BodyTrak: Inferring Full-body Poses from Body Silhouettes Using a Miniature Camera on a Wristband

HYUNCHUL LIM, Cornell University, USA
YAXUAN LI, McGill University, Canada
MATTHEW DRESSA, Cornell University, USA
FANG HU, Shanghai Jiao Tong University, China
JAE HOON KIM, Cornell University, USA
RUIDONG ZHANG, Cornell University, USA
CHENG ZHANG, Cornell University, USA

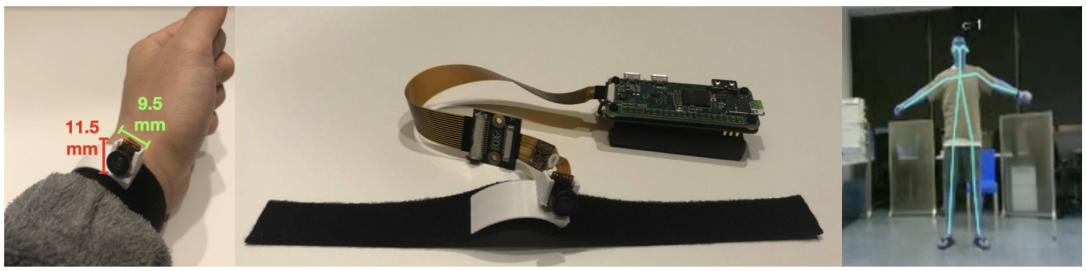

Fig. 1. BodyTrak using a single miniature RGB camera: It is possible to estimate the full body pose using just one small camera with a wide view angle, which is significantly more practical considering the size, weight and battery life. The dimension of this final prototype is 84x36x25 mm with 82g.

In this paper, we present BodyTrak, an intelligent sensing technology that can estimate full body poses on a wristband. It only requires one miniature RGB camera to capture the body silhouettes, which are learned by a customized deep learning model to estimate the 3D positions of 14 joints on arms, legs, torso, and head. We conducted a user study with 9 participants in which each participant performed 12 daily activities such as walking, sitting, or exercising, in varying scenarios (wearing different clothes, outdoors/indoors) with a different number of camera settings on the wrist. The results show that our system can infer the full body pose (3D positions of 14 joints) with an average error of 6.9 cm using only one miniature RGB camera (11.5mm ×9.5mm) on the wrist pointing towards the body. Based on the results, we disscuss the possible application, challenges, and limitations to deploy our system in real-world scenarios.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile devices**.

Authors' addresses: Hyunchul Lim, hl2365o@cornell.edu, hl2365o@cornell.edu, Cornell University, Ithaca, New York, USA, 14850; Yaxuan Li, McGill University, Montreal, Canada, yaxuanli123@gmail.com; Matthew Dressa, Cornell University, Ithaca, USA, mtd67@cornell.edu; Fang Hu, Shanghai Jiao Tong University, Shanghai, China, fanghu777@gmail.com; Jae Hoon Kim, Cornell University, Ithaca, USA, jk2765@cornell.edu; Ruidong Zhang, Cornell University, Ithaca, USA, rz379@cornell.edu; Cheng Zhang, Cornell University, Ithaca, USA, chengzhang@cornell.edu.

2474-9567/2022/9-ART154 $15.00
https://doi.org/10.1145/3552312

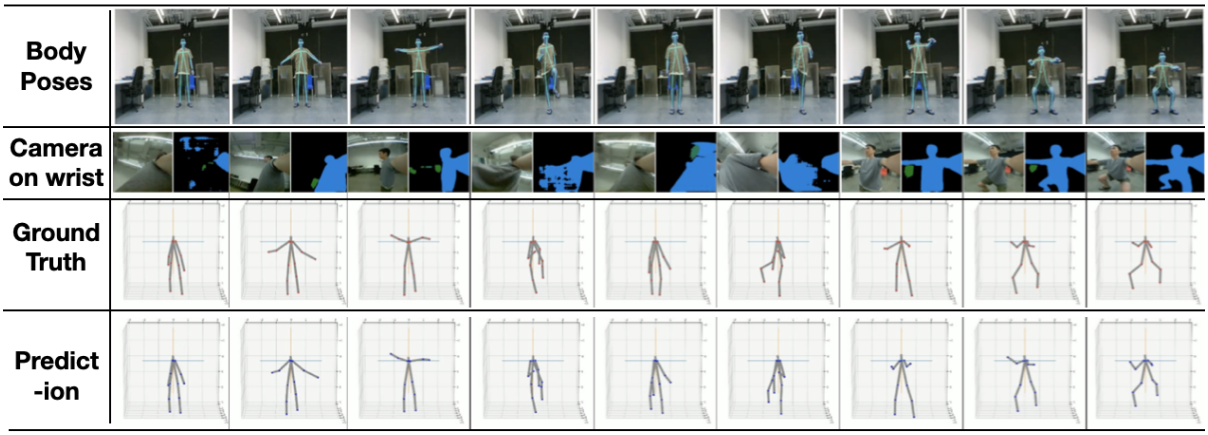

Fig. 2. BodyTrak - body pose estimation using cameras on a wrist. In the first row, the user is performing an activity, and the depth camera is displaying the body posture using skeletal points. The second row of images represents the set of four images captured by a camera on the wristband and their accompanying segmented images. The black in the segmented images is the background and the blue is marked as the user's body. The ground truth rows signify the skeletal figure in correspondence with the depth camera. In the Prediction, the row displays predictive body posture using the images input through the wristband cameras.

Additional Key Words and Phrases: Pose Estimation, Motion Tracking, Wearable Technology, Smart devices

ACM Reference Format:
Hyunchul Lim, Yaxuan Li, Matthew Dressa, Fang Hu, Jae Hoon Kim, Ruidong Zhang, and Cheng Zhang. 2022. BodyTrak: Inferring Full-body Poses from Body Silhouettes Using a Miniature Camera on a Wristband. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 154 (September 2022), 21 pages. https://doi.org/10.1145/3552312

1 INTRODUCTION

Body pose estimation is becoming increasingly important in many fields such as health (e.g., physiology of individuals with physical disorders such as scoliosis and Parkinson's [5][42]), the gaming industry, sports analysis, and even communication studies which can help us understand how we interact with one another through our body language[50] [12] [21]. In order to track the body postures of a person, most of the prior work requires the instrumentation of the environment (e.g., cameras[14], WIFI[27]). However, they may not work well when the user is in motion, outside the indoor setting, or setting up the environment is not feasible.

To address these mobility issues, researchers have developed wearable solutions to estimate body poses. Most of these systems require the users to wear multiple sensors (e.g., IMU) on the body [41, 45], which may not be sustainable or comfortable in real-world settings. The recent advancements have shown the feasibility of using a single wearable form factor, such as a chest-mounted camera[23], a hand-held smartphone[3], a hat-mounted camera [28], or a 360° camera on the wrist[18], to estimate full-body poses. However, these form factors (chest-mount or hat) may not be immediately acceptable or convenient for users to be worn in different daily activities. For instance, chest-mounted devices such as GoPro are acceptable for a group of users in specific contexts. The use of a 360° camera on the wrist [18] can capture full views of the wearer's body for inferring body poses, but it raises privacy concerns by capturing privacy-sensitive, irrelevant information for inferring body poses such as the activity of bystanders and the environment in which the user is situated. Also, the two lenses for 360° cameras may increase the size of the device and the battery consumption, two factors that are important for daily

use. Therefore, in the future eco-system of wearables, it is essential to offer users a variety of wearable sensing technologies to track body poses to decide the technology based on the context.

Compared to other form factors like the chest-mounted device or hat, wrist-mounted wearables is the most popular wearable device on the market. Allowing a wrist-mounted device (e.g., smartwatch) to estimate full body poses can potentially enable a variety of new applications, such as activity recognition, fine-grained exercise tracking, and health sensing. We have already seen some smartwatches embed a miniature RGB camera [1]. Therefore, we develop our research question as:

- *Is it possible to estimate the full body poses using a miniature RGB camera mounted on the wrist?*

To answer this research question, we developed BodyTrak, the AI-powered wrist-mounted sensing technology that can estimate full body poses. It uses cameras on the wrist to capture images of the body. Although these images only contain incomplete body parts (body silhouettes), they are unique and highly informative depending on the arm movements and body poses. Therefore, we derive the *working hypothesis* of BodyTrak: *The incomplete body parts/silhouettes captured by wrist-mounted cameras can be highly informative to infer the full body poses.*

To verify the feasibility of *working hypothesis*, we developed a wristband prototype that can house up to four miniature RGB cameras. This prototype was used to conduct a user study where we had 9 participants perform a list of daily activities or exercises involving a wide variety of body movements. The data was used to identify the optimal camera setting and evaluate its performance. The results showed that BodyTrak could estimate the full-body pose, including the 3D positions of 14 body joints, with an average accuracy of 6.9 cm using one miniature RGB camera (11.5mm × 9.5mm) pointing towards the body. Furthermore, we also conducted additional studies to evaluate how BodyTrak would perform in different real-world scenarios (indoors, outdoors, wearing a different shirt, remounting the device). Based on the results, we discussed the opportunities and challenges of applying BodyTrak in real-world applications.

The contributions of the paper are:

- An intelligent sensing system that can infer 3D positions of 14 body joints (full body poses) from images of *incomplete* body silhouettes captured by a one miniature RGB camera on the wrist.
- A user study with 9 participants to 1) find the optional position for the camera and 2) evaluate the performance of BodyTrak in different activities under different scenarios.
- A discussion on the opportunities and challenges of applying BodyTrak on the future wearables in real-world applications.

