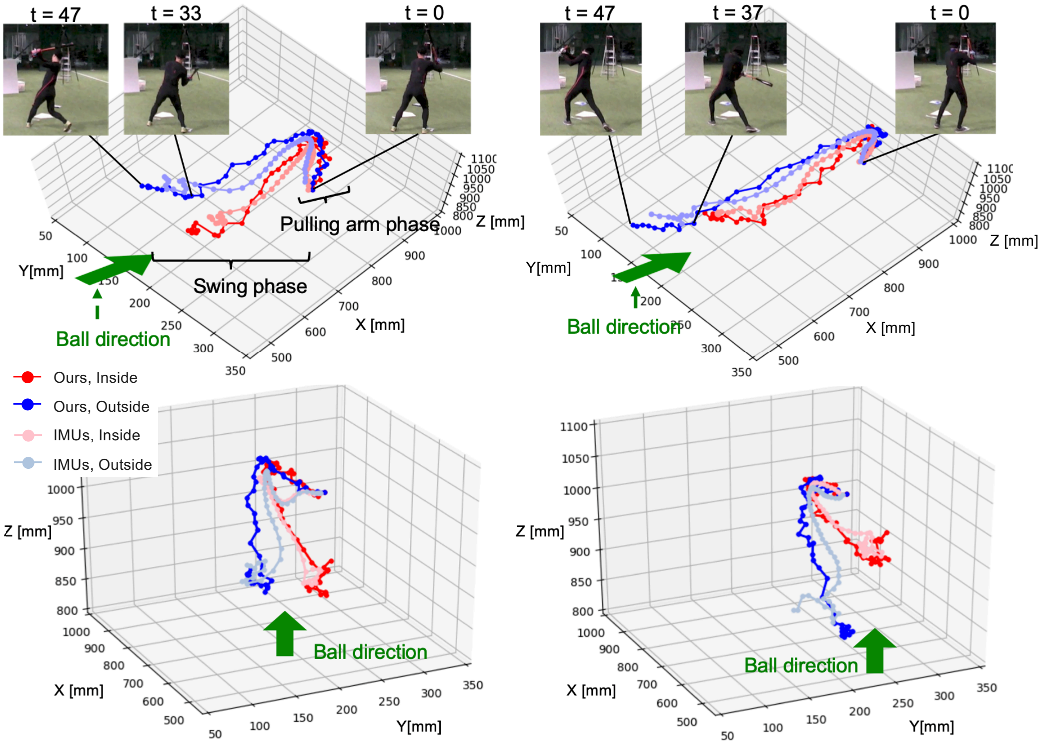


Figure 3.8: Predicted trajectories of the body CoM in dynamic scenes.

a positive direction along the X-axis.

In both the expert and amateur swings, the CoM transitions of the proposed method drew almost the same trajectories as the CoM of the wearable sensor in the pulling arm phase and the swing phase. The mean absolute distance between the two CoM sequences was 25.2 mm. The trajectories against the inside and outside balls were almost the same when the arms were pulled back and gradually split in the swing phase. The lower graphs show the differences in the swings between the expert and amateur. The height (Z-axis) of the CoMs during the expert's swing against the inside pitch was almost the same as that of the outside pitch, while the CoM of the amateur went down during the swing against the outside pitch.

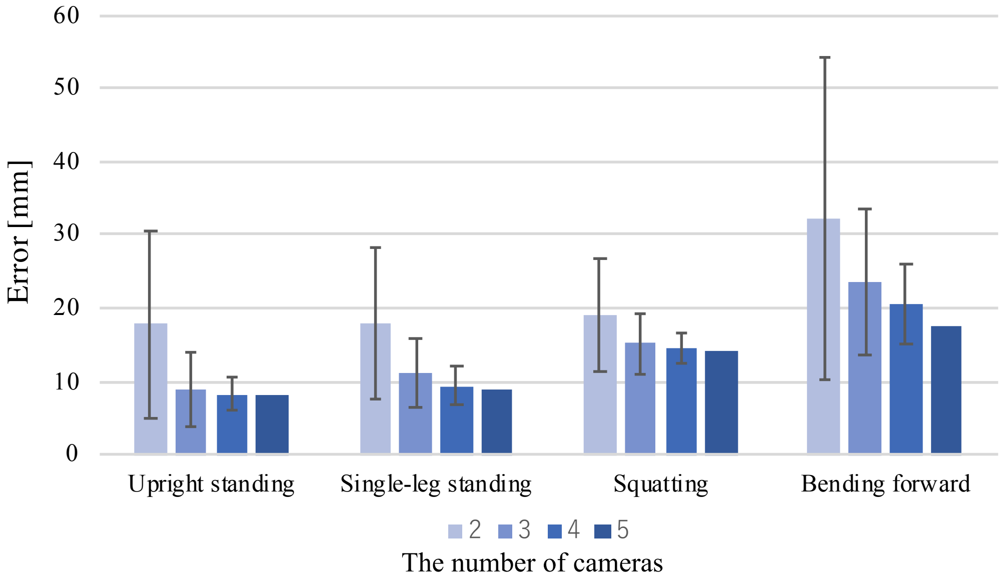


Figure 3.9: Distance between the CoP and the vertically projected CoM, estimated by using different number of cameras.

3.4 Discussion

When estimating the CoM using the proposed method, the number of cameras affects the accuracy of both the 3D human body reconstruction and the 3D kinematic estimation. To examine the relationship between the number of cameras and the estimated CoM accuracy, the CoMs estimated by the proposed method were compared to the CoP positions measured by the force plate, as in the experiment in Section 3.3.1. Since all joints must be detected with two or more cameras, when cameras that do not satisfy the conditions were selected, the undetected joints were complemented manually.

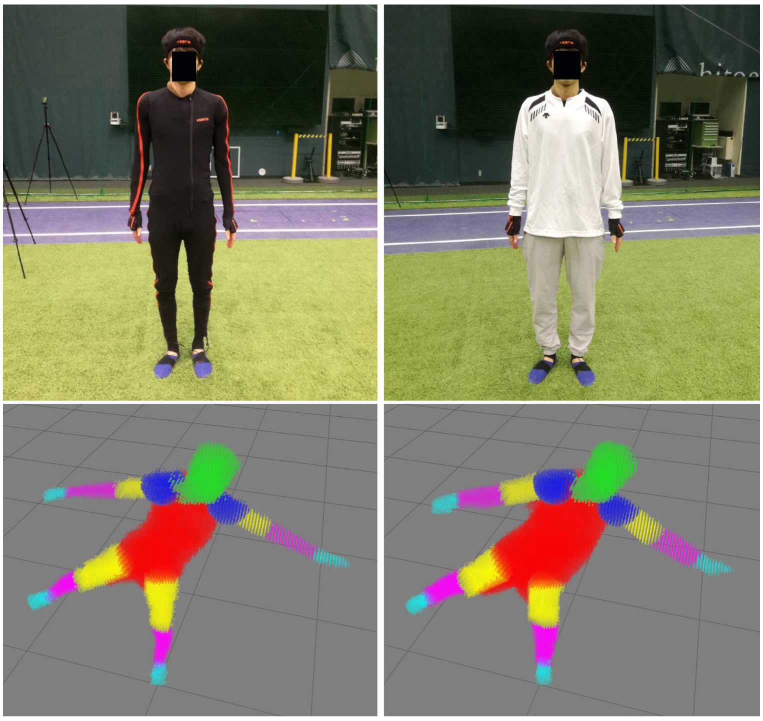
Figure 3.9 shows the error between the CoP and the vertically projected CoM in each static posture when the number of cameras was changed from two to five. Figure 3.9 also depicts the average and standard deviations of each possible combination of cameras. The average error increased gradually as the number of cameras decreased from five to three, but when the CoM was estimated using only two cameras, the CoM precision dropped dramatically. The estimation accuracy of the CoM depends on the accuracy of the human model created with the visual hull. The number and arrangement of cameras in visual-hull-based human motion tracking have been explored by Corraza *et al.* [20]. Modifying camera configurations by referring to their conclusion would improve the accuracy of the CoM estimation.

3.5 Summary

This chapter proposed a novel vision-based CoM estimation algorithm based on multi-view images for sports performance analysis. The key approach of the proposed method was to assign an appropriate weight to each voxel, reconstructed in a visual hull manner. Evaluations with the real data demonstrated that the proposed method could estimate the CoM within errors of approximately 10 mm concerning the CoP compared to the data measured with force plates in static poses. In addition, the proposed method reasonably estimated the 3D trajectory of the CoM in a dynamic scene.

3.5.1 Limitations

The proposed method assumes one subject in a scene. To capture multiple subjects' CoMs, an extension to separate each person in a voxel space is required. Whereas the CNN-based bone estimation [10] can handle multiple



(a) Form-fitting clothes

(b) Loose clothes

Figure 3.10: Comparison of the appearance between the subjects wearing form-fitting and loose clothes.

persons in a single view, the proposed method requires the identification of persons in multiple images, which will necessitate additional efforts.

The proposed method estimates the CoM as the gravity point of a set of voxels. Therefore, clothes may affect the performance since the subject's silhouette changes. Here, additional experiments demonstrated the effects of clothes on the proposed method. As shown in the first row of Figure 3.10, images from subjects wearing both form-fitting and loose clothing were utilized as inputs for the proposed method. The second row of Figure 3.10 demonstrates that the reconstructed 3D model with loose clothes was expanded compared with the subject wearing form-fitting clothes, even when the same subject stood in the same posture. The quantitative results of such cases, utilizing the same configuration introduced in Section 3.3.1, show that the average error when the subjects wore loose clothes was 81% larger than in the case of wearing form-fitting clothes. These results revealed that the proposed method's accuracy degraded when the subject was wearing loose clothes. Therefore, reducing the effect of loose clothing on the method's accuracy remains as future work.

Chapter 4

Human Pose Estimation from IMUs and an RGB Camera

4.1 Introduction

In this chapter, an approach for 3D human pose estimation using a single RGB camera and body-worn IMUs is proposed. RGB cameras and IMUs are utilized for online human pose estimation in real-world settings. IMUs comprise accelerometers, gyroscopes, and magnetometers, which provide measurements of 3D acceleration, angular velocity, and magnetic field, and calculated 3D orientation. The acceleration and orientation of the IMU attached to each body segment help infer human motion [41, 65, 103]. RGB cameras are the most commonly used optical sensors and offer 2D visual information of the environment. Recent image-based human pose estimation methods detect joints of the human body on the image that offer robust 2D human poses [10, 17, 59, 94]. 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; thus, IMUs cannot provide accurate 3D joint positions in the global coordinates. For RGB cameras, it remains difficult to obtain a 3D human pose in the wild using a single view due to the 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 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 CNNs on the images. They have achieved accurate and stable performance in motion capture, but images from multiple viewpoints are required to localize the 3D human position [61, 81, 97, 102, 118].

This chapter presents an optimization-based method for online 3D human pose estimation that resolves the positional ambiguity of the IMU-based posers 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, this work presents 3D positional constraints of foot-ground contact. The timing of the contact is determined from the 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 [10]. The proposed objective function is designed to handle the outlier detection of the joint detector, which resolves the second problem.

The extensive experiments were conducted to evaluate the proposed method using the public 3D dataset TotalCapture [97], which includes all-synchronized videos, IMU data, and ground-truth human pose. The experiments demonstrated that the cost terms incorporated into the proposed objective function contributed to the accuracy and stability of pose estimation.

4.2 Methods

4.2.1 Pose Parameterization and Calibration

The subject's pose is parameterized using a Digital Human Model (DHM) [26] that consists of a 48 degrees of freedom link configuration. The model provides kinematics and the body mesh when the pose including the global translation θ ($\in \mathbb{R}^{51}$) is determined. The proposed method extends the IMU-based motion capture method [65] 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 4.1 shows relations between the coordinates and transformation matrices. The transformations between the global coordinates and the camera coordinates $\mathbf{T}^{\mathrm{GC}}$ is determined using a checkerboard [117]. In this configuration, the checkerboard is placed on the floor. The Z-axis of the global coordinates $(X_{\mathrm{w}}, Y_{\mathrm{w}}, Z_{\mathrm{w}})$, defined by the checkerboard, points in the opposite direction of gravity, and the $Z_{\mathrm{w}} = 0$ plane coincides with the floor. Note that the checkerboard can be removed after the camera is calibrated and

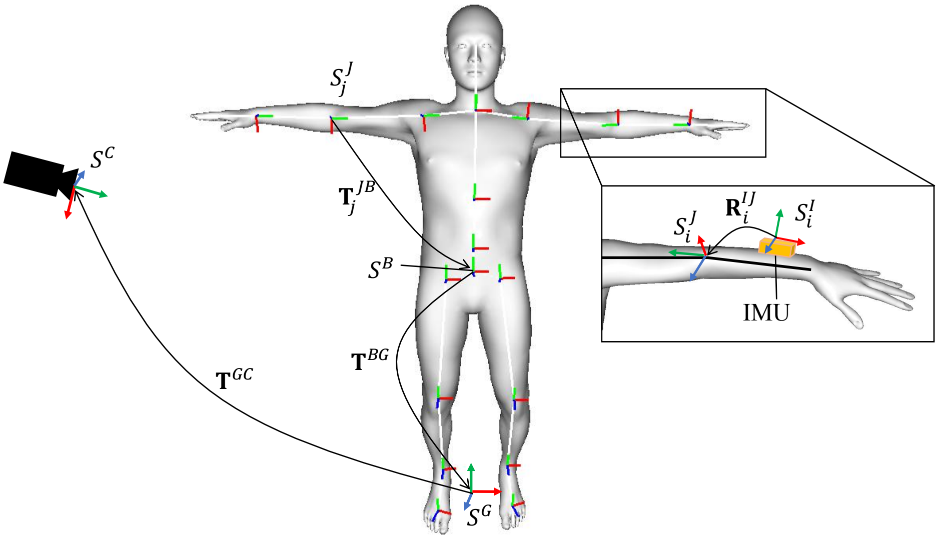


Figure 4.1: Relations among the local coordinate systems.

