

1 **BodyTrak: Inferring Full-body Poses from Body Silhouettes using a Wristband**

2
3 ANONYMOUS AUTHOR(S)

4
5 In this paper, we present BodyTrak, the first smart wristband that can estimate the full body poses in 3D. It uses miniature cameras
6 mounted on the wrist to capture the body silhouettes, which are learned by a customized deep learning model to estimate the 3D
7 positions of 14 joints on arms, legs, torso, and head. We conducted a user study with 9 participants in which each participant performed
8 12 daily activities such as walking, sitting, or exercising, in varying scenarios (wearing different clothes, outdoors/indoors). The results
9 show that our system can infer the full body pose (3D positions of 14 joints) with an average error of 6.89 cm and 6.34 cm, using just
10 one camera and four cameras mounted on the wrist respectively. Based on the results, we discuss the possible application, challenges,
11 and limitations to deploy our system in real-world scenarios.

12
13 CCS Concepts: • Computer systems organization → Embedded systems; Redundancy; Robotics; • Networks → Network reliability.

14
15 Additional Key Words and Phrases: Pose Estimation, Motion Tracking, Wearable Technology, Smart devices

16
17 ACM Reference Format:

18 Anonymous Author(s). 2018. BodyTrak: Inferring Full-body Poses from Body Silhouettes using a Wristband. In *Woodstock '18: ACM
19 Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/1122445.1122456>

20
21 **1 INTRODUCTION**

22
23 Body pose estimation is becoming increasingly important in many fields such as health (e.g., physiology of individuals
24 with physical disorders such as scoliosis and Parkinson's [4][38]), the gaming industry, sports analysis, and even
25 communication studies which can help us understand how we interact with one another through our body language[45]
26 [11] [18]. Previous research has used methods such as external sensors placed within a room, depth cameras. It works for
27 some applications, e.g., motion games at home. However, these solutions are limited in terms of mobility, not allowing
28 reconstructing the body in the field.

29
30 To address these mobility issues, researchers have used wearable solutions to estimate body poses. Most of these
31 systems require the users to wear multiple sensors (e.g., IMU) on the body [37, 41], which are less practical in the real-
32 world setting. The recent advancements show that using a single form factor, such as a chest-mounted camera[20], a hand-
33 held smartphone[2] or a hat-mounted camera,[25], can also estimate full-body poses with encouraging performances.
34 However, these form factors (chest-mount or hat) may not be immediately acceptable or convenient for users to be
35 worn in different daily activities. For instance, chest-mounted devices such as GoPro are acceptable for a group of users
36 in specific contexts. However, it is still not yet be acceptable to be worn on a daily basis for many. Therefore, in the
37 future eco-system of wearables, it is essential to offer users a variety of wearable sensing technologies to track body
38 poses to decide the technology based on the context.

39
40 Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not
41 made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components
42 of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to
43 redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