2 RELATED WORK

Our work is to estimate full body poses in 3D using a wrist band with cameras. In this section, we first review the literature related to estimating 3D full-body poses using Non-wearable and wearable-based technologies. Then, we discuss the previous research using wrist-mounted cameras to estimate body poses.

2.1 Estimating Full Body Poses using Non-wearable Devices

Applying external devices is one of the most typical methods of capturing body posture with either passive or active sensing technology. Devices with high portability and compact dimensions, such as the Microsoft Kinect [11] and Intel RealSense [10], have proved their effectiveness in motion capture. Due to the advances of the RGBD camera, segmentation of objects from backgrounds and the continuous tracking of body movement can be accomplished by these devices without the use of any markers, which are commonly used for 3D animation[16]. In commercial high-quality motion capture systems such as OptiTrack [24] and Vicon [36], several retroreflective markers are usually placed on the rigid body where external cameras can easily capture them. Similarly, fiducials

[1] https://www.gadgetreview.com/best-smartwatch-camera

Table 1. Comparison with Other Previous work. We note that it is not appropriate to directly compare which system performs better based on accuracy because the evaluation methods are different.

Work	Position of the Sensor	Body Estimation	Sensor	Tracking Moment	Activity	*Accuracy
xR-EgoPose[48]	VR headset	Full body	Fish-eye camera	Always on	9	5.82cm
Mo2cap2[52]	Cap-mounted	Full body	Fish-eye camera	Always on	8	6.14cm
Monoeye[22]	Chest	Full body	Fish-eye camera	Always on	3	5.0cm
You2Me[39]	Chest	Full body	GoPro camera	Always on	4	8.6cm
Pose-on-the-Go [3]	Smart Phone	Full body	Depth camera IMU	When holding the phone	6	<25cm
Real-time arm[33]	Wrist	An upper arm	IMUs	Always on	17	10.53cm for elbow 12.94cm for wrist
I am a smartwatch [44]	Wrist	An upper arm	IMU	Always on	10	9.2cm
Ryosuke et al [18]	Wrist	Full body	A 360° camera	Always on	4	11.5cm
BodyTrak	**Wrist**	**Full body**	**A RGB camera**	**Always on**	**13**	**6.90cm**

are also used as a standard marker in optical measurement [34] while PhaseSpace [26] and VIVE [49] apply active markers. Thanks to the flourishing of computer vision and machine learning, researchers are able to reconstruct human poses simply using RGB images, which is present in works such as HybridTrak [54], DensePose [15], PoseNet [29], and OpenPose [7]. In addition to utilizing cameras, projects utilizing inertial sensors [41], acoustics sensors [2, 13, 53], RF [55] and magnetic fields [25] have also demonstrated impressive performances in human pose estimation.

However, non-wearable devices with either optical or other sensors, such as RF [55] and magnetic field [25], are sometimes not practical or even unavailable in the wild. Systems requiring markers typically involve a specific working space, while other devices like the Kinect requires users to perform in front of the device. These limitations restrict the user's mobility and their applications in real-world scenarios.

2.2 Estimating Full Body Poses Using Wearable Devices

To overcome the constraints of the external non-wearable devices, researchers developed wearable technology to track body poses. One method is to attach multiple Internal measuring units (IMU) to the body to estimate the body postures [41, 45]. The recent work places fish-eye cameras on the hat[52], VR headset[48], or chest [22, 39] to infer the full body poses. More recent work, Post-on-the-Go [3] allows the user to hold a smartphone in hand to estimate the full body poses from a variety of built-in sensors. However, the above form factor such as chest-band, hat, or holding smartphone, may not always be available to users in daily activities.

It is important to provide users with a variety of wearable sensing solutions to track body poses so that they can choose the wearable technology based on the applications or contexts. BodyTrak offers such a new sensing solution to continuously estimate full body poses using a wrist-mounted camera,

2.3 Activity Recognition Using Wrist-mounted Cameras

It is becoming increasingly common to see cameras integrated into commercial wearable watches, which allow for easy communication and the capture of active moments. For example, WristCam, the first-ever camera for Apple Watch, is a wearable, unobtrusive smartwatch with dual cameras [2]. Researchers have explored using wrist-mounted cameras to estimate hand poses[19, 30] or emotional state [40]. The most recent work [18] demonstrated using a 360° camera on the wrist to estimate the full body poses with an accuracy of 11.5 centimeters. The 360°

[2]https://www.wristcam.com/

camera was designed to capture possible entire body silhouette images for full-body pose estimation. However, this approach might bring about privacy concerns since a 360° camera captures privacy-sensitive, irrelevant information for the reconstruction such as the activity of bystanders and the environment in which the user is situated. Also, the 360° camera at the current dimension and size may not be immediately practical to be integrated to wrist-mounted wearables, as it contains two cameras with an ultra-wide lens, consuming more battery. In contrast, BodyTrak shows that the partial silhouette body images captured by a wrist-mounted miniature RGB camera (11.5 mm by 9.5 mm) is enough to estimate the full body poses with an accuracy of 6.90 cm. Although the 360° camera sensors can be very small, the use of one camera rather than two for the 360° camera module inherently makes wrist-mounted hardware much cheaper, lighter, smaller, consuming-less battery and less power to process the data without compromising any performance compared to [18]. We believe that BodyTrak could provide a privacy-sensitive solution approach and the practical implementation to continuously track body poses, considering the optimal position for the camera, device size, and battery consumption.

To facilitate the comparison between BodyTrak and other work, we present Table 1, where we listed the key settings and performance of related wearable-based pose tracking systems. please note that because different projects use different ground-truth acquisition methods, and normalization methods on measuring body skeletons, comparing the performance between projects may not be fair and possible. For instance, BodyTrak uses a depth camera as the ground truth acquisition method, which may introduce larger errors in capturing the body postures compared to other motioncap systems. What we provide in the table is a reference point.

3 THEORY OF OPERATION

Fig. 3. Research Idea. The body silhouettes images captured by the camera on the wrist are unique from each human motion. We believe that these different images are highly informative to infer full-body poses.

Under conventional CV methods for 3D full-body posture, a full-body image is needed to estimate body posture. However, capturing the full body images with wrist-mounted cameras such as [18] brings about privacy concerns since the captured images inevitably include the information of bystanders and the environment where the wearer is situated. Unlike the conventional CV method, we believe that incomplete body parts in the images can also be highly informative to infer the complete body poses, reducing privacy concerns (e.g., limiting information of bystanders and the environments in the view). In other words, how the body parts were occluded provides rich information on the body pose. For instance, intuitively, if certain body parts (e.g., right arms) are not captured in the images, this occlusion already significantly limits the search space for all possible body poses, making an accurate prediction. We observed that the partial body images/silhouettes captured from the wrist are unique on different body poses, as shown in Fig. 3. Furthermore, recent work has already demonstrated that using partial body images/silhouettes to estimate hand poses[20, 51] and facial expressions [8, 9]. Therefore, we believe that it is possible for AI to learn the incomplete body silhouettes captured from the wrist to infer the full body poses on the arm, legs, torso, and head.

4 DESIGN SPACE OF BODY POSES

To verify the idea of BodyTrak, the first step is to consider the design space of body poses so that the design and evaluation of BodyTrak would include the full range of all body movements. Based on human movement, the two main factors we considered in our design space are upper body movement and lower body movement.

- Upper Body Movement: Since our arm moves in various ways, we have divided upper arm movement into two categories: one arm and two arm movements. Arm movement is separated into the left arm or the right arm. This is to account for users wearing the wristband on one side of their arm. In this case, only one arm is moving in the activity, either the right or the left. Also, we divided the two-arm movements into two dimensions: synchronous and asynchronous movements of both arms in which both arms are in motion at the same time or each arm is in motion at a different time.
- Lower Body Movement: We have divided lower body movement into no movement and movements. No movement (noted as Movement (x) in Fig. 3) means that the user is standing. Movement (noted as Movement (O) in fig. 3) refers to the motion of the legs.

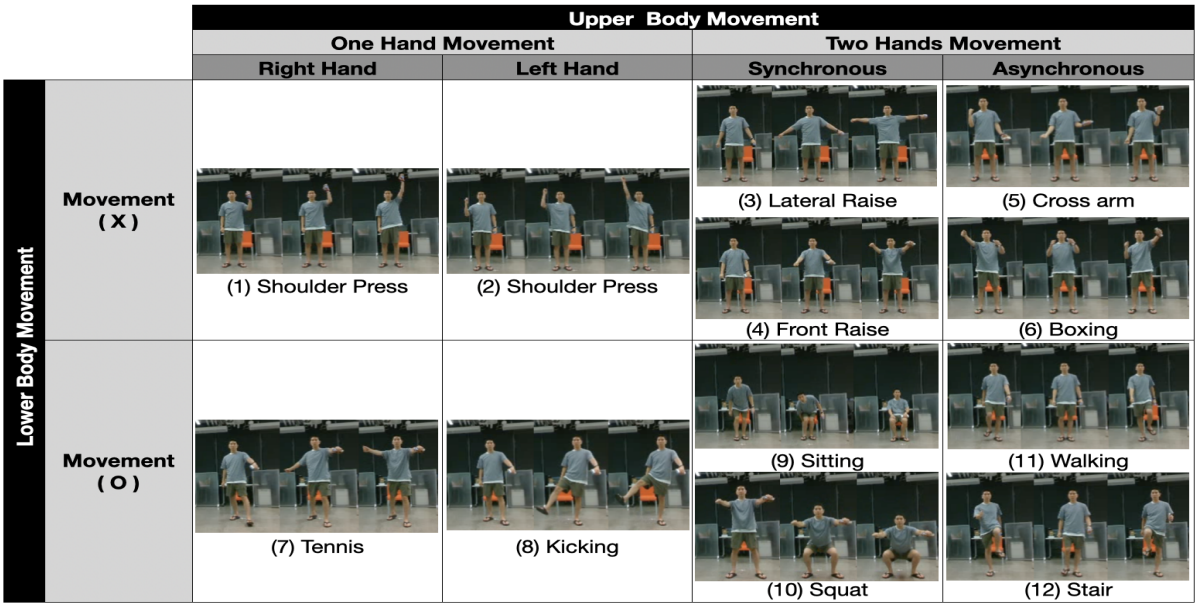

Fig. 4. Design Space. A matrix of 12 body movements that cover the full range of body movements.