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

$$\mathbf{R}_i^{\mathrm{IJ}} = \mathbf{R}_i^{\mathrm{J}}(\theta_0) \cdot (\mathbf{R}_i^{\mathrm{I}}(t_0))^{-1}, \tag{4.1}$$

where $\mathbf{R}_i^{\mathrm{I}}(t_0)$ represents the i-th IMU sensor orientation in the global coordinates when the subject takes the calibration pose, and $\mathbf{R}_i^{\mathrm{J}}(\theta_0)$ 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 θ_0 represent the frame and pose parameter of the calibration pose, respectively. As illustrated in Figure 4.1, $\mathbf{R}_i^{\mathrm{J}}(\theta_0)$ 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 $\mathbf{T}_j^{\mathrm{JB}}(\theta_0)$ and $\mathbf{T}^{\mathrm{BG}}(\theta_0)$. $\mathbf{T}_j^{\mathrm{JB}}(\theta_0)$ denotes the transformation from S_j^{J} to the body coordinates S^{B}. In the proposed method, S^{B} is defined to correspond with the local coordinates of the pelvis joint of the human model. The transformation $\mathbf{T}_j^{\mathrm{JB}}(\theta_0)$ can be obtained from the forward kinematics of predefined link configuration of the model. The transformation from the body coordinates to the global coordinates, $\mathbf{T}^{\mathrm{BG}}(\theta_0)$, is determined by the position and orientation of the subject taking the calibration pose.

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 foot stamp is applicable because, for the camera, the timing of the cue is obtained from the motion of the ankle joints detected on the image, and for the IMUs, the timing can be calculated from the acceleration measurements of the IMU attached to the feet. The synchronization should be performed after the calibration pose.

4.2.2 Full-Body Pose Optimization

The proposed method follows the paradigm of constraint-based motion tracking. More specifically, the method minimizes the following total cost function composed of multiple cost terms on a per-frame basis.

$$E(\theta) = E_O(\theta) + \lambda_{RoM} E_{RoM}(\theta) + \lambda_P E_P(\theta) + \lambda_G E_G(\theta), \tag{4.2}$$

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 foot-ground contact points, respectively. These positional error terms are designed to stably estimate the human pose in an under-constrained environment. Every term is weighted by a corresponding weight. The quasi-Newton algorithm [23] is applied to solve the optimization problem.

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 coordinate. Using the transformation matrix from the sensor coordinates to the joint coordinates $\mathbf{R}_i^{\mathrm{IJ}}$ (Equation (4.1)), the cost $E_O(\theta)$ can be expressed as

$$E_O(\theta) = \sum_{i=1}^{N_{\mathrm{I}}} \|\mathbf{R}_i^{\mathrm{IJ}} \cdot \mathbf{R}_i^{\mathrm{I}} - \mathbf{R}_i^{\mathrm{J}}(\theta)\|_{\mathrm{F}}^2, \tag{4.3}$$

where $\mathbf{R}_i^{\mathrm{I}}$, and $\mathbf{R}_i^{\mathrm{J}}(\theta)$ 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, $E_{RoM}(\theta)$, adds cost when the joint angle exceeds or falls short of the range of motion (RoM) ψ. ψ 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

$$e_{RoM}(\phi(\theta), \psi) = \sum_{k \in \{r,p,y\}} \left\{ \begin{array}{ll} \rho((\phi_k(\theta) - \psi_k^{\min})^2) & (\phi_k(\theta) < \psi_k^{\min}) \\ \rho((\phi_k(\theta) - \psi_k^{\max})^2) & (\phi_k(\theta) > \psi_k^{\max}) \\ 0 & (otherwise) \end{array} \right. , \tag{4.4}$$

where $\phi_k(\theta)$ represents the estimated rotation around the k-axis of the joint. $\rho(\cdot)$ is a loss function detailed in Section 4.2.2. Then, the RoM cost for the entire body can be computed by

$$E_{RoM}(\theta) = \sum_{j=1}^{N_\mathrm{J}} e_{RoM}(\phi^{(j)}(\theta), \psi^{(j)}), \tag{4.5}$$

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. In the proposed approach, the RoM defined in the commercial Digital Human Model [26] is adopted.

Image-Based Constraints

$E_P(\theta)$ constrains positional differences between keypoints on an image $\mathbf{p}^\mathrm{C}$ detected by a CNN-based 2D pose estimator [10] and corresponding 3D joint positions projected onto the image $\hat{\mathbf{p}}^\mathrm{C}$. The 3D point of the solved model in the body coordinates $\hat{\mathbf{P}}^\mathrm{B}$ can be projected to the camera coordinates by

$$\hat{\mathbf{p}}^\mathrm{C}(\theta) = \mathbf{T}^\mathrm{GC}\mathbf{T}^\mathrm{BG}(\theta_0)\hat{\mathbf{P}}^\mathrm{B}(\theta), \tag{4.6}$$

where $\hat{\mathbf{P}}^\mathrm{B}(\theta)$ denotes the 4D column vector, which represents the 3D joint position in a homogeneous coordinate system. $\mathbf{T}^\mathrm{GC}$ and $\mathbf{T}^\mathrm{BG}(\theta_0)$ are the 4×3 translation matrices described in Section 4.2.1.

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, the proposed method extends 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

$$w_p = \begin{cases} \exp(-\frac{d_p^2}{2s^2k_p^2}) & (d_p \leq \beta_d s k_p) \\ 0 & (otherwise) \end{cases}, \tag{4.7}$$

where $p\,(1 \leq p \leq N_P)$, β_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 the experiments. Here, 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) [88], 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

$$E_P(\theta) = \sum_{p=1}^{N_{\mathrm{P}}} \rho(w_p c_p^{\mathrm{im}} \|\mathbf{p}_p^{\mathrm{C}} - \hat{\mathbf{p}}_p^{\mathrm{C}}(\theta)\|_{\mathrm{F}}^2), \tag{4.8}$$

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

In the 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, this method presents the foot-ground contact cost term $E_G(\theta)$. The proposed method detects foot-ground contact from the IMU acceleration attached to the feet, as with the conventional works for IMU-based human pose estimation and pedestrian dead-reckoning [35, 107, 111]. Fusing IMU acceleration and positional measure-

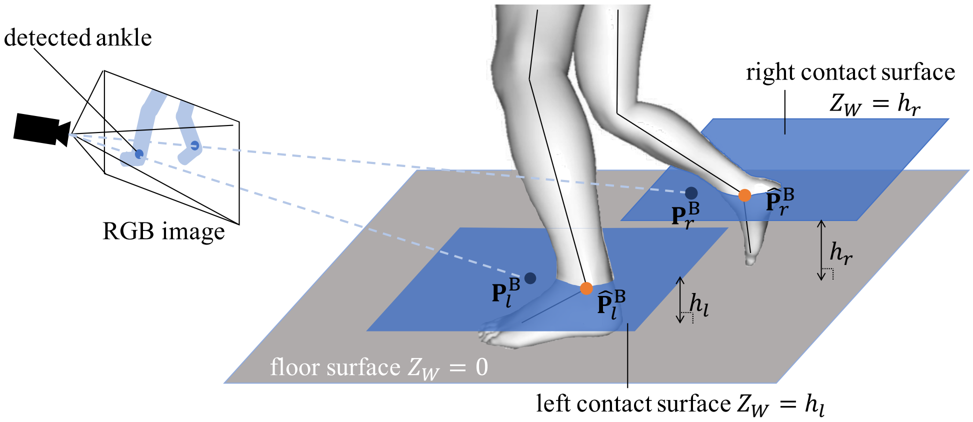

detected ankle
RGB image
right contact surface
$Z_W = h_r$
$\mathbf{P}_r^{\mathrm{B}}$
$\hat{\mathbf{P}}_r^{\mathrm{B}}$
h_r
$\mathbf{P}_l^{\mathrm{B}}$
$\hat{\mathbf{P}}_l^{\mathrm{B}}$
h_l
floor surface $Z_W = 0$
left contact surface $Z_W = h_l$

Figure 4.2: Visualization of the ground contact constraint.

ment from the camera, E_G minimizes the distance between foot position and ground contact point.

The cost terms are defined as depicted in Figure 4.2. Let $\hat{\mathbf{P}}_g^{\mathrm{B}}(\theta)$, where $g \in \{left\_foot, right\_foot\}$ is the left or right ankle position of the estimated model, and let $\mathbf{P}_g^{\mathrm{B}}$ 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 4.2.1. The confidence score c_g^{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

$$\begin{aligned} E_G(\theta) = &\ \textstyle\sum_g \rho(c_g^{\mathrm{G}} w_g c_g^{\mathrm{im}} \|\mathbf{P}_g^{\mathrm{B}} - \hat{\mathbf{P}}_g^{\mathrm{B}}(\theta)\|_{\mathrm{F}}^2), \\ \text{where} \quad c_g^{\mathrm{G}} = &\ \delta + \begin{cases} \beta_{\mathrm{G}}/\|\mathbf{a}_g\|_{\mathrm{F}}^2 & (\beta_{\mathrm{G}}/\|\mathbf{a}_g\|_{\mathrm{F}}^2 \leq 1) \\ 1 & (otherwise) \end{cases}, \end{aligned} \tag{4.9}$$

where $\mathbf{a}_g$ and β_{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, $\beta_{\mathrm{G}} = 5$ and $\beta_{\mathrm{G}}/\|\mathbf{a}\|$ was calculated using $\beta_{\mathrm{G}}/(\|\mathbf{a}\| + \epsilon)$, $\epsilon = 1.0 \times 10^{-6}$ to avoid zero division. A parameter δ takes 1 when the lowest mesh of g is lower than that of the other foot, and 0 otherwise. The weight w_g is also multiplied for handling outlier detection of foot keypoints. In the proposed 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 the extremely large error is that when the distance from the camera to the subject is large and the camera position is relatively low, the small 2D position error of detected joints on the image causes huge error in the 3D space.

4.3 Experiments

4.3.1 Dataset

The quantitative experiments are performed to evaluate the performance of the proposed approach on the 3D human pose dataset TotalCapture [97]. TotalCapture provides 60 fps of all-synchronized IMU data, HD videos from fixed cameras, and ground-truth human pose measured by the marker-based optical motion capture system. A total of 13 IMUs are attached to the head, sternum, pelvis, upper and lower limbs, and feet. The presented method uses acceleration and orientation of IMUs and an image sequence from a single camera. Note that the data measured with the optical motion capture system are not used for the proposed 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 reconstructed from the optical motion capture data has a different definition of the link structure from that of DHM. 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, the joint position and orientation of DHM are determined so that the Vicon 57-point markers defined in advance on the DHM surface match the marker positions measured by the optical motion capture [26], and used as the ground truth in this experiment.

The proposed method was evaluated following the standard evaluation protocol defined in [97]. 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 the dataset and S2-ROM3 (S2-R3), S3-F1, and S5-F1 are used instead. The limitations on the scenes where the proposed method is effective will be mentioned in Section 4.4.1.

4.3.2 Implementation Details

In the experiments, a human model is generated statistically from the height and weight of the subject, which is offered by DHM software [26]. Before starting the pose estimation, the subject took T-pose as a calibration pose. During the calibration pose, the global coordinates $(X_{\mathrm{W}}, Y_{\mathrm{W}}, Z_{\mathrm{W}})$ is defined so that the subject stands on the plane at $Z_{\mathrm{W}} = 0$. For the model of the 2D joint detector used in image-based constraints, the proposed method utilized the weights of the public pretrained model [10]. No additional training or finetuning is conducted.

The weighting parameter controls the contribution of each cost term to the overall cost (Equation (4.2)). 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.

4.3.3 Contribution of the Proposed Cost Terms

Ablation studies 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 the 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 4.3(a) represents the per-frame mean Euclidean dis-

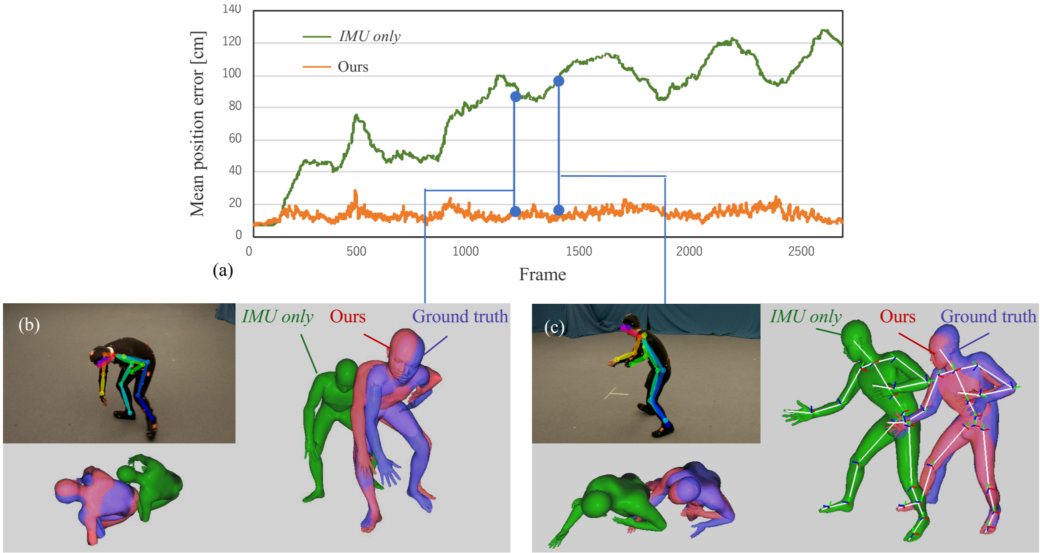


Figure 4.3: Visualization of the accumulated per joint position error.

tance between the solved pose and ground-truth. Figure 4.3(b) and (c) visualize the output of the 2D joint detector [10], and the human models colored in green, red, and blue represent the 3D human pose solved by the *IMU only* method [65], the proposed method, and the marker-based optical motion capture system (ground-truth), respectively. The estimated 2D joints and 3D models in (b) and (c), respectively capture the same frame in the same scene.

Figure 4.3(a) and the human model visualization revealed that the proposed approach using a single camera prevented the accumulation of position error. The right foot in (c) is self-occluded and the misdetection occurred; however, the proposed 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 4.1: 3D position error (cm) and orientation error (degrees) on Total-Capture dataset.