Based on these considerations, we created the 4 × two design spaces (see. Figure 4), allowing BodyTrak to explore various body postures. In each category, we found daily activities from previous 3D full-body reconstruction works [23] (walking, boxing, kicking, sitting) and common daily exercises (e.g., squat, front raise). This design space ensures that users perform the full breadth of human motion by completing at least one activity from each category. Finally, we chose 13 activities, including standing to evaluate our system. Standing was included to account for the occlusion gestures as mentioned in section 3 (e.g., Right shoulder press and lateral raise).

5 SYSTEM DESIGN AND IMPLEMENTATION

In this section, we present the design and implementation of BodyTrak, which consists of four parts: the hardware prototype, image segmentation and camera setting, 3D skeleton as the ground-truth for a full-body pose, and the customized deep learning pipeline (see. Fig. 5), as detailed below.

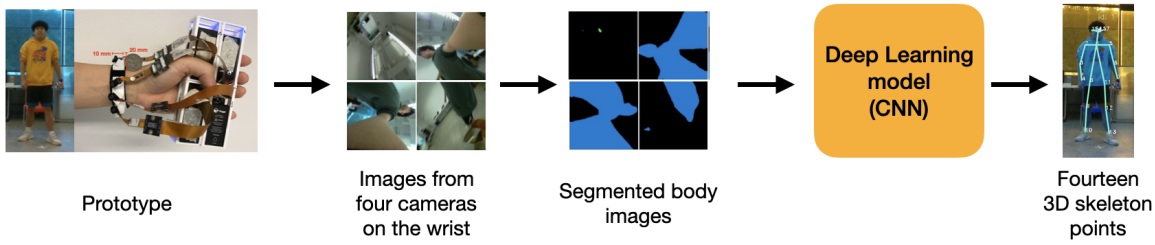

Fig. 5. System Overview of BodyTrak. The prototype is put onto the user. They then perform 13 activities. While the activities are being performed the cameras are capturing images of the body. These images are then segmented and fed into a CNN which outputs an estimated body posture for each activity, using 14 skeleton points.

5.1 Prototype

5.1.1 Hardware. Our hardware prototype consists of three main components, the wristband with miniature RGB cameras, a depth camera for tracking the ground truth of 3D body poses, and a PC for data processing. In order to determine the best setting for miniature cameras on the wrist, we design a 3D printed wristband, which can house up to four miniature RGB cameras (i.e. b006605 RGB Arducam). The camera array is used to collect data to identify the optimized camera setting (e.g., camera positions, combinations). Each camera has a dimension of 11.5mm × 9.5mm and a fixed FOV of 160° and is connected to a Raspberry Pi Zero using cables, as shown in figure 5. An external power bank powers the Raspberry Pi and cameras. These cameras will capture images with the resolution of 400 x 400 at 5 FPS, which are sent and saved on a laptop via WiFi.

The Raspberry Pi Zero and battery were placed in a 3D printed box measured at 168mm × 72mm × 25mm and weighed in at 350g, which are held in the user's hand during the study. We noted that this prototype is designed

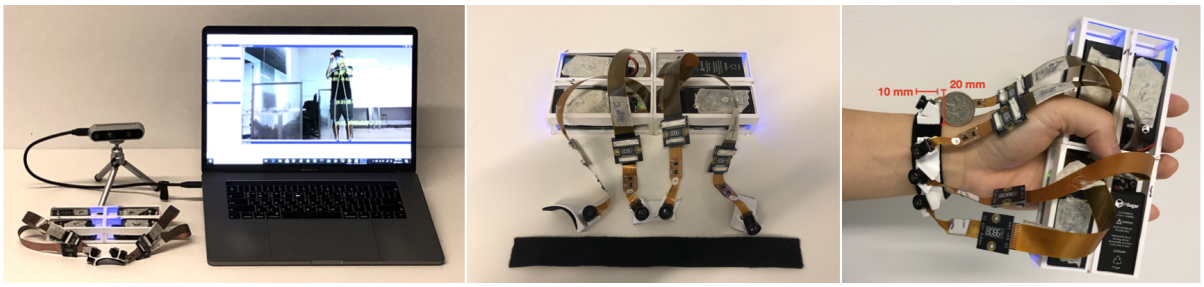

Fig. 6. Experimental Setup. Our hardware prototype consists of three main components, the wristband with 4 miniature RGB cameras, a depth camera for tracking the ground truth of 3D body poses, and a PC for data processing. We note that the goal of the prototype is to find the best position for one RGB camera for evaluation. Please, see our final prototype using one miniature RGB camera in Fig 1.

to find the best position for one miniature RGB camera for evaluation. We asked the participants to hold the data collection device in hand instead of wearing them on the body (e.g., armband) because these devices can occlude the body images observed from the wrist if worn on the body. In order to simulate a real-world scenario where the camera is embedded into a smartwatch, we decided to ask the participants to hold the Raspberry Pi Zero in hand so that the cameras on the wrist would not capture any part of the device. All participants reported that our prototype did not affect their movement during the experiments.

Lastly, we used Intel's RealSense depth camera[3] to record the ground truth of the 3D human body pose. The application, i.e., Skeleton Tracking SDK for Intel® RealSense™ Depth, was employed to get 18 skeleton points on the body, such as the shoulder, elbow, and knees (See Fig. 5). A PC was used for data processing. We built an Ethernet network with high bandwidth between the Raspberry Pi boards and the PC to guarantee transmission robustness and avoid frame dropping. By using timestamps, we synchronized the depth camera images relative to those of the Raspberry Pis.

5.2 Body Segmentation and Camera Setting

5.2.1 Segmentation Technique. The idea of BodyTrak is to use a deep learning model to learn the partial body/silhouette images to estimate 3D full body pose. The critical first step is to segment the body silhouette from the captured images. Many previous machine learning systems have demonstrated reliable performances on this task for human body segmentation [35, 38]. However, it was challenging for existing body segmentation techniques to segment our partial body parts from the background. This is because when using existing techniques, body parts such as the head or torso serve as reference points to infer a body part or position. When conducting segmentation using partial body images, we might lose these fundamental body parts as reference points, as shown in Fig. 3. After applying several body segmentation techniques such as PixelLib [4], Pose2Seg[5], and CDCL [31], we decided to use a well-known pre-trained model named FCN-ResNet101 [35][6] as our image semantic segmentation method, which demonstrated reliable performances in our experiments (e.g., different cloth and indoor/outdoor setting) as shown in Fig 7.

To use FCN-ResNet101, we first normalize the size of our input images to 400×400. Then, we segment the human body from the background using FCN-ResNet101, where we decides whether each pixel belongs to the background or the human body. Although the segmentation was relatively stable, it is not perfect. We found it occasionally miss-segmented body parts. In our experiment, we used all images regardless of the segmentation

[3]https://www.intelrealsense.com/depth-camera-d435/
[4]https://github.com/ayoolaolafenwa/PixelLib
[5]https://github.com/erezposner/Pose2Seg
[6]https://pytorch.org/hub/pytorch_vision_fcn_resnet101/

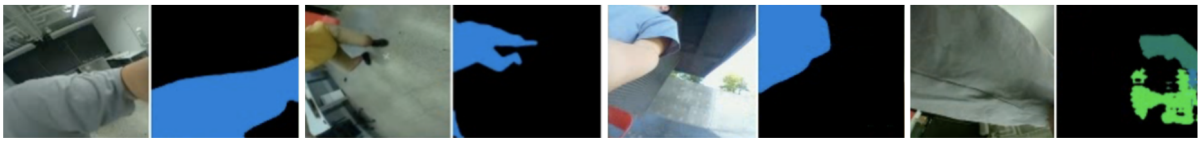

Fig. 7. Body Segmentation. The images provided ensure that we can visually distinguish between properly and poorly segmented images. The blue in the segmented images represent the user's body and the black indicates the background. In image (b) the user is performing a different posture and wearing different clothing. In the outdoor environment (c), segmented images are displayed when the user is performing the activity in an outdoor setting.

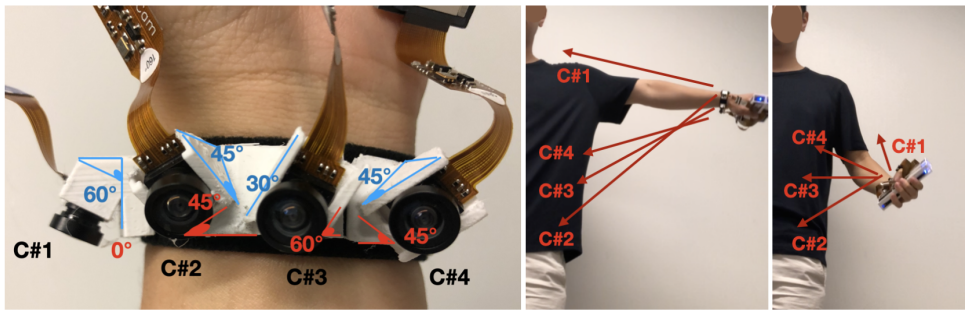

Fig. 8. Camera Setting. In this figure is the camera arrangement, along with their angular measurements. The x plane runs horizontally across the wrist, the y plane runs vertically across the wrist, and the z plane runs away from the wrist. In blue is the rotational degree of the camera on the y-z plane. In blue is rotational degree of the camera on the x-y plane. In these other images the participant demonstrates how the wristband is held and the directions in which the cameras point.

results to train and test the performance on BodyTrak. Therefore, if the body segmentation in the images improves, the performance of BodyTrak can potentially be even better. What we present in this paper is just a baseline.

5.2.2 Camera Setting: Position Arrangement. We place four miniature cameras on the wrist to explore the best camera setting. An optimized camera setting (number, position and orientation) is the key for this system to accurately estimate full body posture. By experimenting with our research team, we aimed to find the best camera setting to capture maximum information on body poses. After conducting a pilot study on segmentation, we determined the camera settings using the following criteria.

- In order to reliably segment the body, cameras are better to be set pointing to the head or torso in combination with other body parts such as arm and leg.
- Lower body images are important to compliment the information we get from the upper body. We need to arrange the cameras to capture this part of the body.

Based on these criteria, we investigated the camera arrangement considering 1). the range of arm movement such as holding up or folding the arm. 2). the natural rotation of the wrist reaching up to 150° [43] and 3) our cameras' FoV is 160°. As a result, the first camera (See. Fig. 8. C#1) is empirically placed on the side of the arm at 60° perpendicular to the arm. This ensures that we can capture more body information, including head pose, when users stretch their arms without folding. The other three cameras (see. Fig. 8 C#2, C#3, and C#4) are positioned side by side on the inner part of the arm considering the rotation of the wrist when the arm is folded. The positions and arrangement were empirically decided based on the preliminary experiments on researchers. Here, we titled and rotated the cameras as seen in Fig. 8 (the tilt of cameras is marked in blue and rotation in red). The purpose of titling the cameras, from 30° to 60°, is to make each camera pointing more towards the body. In addition, all three cameras excepting C#1 were rotated from 45° to 60°, which was helpful to capture the torso with lower/upper body for proper segmentation. In Section 6.3, we will discuss the impact of camera settings on the performance.

5.3 Ground Truth Acquisition

We used the depth camera (Intel RealSense) and associated body Skeleton tracking application to capture the groundtruth of body poses, including 3D positions of 18 body joints on eyes, ears, nose, arms, torso, and legs Fig.9(a). We exclude eyes and ears with only the nose point representing head position. Thus, we have 14 skeleton points as ground-truth for the full body poses. As each person has a different height and body size, we normalized

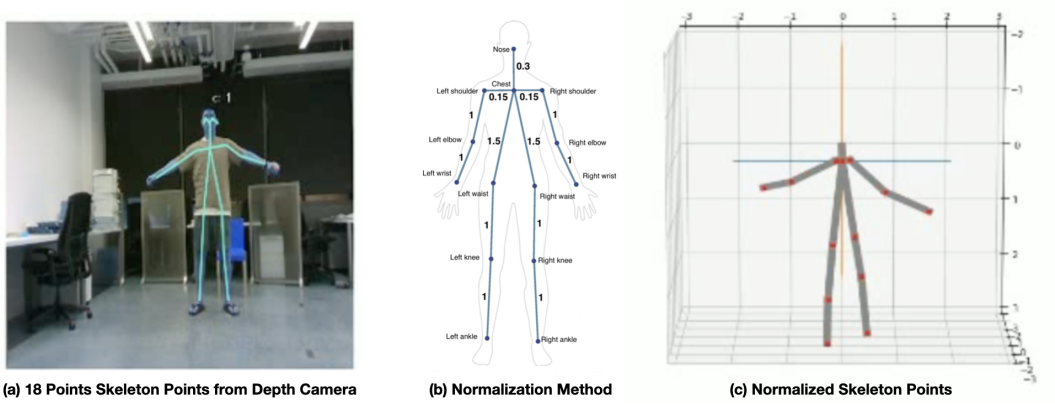

Fig. 9. Ground Truth Acquisition. Image (a) represents the ground truth that is displayed when using the RealSense depth camera. In image, (b) we depict the normalization values for each body part. In figure (c) the image displays the normalized skeleton after it is passed through the deep learning pipeline.

skeleton information before feeding it into a user-independent deep learning model. As shown in Fig. 9(c), we normalized the skeleton so that the shoulder is parallel to the YZ plane and the body center (chest joint) is at the origin of the coordinate. We then normalized the length of the body with values as shown in Fig.9(b). Here we label the coordinate of joint j, the coordinate x, y, z of j after normalization is calculated with the equation:

$$\text{Normalized}_{j,p} = \text{NormLength}_i \times \frac{\text{Raw}_{j,p}}{\text{RawLength}_i} + \text{Normalized}_{j-1,p}$$

where p equal to either x, y, or z, RawLength_i and NormLength_i is the length of the corresponding trunk before and after normalization, and $\text{Normalized}_{j-1,p}$ is the coordinate of the last normalized joint, where we start normalization from Chest. For example, with raw coordinate z of left wrist,

$$\text{Normalized}_{\text{left wrist},z} = 1 \times \frac{\text{Raw}_{\text{left wrist},z}}{\text{RawLength}_{\text{left wrist to left elbow}}} + \text{Normalized}_{\text{left elbow},z}$$

As we discussed in the previous section, we admit that using a depth camera (Intel RealSense) to acquire the ground truth may not be as accurate as using other MotionCap system. A more accurate MotionCap system may potentially futher improve the performance of BodyTrak.

5.4 Deep Learning Pipeline

5.4.1 Network Architecture. Convolutional neural networks have demonstrated promising performance in dealing with 2D image tasks such as classification, retrieval, and segmentation as compared with other traditional machine learning algorithms [17]. Another significant reason for attempting to solve this problem through a deep learning model is that mapping partial body images to locomotion is not straightforward. Directly detecting the connection is even challenging for human eyes. Thus, we hypothesize that a more complex machine learning model, such as deep learning, would be able to identify the hidden connections especially using 2D images as input data. We developed our model with four branches for four image streams captured by each camera. Each branch consists of four 3×3 convolutional blocks followed with 2×2 Max Pooling layer as shown in Fig.10. We then conduct late fusion by concatenating the output of fully connected layer in each branch and add final fully connected layer to obtain the estimation of 14 joint coordinates.

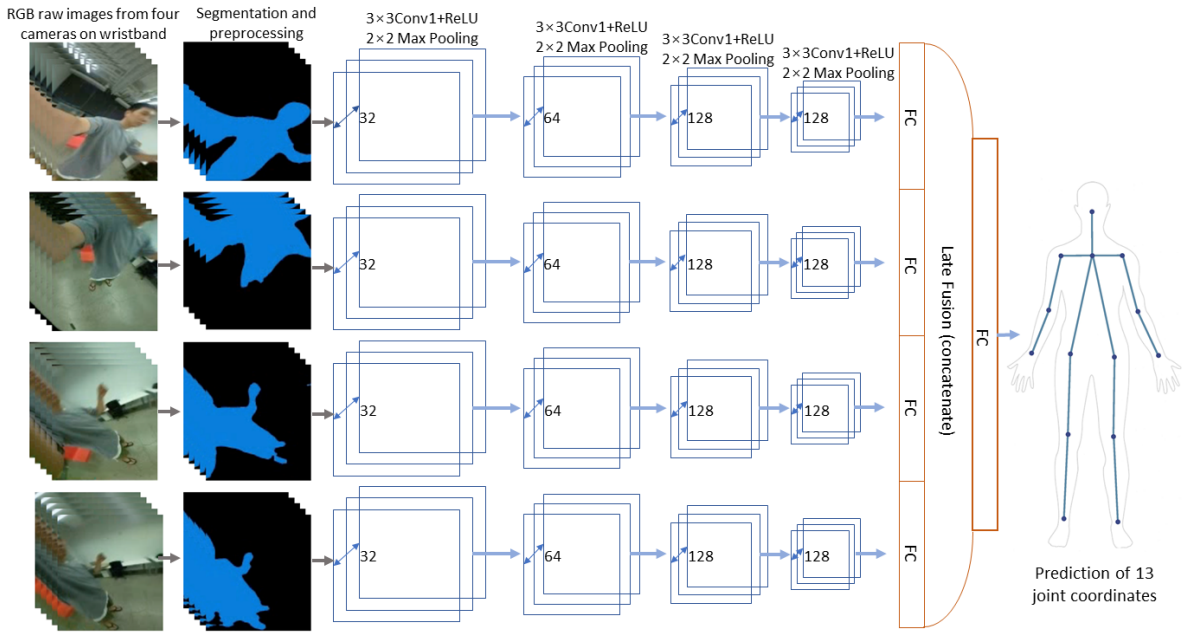

Fig. 10. Deep Learning Model Structure

5.4.2 Model Training. Using the Adam loss function, our model is trained to predict 42 parameters (i.e., 14 joint points × 3 coordinates (x, y, and z)). The training was stopped when a monitored loss had not improved using ten patience. In addition, we reduced the learning rate (factor=0.2, patience=5, _lr=0.001) using five patience when the loss has stopped improving, which works well when learning stagnates. We only kept the model that has achieved the best performance before stopping. Our model is trained for an average 82.4 epochs (SD = 21.4) on different training sessions for all experiments.