Scenes		**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)																
RGB only [68]	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
IMU only [65]	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
Pose constraint	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
Uni-weight	**19.6**	**14.8**	**11.9**	**11.5**	**9.22**	7.37	15.3	**10.1**	14.3	**15.7**	13.8	**14.6**	**14.9**	46.7	17.5	15.8
This work	20.2	15.6	12.2	12.2	10.2	**7.32**	**15.2**	12.5	**11.1**	16.3	**12.3**	14.7	16.0	**10.0**	**16.9**	**13.5**
Mean orientation error (degrees)																
IMU only [65]	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
This work	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

Table 4.1 summarizes the quantitative results for pose estimation using the position error metric. The state-of-the-art 3D human pose estimation method using only a single RGB camera is referred to as *RGB only* [68]. A baseline method, termed *Pose constraint*, estimates the human pose by minimizing the objective function of the proposed method without the foot-ground constraint, i.e., *Pose constraint* minimizes the cost function composed of $E_O(\theta)$, $E_{RoM}(\theta)$, and $E_P(\theta)$. The results revealed that the foot-ground contact cost term $E_G(\theta)$ improves the positional error. Another baseline method, termed *Uni-weight*, optimizes the pose by Equation (4.2), but adaptive weight w_p is fixed to 1. Meanwhile, the proposed cost function calculates w_p according to Equation (4.7). Although the mean error of the proposed method in the 15 scenes was smallest, *Uni-weight* estimated the human pose with the highest accuracy in more than half of the test scenes. Especially in Walking 2 (W2), *Uni-weight* outperformed the proposed approach in four out of five 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

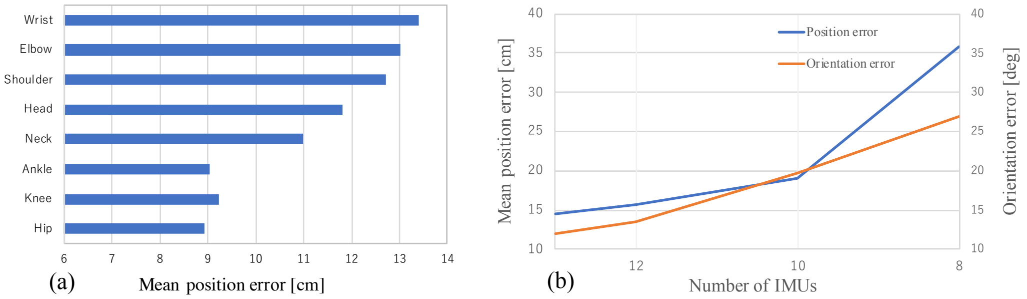


Figure 4.4: (a) Mean per-joint positional error of the human motion capture on all the test scenes. (b) Mean 3D position and orientation errors on subjects S3-F1 and S4-F3 with 8 to 13 IMUs.

of the joints on an image due to the self-occlusion or unusual posture of the subject (included in Freestyle 3 and Acting 3).

The mean orientation error of joints w.r.t. the pelvis joint coordinates is shown in the bottom of Table 4.1. The error of *IMU only* and the proposed method were 8.75 degrees and 8.83 degrees, respectively. Also, the mean joint position errors w.r.t. the pelvis joint coordinates were 6.72 cm and 6.74 cm, respectively. No significant differences in both positional and orientational errors were observed, which suggests that the vision-based cost terms give a small effect on posture estimation.

The effect of the ground contact cost term can be observed in Figure 4.4(a). It represents the per-joint position error of the 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.

The proposed method can easily be extended to use multi-view cameras

Table 4.2: 3D position error (cm) on the multi-camera setting.

Scenes	**S1-F3**	**S2-R3**	**S3-F1**	**S4-F3**	**S5-F1**	**Mean**
Trumble *et al.* [97]	9.4	9.3	13.6	11.6	10.5	10.9
Malleson *et al.* [61]	7.4	**3.9**	6.7	6.4	7.0	6.3
This work (multi-view)	**6.25**	5.66	6.70	**6.32**	**5.91**	**6.17**

by adding the image-based cost function $E_P(\theta)$ and $E_G(\theta)$ for each camera and simultaneously minimize the total cost. The experiments using 8 cameras and 13 IMUs are conducted. The state-of-the-art approach for 3D motion capture that infers both joint position and orientation from IMUs and multiple images [61] extracted several images from TotalCapture to test their approach. The performance of the proposed approach was compared with [61, 97] on the same scenes as the test set of [61], excluding the scenes where the subject jumped.

As shown in Table 4.2, in several scenes, the present method outperformed the conventional approaches that optimize the pose parameter to reconstruct human motion. In the scene where the proposed approach was inferior in the 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 perform well. The experiments demonstrated that the proposed ground contact constraint contributed to improving the accuracy of 3D human pose estimation in the multi-view camera settings as well as single-camera settings when the floor plane was pre-defined and the foot can be detected from the camera.

4.3.4 The Number of IMUs

Wearing many IMUs takes time and hampers the subject's range of motion. Towards the real-world use of the proposed method, the relation between the accuracy of the pose estimation and the number of IMUs were investigated. The experiments were conducted with (1) 13 IMUs: full set as described in Section 4.3.1, (2) 12 IMUs: full set without a 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 4.4(b).

The decrease of the IMUs largely affects the accuracy of both position and orientation. It would be because the proposed single-camera approach does not constrain joint positions other than the foot in three dimensions. In the experiments on *IMU only* and *Pose constraint*, 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.

4.4 Summary

This chapter presented the first online approach to estimate the 3D human pose fusing IMUs and a single camera. To constrain the position of the solved model in three dimensions, the cost term was proposed to optimize the timing and position of foot grounding. This work handled 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.

4.4.1 Limitations

To calculate the confidence of foot grounding, Equation (4.9) assumes that at least one of the feet is grounded. Therefore, the accuracy of the proposed approach degrades in a sequence in which a subject lifts both feet off the ground for a long time, such as by jumping. It is confirmed from the experiment on S5-F1 which included side-skip steps that the short period of feet takeoff does not seriously affect the accuracy. This limitation will be overcome by inferring ground contact confidence from visual context and IMU data.

Since the present method assumes that the subject walks on a flat floor, it does not support pose estimation when the subject gets on a step or goes up the stairs. This limitation should be alleviated by reconstructing a 3D model of the environment around the subject and defining the floor surface according to the 3D model before the measurements.

Chapter 5

Automation of IMU calibration

5.1 Introduction

IMUs are a prominent option for analyzing human motion. Body-worn IMUs can be used to estimate rotational and, sometimes, translational motion of the attached segment, which help estimate the required motion parameters. As the sensors operate at a high frame rate with low latency, they can be introduced in real-time applications for motion analysis, such as full-body motion capture [41, 61, 65] and navigation [31, 44]. Furthermore, recent technological advances have dramatically reduced the size and price of IMUs, making them the most promising technology for the continuous tracking of human movements in daily life [84, 95, 96]. Because of recent improvements that have enabled easy configuration, non-expert (but trained) users can collect motion data with IMUs. A clinic's doctors or their assistants can use the inertial sensors to track patients' motions to assist in rehabilitation or disease diagnosis [11, 67, 120]. Some studies have collected data from many participants wearing IMUs during everyday life for an action recognition task [36, 50, 87].

For a detailed and robust motion analysis, many IMU-based applications derive data from multiple sensors mounted on multiple body segments. The conventional approach to gait analysis attaches six IMUs to the upper and lower legs and feet [89]. Some IMU-based full-body motion analyses require more than 10 inertial sensors to track one subject [46,61]. Such configurations are prone to errors because each sensor must be attached to a predefined body segment. If an IMU is mounted on the wrong segment, remeasurement will be required. After the measurement operator carefully attached the IMUs, the relative orientation between the IMUs and the attached joints must be calculated. Many operators use a calibration pose (e.g., T-pose: standing upright with hands open at the sides) to obtain the relative orientation, which requires prior knowledge of manipulation and skills for operation. These problems can be an obstacle for general users' ability to measure motion with IMUs. Hence, a technique to identify the segment to which each sensor is attached and to calculate the relative orientation based on the sensor signals is desired, as it would make IMU attachment easier and quicker. These identification and orientation calculation tasks are called IMU-to-segment (I2S) assignment and I2S orientation alignment, respectively [123]. In this thesis, these two assignment and alignment procedures are collectively referred to as IMU calibration.

This chapter addresses both I2S assignment: the task of classifying IMU data into classes corresponding to the body segments on which IMUs are mounted and I2S orientation alignment: the task to obtain the relative orientation between each IMU and the attached joint. When a measurement operator uses the IMU calibration framework proposed in this chapter, although only one IMU needs to be attached to the predetermined segment, the other IMUs can be mounted on arbitrary segments because the framework automatically assigns the sensors to the segments to which they are attached. Further, the operator does not need to make the subject take a calibration pose because the calibration framework predicts the relative orientation based solely on the sensors' measurements during a few seconds of walking.

The classical approaches to IMU calibration involve manually designing features for discriminating IMU placements [3, 55, 104]. Recent work has proposed extraction for features using deep neural networks (DNNs) [123]. Although these approaches have achieved high assignment accuracy in well-controlled settings (e.g., the approximate angle of the sensor to the segment in the test set is the same as those of the training set), their accuracy has decreased in trials that did not meet these conditions.

To mitigate these limitations and robustly perform the IMU calibration, the proposed approach merges features across all body-worn IMUs and learns the global dependencies between these IMUs. Unlike conventional methods that assign and align sensors one by one, the proposed approach calibrates all body-worn IMUs at once through the DNNs. The proposed model assigns and aligns each IMU based on a global feature that represents the motions of all sensor-attached segments of a body. In addition, the model learns the dependency relationships between IMUs, which enables it to perform calibration based on data from relevant IMUs (e.g., IMUs attached to the adjacent segment). To implement this feature fusion and dependency learning, a new

DNN architecture that incorporates a global feature generation module and an attention-based mechanism is presented.

The proposed method was experimentally evaluated using synthetic and real datasets in three sensor configurations. The results demonstrated that the proposed approach significantly outperformed those of the conventional work and baselines in assignment and alignment accuracy. The ablation studies and attention maps generated by the intermediate layer of the proposed model suggested that the present model captured the dependency relationships between IMUs. The results obtained with the real IMU dataset validated the robustness of the proposed method. The contributions of this work are summarized as follows:

- This chapter proposes a novel IMU calibration model that generates a global feature representing the motion of all body segments to which IMUs are attached and learns pairwise dependencies between the IMUs.
- Ablation studies demonstrate that merging features extracted from multiple body-worn IMUs can benefit the identification of a segment where each IMU is mounted.
- The extensive evaluation shows that the proposed method outperforms the conventional and baseline methods in three sensor configurations on synthetic and real public datasets.

Note that this chapter first addresses the I2S assignment in the following sections. Subsequently, the differences between the proposed method for I2S assignment and I2S orientation alignment are mentioned because these two tasks share the core of the problem.

5.1.1 Global Feature Extraction

The proposed module to generate a global feature that represents the motion of all segments to which IMUs are attached is inspired by a technique used in point cloud semantic segmentation: the task of separating a point cloud into multiple regions according to the semantic meanings of points [34]. Because a 3D point in a point cloud, which has only positional data, has little information, recently developed approaches have successfully handled point clouds by aggregating local features and obtaining global features [29,38,82]. The feature aggregation module incorporated in the proposed model allows the model to use the global motion of the body segments for the assignment of IMUs.

Pointnet [82] is the pioneering work in applying neural networks to learn over general point sets. It takes raw point clouds as input and obtains a global feature through a pooling layer that follows individual feature extractors composed of a simple multi-layer-perceptron (MLP). The pooling aggregator is widely used in various tasks against various data structures [40,52,109] due to its simple implementation and the permutation invariance of the inputs. The proposed assignment model generates a global feature using the pooling aggregator to merge individual features from the IMU data that are input in random order.

5.1.2 Attention Mechanism

Attention-based neural networks have been successfully applied to a wide variety of fields, such as natural language [24,99], image [12,25], and speech [18] processing. The studies report that learning the dependencies among the intermediate features through the attention mechanism improves recognition accuracy. The learned attention also helps interpret the reasoning behind the machine prediction and improves the explainability of the DNN models [16, 110].

Transformer is one of the most promising approaches for learning global dependencies using the attention mechanism [99]. Transformer has been proposed for use in the task of natural language processing and has been quickly adopted for a variety of tasks, such as image classification [25] and object detection [12]. The self-attention operator in Transformer explores the dependencies of input feature vectors. The proposed method incorporates the Transformer encoder into the presented model to obtain the dependency relationships between body-worn IMUs. The attention mechanism is expected to capture the pairwise dependencies of the sensors, which enables the assignment of an IMU that relies on the features extracted from the dependent IMUs.

5.2 Methods

5.2.1 Problem Setting

I2S assignment method identifies a segment to which each IMU is mounted based only on the IMU signals without relying on external sensors. A DNN-based model to learn the discriminant features and classify the IMUs into the attached segments is constructed. In the proposed framework, a user processes the assignment following the three steps below:

1. The user selects a root IMU from a set of IMUs to be mounted and attaches it to the predetermined root segment of a subject.
2. The user mounts the remaining IMUs on the defined position of the arbitrary body segments of the subject.
3. The proposed model provides assignment predictions using the data from all body-worn IMUs while the subject walks for a few seconds.

In this work, only one sensor is placed on the predetermined segment, which dramatically reduces the risk of misplacement and the effort required from the user to attach the sensors. Unlike with the conventional methods [123], the user can mount IMUs at any angle. The position of the sensors needs to be known (e.g., an arbitrary sensor should be mounted on the middle of a bone); however, this constraint is satisfied in most practical situations [104]. The role of the root IMU and the difference in assignment accuracy depending on the selected segment as a root are mentioned in Sections 5.2.3 and 5.6.2, respectively. When 15 sensors are mounted on different segments, the I2S

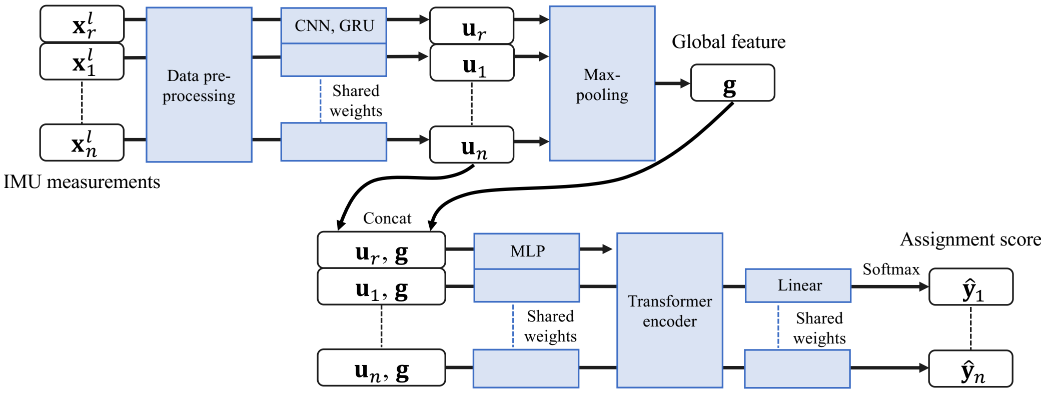


Figure 5.1: Overview of the proposed I2S assignment framework.

assignment can be regarded as a task to classify the sequence data of 14 IMUs into 14 classes associated with the segments, except for the root segment.