6 EVALUATION

6.1 Procedure

6.1.1 Participant. We recruited 9 participants (6 Male, Mean = 26.33, SD =5.39) from the university campus to evaluate the system. All participants were right-handed and wore our hardware prototype on their left wrist, which follows the tradition that people wear smartwatches on the non-dominant hand.

6.1.2 User Study Procedure. Before data collection started, participants were asked to watch a tutorial video that consisted of the researcher performing all 12 activities as shown in our design space (See figure 4). This helped participants to get familiar with the activities they needed to perform. At the end of the training video, the researcher handed the participant the 3D printed box filled with the raspberry pis and power sources and wrapped the Velcro portion of the device on the left hand. After mounting the prototype on the participant, they were instructed to follow the video instructions on the monitor to perform the 12 activities in four sessions.

6.1.3 Data Collection. We have four sessions for data collection. In each session, participants were asked to perform 12 activities. We intentionally asked the participants to slowly perform these activities such that the depth camera could capture the ground truth reliably. In the first session, the participants repeated each activity

Table 2. Results of Within Session Evaluation. (cm for Mean Per Join Position Error (MPJPE) and percentage in parentheses for the 3D percentage of correct keypoints (PCK)) R: Right, L: Left; Ch: Chest, S: Shoulder, E: Elbow, Wr: Wrist, Wa: Waist, K: knee, A: Ankle.

Moition Type	Avg	Nose	Ch	RS	RE	RWr	LS	LE	LWr	RWa	RK	RA	LWa	LK	LA
Average	6.34 (84.2)	1.95 (100)	0.00 (100)	1.12 (100)	9.44 (74.7)	14.9 (58.3)	1.18 (100)	9.16 (75.5)	13.5 (63.5)	5.36 (88.3)	5.36 (83.9)	6.87 (82.6)	5.65 (86.4)	6.41 (83.6)	6.71 (83.0)
Standing	4.21 (88.9)	1.37 (100)	0.00 (100)	0.74 (100)	7.00 (77.9)	10.12 (66.8)	0.91 (100)	5.43 (83.9)	6.18 (78.9)	4.12 (89.8)	4.05 (90.9)	4.49 (89.3)	4.04 (92.4)	4.44 (87.8)	4.55 (86.7)
Right Shoulder Press	4.03 (87.4)	1.36 (100)	0.00 (100)	0.78 (100)	9.51 (68.9)	12.91 (57.8)	0.84 (100)	6.12 (79.4)	7.11 (77.5)	4.17 (90.3)	4.33 (88.8)	4.47 (89.0)	3.21 (93.9)	4.16 (88.8)	4.11 (90.1)
Left Shoulder Press	4.26 (89.2)	1.60 (99.9)	0.00 (100)	0.89 (100)	5.61 (84.9)	8.12 (73.8)	0.83 (100.0)	7.01 (77.6)	10.45 (66.5)	3.88 (94.5)	4.43 (89.7)	4.44 (89.9)	3.85 (92.7)	4.28 (89.5)	4.38 (89.8)
Lateral Raise	4.59 (87.2)	1.65 (100)	0.00 (100)	0.83 (100)	7.06 (78.5)	11.17 (64.7)	1.05 (100)	7.64 (76.4)	11.50 (64.4)	4.15 (89.7)	4.71 (89.4)	4.86 (86.5)	4.09 (89.9)	4.57 (88.2)	4.96 (93.8)
Front Raise	4.76 (84.5)	1.57 (100)	0.00 (100)	0.98 (100)	7.20 (73.7)	12.80 (58.5)	1.03 (100)	8.08 (70.8)	12.39 (58.2)	4.01 (90.2)	4.75 (88.0)	5.01 (85.6)	4.21 (88.8)	5.11 (85.4)	5.02 (84.5)
Cross Arm	4.34 (86.4)	1.36 (100)	0.00 (100)	0.81 (100)	6.95 (77.4)	14.53 (54.4)	0.90 (100)	6.85 (78.3)	11.47 (61.4)	3.66 (93.4)	4.38 (90.7)	4.61 (88.0)	4.24 (89.8)	4.54 (89.4)	4.78 (87.4)
Boxing	5.26 (81.4)	1.76 (99.9)	0.00 (100)	1.20 (100)	11.29 (64.4)	15.41 (45.7)	1.07 (100)	10.31 (66.8)	12.66 (59.2)	4.85 (85.4)	5.39 (81.8)	5.64 (80.7)	5.08 (86.3)	5.01 (86.7)	5.47 (83.4)
Tennis	6.28 (80.3)	1.77 (99.9)	0.00 (100)	1.11 (100)	10.04 (67.7)	14.68 (53.4)	1.03 (100)	8.80 (71.7)	13.84 (55.8)	5.20 (85.7)	6.51 (78.8)	6.61 (79.1)	5.62 (79.7)	6.58 (78.1)	7.10 (76.4)
Kicking	6.76 (80.7)	1.80 (100)	0.00 (100)	1.04 (100)	7.19 (76.8)	12.55 (58.7)	1.05 (100)	8.19 (74.1)	12.94 (58.5)	5.63 (83.7)	7.84 (75.2)	9.59 (69.9)	6.14 (78.0)	6.67 (78.7)	7.53 (76.2)
Sitting	4.74 (88.2)	1.69 (100)	0.00 (100)	0.86 (99.9)	5.29 (84.3)	7.41 (73.0)	0.76 (100)	5.23 (83.8)	6.87 (77.7)	3.90 (94.5)	5.72 (81.8)	5.33 (84.6)	4.63 (87.8)	5.55 (82.8)	5.13 (85.4)
Squat	6.12 (79.3)	1.76 (100)	0.00 (100)	1.02 (99.9)	9.18 (70.5)	18.70 (35.7)	1.10 (100)	7.63 (74.9)	15.16 (46.8)	4.90 (86.3)	6.08 (80.5)	6.45 (78.8)	5.68 (80.9)	7.03 (77.9)	6.65 (78.5)
Walking	5.94 (81.8)	1.70 (99.9)	0.00 (100)	1.02 (100)	8.06 (73.8)	12.25 (58.7)	1.02 (100)	8.01 (74.4)	11.37 (63.9)	4.89 (80.4)	5.78 (80.8)	6.41 (79.1)	5.62 (83.5)	6.26 (78.9)	7.15 (73.9)
Stepping a Stair	7.19 (79.9)	1.86 (100)	0.00 (100)	1.03 (100)	8.59 (72.8)	13.43 (56.9)	1.24 (100)	9.20 (70.0)	13.35 (57.5)	5.30 (84.8)	7.56 (74.8)	7.91 (73.7)	6.03 (79.5)	7.46 (75.4)	7.75 (74.0)

ten times. In the later three sessions, the participants repeated each activity three times. All activities were randomly ordered.

The first session evaluated the system performance within the same session where the device was not remounted. In the second session, we asked the participant to remount the wristband to evaluate how the performance would be impacted after remounting. In the third session, the participants were asked to wear a different shirt to evaluate how would BodyTrak perform if the cloth was different. In the fourth session, the participants performed the activities in an outdoor setting to investigate how our system operates in outdoor environments where lighting and background vary.

6.2 Result

In this section, we reported the performance of BodyTrak in different settings, including without remounting, after remounting, after changing the cloth, and outdoor environments. We used the images captured from all four cameras in the following experiments. Also, we report the performance by using all possible combinations of cameras to propose the best camera setting.

6.2.1 Evaluation Matrix. We calculated the performance, i.e., accuracy, by the following two ways: 1) Mean Per Joint Position Error (MPJPE) and 2) the percentage of correct keypoints (PCK) accuracy. MPJEP is widely

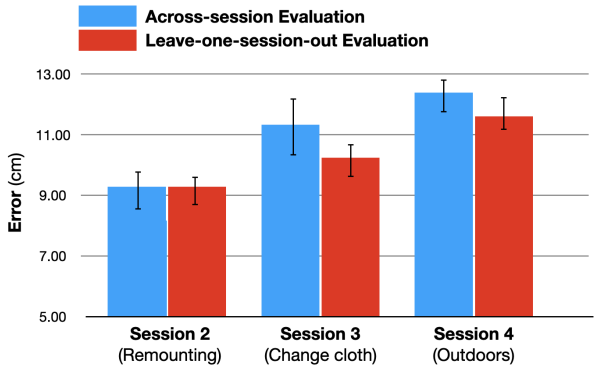

Fig. 11. Comparison on Performance between Across-session and Leave-one-session-out. In this figure, the blue bars indicate the error when data from the Across-session evaluation was included. The red bars indicate the error from the Leave-one-session-out evaluation. As demonstrated in this figure by including more data (i.e. Leave-one-session-out evaluation), our performance improved.

used in 3D pose estimation while PCK is more expressive and robust than MPJPE, revealing individual joint mispredictions more strongly [37]. MPJPE is calculated by taking the Euclidean distance between the predicted 3D joints and the true value of the 3D joints from the ground truth. Then it is calculated by averaging over the sequence. This distance is then scaled to centimeters (cm) based on each participant's arm (Mean = 22.78 cm, SD = 1.4 cm). Also, we calculate 3D PCK performance by picking a threshold of 150mm as suggested in [37]. It measures if the Euclidean distance between predicted 3D joints and the true 3D joint values from ground truth is within 150mm. We report MPJPE and 3D PCK over all 14 body joints and 12 poses. For reporting the accuracy of each pose, a sequence was manually divided.