5.2.2 Method Overview

The proposed I2S assignment framework, as illustrated in Figure 5.1, consists of data preprocessing, IMU-wise feature extraction, global feature generation, and attention learning modules. The proposed model takes as input the accelerations and angular velocities of n target IMUs to be classified and one root IMU and provides n predicted classification scores associated with all segments except the root. Note that the data from the root IMU is placed at the top of the input matrix; however, the data from the n target IMUs are stored in the input matrix in a random order to train the model for the assignment task.

In the data preprocessing module, accelerations and angular velocities in the sensor-local coordinates are converted to the root sensor coordinates, and noise is added to the accelerations for data augmentation. Then, the discriminant features are extracted from the IMU signals in a one-by-one manner, and these features are merged in the global feature generation module. In the final step, pairwise dependencies between the IMUs are learned in the Transformer encoder [99], and the model then provides classification scores through a linear transformation with softmax activation.

5.2.3 Data Preprocessing

Coordinate transformation and data augmentation are performed in the data preprocessing modules for better generalization and convergence of the proposed assignment model. In this section and Figure 5.2, the accelerations, angular velocities, and orientations refer to the values at a specific time step t $(1 \leq t \leq T)$, where T is the window size of the IMU data; however, the notation of time step t is eliminated for simplicity.

At first, the raw sensor signals w.r.t. the sensor-local coordinates F_S^i $(1 \leq i \leq n)$, where n is the number of IMUs to be assigned, are transformed into the root sensor coordinate frame F_R. The transformation makes the inputs invariant to the walking direction of the subject; this means the representation of the sensor signals can be the same when the subject is walking north and south, which facilitates the training of the model. The transformation matrix $\mathbf{R}_{RS}^i$ that maps F_S^i to F_R can be obtained via

$$\mathbf{R}_{RS}^i = \mathbf{R}_{WR}{}^{\mathsf{T}}\mathbf{R}_{WS}^i, \tag{5.1}$$

where $\mathbf{R}_{WR}$ and $\mathbf{R}_{WS}^i$ represent the orientation of the root sensor and the i-th sensor w.r.t. the world coordinate frame F_W, respectively. Figure 5.2 depicts an example of coordinate transformation when the lower back is chosen as a

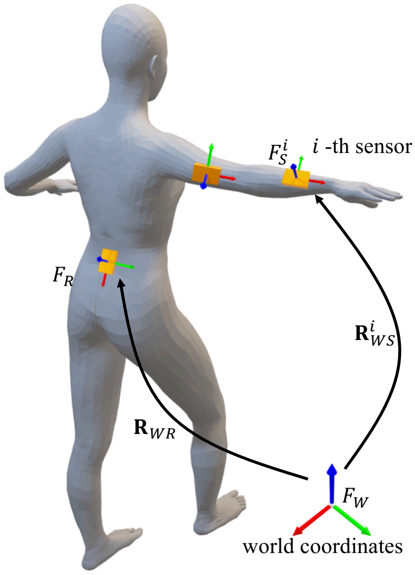


Figure 5.2: Relations among the coordinate systems.

root segment. Then, 3D acceleration $\mathbf{a}_i$ w.r.t. F_R is calculated by a simple dot product with $\mathbf{R}_{RS}^i$ and the sensor-local acceleration $\mathbf{a}_i^l$ expressed as

$$\mathbf{a}_i = \mathbf{R}_{RS}^i \mathbf{a}_i^l. \tag{5.2}$$

Given $\mathbf{R}_{RS}^i$ and the sensor-local angular velocity, 3D angular velocity $\boldsymbol{\omega}_i$ w.r.t. F_R is obtained by applying the classical method [6].

Data augmentation is executed to avoid over-fitting and to stabilize the performance of the trained model. Following the methods of successful studies that have applied DNNs to IMU data [78, 123], the sensor signals are augmented by adding zero-mean Gaussian noise to the accelerations. The i-th IMU data after the above data preprocessing is referred to as $\mathbf{x}_i \in \mathbb{R}^{T\times6}$, which stacks T frames of $\mathbf{a}_i$ and $\boldsymbol{\omega}_i$.

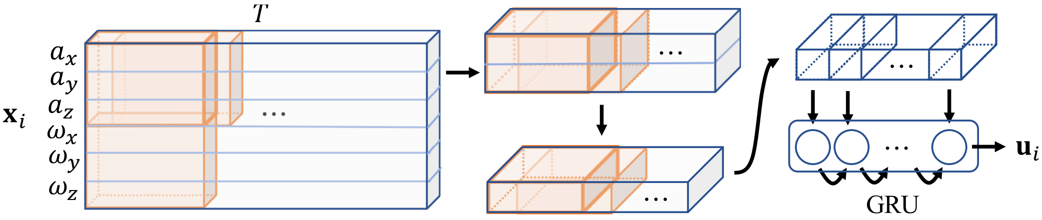


Figure 5.3: Illustration of the proposed convolution operator.

5.2.4 IMU-Wise Feature Extraction and Feature Aggregation

The proposed DNN-based assignment model starts with IMU-wise feature extraction. Inspired by the conventional architectures applied to IMU accelerations and angular velocities [78, 123], the feature extractor is constructed with CNN layers and a recurrent network layer.

The main difference between previous work and the proposed convolution operators is the step-by-step change in kernel size for each CNN layer. Figure 5.3 illustrates the proposed convolution operator. The orange boxes in the blue blocks represent the convolution kernels. The input (a_x, a_y, a_z) and $(\omega_x, \omega_y, \omega_z)$ represent the accelerations $\mathbf{a}_i$ and angular velocities $\boldsymbol{\omega}_i$, respectively. The kernel size and strides of the first convolution along the height are three. This operator explicitly extracts features from accelerations and angular velocities separately, and the next convolution layer with kernel height $k_h = 2$ fuses both features. Another convolution layer follows to acquire deeper merged features. This feature extraction architecture is inspired by those in the previous literature that report high recognition accuracy in multi-modal fusion tasks using multi-stream feature extraction and fusion modules [15,51]. In the proposed model, batch normalization and nonlinear activation follow each convolution operation. ReLU activation $\rho(\cdot)$ is used for the activation function that computes $\rho(x) = \max(0, x)$.

The recurrent units are incorporated after the convolution layers. The method adopts gated recurrent units (GRU) [19] following the results presented in the previous work that performed I2S assignments [123]. The feature map from the last CNN layer $\mathbf{m} \in \mathbb{R}^{T_L \times d_L}$ is divided into T_L one-dimensional features $\mathbf{m}_j \in \mathbb{R}^{d_L}$, where $(1 \leq j \leq T_L)$. The feature $\mathbf{m}_j$ is recurrently processed by GRU, and the output at the last time step T_L is returned. Finally, the introduced module provides the IMU-wise feature representation $\mathbf{u}_i$, which is extracted from $\mathbf{x}_i$.

The IMU-wise features individually extracted by the CNNs and the recurrent layer are aggregated to generate a global feature that represents the global motion of the segments to which the IMUs are attached. The architecture chosen for feature merging follows the recent success of the pooling aggregator proposed in [82]. The aggregated feature $\mathbf{g}$ is described as

$$\mathbf{g}(p, q) = \max(\mathbf{u}_{\mathrm{r}}(\mathrm{p}, \mathrm{q}), \mathbf{u}_{1}(\mathrm{p}, \mathrm{q}), \cdots, \mathbf{u}_{\mathrm{n}}(\mathrm{p}, \mathrm{q})), \tag{5.3}$$

where $\mathbf{g}(p, q)$ and $\mathbf{u}_i(p, q)$ denote the values of $\mathbf{g}$ and $\mathbf{u}_i$ at position (p, q), respectively, and $\mathbf{u}_r$ represents the feature extracted from the root IMU data. The global feature $\mathbf{g}$ forms the same shape as $\mathbf{u}_i$. The features $\mathbf{g}$ and $\mathbf{u}_i$ are concatenated to describe the feature of the i-th IMU, which contains the global feature extracted from all the body-mounted IMUs.

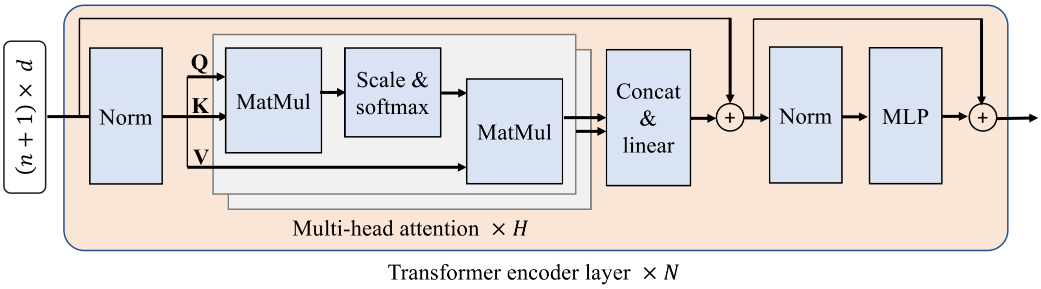


Figure 5.4: Architecture of the Transformer encoder layer.

5.2.5 Attention-Based Architecture

Transformer learns the dependency relationships between the feature vectors and obtain discriminant feature representations [99]. The IMU-wise features concatenated with the global feature $(\mathbf{u}_i, \mathbf{g})$ are projected to d-dimensional vectors through the MLP with d nodes. The $(n+1)$ d-dimensional features form a matrix $\mathbf{U} \in \mathbb{R}^{(n+1)\times d}$, which is input to the Transformer layer, as shown in Figure 5.4.

The architecture within the attention learning layer is designed to be similar to that of the original Transformer encoder [99]; however, there are two differences between the original and the proposed model. One is the position at which layer normalizations (LNs) are applied. LNs are applied before the multi-head attention module and before MLP, following the method used by recent works that modified the Transformer and improved its recognition accuracy [25,75]. Another difference is the lack of position embeddings because the proposed model solves an assignment problem that assumes the order of the input is unknown.

A given input $\mathbf{U}$ to the attention learning module is normalized by LN. The normalized $\mathbf{U}$ is projected H times into queries $\mathbf{Q}_h \in \mathbb{R}^{(n+1)\times d_k}$, keys

$\mathbf{K}_h \in \mathbb{R}^{(n+1)\times d_k}$, and values $\mathbf{V}_h \in \mathbb{R}^{(n+1)\times d_v}$ by three learnable matrices $\mathbf{W}_h^q, \mathbf{W}_h^k \in \mathbb{R}^{d\times d_k}$, and $\mathbf{W}_h^v \in \mathbb{R}^{d\times d_v}$, where $1 \leq h \leq H$. Using $\mathbf{Q}_h$ and $\mathbf{K}_h$, the attention matrix $\mathbf{A}_h$ is calculated by

$$\mathbf{A}_h = \mathrm{softmax}\left(\frac{\mathbf{Q}_\mathrm{h}\mathbf{K}_\mathrm{h}}{\sqrt{\mathrm{d}_\mathrm{k}}}\right). \tag{5.4}$$

The H outputs from the multi-head attention, $\mathbf{A}_h\mathbf{V}_h$ are concatenated, linearly projected, and undergo LN. Then, the layer produces an output of the same shape as the input through the IMU-wise MLP. The residual connections are applied before the second LN operator and after the IMU-wise MLP. The attention-based module is composed of a stack of N identical attention learning layers. From the output of the last layer, n feature vectors (except that of the root IMU) are linearly projected with softmax activation, resulting in n probabilities $\hat{\mathbf{y}}_i \in \mathbb{R}^n$.

In the training phase, the model is trained using the cross-entropy loss between $\hat{\mathbf{y}}_i$ and the one-hot true label $\mathbf{y}_i \in \mathbb{R}^n$ which is associated with the input $\mathbf{x}_i$ as an objective function. The proposed model is trained in an end-to-end manner.

In the test phase, it was experimentally revealed that defining an objective function from the probability distribution $\hat{\mathbf{y}}_i$ and assigning the IMUs to maximize the function improves the accuracy, rather than classifying them directly into the segment indicated by the maximum value of $\hat{\mathbf{y}}_i$. Specifically, the prediction matrix $\mathbf{Y} \in \mathbb{R}^{n\times n}$ is defined by

$$\mathbf{Y}^\mathsf{T} = (\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \cdots, \hat{\mathbf{y}}_n). \tag{5.5}$$

Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ be a boolean matrix, where $\mathbf{B}(i, j) = 1$ if row i is assigned to column j. Only one of the elements in a row has 1, and the others must have 0. Then, the assignment algorithm seeks $\hat{\mathbf{B}}$ by solving the following optimization:

$$\hat{\mathbf{B}} = \arg\max_{\mathbf{B}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{B}(i,j)\mathbf{Y}(i,j). \tag{5.6}$$

The objective function is optimized using the 2D rectangle assignment algorithm [21] implemented in the SciPy library [100]. In the experiments, this optimization was applied to the proposed method and all comparison approaches, which contributed to the improved accuracy of all methods, including the conventional method.

5.3 Experimental Setup

5.3.1 Implementation Details

The left three blocks in Figure 5.5 illustrate the architecture and hyperparameters of the proposed model. The architecture of each block is detailed in Section 5.2. The algorithm based on the Tree-structured Parzen Estimator was used to seek the hyperparameter values, such as the learning rate, the batch size, and the number of kernels and GRU nodes. The dataset is divided into training, validation, and test set (see Appendix A.1 for details); the validation set was then used for parameter tuning, and the values found are described in Appendix A.2. The parameters are fixed through all the experiments.

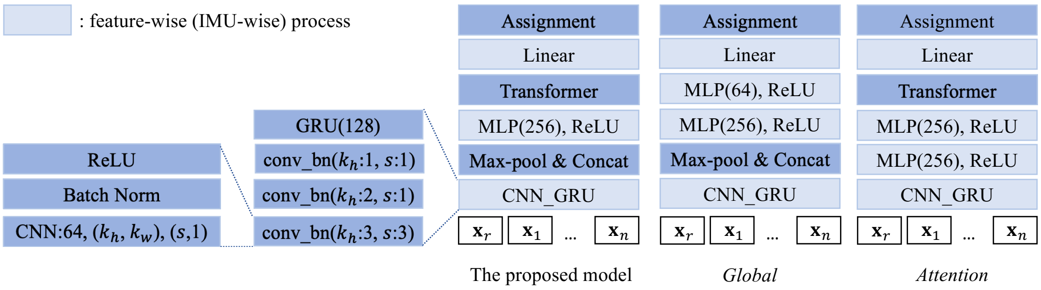


Figure 5.5: The architecture and the hyperparameters of the networks.

5.3.2 Baselines

The assignment accuracy of the proposed model was compared to that of the conventional method [123], referred to as *one-by-one*, which applied DNN to identify IMU placement and infer the I2S orientation alignment of the IMU in a one-by-one manner. Since this section focuses on the I2S assignment, the branch layers for the alignment in *one-by-one* were pruned.

To validate the contribution of the feature aggregation module and the attention-based mechanism, two baseline methods were implemented. The two models, *Global* and *Attention*, are depicted as the right two blocks in Figure 5.5. *Global* is composed of IMU-wise feature extraction and global feature aggregation by the max-pooling layer. *Global* is a model made by removing the attention-based learning module from the proposed architecture. In contrast, *Attention* handles the features extracted from each IMU data to learn the dependency relationships without aggregating the IMU-wise features. The hyperparameters, the dataset division, and the coordinate frame of the input are consistent for the proposed, conventional, and baseline models across all the experiments.

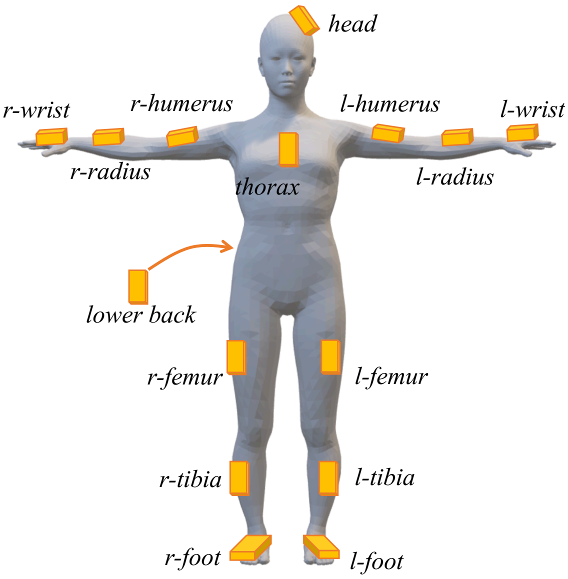


Figure 5.6: Sensor placement in the full-body configuration.

5.3.3 Dataset

The performance of the presented approach was quantitatively evaluated on the synthetic and real IMU datasets: CMU-MoCap [2] and TotalCapture [97]. The sensor arrangement of the CMU-MoCap is shown in Figure 5.6. Assuming that the proposed framework is utilized not only for full-body motion analysis but also for the measurement of body parts, the model was evaluated on lower-, upper-, and full-body configurations. The sensor placements are defined as follows:

- lower body (7): *lower back*, *l-femur*, *r-femur*, *l-tibia*, *r-tibia*, *l-foot*, and *r-foot*
- upper body (9): *head*, *thorax*, *lower back*, *l-humerus*, *r-humerus*, *l-radius*, *r-radius*, *l-wrist*, and *r-wrist*

- full body (15): segments on both lower and upper body (*lower back* is duplicated),

where *l*- and *r*- represent left and right body segments, and the figures in $(\cdot)$ denote the number of the segments. Then, since the root segment is determined a priori, the I2S assignment in lower-, upper-, and full-body configurations can be regarded as the task of classifying the time-series signals of the IMUs into 6, 8, and 14 classes, respectively. The segment *lower back* was selected as a root segment through all experiments, excluding Section 5.6.2.

CMU-MoCap is the public human motion dataset captured with the marker-based optical motion capture system [2]. The synthetic IMU data was generated assuming that the IMU was attached to the segments of the body measured in CMU-MoCap. The generation algorithm is described in Appendix A.3. In the simulation dataset, 42 subjects performing different walking styles are selected, which are used in [123]. The models were trained with IMU signals from 26 subjects in the training set and 7 subjects in the validation set, and they were tested with the remaining 9 subjects' data (detailed in Appendix A.1).

TotalCapture is a public dataset providing 60 fps of all-synchronized IMU data, HD videos, and ground-truth human poses measured by the marker-based optical motion capture system [97]. Since the proposed approach uses only IMU signals for the I2S assignment, real IMU data were utilized for the training and evaluation of the models. The number of IMUs was 13, and the sensor arrangement was the same as with CMU-MoCap, with the *l-wrist* and *r-wrist* sensors removed. TotalCapture has five subjects with a variety of motions measured. The walking scenes including three subjects' data were used for training, one subject's data was used for validation, and the rest was for testing. The period during which the subjects took a calibration pose (the first and last two seconds) and walked backward were manually removed from the dataset. TotalCapture is a challenging dataset in three aspects. First, the number of subjects in the training data is small, which easily causes over-fitting. Second, it contains a variety of walking styles, including many twists and turns and slow and fast walking. Finally, the positions and angles of the sensors attached to the body change slightly depending on the subject because TotalCapture is not a dataset intended for evaluating I2S assignment but for pose estimation. Through the experiments on TotalCapture, the versatility of the proposed method was evaluated.

The window size of the input IMU data was two seconds (i.e., the number of frames $T = 120$ in 60 fps input data), and the windows were always shifted by 0.25 seconds. CMU-MoCap and TotalCapture provide 120 fps and 60 fps IMU signals, respectively, and they are used at the original frame rate.

5.4 Results

5.4.1 Assignment Accuracy

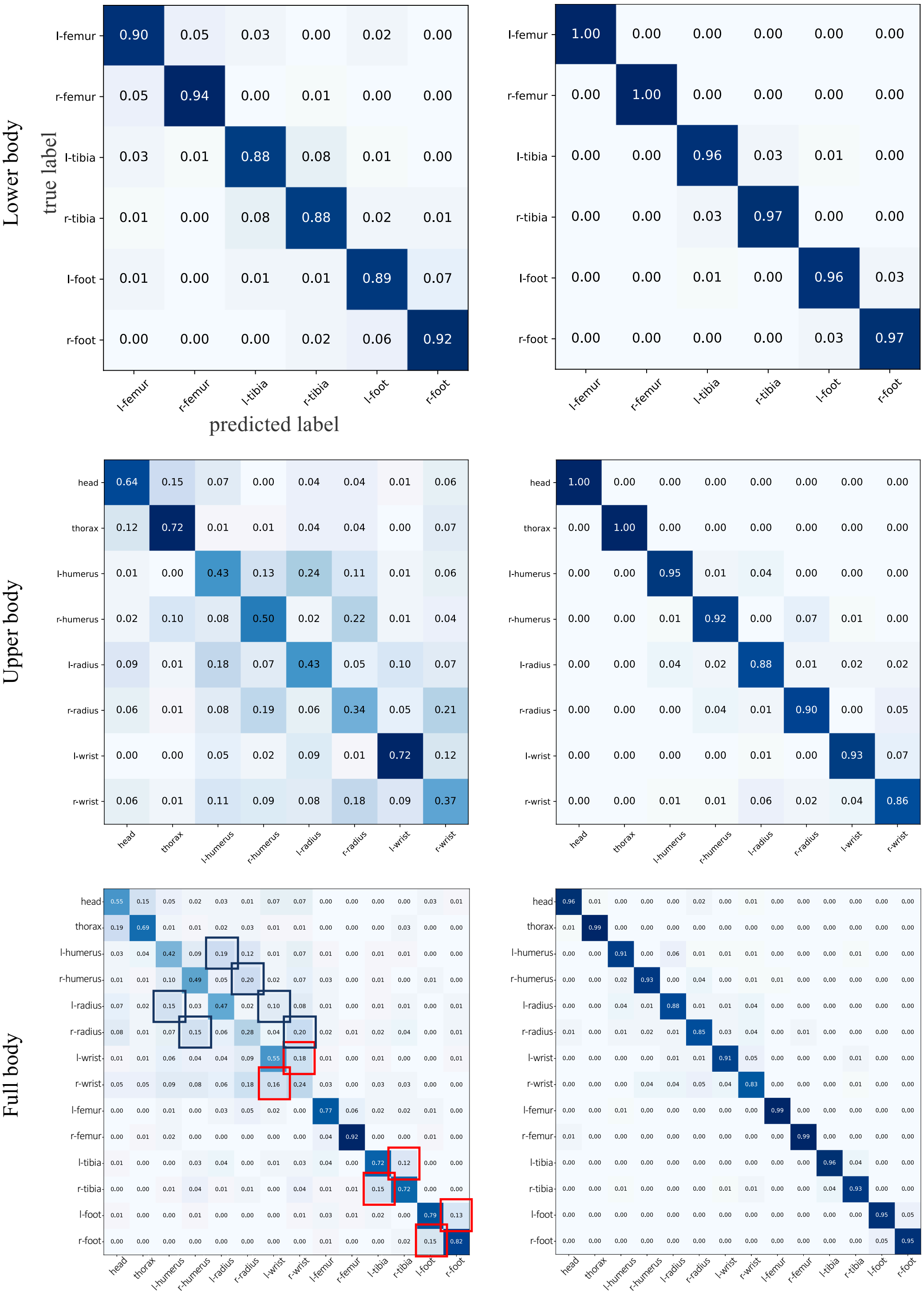
The experimental results obtained using the setup described in Section 5.3 are shown in Table 5.1. As seen in this table, the proposed method outperformed the other methods on both datasets for three configurations of sensor attachment, showing that I2S assignment training in the proposed approach yields better feature representations to discriminate the segment to which

Table 5.1: Assignment accuracy on the two datasets in the three configurations. All figures represent percentages.

	CMU-MoCap [2]			**TotalCapture** [97]		
	lower	upper	full	lower	upper	full
One-by-one [123]	90.0	51.7	60.2	93.6	80.2	83.1
Global	97.3	81.8	88.1	96.6	89.9	89.9
Attention	97.6	89.7	91.9	91.7	91.0	90.1
This work	**97.8**	**93.0**	**93.1**	**96.7**	**93.5**	**91.6**

each IMU is attached.

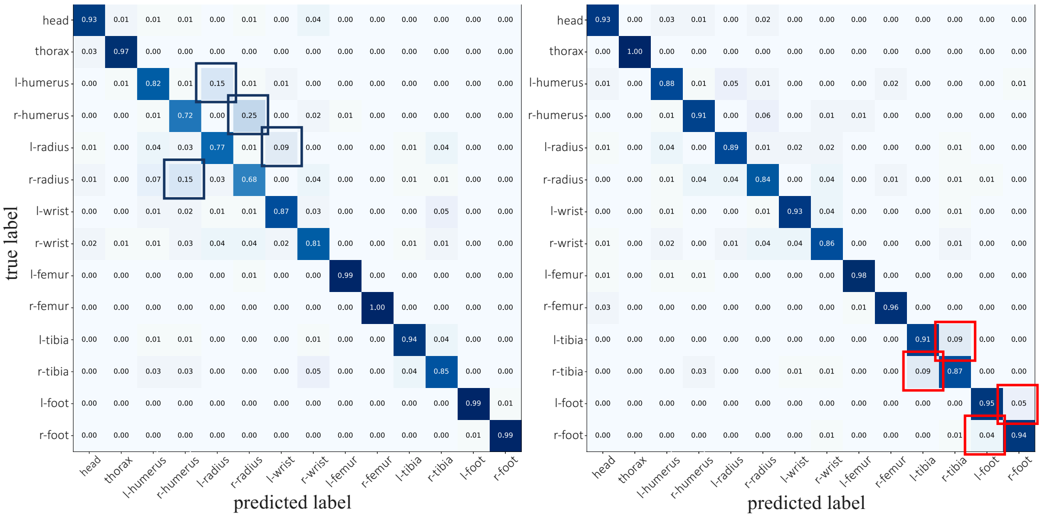
The assignment results on the CMU-MoCap [2] are visualized using confusion matrices in Figure 5.7. The matrices show that the assignment errors are caused by two main types of mistakes: left/right switch (l/r switch) and intra-limb misassignment (intra-misassignment). The l/r switch indicates an incorrect assignment to the opposite side of the attached segments (e.g., the IMU mounted on the *l-wrist* is classified into the *r-wrist* class). The intra-misassignment denotes that the IMU attached to a part of the limb is misclassified to another part of the same limb (e.g., the IMU mounted on the *l-wrist* is assigned to the *l-radius* or *l-humerus* class). Some of the l/r switches and intra-misassignments are highlighted in the confusion matrix at the lower left part of Figure 5.7 with red and blue squares, respectively. The figure shows that the proposed method reduced both mistakes and significantly improved the assignment accuracy.



(a) *One-by-one* [123]

(b) Ours

Figure 5.7: Some results on CMU-MoCap [2] in terms of confusion matrices. The red and blue rectangles on the lower-left confusion matrix highlight the left/right switches and intra-limb misassignments, respectively.