6.2.2 Within Session Evaluation. At first, we conducted a within-session evaluation on BodyTrak using the data from the first session as the training and testing data. Specifically, we used the first 8 instances of the 12 activities (based on chronicle order) as the training data and the last 2 instances as the testing data. Please note that the collected data was not shuffled when splitting the training and testing set. The results show that BodyTrak achieved an average of 6.34 cm (SD = 0.61 cm) or 84.2 % 3DPCK (SD = 1.87%), indicating that the body silhouettes captured using wrist-mounted cameras are informative to estimate body pose as shown in Fig. 2.

Similar to other works [22, 39], the joints on the wrist and elbow showed the worst performance among all 14 body joints. This is not surprising because the wrist and elbow have the largest moving distance compared to the body joints on other parts. Furthermore, as we expected, the error in the positions of body joints on the right arm and wrist was larger than on the left. Because the device was worn on the left, which naturally would capture more information about the left half of the body. The detailed results were presented in Table 2.

6.2.3 Across-session Evaluation. In the second experiment, we evaluated how would BodyTrak perform when the environment contexts changed, including remounting the device, changing the cloth, and outdoor environments.

In this experiment, the model was the same as the one used in the within-session experiment, where the training data set came from the first eight instances in the first session. The testing data came from the second (remounting), third (change cloth), and fourth session(outdoor), respectively. The results showed that Bodytrak achieved a average accuracy of 9.31 cm (SD = 1.07 cm), 11.16 cm (SD = 1.03 cm), and 12.33 cm (SD = 1.27 cm), respectively in these three testing sessions. For 3DPCK performance, it achieved 71.3% (SD = 1.97%), 66.2% (SD

Table 3. Results of User Study under different conditions. (cm for Mean Per Join Position Error (MPJPE) and percentage in parentheses for the 3D percentage of correct keypoints (PCK))

Evaluation Type	**Across-session**			**Leave-one-session-out**		
Testing Session	Session 2	Session 3	Session 4	Session 2	Session 3	Session 4
Avg.	9.31 (71.3)	11.14 (66.2)	12.33 (63.4)	9.30 (72.4)	10.12 (69.0)	11.33 (64.8)
Nose	2.62 (100)	2.96 (100)	3.70 (99.8)	2.78 (100)	2.96 (100)	3.28 (99.9)
Chest	0.00 (100)	0.00 (100)	0.00 (100)	0.00 (100)	0.02 (99.9)	0.00 (100)
Right Shoulder	1.39 (100)	1.66 (100)	2.21 (99.9)	1.57 (100)	1.75 (100)	2.17 (99.9)
Right Elbow	13.05 (62.8)	15.93 (52.2)	16.93 (47.5)	13.24 (62.0)	14.20 (58.9)	15.89 (45.9)
Right Wrist	23.03 (8.5)	27.03 (4.31)	28.94 (6.45)	22.61 (10.2)	24.03 (7.1)	26.94 (8.2)
Left Shoulder	1.64 (100)	1.75 (100)	2.23 (99.9)	1.75 (100)	1.77 (100)	2.08 (100)
Left Elbow	12.59 (63.2)	15.18 (54.9)	16.88 (48.2)	12.57 (62.8)	13.79 (52.4)	15.19 (51.2)
Left Wrist	20.00 (16.2)	24.67 (7.8)	27.37 (5.4)	18.83 (39.2)	20.84 (15.4)	23.85 (7.9)
Right Waist	7.63 (80.2)	8.35 (78.9)	9.99 (72.1)	8.00 (78.0)	8.29 (77.6)	9.33 (73.4)
Right Knee	9.64 (73.3)	11.26 (68.8)	12.25 (63.4)	9.50 (74.2)	10.41 (69.8)	10.92 (69.3)
Right Ankle	10.48 (70.1)	12.73 (62.3)	14.29 (58.1)	10.25 (70.8)	11.55 (66.2)	13.01 (61.8)
Left Waist	8.57 (78.1)	10.72 (69.1)	11.99 (62.1)	9.07 (74.3)	9.60 (73.1)	11.26 (64.1)
Left Knee	9.29 (76.1)	11.43 (66.8)	12.41 (62.8)	9.56 (71.9)	10.68 (63.3)	11.82 (64.7)
Left Ankle	10.26 (69.9)	12.63 (61.1)	13.37 (61.8)	10.43 (70.3)	11.82 (66.1)	12.88 (60.9)

= 2.21%), and 63.4% (SD = 2.10%), respectively in the three testing sessions. Although the overall accuracy was worse than within-session performance, it indicates the potential of applying BodyTrak in real-world settings.

6.2.4 Leave-one-session-out Evaluation. One possibility on why the performance was worse across sessions is that the training data was too small or limited. As a result, the model has not seen enough variance of the data in different settings. In order to investigate this issue, we conducted the third experiment: Leave-one-session-out. We used 3 sessions as the training data, and one session as the testing data. This process was repeated four times so that each session was used as a testing session once. The results showed that BodyTrak achieved an average accuracy of 9.30 cm (SD = 0.75 cm), 10.12 cm (SD = 1.17 cm) and 11.33 cm (SD = 1.11 cm), when session 2 (Remounting), for session 3 (Cloth) and session 4 (Outdoors) were used as testing sessions respectively. For 3DPCK performance, it achieved 74.5% (SD = 2.23%), 68.9% (SD = 2.17%), and 63.1% (SD = 1.93%), respectively in the three sessions mentioned above. Compared to the second experiment, leave-one-session-out improved the performance by around 1 cm in each session (See. Fig 11). Also, leave-one-session-out leads to 1.1% 2.8% 3DPCK improvement on each over across session. It indicates that if larger and more diverse training data can be collected, the performance of BodyTrak can be further improved especially when it was used in different settings. Table 3 summarized the result in detail.

6.3 Comparison of Different Camera Settings

Here, we would like to explore whether fewer cameras can solve the problem in consideration of the real-world application, size, and battery capacity. Thus, we further analyzed how the number of cameras has affected the performance. We used all possible combinations of cameras to conduct the within-session experiment and across-session experiments. The results (detailed in Table 4) showed that camera #1 is the most informative camera. If only using the data from camera #1, the performance of BodyTrak can still achieved an average accuracy of 6.9 cm (82.1 % 3DPCK) in within-session experiment, compared to 6.3 cm (84.2 % 3DPCK) using all

Table 4. The Impact of Camera Settings (cm for Mean Per Join Position Error (MPJPE) and percentage in parentheses for the 3D percentage of correct keypoints (PCK))

Evaluation Type		**Within Session**	**Accross Session**		
Setting	**Camera**	**Session 1**	**Session 2**	**Session 3**	**Session 4**
One Camera	c#1	6.90 (82.1)	11.74 (64.2)	12.92 (60.1)	13.49 (62.8)
	c#2	6.89 (83.3)	13.37 (61.2)	13.50 (61.4)	14.53 (58.0)
	c#3	7.25 (79.9)	12.91 (62.1)	13.41 (61.6)	15.11 (55.2)
	c#4	7.41 (78.2)	14.05 (60.8)	14.42 (59.1)	14.42 (59.9)
Two Cameras	c#1, c#2	6.69 (82.0)	12.25 (63.7)	13.03 (62.1)	13.80 (60.3)
	c#1, c#3	6.82 (82.7)	10.68 (71.3)	11.98 (64.5)	14.27 (60.3)
	c#1, c#4	6.99 (81.7)	12.65 (63.3)	13.03 (62.2)	14.27 (60.2)
	c#2, c#3	6.79 (82.2)	10.84 (70.1)	13.05 (62.3)	13.05 (62.2)
	c#2,c#4	6.98 (80.7)	11.49 (68.5)	13.48 (61.9)	13.48 (61.0)
	c#3,c#4	7.25 (78.8)	13.90 (60.2)	14.69 (57.9)	14.69 (56.9)
Three Cameras	c#1,c#2,c#3	6.46 (83.8)	11.79 (63.8)	12.80 (60.1)	13.64 (68.0)
	c#1,c#2,c#4	6.51 (82.7)	10.51 (71.0)	12.12 (63.7)	12.94 (62.8)
	c#1,c#3,c#4	6.63 (82.3)	10.02 (72.8)	11.59 (64.2)	12.71 (62.1)
	c#2,c#3,c#4	6.63 (82.5)	13.17 (62.8)	13.71 (61.0)	14.10 (60.1)
Four Cameras	c#1,c#2,c#3,c#4	6.34 (84.2)	9.30 (74.5)	11.16 (68.9)	12.33 (63.1)

four cameras. Our interpretation is that camera 1 with a wide view angle, may cover on average more information on the upper body and lower body.

6.4 An Additional User Study with the Final Hardware Prototype

This result showed that using one camera can estimate the full body poses with an accuracy of 6.9 cm. However, the prototype used in the first study was designed to house 4 cameras, which helped identify the best camera position #1.

Based on the optimal camera position identified in the study, we further design another hardware prototype that only house one camera at position #1, as shown in 1. The camera is connected to a Raspberry Pi zero for data collection. In order to verify that the final prototype provides a similar performance as the original one, we conducted a follow-up study with four participants. The study procedure, evaluation protocol, and algorithms are similar to the previous study. The result showed that using the final prototype, BodyTrak achieved 6.96 cm in within-session experiment and 11.34cm (SD = 0.87 cm), 13.12cm (SD= 1.28 cm), 13.51cm (SD = 0.98 cm) for session 2, 3 ,4 respectively.For 3DPCK performance, it achieved 81.7% (SD = 0.96%), 64.9% (SD = 1.01%), 61.2% (SD = 0.77%), and 62.9% (SD = 0.81%), respectively in the four sessions mentioned above. The performance is similar to the result from the first prototype that housed four cameras. It demonstrated that using one miniature RGB camera on the wrist can estimate the full body poses. As we have discussed previously, many existing commodity smartwatches already have a built-in camera. By slightly changing the position and orientation of the camera, these built-in cameras can potentially be repurposed to track full-body poses on a commodity smartwatch.