(a) *Global*

	head	thorax	l-humerus	r-humerus	l-radius	r-radius	l-wrist	r-wrist	l-femur	r-femur	l-tibia	r-tibia	l-foot	r-foot
head	0.93	0.01	0.01	0.01	0.00	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
thorax	0.03	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
l-humerus	0.00	0.01	0.82	0.01	0.15	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r-humerus	0.00	0.00	0.01	0.72	0.00	0.25	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00
l-radius	0.01	0.00	0.04	0.03	0.77	0.01	0.09	0.00	0.00	0.00	0.01	0.04	0.00	0.00
r-radius	0.01	0.00	0.07	0.15	0.03	0.68	0.00	0.04	0.00	0.00	0.01	0.01	0.00	0.00
l-wrist	0.00	0.00	0.01	0.02	0.01	0.01	0.87	0.03	0.00	0.00	0.00	0.05	0.00	0.00
r-wrist	0.02	0.01	0.01	0.03	0.04	0.04	0.02	0.81	0.00	0.00	0.01	0.01	0.00	0.00
l-femur	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00
r-femur	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
l-tibia	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.94	0.04	0.00	0.00
r-tibia	0.00	0.00	0.03	0.03	0.00	0.00	0.00	0.05	0.00	0.00	0.04	0.85	0.00	0.00
l-foot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.01
r-foot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.99

(b) *Attention*

	head	thorax	l-humerus	r-humerus	l-radius	r-radius	l-wrist	r-wrist	l-femur	r-femur	l-tibia	r-tibia	l-foot	r-foot
head	0.93	0.00	0.03	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
thorax	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
l-humerus	0.01	0.00	0.88	0.01	0.05	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01
r-humerus	0.00	0.00	0.01	0.91	0.00	0.06	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
l-radius	0.01	0.00	0.04	0.00	0.89	0.01	0.02	0.02	0.00	0.00	0.00	0.01	0.00	0.00
r-radius	0.00	0.00	0.01	0.04	0.04	0.84	0.00	0.04	0.00	0.01	0.00	0.01	0.01	0.00
l-wrist	0.00	0.00	0.01	0.00	0.00	0.01	0.93	0.04	0.00	0.00	0.00	0.01	0.00	0.00
r-wrist	0.01	0.00	0.02	0.00	0.01	0.04	0.04	0.86	0.00	0.00	0.00	0.01	0.00	0.00
l-femur	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00
r-femur	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.96	0.00	0.00	0.00	0.00
l-tibia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.09	0.00	0.00
r-tibia	0.00	0.00	0.00	0.03	0.00	0.00	0.01	0.01	0.00	0.00	0.09	0.87	0.00	0.00
l-foot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	0.05
r-foot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.04	0.94

Figure 5.8: Comparison between the two baselines on the CMU-MoCap [2] in the full-body configuration.

5.4.2 Ablation Studies

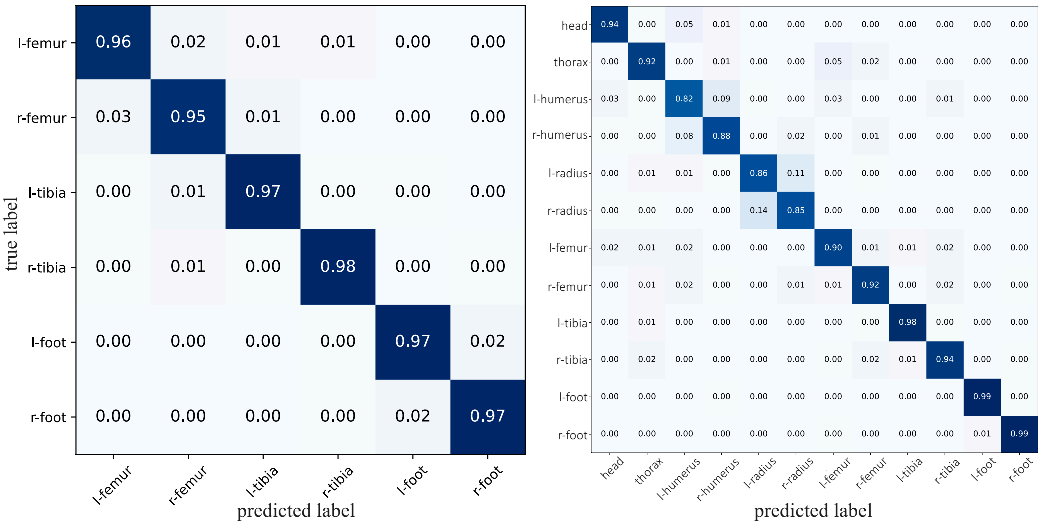
To analyze the contribution of each module in the proposed model to mitigate the l/r switch and intra-misassignment problems, the confusion matrices of *Global* and *Attention* are visualized in Figure 5.8 and computed the error rate caused by each mistake. On the CMU-MoCap dataset, the average l/r switch rates (the number of l/r switches divided by the total number of assignments) and the intra-misassignment rates for all three configurations were 2.2% and 5.5% for *Global*, and 3.1% and 2.4% for *Attention*. The lower l/r switch rates of *Global* and the lower intra-misassignment rate of *Attention* can be observed in the confusion matrices shown in Figure 5.8 as well.

The results suggest that the global feature aggregation alleviates the l/r switch problem. This could be because the aggregation allows the network to model the motion of all body segments and capture the motion of each IMU relative to the global body motion, thus enabling the model to discriminate between left and right. The results also suggest that the attention module reduces intra-misassignment errors. This could be because the model with the attention learning architecture classifies the IMU data with consideration of the information from the relevant IMUs, such as IMUs attached to adjacent and opposite segments. For example, as can be seen in Figure 5.11(b) (see Section 5.1.2 for an explanation of the figure), when assigning an IMU mounted on *l-tibia*, the self-attention architecture devotes much attention to *l-femur*, *l-foot*, and *r-tibia*. The assignment prediction relying on the IMUs on the segments in the same limb should prevent intra-misassignment.

5.4.3 Results on a Challenging Dataset

The results on the TotalCapture dataset [97], as presented in Table 5.1 and Figure 5.9, revealed that the proposed approach is robust to different walking styles and slight changes in the IMU positions depending on the subjects. The proposed model took the same period of data as an input regardless of the change in walking speed, but the method achieved high accuracy in all the sensor configurations.

The accuracy in assigning the arm segments was lower than that of the other segments for two possible reasons. One is a variety of movements not found in a normal gait in the training dataset, such as touching a head and face and raising clenched fists. The other is that the subject in the test set walked without moving his arms for a few seconds. The trained model could not distinguish between the IMU movement on the arms, head, and chest in the scene. Specifically, the mean assignment accuracy in the three seconds



(a) lower body

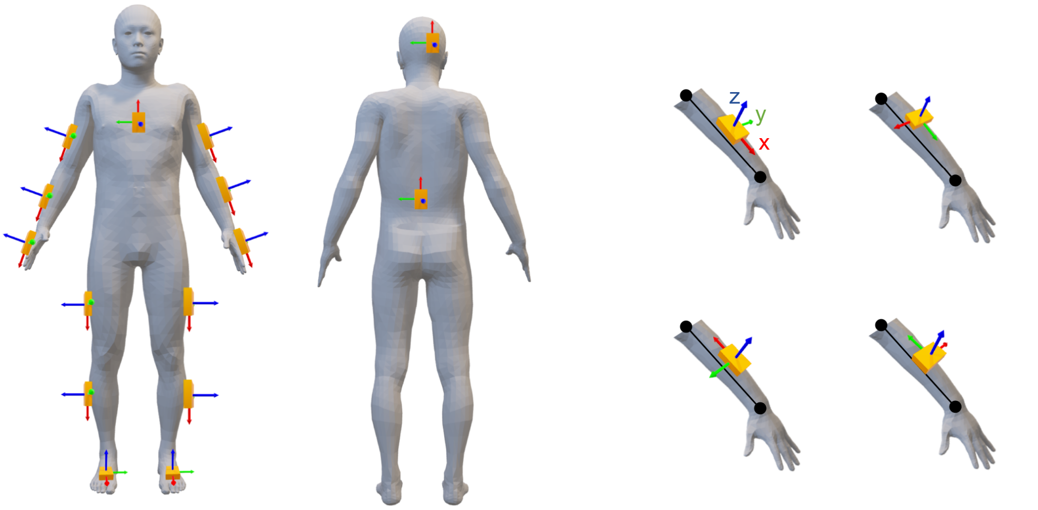
(b) full body

Figure 5.9: Assignment accuracy of the proposed method on the TotalCapture dataset [97] in terms of confusion matrices.

of the test scene in which the subject walked slowly without waving his arms (from 55 to 58 seconds in S5-W2 in TotalCapture [97]) was 61.1% in the full-body setting.

5.5 IMU-to-Segment Orientation Alignment

IMU calibration in this thesis consists of two tasks: I2S assignment and I2S orientation alignment. I2S orientation alignment is a task to obtain the relative orientation between the IMU and the bone to which the IMU is attached. The estimated relative orientation is essential for human motion capture, as is used in Equation (4.1). In this work, the I2S assignment method proposed and evaluated in previous sections is extended to predict the relative orientation of IMUs. The remainder of this section contains the description of the proposed method and evaluation of I2S orientation



(a) Sensor configuration

(b) Four categories of I2S orientation alignment

Figure 5.10: Setups of IMU-to-Segment orientation alignment.

alignment.

Problem Setting. The problem setting of this work follows that of I2S assignment mentioned in Section 5.2.1; however, this task adds one constraint to the attachment of the IMUs: each IMU is mounted on the segment so that its sensor-local x-axis is parallel or vertical to the corresponding bone. The z-axis of the IMU remains perpendicular to the bone. Figure 5.10(b) depicts this setting. In this setting, I2S orientation alignment can be regarded as a task of classifying each IMU's orientation into four categories that indicate the angles of its x-axis to the bone of 0°, 90°, 180°, and 270°. This assumption improves the convergence of the model but does not significantly increase the burden on the measurement operator.

5.5.1 Method

In contrast to I2S assignment, the information from the sensors attached to other segments does not help to align the orientation of an IMU, since I2S orientation alignment predicts the relative orientation of the IMU to the corresponding (attached) joint. Therefore, the proposed method computes the sensor orientation one by one.

The proposed network architecture is very similar to the IMU-wise feature extraction module of the I2S assignment network, i.e., three CNN layers are followed by a GRU. The only difference is that the data from each modality is convolved independently and concatenated after the convolution because this model takes orientation matrices as the inputs and their dimensions are different from the other modalities (acceleration and angular velocity). The features extracted from CNNs and GRU are linearly transformed with softmax activation, which represent the model predictions.

5.5.2 Experiments

The accuracy of the proposed I2S orientation alignment method was compared with the baseline method that learns I2S assignment and alignment simultaneously. In the baseline method, termed *Simultaneous*, the latent features from the penultimate layer of the assignment model are concatenated with the extracted feature vectors in the alignment module. The proposed and baseline models are trained with the same hyperparameters as the I2S assignment model described in Appendix. A.2. The experiments are performed on CMU-MoCap [2] in the same dataset division as mentioned in

Table 5.2: I2S assignment and orientation alignment accuracy on CMU-MoCap in the three configurations. All figures represent percentages.

	Lower body		**Upper body**		**Full body**	
	assign	align	assign	align	assign	align
Section 5.2	**97.8**	-	**93.0**	-	**93.1**	-
Simultaneous	96.7	**99.6**	86.8	88.4	90.2	97.6
Section 5.5	-	**99.6**	-	**97.4**	-	**98.6**

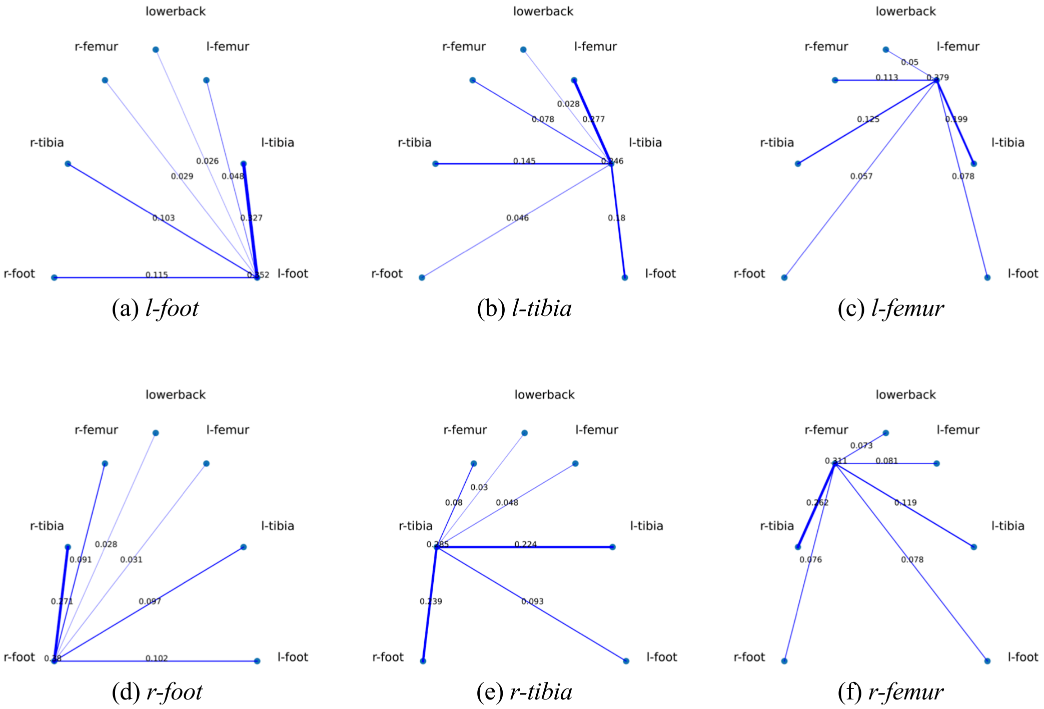
Appendix A.1. As shown in Figure 5.10(a), the virtual IMUs are mounted at the positions to which general IMU-based human motion measurements attach.

As shown in Table 5.2, the proposed model that trained IMU assignment and alignment separately achieved higher accuracy than *Simultaneous*, and achieved 98.6% in full-body configuration. The baseline model degraded the assignment accuracy as well. It suggests that *Simultaneous* could not share the discriminative features that contribute to both I2S assignment and alignment.

5.6 Discussion

5.6.1 Attention Maps Visualization

An attention mechanism can be used to improve the explainability of deep learning models [110, 115, 116]. Explainability, in this context, refers to a better understanding for humans of why the models behave as they do. The explainability of a model helps users make decisions based on the model and allows researchers to understand what input and intermediate features affect



(a) *l-foot*

(b) *l-tibia*

(c) *l-femur*

(d) *r-foot*

(e) *r-tibia*

(f) *r-femur*

Figure 5.11: Visualization of the self-attention matrices of the proposed model. The figures on the lines denote the attention scores.

the results of the model. The attention learning architecture used in the proposed model can capture the pairwise relationships between the IMUs and explain what dependencies the predicted assignments rely highly on.

To visualize the dependencies between the IMUs, the mean attention matrix was computed, which represents the average of the self-attention matrix (calculated by Equation (5.4)) from all the H heads and all the N Transformer encoder layers. Each row of the attention matrix represents the dependencies between the IMUs associated with the columns. The pairwise dependencies are separately visualized for each body segment in Figure 5.11. A high attention score suggests high dependency. For example, Figure 5.11(a) describes the degree of dependence on the IMU attached to each segment

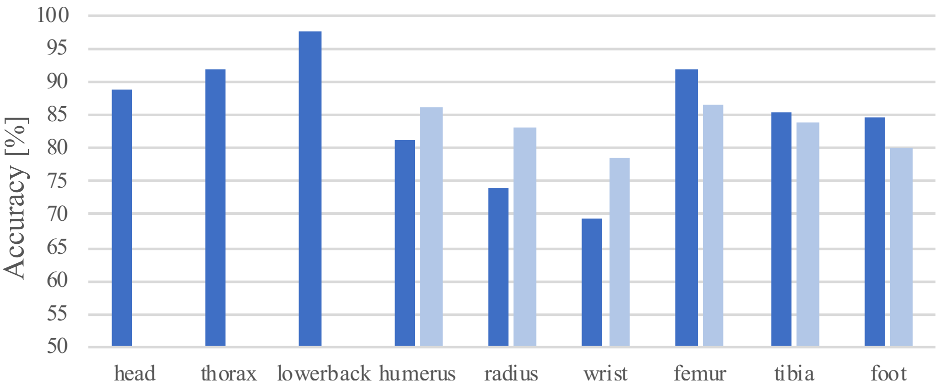


Figure 5.12: Assignment accuracy depending on the root segment in full-body configuration. The graphs in lighter blue represent the right side of the segment (e.g., *r-humerus* and *r-radius*).

when the model performed the assignment for the IMU mounted on *l-foot*. For all the segments from (a) to (f) in Figure 5.11, it can be seen that much attention is devoted to the adjacent segments and the opposite segments even in the test phase, during which the model has no prior knowledge of which segment each IMU is mounted on.

5.6.2 Root Segment Selection

The accuracy of the I2S assignment according to the root segment is shown in Figure 5.12. The results suggest that the segments that stably and faithfully follow the body orientation (e.g., *lower back*, *thorax*, *head*, and *femur*) are suitable for the root. In contrast, when the segments on the arms that have great freedom of movement during walking were chosen, the assignment accuracy decreased significantly.

5.7 Summary

This chapter presented an approach that identifies the segment on which each IMU is mounted by merging the features of all the body-worn IMUs and by learning the dependency relationships between the sensors. A pooling aggregator was incorporated to obtain a feature that represents the global motion of the body. In addition, a self-attention learning architecture was implemented to allow the model to perform an IMU assignment, relying on the signals from the relevant IMUs. The proposed model was quantitatively evaluated on simulated and real IMU datasets, which validated our method, showing that it accurately and robustly performed the I2S assignment and orientation alignment. Ablation studies suggested that the global feature fusion and attention mechanism reduced left/right switches and intra-limb misassignments.

5.7.1 Limitations

The present I2S assignment framework assumes that the sensor configuration is known a priori and that one of the sensors is placed on the predetermined segment. These limitations do not significantly impair practicality; however, further studies to relax them are needed.

Chapter 6

Conclusion and Outlook

This thesis proposed two approaches for motion capture using visual and inertial sensors and an IMU calibration framework for real-world motion measurements. This chapter restates the contributions and speculates on promising directions for future work.

This thesis first proposed a novel multi-view motion capture approach that reconstructs weighted voxels of the human body. The method combines the back-projection of the 2D body keypoints and silhouettes. Focusing on the 3D position of the body CoM, extensive experiments were performed, which validated the effectiveness of weighting each segment. The experiment considering the use in a sports field verified the applicability of the proposed method for tracking CoM in dynamic postures.

Towards resolving a limitation of current visual-inertial motion capture systems that require multi-view cameras to localize the human body, this thesis has also contributed the world's first method for 3D human pose estimation with a single-view camera and multiple body-worn IMUs, which optimizes the position and timing of foot-ground contact. Given a synchronized RGB video and signals from body-worn IMUs, the proposed method optimizes pose parameters of the human model by minimizing IMU-based cost terms and image-based ones that penalize the discrepancy between the foot-ground contact points of the optimized model and that of visual estimation. The experiments demonstrated that the proposed method outperforms the image-based and IMU-based 3D pose estimation methods in the joint position accuracy. The experiments further showed that the proposed method is comparable to the state-of-the-art methods in a multi-view and multi-IMU setting.

Lastly, this thesis proposed a framework for IMU calibration, aiming to resolve the complexities of the preprocessing steps of the IMU-based human motion measurements, including the secondly proposed method in this thesis. This work contributes to extensive IMU-based motion analyses because it relieves the procedures of the measurement operator. The proposed framework identifies the body segment on which each IMU is mounted and computes the relative orientation between the sensor and the attached bone from the IMU signals during a few seconds of a walk. The experiment demonstrated that the proposed approach significantly outperformed the conventional and baseline approaches.

Future Work. This thesis proposed three approaches for motion capture in the real world. This section describes a few interesting directions for future research opportunities based on this thesis.

Exploiting the measured motion can be an extension of the work performed in this thesis. The analyses of the captured human movements would lead to the development of various applications, such as surgical diagnostics, rehabilitation methods, motion assist devices, and user applications for health care, which may contribute to improving the quality of life of many people. In addition, exploitation of the captured motion by extracting contextual information should be explored. For example, since the second proposed method captures human motion using a monocular camera and body-worn IMUs, features from the estimated human pose and images would enable fine-grained action recognition and motion forecasting.

Capturing human motion with sparse sensors is challenging but desirable. Extending the proposed method to be used in a measurement system consisting of only a few sensors is worth exploring. One approach to reducing the sensors would be adding geometric or statistical constraints to optimize the human motion parameters. Another solution would be to train models using data obtained from measurements with dense sensor configurations. For example, the data obtained with the CoM estimation approach proposed in this thesis would be applicable to train a model that predicts the CoM position using a monocular camera. In addition to the abovementioned studies, relaxing the limitations described in each chapter is included in future work.

Appendix A

Implementation Details of IMU Calibration

A.1 Division of the Dataset

The data in a dataset are divided into training, validation, and test sets. In this paper, both the synthetic dataset CMU-MoCap [2] and the real dataset TotalCapture [97] are divided into the three sets on the basis of the subject (i.e., all the trials (scenes) of a subject are put into one of the three sets). The specific division of each dataset is summarized in Table A.1. In CMU-MoCap [2], subjects performing simple "walking" and having at least 600

Table A.1: Dataset division. The figures represent the ID of the subjects.

	Train	**Validation**	**Test**
CMU-MoCap [2]	2, 6, 7, 10, 15, 16, 32, 36, 37, 38, 39, 45, 81, 91, 93, 103, 104, 105, 114, 120, 132, 133, 139, 141, 143, 144	3, 8, 43, 69, 113, 136, 137	5, 12, 26, 27, 29, 46, 49, 55, 111
TotalCapture [97]	1, 2, 3	4	5

frames in every scene are chosen as a test set.

A.2 Hyperparameters for Model Training

Figure 5.5 visualizes the architecture and parameters of the network.

In IMU-wise feature extraction, three CNNs with different kernel sizes (3,k_w), (2,k_w), and (1,k_w), where k_w represents the kernel width, were utilized in this order. The strides of these kernels were (3,1), (1,1), and (1,1), respectively: k_w varies to scale the size of the convolution operator, depending on the input frame T. Specifically, k_w was set to $[T/15 + 1]$ in the experiments. The number of nodes in GRU following the CNNs was 128. After the max pooling and the concatenations of the vectors, MLP with the number of nodes $d = 256$ mapped each feature to be the input of the Transformer encoder layer [99]. The hyperparameters in the Transformer encoder were as follows: The embedding dimensions of the query d_k, key d_k, and value d_v were 256. The number of the MLP nodes after the second LN operator was set to 768. The number of the attention heads H and of encoder layers, N was fixed to 4.

The presented network was trained for 1000 epochs with a batch size of 128. Early stopping with patience 400 was performed, and the model that achieved the lowest loss on the validation set was utilized for the test. RMSProp with a fixed learning rate of 0.001 was applied to optimize the model.

A.3 Simulated Data Generation

The public human motion dataset named CMU-MoCap [2] provides much 3D kinematics data which are measured using the optical motion capture system. The human joint position $\mathbf{p}_{WJ}^{t}$ and orientation $\mathbf{R}_{WJ}^{t}$ w.r.t. the world coordinates F_W at time step t are utilized. A virtual IMU is attached to a bone by defining a rotation matrix $\mathbf{R}_{JS}$ and translation vector $\mathbf{t}_{JS}$, which represent the orientation and position of the virtual sensor w.r.t. the joint coordinate frame F_J. They are kept fixed in F_J during movement, assuming that the IMU is attached to a rigid human body and its motion is perfectly linked to the associated joint motion. In the experiments, the IMU was placed at the midpoint of the bone, and the joint position data was preprocessed with a zero-lag Butterworth filter [8] of order 8 and a cutoff frequency of 10 Hz, following the previous work [123]. Then, the IMU position $\mathbf{p}_{WS}^{t}$ and orientation $\mathbf{R}_{WS}^{t}$ w.r.t. F_W at time step t were obtained via

$$\mathbf{p}_{WS}^{t} = \mathbf{p}_{WJ}^{t} + \mathbf{R}_{WJ}^{t}\mathbf{t}_{JS} \tag{A.1}$$

$$\mathbf{R}_{WS}^{t} = \mathbf{R}_{WJ}^{t}\mathbf{R}_{JS}. \tag{A.2}$$

The angular velocity of the IMU w.r.t. F_W is computed by [6] using the sensor orientation in the current frame R_{WS}^{t} and the next frame $R_{WS}^{(t+1)}$. The IMU acceleration at t time step $\mathbf{a}_{WS}^{t}$ w.r.t. F_W is calculated by

$$\mathbf{a}_{WS}^{t} = \frac{\mathbf{p}_{WS}^{(t+1)} - 2\mathbf{p}_{WS}^{t} + \mathbf{p}_{WS}^{(t-1)}}{\Delta t^2}, \tag{A.3}$$

where Δt denotes the period of the time step. Since all the IMU accelerations and angular velocities are transformed to the root sensor coordinate before they are input to the assignment model, these values are invariant to the

IMU orientations w.r.t. F_J. Therefore, unlike the previous work [123], this work did not generate IMU data with various orientations relative to F_J.

References

[1] GIMP: GNU IMAGE MANIPULATION PROGRAM. `https://www.gimp.org`, visited on 01/12/2018.

[2] CMU Graphics Lab Motion Capture Database. `http://mocap.cs.cmu.edu`, visited on 03/15/2020.

[3] Navid Amini, Majid Sarrafzadeh, Alireza Vahdatpour, and Wenyao Xu. Accelerometer-based on-body sensor localization for health and medical monitoring applications. *Pervasive and mobile computing*, 7(6):746–760, 2011.

[4] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019.

[5] Harshada Badave and Madhav Kuber. Head pose estimation based robust multicamera face recognition. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 492–495, 2021.

[6] Frank N. Barnes. Stable member equations of motion for a three-axis gyro stabilized platform. *IEEE Transactions on Aerospace and*

Electronic Systems, (5):830–842, 1971.

[7] Richard N. Baumgartner, W. Cameron Chumlea, and A. F. Roche. Estimation of body composition from bioelectric impedance of body segments. *The American journal of clinical nutrition*, 50(2):221–226, 1989.

[8] Stephen Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

[9] Song Cao, Kan Chen, and Ram Nevatia. Activity recognition and prediction with pose based discriminative patch model. In *2016 IEEE Winter Conference on Applications of Computer Vision*, pages 1–9, 2016.

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[11] Carlotta Caramia, Diego Torricelli, Maurizio Schmid, Adriana Munoz-Gonzalez, Jose Gonzalez-Vargas, Francisco Grandas, and Jose L. Pons. Imu-based classification of parkinson's disease from gait: A sensitivity analysis on sensor location and feature selection. *IEEE journal of biomedical and health informatics*, 22(6):1765–1774, 2018.

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020.

[13] Justin Carpentier, Mehdi Benallegue, Nicolas Mansard, and Jean-Paul Laumond. Center-of-mass estimation for a polyarticulated system in contact—a spectral approach. *IEEE Transactions on Robotics*, 32(4):810–822, 2016.

[14] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019.

[15] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171, 2017.

[16] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 765–774, 2019.

[17] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[18] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recog-

nition. *arXiv preprint arXiv:1506.07503*, 2015.

[19] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[20] Stefano Corazza, Lars Mündermann, Emiliano Gambaretto, Giancarlo Ferrigno, and Thomas P. Andriacchi. Markerless motion capture through visual hull, articulated ICP and subject specific model generation. *International journal of computer vision*, 87(1):156–169, 2010.

[21] David F. Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.

[22] Paolo De Leva. Adjustments to zatsiorsky-seluyanov's segment inertia parameters. *Journal of biomechanics*, 29(9):1223–1230, 1996.

[23] John Dennis and Jorge Moré. Quasi-newton methods, motivation and theory. *SIAM Review*, 19(1):46–89, 1977.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[26] Yui Endo, Mitsunori Tada, and Masaaki Mochimaru. Dhaiba: development of virtual ergonomic assessment system with human models. In *Proceedings of The 3rd International Digital Human Symposium*, 2014.

[27] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[28] Zeyu Fu, Federico Angelini, Jonathon Chambers, and Syed Mohsen Naqvi. Multi-level cooperative fusion of gm-phd filters for online multiple human tracking. *IEEE Transactions on Multimedia*, 21(9):2277–2291, 2019.

[29] Liuhao Ge, Zhou Ren, and Junsong Yuan. Point-to-point regression pointnet for 3d hand pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 475–491, 2018.

[30] Alejandro González, Mitsuhiro Hayashibe, Vincent Bonnet, and Philippe Fraisse. Whole body center of mass estimation with portable sensors: Using the statically equivalent serial chain and a kinect. *Sensors*, 14(9):16955–16971, 2014.

[31] Rodrigo Gonzalez and Paolo Dabove. Performance assessment of an ultra low-cost inertial measurement unit for ground vehicle navigation. *Sensors*, 19(18):3865, 2019.

[32] Caroline Göpfert, Mikko V. Pohjola, Vesa Linnamo, Olli Ohtonen, Walter Rapp, and Stefan J. Lindinger. Forward acceleration of the centre of mass during ski skating calculated from force and motion capture data. *Sports Engineering*, 20(2):141–153, 2017.