7 DISCUSSIONS

7.1 Privacy Issue

A wearable camera will always raise concerns about privacy. As BodyTrak does not intend to capture full views of the body, we think BodyTrak has a clear advantage in protecting the privacy of the user over [18]. A 360-degree

camera can capture more information than a miniature RGB camera used in our system. However, the use of a 360-degree camera raises privacy concerns, capturing not only privacy-sensitive but also irrelevant information such as the activity of bystanders and the environment in which the wearer is situated. For example, the frequency with which bystanders around the wearer are seen and the amount of sensitive information might be much more than those of BodyTrak. Inherently, a 360-degree camera captures more surrounding information by about 66% compared to a 160-degree camera. In comparison, our system significantly mitigates privacy concerns in two ways. First, our system is designed to point the camera only towards the wearer's body, limiting the background environment in the view. Second, our system does not need to capture the full body, limiting the wearer's body in the view as well. However, our system might still include the bystanders' information in a few scenarios, e.g., when the bystanders are next to the wearer within about 1m or when the wearer is raising the wrist over the elbow position. However, these scenarios would not happen frequently in daily activities. Also, to address this issue, one solution is to extract features on the fly. Instead of saving the raw RGB images, we can only segment the images on the device (e.g., smartphone), containing only the body silhouettes. In this case, even if the deep learning model is trained or running on the cloud, transmitting the segmented images would not contain much of the sensitive private information of the user.

7.2 Body Segmentation

For body image segmentation, we only use the pre-trained version of the segmentation model, FCN-ResNet101. However, this model might not work well with images that did not include a part of a human body, e.g., head and leg. However, the images captured by the 160-degree camera can include the combination of the parts of a head, arm, or legs, which allows the pre-trained model to succeed in segmenting the partial body. We found that the key to reliable segmentation by FCN-ResNet101 is to have the user's head or torso be sufficiently framed in the camera. Although the segmentation was relatively stable, it is not perfect. One possible way to improve the performance of the segmentation is to train the model with a new dataset. For example, we can collect the data with participants wearing the device in a clear background (e.g. green chroma key) and then synthesize the data with the different background images. Also, like [18, 22], a large-scale synthetic dataset can be rendered by humanoid models having a virtual camera on the wrist.

7.3 Performance Comparison

We believe that it is challenging to directly compare which system performs better based on accuracy because the system design principle, evaluation methods, and user study procedures are different (e.g., different input images, DL model design, network training, and training dataset). Multiple factors in the system design may contribute to better performance compared to [18]. We assume that one factor is input images for DL models. The input images for 3d body pose estimation come from different angles. [18] placed the 360 camera on the wrist closer to the back of the hand. It captures a lot of surrounding information with 360 camera, which may increase the difficulty of segmenting the human body from the background. On the other hand, our captured images are largely about the torso and head, which may be easier for segmentation. Furthermore, the complexity of the models might contribute to the difference in performance as well. Compared to [18], which uses a more completed model (ResNet + BiLSTM + MLP), we developed a relatively simple CNN model to infer 3d body poses. Since the body silhouette image is binary, we chose a less complex model. In practice, we observed that our simpler model may help the model to be more generalizable and reduce the risk of potential over-fitting, given relatively smaller training datasets.

7.4 User-independent Model

The experiment above was all conducted using user-dependent models, where the training and testing data came from the same participant. This means, a new user has to provide training data before using the technology, which may not be preferred from the user experience perspective. Therefore, we conducted one more experiment to evaluate how would BodyTrak was trained and tested using user-dependent models. Thus, we conducted a leave-one-participant-out evaluation, where we used all data from 8 participants to train the model, and then used sessions 2,3,4 as the testing data respectively. This process was repeated 9 times so that each participant's data was used as the testing data once. The results showed that the average error of estimating 14 body joints positions were 11.37cm (SD = 1.47 cm), 11.70cm (SD = 2.11 cm), 13.10cm (SD = 1.58 cm) for session 2, 3 ,4 respectively. The performance was worse compared to the results from user-dependent models, which is expected. Because participants' body shape, cloth, and hair, are likely different from each other, which leads to different body silhouettes even if the body pose is the same. However, the performance is also encouraging, which indicates the possibility of building a user-independent model in the future, especially if significant larger training data with more participants can be collected. Another possibility of addressing this issue is to generate synthetic training data to simulate different body shapes, cloths, hairs, and backgrounds, as demonstrated in [22]. This would greatly improve the diversity in the training data without the need of collecting data from participants. We will leave this to future investigation.

7.5 Real World Application

With the development of commercial smartwatches with cameras such as WristCam[7], it is possible that our system can be integrated into a commercial smartwatch in the future. In this section, we discuss several concerns regarding the deployment of our system in real life. These topics are divided into the future areas of application for BodyTrak and the hardware considerations for deploying this system.

7.5.1 Areas of Application. By using 3D body reconstruction in a smartwatch we can get important information about human behavior. Notably, this technology can make an improvement in recognizing detailed daily activities using estimated body poses e.g., stepping on stairs, walking, running, or different types of exercises like tennis, boxing, and so on [6, 46]. By maintaining a record of these detailed daily activities with BodyTrak, it could provide workout progress and offer suggestions for improving the wearer's health. For AR/VR applications, estimated body poses could contribute to making immersive VR environments. For example, the body poses can be used for digital characters in the virtual environment to recognize and automatically reproduce the same pose as a person who wears BodyTrak.

7.5.2 Hardware Considerations. Overall, BodyTrak shows reliable performance using a one miniature camera on the wrist. we believe that our proposed system using one miniature camera is more practical than using a 360-degree camera [18]. There seems no 360-degree camera available on any wearable wrist-mounted device. If we place any similar 360-degree camera on a wristband as [18] or the above smartphones, the wristband would become heavy and impossible to be worn in daily activities. Also, Laying the sensor flat on the skin will automatically lose half of the view angle because the skin will block it. Putting another camera used by the Protruly[8] on the other side will increase the view angle. However, this is not necessary for improving performance and will only increase device size, cost, and power consumption. Even if eventually, technology revolution would allow cameras to be miniaturized and battery capacity is no longer an issue on a smartwatch, our system still

[7] https://www.wristcam.com/
[8] https://www.gsmchoice.com/en/catalogue/protruly/darlingvrd7/

provides a feasible solution before that days arrive, which we believe will not be soon. Instead, it has been widely demonstrated that one miniature RGB camera can be added to a regular smartwatch [9].

Besides the weight and size, our system of using one camera also shows the advantage of energy consumption and requires less power to process the data without compromising any performance (if not better). Nevertheless, one important challenge of deploying BodyTrak on a commodity smartwatch is how to integrate the hardware into the smartwatch in a battery-sustainable manner. In our prototype, we implement the system with Raspberry Pi Zero and RGB cameras. The no-load power consumption of Raspberry Pi Zero is about 80 mA (0.4W) and 140 mA for the camera module in video mode measured by PowerJive USB Power Meter [4]. With further optimization of the power consumption by using a customized board, the power consumption is more manageable, especially if only using one camera.

In terms of computational burden, in our study, we recorded the images and then processed all the data offline on a workstation. However, moving forward, the implementation of our system can be altered by applying one of the following settings. The first option is to transmit all data to a cloud or remote server using wireless communication methods such as WiFi. The major issue in this setting is the data transmission speed and computation power of cloud/remote computing. The second option is to deploy the machine learning model on the edge device such as a smartphone. With the advancement of smartwatch processors and GPUs, faster predictions on edge devices is possible in the future.

7.6 Limitations and Future Work

BodyTrak has demonstrated the feasibility of 3D full-body reconstruction using a wristband, showing a good performance over varying scenarios. However, this is just the first step toward a smartwatch that can track full-body poses in free-living conditions. There is apparent room for improvement.

First, our training data set is relatively small. If we include a larger dataset with diverse users and activities, the performance of the system is likely to be improved, especially across sessions, and in user-independent models.

Second, the current model is a CNN. Given the body, posture contains rich temporal information. A time-series deep learning model, i.e., LSTM, can potentially further improve the performance.

Third, if we add IMU data on the wristband (also widely available on smartwatches), using fusing the data from IMU and cameras, can potentially further improve the estimation performance.

Fourth, the camera arrangement for the study to find the optimal position is empirically decided. To systemically explore the position, the post-hoc study by cropping the images from a larger view in [18], is feasible. However, this type of simulation does not always provide accurate estimation compared to data collected in the real world. For instance, [18] captured 360 degree images on the wrist. However, the camera setting and position would result in variance in the images between cropped images compared to true images captured with a single RGB camera.

Fifth, the ultimate goal of our research is to understand human behavior based on human action recognition (HAR) for various fields such as health and augmented reality. In doing this, we believe that inferred 3D skeleton values from BodyTrak could be utilized for HAR with recent machine/deep learning techniques. Recent work [1] has developed HAR models using Kinematics Posture Feature (KPF) extraction from 3D joint positions based on skeleton data for action recognition. [32, 47] also employ deep convolutional neural network (DCNN) models, such as ResNet18 and MobileNetV2, for skeleton-based posture recognition. Although The classification on 12 poses from our design is beyond the scope of this paper, we will explore this for future work.

Lastly, the ground truth acquired by RealSense depth camera was relatively stable. It still includes errors and noise occasionally. If we have a more accurate and reliable ground truth (MotionCap system), the performance can potentially be improved too. Besides, using a depth camera as the ground truth acquisition method also

[9] https://www.gadgetreview.com/best-smartwatch-camera

limited the activities that the participants can perform. For instance, the participants can not actually walk or run in large distances. Otherwise, the depth camera would not capture the full body images, which would lead to inaccurate estimation of body poses. Therefore, one important next step is to use more accurate ground-truth acquisition methods to improve the performance of the system under more challenging scenarios and contexts.

8 CONCLUSION

This paper presents BodyTrak, an intelligent sensing method that can estimate full body poses including the 3D positions of 14 body joints on legs, head, arms, and torso using a miniature wrist-mounted RGB camera. By pointing the camera towards the body, it can capture body silhouettes, which are learned by a customized deep learning model to infer the full body poses. A user study demonstrated that it can track full-body poses with an average of 6.89 cm. Given the miniature RGB camera is already embedded in many smartwatches, BodyTrak has the potential to be deployed in future smartwatches. We discuss the opportunities and challenges associated with deploying our system and implementing it in real-world scenarios.

REFERENCES

[1] Md Atiqur Rahman Ahad, Masud Ahmed, Anindya Das Antar, Yasushi Makihara, and Yasushi Yagi. 2021. Action recognition using kinematics posture feature on 3D skeleton joint locations. *Pattern Recognition Letters* 145 (2021), 216–224.

[2] Karan Ahuja, Andy Kong, Mayank Goel, and Chris Harrison. 2020. Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Devices Ecosystems.. In *UIST*. 1121–1131.

[3] Karan Ahuja, Sven Mayer, Mayank Goel, and Chris Harrison. 2021. Pose-on-the-Go: Approximating User Pose with Smartphone Sensor Fusion and Inverse Kinematics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[4] Amonzon. [n.d.]. Musou USB Safety Tester,USB Digital Power Meter Tester Multimeter Current and Voltage Monitor DC 5.1A 30V Amp Voltage Power Meter, Test Speed of Chargers, Cables, Capacity of Power Banks,Black. [EB/OL]. https://www.amazon.com/Musou-Digital-Multimeter-Chargers-Capacity/dp/B071214RD8 Accessed Oct 4, 2020.

[5] Rozilene Maria C Aroeira, B Estevam, Antônio Eustáquio M Pertence, Marcelo Greco, and João Manuel RS Tavares. 2016. Non-invasive methods of computer vision in the posture evaluation of adolescent idiopathic scoliosis. *Journal of bodywork and movement therapies* 20, 4 (2016), 832–843.

[6] Carlijn VC Bouten, Karel TM Koekkoek, Maarten Verduin, Rens Kodde, and Jan D Janssen. 1997. A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE transactions on biomedical engineering* 44, 3 (1997), 136–147.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[8] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, Francois Guimbretiere, and Cheng Zhang. 2021. NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–31.

[9] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-Mounted Miniature Cameras. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 112–125.

[10] Intel Corporation. 2021. RealSense. In *https://www.intelrealsense.com/*.

[11] Microsoft Corporation. 2021. Microsoft Kinect.. In *https://en.wikipedia.org/wiki/Kinect*.

[12] Rita Cucchiara, Costantino Grana, Andrea Prati, and Roberto Vezzani. 2004. Probabilistic posture classification for human-behavior analysis. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 35, 1 (2004), 42–54.

[13] Amit Das, Ivan Tashev, and Shoaib Mohammed. 2017. Ultrasound based gesture recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 406–410.

[14] Mohamed El Amine Elforaici, Ismail Chaaraoui, Wassim Bouachir, Youssef Ouakrim, and Neila Mezghani. 2018. Posture recognition using an RGB-D camera: exploring 3D body modeling and deep learning approaches. In *2018 IEEE life sciences conference (LSC)*. IEEE, 69–72.

[15] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7297–7306.

[16] Samuel Gandang Gunanto et al. 2016. 2D to 3D space transformation for facial animation based on marker data. In *2016 6th International Annual Engineering Seminar (InAES)*. IEEE, 1–5.

[17] Samer Hijazi, Rishi Kumar, and Chris Rowen. 2015. Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA* (2015), 1–12.

[18] Ryosuke Hori, Ryo Hachiuma, Hideo Saito, Mariko Isogawa, and Dan Mikami. 2021. Silhouette-Based Synthetic Data Generation For 3D Human Pose Estimation With A Single Wrist-Mounted 360° Camera. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1304–1308.

[19] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on Wrist. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 71 (June 2020), 24 pages. https://doi.org/10.1145/3397306

[20] Fang Hu, Peng He, Songlin Xu, Yin Li, and Cheng Zhang. 2020. FingerTrak: Continuous 3D hand pose tracking by deep learning hand silhouettes captured by miniature thermal cameras on wrist. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–24.

[21] Xinyue Huang and Adriana Kovashka. 2016. Inferring Visual Persuasion via Body Language, Setting, and Deep Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[22] Dong-Hyun Hwang, Kohei Aso, and Hideki Koike. 2019. MonoEye: Monocular Fisheye Camera-based 3D Human Pose Estimation. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 988–989.

[23] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. 2020. MonoEye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 98–111.

[24] NaturalPoint Inc. 2021. OptiTrack. In *http://optitrack.com*.

[25] Northern Digital Inc. 2021. trakSTAR. In *https://www.ndigital.com/msci/products/drivebay-trakstar/*.

[26] PhaseSpace Inc. 2021. PhaseSpace. In *https://phasespace.com/*.

[27] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[28] Shian-Ru Ke, LiangJia Zhu, Jenq-Neng Hwang, Hung-I Pai, Kung-Ming Lan, and Chih-Pin Liao. 2010. Real-time 3D human pose estimation from monocular view with applications to event detection and video gaming. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 489–496.

[29] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*. 2938–2946.

[30] David Kim, Otmar Hilliges, Shahram Izadi, Alex D Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 167–176.

[31] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. 2020. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 3 (2020), 1066–1078.

[32] Jianbo Liu, Ying Wang, Yongcheng Liu, Shiming Xiang, and Chunhong Pan. 2020. 3D PostureNet: A unified framework for skeleton-based posture recognition. *Pattern Recognition Letters* 140 (2020), 143–149.

[33] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time arm skeleton tracking and gesture inference tolerant to missing wearable sensors. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 287–299.

[34] ALT LLC. 2021. Antilatency. In *https://antilatency.com/*.

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[36] Vicon Motion Systems Ltd. 2021. Vicon. In *https://vicon.com/*.

[37] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*. IEEE, 506–516.

[38] Greg Mori, Xiaofeng Ren, Alexei A Efros, and Jitendra Malik. 2004. Recovering human body configurations: Combining segmentation and recognition. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Vol. 2. IEEE, II–II.

[39] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. 2020. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9890–9900.

[40] Jaime A Rincon, Angelo Costa, Paulo Novais, Vicente Julian, and Carlos Carrascosa. 2018. Intelligent wristbands for the automatic detection of emotional states for the elderly. In *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 520–530.

[41] Daniel Roetenberg, Henk Luinge, and Per Slycke. 2009. Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep* 1 (2009), 1–7.

[42] J Roggendorf, S Chen, S Baudrexel, S Van De Loo, C Seifried, and R Hilker. 2012. Arm swing asymmetry in Parkinson's disease measured with ultrasound based motion analysis during treadmill gait. *Gait & posture* 35, 1 (2012), 116–120.

[43] Ralf Schmidt, Catherine Disselhorst-Klug, Jiri Silny, and Günter Rau. 1999. A marker-based measurement procedure for unconstrained wrist and elbow motions. *Journal of biomechanics* 32, 6 (1999), 615–621.

[44] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a smartwatch and i can track my user's arm. In *Proceedings of the 14th annual international conference on Mobile systems, applications, and services*. 85–96.

[45] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K. Hodgins. 2011. Motion Capture from Body-Mounted Cameras. *ACM Trans. Graph.* 30, 4, Article 31 (July 2011), 10 pages. https://doi.org/10.1145/2010324.1964926

[46] Christina Strohrmann, Holger Harms, Cornelia Kappeler-Setz, and Gerhard Troster. 2012. Monitoring kinematic changes with fatigue in running using body-worn sensors. *IEEE transactions on information technology in biomedicine* 16, 5 (2012), 983–990.

[47] Nusrat Tasnim, Md Islam, Joong-Hwan Baek, et al. 2020. Deep learning-based action recognition using 3D skeleton joints information. *Inventions* 5, 3 (2020), 49.

[48] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. 2019. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7728–7738.

[49] Vive. 2021. HTC VIVE.. In *https://www.vive.com/*.

[50] Kathan Vyas, Rui Ma, Behnaz Rezaei, Shuangjun Liu, Michael Neubauer, Thomas Ploetz, Ronald Oberleitner, and Sarah Ostadabbas. 2019. Recognition of atypical behavior in autism diagnosis from video using pose estimation over time. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[51] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M Kitani. 2020. Back-Hand-Pose: 3D Hand Pose Estimation for a Wrist-worn Camera via Dorsum Deformation Network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1147–1160.

[52] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. 2019. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics* 25, 5 (2019), 2093–2101.

[53] Jackie Yang, Gaurab Banerjee, Vishesh Gupta, Monica S Lam, and James A Landay. 2020. Soundr: Head Position and Orientation Prediction Using a Microphone Array. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[54] Jackie Yang, Tuochao Chen, Fang Qin, Monica S Lam, and James A Landay. 2022. HybridTrak: Adding Full-Body Tracking to VR Using an Off-the-Shelf Webcam. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

[55] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.