[33] Liangjie Guo and Shuping Xiong. Accuracy of base of support using an inertial sensor based motion capture system. *Sensors*, 17(9):2091, 2017.

[34] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2021.

[35] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4318–4329, 2021.

[36] Fasih Haider, Fahim A. Salim, Dees B. W. Postma, Robby van Delden, Dennis Reidsma, Bert-Jan van Beijnum, and Saturnino Luz. A super-bagging method for volleyball action recognition using wearable sensors. *Multimodal Technologies and Interaction*, 4(2):33, 2020.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[38] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Learning semantic segmentation of large-scale point clouds with random sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[39] Fuyang Huang, Ailing Zeng, Minhao Liu, Qiuxia Lai, and Qiang Xu. Deepfuse: An imu-aware network for real-time 3d human pose estima-

tion from multi-view image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 429–438, 2020.

[40] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[41] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018.

[42] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[43] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019.

[44] Faisal Jamil, Naeem Iqbal, Shabir Ahmad, and Do-Hyeun Kim. Toward accurate position estimation using learning to prediction algorithm in indoor navigation. *Sensors*, 20(16):4410, 2020.

[45] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *European Conference on Computer Vision*, pages 718–734, 2020.

[46] Tomoya Kaichi, Tsubasa Maruyama, Mitsunori Tada, and Hideo Saito. Resolving position ambiguity of imu-based human pose with a single rgb camera. *Sensors*, 20(19):5453, 2020.

[47] Tomoya Kaichi, Tsubasa Maruyama, Mitsunori Tada, and Hideo Saito. Learning sensor interdependencies for imu-to-segment assignment. *IEEE Access*, 9:116440–116452, 2021.

[48] Tomoya Kaichi, Shohei Mori, Hideo Saito, Kosuke Takahashi, Dan Mikami, Mariko Isogawa, and Hideaki Kimata. Estimation of center of mass for sports scene using weighted visual hull. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1809–1815, 2018.

[49] Christoph Kalkbrenner, Steffen Hacker, Maria-Elena Algorri, and Ronald Blechschmidt-Trapp. Motion capturing with inertial measurement units and kinect. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, volume 1, pages 120–126, 2014.

[50] Parinaz Kasebzadeh, Gustaf Hendeby, Carsten Fritsche, Fredrik Gunnarsson, and Fredrik Gustafsson. Imu dataset for motion and device mode classification. In *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8, 2017.

[51] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 529–535, 2017.

[52] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy flow networks: deep sets for particle jets. *Journal of High Energy Physics*, 2019(1):121, 2019.

[53] Dimitrios Kotiadis, Hermanus J. Hermens, and Petrus H. Veltink. Inertial gait phase detection for control of a drop foot stimulator: Inertial sensing for gait phase detection. *Medical engineering & physics*, 32(4):287–297, 2010.

[54] Jogendra Nath Kundu, Siddharth Seth, M. V. Rahul, Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11312–11319, 2020.

[55] Kai Kunze, Paul Lukowicz, Holger Junker, and Gerhard Tröster. Where am i: Recognizing on-body positions of wearable sensors. In *International Symposium on Location-and Context-Awareness*, pages 264–275, 2005.

[56] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[57] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37, 2016.

[58] Xiaobai Liu, Yuanlu Xu, Lei Zhu, and Yadong Mu. A stochastic attribute grammar for robust cross-view human tracking. *IEEE transactions on circuits and systems for video technology*, 28(10):2884–2895, 2017.

[59] Diogo C. Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019.

[60] Sumit Majumder and M. Jamal Deen. Wearable imu-based system for real-time monitoring of lower-limb joints. *IEEE Sensors Journal*, 21(6):8267–8275, 2020.

[61] Charles Malleson, Andrew Gilbert, Matthew Trumble, John Collomosse, Adrian Hilton, and Marco Volino. Real-time full-body motion capture from video and IMUs. In *2017 International Conference on 3D Vision (3DV)*, pages 449–457, 2017.

[62] Andrea Mapelli, Matteo Zago, Laura Fusini, Domenico Galante, Andrea Colombo, and Chiarella Sforza. Validation of a protocol for the estimation of three-dimensional body center of mass kinematics in sport. *Gait & posture*, 39(1):460–465, 2014.

[63] Worthy N. Martin and Jagdishkumar Keshoram Aggarwal. Volumetric descriptions of objects from multiple views. *IEEE transactions on pattern analysis and machine intelligence*, (2):150–158, 1983.

[64] J. Martinez, Rayat Hossain, J. Romero, and J. Little. A simple yet effective baseline for 3d human pose estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017.

[65] Tsubasa Maruyama, Mitsunori Tada, and Haruki Toda. Riding motion capture system using inertial measurement units with contact constraints. *International Journal of Automation Technology*, 13(4):506–516, 2019.

[66] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[67] Nastaran Mohammadian Rad, Twan Van Laarhoven, Cesare Furlanello, and Elena Marchiori. Novelty detection using deep normative modeling for imu-based abnormal movement monitoring in parkinson's disease and autism spectrum disorders. *Sensors*, 18(10):3533, 2018.

[68] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. 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*, pages 10133–10142, 2019.

[69] Gyeongsik Moon, Heeseung Kwon, Kyoung Mu Lee, and Minsu Cho. Integralaction: Pose-driven feature integration for robust human action recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2021.

[70] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer*

Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, pages 548–564, 2020.

[71] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019.

[72] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017.

[73] Lars Mundermann, Stefano Corazza, Ajit M. Chaudhari, Eugene J. Alexander, and Thomas P. Andriacchi. Most favorable camera configuration for a shape-from-silhouette markerless motion capture system for biomechanical analysis. In *Videometrics VIII*, volume 5665, page 56650T, 2005.

[74] Bijan Najafi, Jacqueline Lee-Eng, James S. Wrobel, and Ruben Goebel. Estimation of center of mass trajectory using wearable sensors during golf swing. *Journal of sports science & medicine*, 14(2):354, 2015.

[75] Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*, 2021.

[76] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. 3d human pose estimation with 2d marginal heatmaps. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1477–1485, 2019.

[77] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3467–3475, 2017.

[78] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[79] Yi-Chung Pai and James Patton. Center of mass velocity-position predictions for balance control. *Journal of biomechanics*, 30(4):347–354, 1997.

[80] Melissa T. Parks, Zhuo Wang, and Ka-Chun Siu. Current low-cost video-based motion analysis options for clinical rehabilitation: a systematic review. *Physical therapy*, 99(10):1405–1425, 2019.

[81] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 663–670, 2010.

[82] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 652–660, 2017.

[83] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4342–4351, 2019.

[84] Sen Qiu, Zhelong Wang, Hongyu Zhao, and Huosheng Hu. Using distributed wearable sensors to measure and evaluate human lower limb motions. *IEEE Transactions on Instrumentation and Measurement*, 65(4):939–950, 2016.

[85] Manju Rana and Vikas Mittal. Wearable sensors for real-time kinematics analysis in sports: a review. *IEEE Sensors Journal*, 21(2):1187–1207, 2020.

[86] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[87] Philipp V. Rouast, Hamid Heydarian, Marc T. P. Adam, and Megan E. Rollo. Oreba: A dataset for objectively recognizing eating behavior and associated intake. *IEEE Access*, 8:181955–181963, 2020.

[88] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 369–378, 2017.

[89] Thomas Seel, Jörg Raisch, and Thomas Schauer. Imu-based joint angle measurement for gait analysis. *Sensors*, 14(4):6891–6909, 2014.

[90] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 53–70, 2020.

[91] Alison L. Sheets, Geoffrey D. Abrams, Stefano Corazza, Marc R. Safran, and Thomas P. Andriacchi. Kinematics differences between the flat, kick, and slice serves measured using a markerless motion capture method. *Annals of biomedical engineering*, 39(12):3011, 2011.

[92] Takeshi Shimmura, Ryosuke Ichikari, Takashi Okuma, Hiroyuki Ito, Kei Okada, and Tomomi Nonaka. Service robot introduction to a restaurant enhances both labor productivity and service quality. *Procedia CIRP*, 88:589–594, 2020.

[93] Leonid Sigal and Michael J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 120(2), 2006.

[94] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019.

[95] Luke Sy, Michael Raitor, Michael Del Rosario, Heba Khamis, Lauren Kark, Nigel H. Lovell, and Stephen J. Redmond. Estimating lower limb kinematics using a reduced wearable sensor count. *IEEE Transactions on Biomedical Engineering*, 68(4):1293–1304, 2021.

[96] Salvatore Tedesco, John Barton, and Brendan O'Flynn. A review of activity trackers for senior citizens: Research perspectives, commercial landscape and the role of the insurance industry. *Sensors*, 17(6):1277, 2017.

[97] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, volume 2, page 3, 2017.

[98] Athanasios Tsitsoulis and Nikolaos G. Bourbakis. A methodology for extracting standing human bodies from single images. *IEEE Transactions on Human-Machine Systems*, 45(3):327–338, 2015.

[99] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[100] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[101] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[102] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and IMUs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1533–1547, 2016.

[103] Timo Von Marcard, Bodo Rosenhahn, Michael J. Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, volume 36, pages 349–360, 2017.

[104] Dirk Weenk, Bert-Jan F. Van Beijnum, Chris T. M. Baten, Hermie J. Hermens, and Peter H. Veltink. Automatic identification of inertial sensor placement on human body segments during walking. *Journal of neuroengineering and rehabilitation*, 10(1):1–9, 2013.

[105] Christian M. Welch, Scott A. Banks, Frank F. Cook, and Pete Draovitch. Hitting a baseball: A biomechanical description. *Journal of orthopaedic & sports physical therapy*, 22(5):193–201, 1995.

[106] Pierre-Brice Wieber. Holonomy and nonholonomy in the dynamics of articulated motion. In *Fast motions in biomechanics and robotics*, pages 411–425, 2006.

[107] Yuan Wu, Haibing Zhu, Qingxiu Du, and Shuming Tang. A pedestrian dead-reckoning system for walking and marking time mixed movement using an shss scheme and a foot-mounted imu. *IEEE Sensors Journal*, 19(5):1661–1671, 2018.

[108] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019.

[109] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 244–253, 2018.

[110] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Towards explainable human pose estimation by transformer. *arXiv preprint arXiv:2012.14214*, 2020.

[111] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 40(4), 08 2021.

[112] Xiaoping Yun, Eric R. Bachmann, Hyatt Moore, and James Calusdian. Self-contained position tracking of human movement using small inertial/magnetic sensor modules. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 2526–2533, 2007.

[113] Vladimir M. Zatsiorsky and Deborah L. King. An algorithm for determining gravity line location from posturographic recordings. *Journal of biomechanics*, 31(2):161–164, 1997.

[114] Dan Zecha, Moritz Einfalt, Christian Eggert, and Rainer Lienhart. Kinematic pose rectification for performance analysis and retrieval in sports. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1791–1799, 2018.

[115] Yingying Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1089–1097, 2019.

[116] Yuanxing Zhang, Pengyu Zhao, Yushuo Guan, Lin Chen, Kaigui Bian, Lingyang Song, Bin Cui, and Xiaoming Li. Preference-aware

mask for session-based recommendation with bidirectional transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3412–3416, 2020.

[117] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[118] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. 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*, pages 2200–2209, 2020.

[119] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 129(3):703–718, 2021.

[120] Hongyu Zhao, Zhelong Wang, Sen Qiu, Yanming Shen, and Jianjun Wang. Imu-based gait analysis for rehabilitation assessment of patients with gait disorders. In *2017 4th International Conference on Systems and Informatics (ICSAI)*, pages 622–626, 2017.

[121] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *arXiv preprint arXiv:2012.13392*, 2020.

[122] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. HybridFusion: real-time performance capture using a

single depth sensor and sparse IMUs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.

[123] Tobias Zimmermann, Bertram Taetz, and Gabriele Bleser. Imu-to-segment assignment and orientation alignment for the lower body using deep learning. *Sensors*, 18(1):302, 2018.

[124] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021.

Achievements

Journal

1. Tomoya Kaichi, Tsubasa Maruyama, Mitsunori Tada, and Hideo Saito. Learning Sensor Interdependencies for IMU-to-Segment Assignment, *IEEE Access*, 9:116440-116452, 2021.

2. Tomoya Kaichi, Tsubasa Maruyama, Mitsunori Tada, and Hideo Saito. Resolving Position Ambiguity of IMU-Based Human Pose with a Single RGB Camera, *Sensors*, MDPI, 20(19):5453, 2020.

3. Tomoya Kaichi, Shohei Mori, Hideo Saito, Kosuke Takahashi, Dan Mikami, Mariko Isogawa, and Kusachi Yoshinori. Image-based center of mass estimation of the human body via 3D shape and kinematic structure, *Sports Engineering*, 22(3):17, 2019.

Journal (in Japanese)

1. Tomoya Kaichi, Shohei Mori, Hideo Saito, Junichi Sugano, and Hideyuki Adachi. Visual Inspection by Capturing a Rotating Industrial Part, *Journal of the Japan Society for Precision Engineering*, 12(83):1184-1192, 2017.

International Conference

1. Tomoya Kaichi and Yuko Ozasa. A Hyperspectral Approach for Unsupervised Spoof Detection with Intra-sample Distribution, *IEEE International Conference on Image Processing (ICIP)*, pages 839-843, 2021.

2. Tomoya Kaichi, Toshiki Kikuchi, and Yuko Ozasa. One-shot Light Source Searching with Neural Networks, *IIEEJ International Conference on Image Electronics and Visual Computing (IEVC)*, 2019.

3. Tomoya Kaichi, Shohei Mori, Hideo Saito, Kosuke Takahashi, Dan Mikami, Mariko Isogawa, and Hideaki Kimata. Estimation of Center of Mass for Sports Scene Using Weighted Visual Hull, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1809-1815, 2018.

4. Tomoya Kaichi, Shohei Mori, Hideo Saito, Junichi Sugano, and Hideyuki Adachi. Multi-view Surface Inspection Using a Rotating Table, *IS&T International Symposium on Electronic Imaging*, 9:278, 2018.

Award

1. Japan Statistical Society Award (University Student and General Applicants), Statistical Data Analysis Competition, 2018.

2. Student Encouragement Award of the 79th IPSJ National Convention, 2017.

3. First Prize at VRSJ Tracking Competition, 2016.