44
45 © 2018 Association for Computing Machinery.
46 Manuscript submitted to ACM

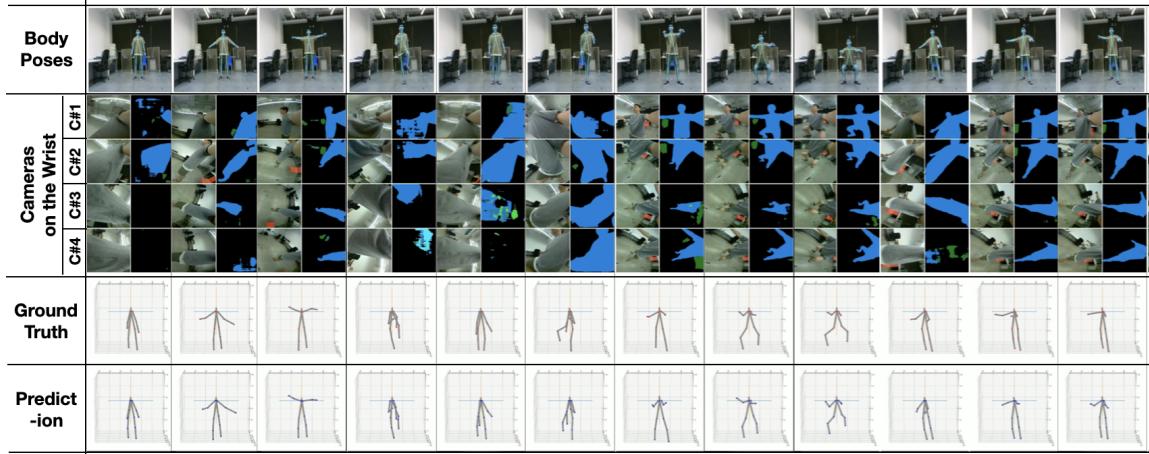


Fig. 1. Body Pose Estimation results using BodyTrak. In the first row, the user is performing an activity, and depth camera is displaying the body posture using skeletal points. The second row of images represent the set of four images captured by each camera on the wristband and their accompanying segmented 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 row displays predictive body posture using the images input through the wristband cameras.

Considering this form factor limitation, we explore the feasibility of estimating the full-body pose using a wristband in this paper. Compared to other form factors like the chest-mounted device or hat, the wristband is arguably a more common form factor and is currently widely used. Thus, we form the research question of the paper as:

- Is it possible to estimate the full body poses by only using a wrist-mounted device?

To answer this research question, we developed BodyTrak, the first 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 with 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 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 and 6.34 cm using one camera and four cameras, respectively. 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:

- The first wrist-mounted device that can estimate the full body poses.
- A customized deep learning pipeline which can infer the 3D positions of 14 body joints (full body poses) from the image of body silhouettes/incomplete body parts captured by wrist-mounted cameras.
- A user study with 9 participants to evaluate the performance of BodyTrak when the user performs different activities under different scenarios.

- 105 • A discussion on the opportunities and challenges of applying BodyTrak on the future wrist-mounted devices in
 106 real-world applications.
 107

108 2 RELATED WORK

110 Our work is to estimate full body poses in 3D using a wrist band with cameras. In this section, we first review the
 111 literature related to estimating 3D full-body poses using Non-wearable and wearable-based technologies. Then, we
 112 discuss the previous research using wrist-mounted cameras to estimate body poses.
 113

115 2.1 3D Full Body Reconstruction using Non-wearable Devices

116 Applying external devices is one of the most typical methods of capturing body posture with either passive or active
 117 sensing technology. Devices with high portability and compact dimensions, such as the Microsoft Kinect [10] and Intel
 118 RealSense [9], have proved their effectiveness in motion capture. Due to the advances of the RGBD camera, segmentation
 119 of objects from backgrounds and the continuous tracking of body movement can be accomplished by these devices
 120 without the use of any markers, which are commonly used for 3D animation[14]. In commercial high-quality motion
 121 capture systems such as OptiTrack [21] and Vicon [32], several retroreflective markers are usually placed on the rigid
 122 body where external cameras can easily capture them. Similarly, fiducials are also used as a standard marker in optical
 123 measurement [30] while PhaseSpace [23] and VIVE [44] apply active markers. Thanks to the flourishing of computer
 124 vision and machine learning, researchers are able to reconstruct human poses simply using RGB images, which is
 125 present in works such as DensePose [13], PoseNet [26], and OpenPose [6]. In addition to utilizing cameras, projects
 126 utilizing acoustics sensors [1, 12], RF [50] and magnetic fields [22] have also demonstrated impressive performances in
 127 human pose estimation.
 128

129 However, non-wearable devices with either optical or other sensors are sometimes not practical or even unavailable
 130 in some scenarios. Systems requiring markers typically involve a specific working space, while other devices like
 131 the Kinect requires users to perform in front of the device. These limitations restrict the user's mobility and their
 132 applications in real world scenarios.
 133

138 2.2 3D Full Body Reconstruction using Wearable Devices

141 Work	142 Position of the Sensor	143 Body Estimation	144 Sensor	145 Tracking Moment	146 Activity	147 Accuracy
xR-EgoPose[43]	VR headset	Full body	Fish-eye camera	Always on	9	5.82cm
Mo2cap2[49]	Cap-mounted	Full body	Fish-eye camera	Always on	8	6.14cm
Monoeye[19]	Chest	Full body	Fish-eye camera	Always on	3	5.0cm
You2Me[34]	Chest	Full body	GoPro camera	Always on	4	8.6cm
Pose-on-the-Go [2]	Smart Phone	Full body	Depth camera IMU	When holding the phone	6	<25cm
Real-time arm[29]	Wrist	An upper arm	IMUs	Always on	17	10.53cm for elbow 12.94cm for wrist
I am a smartwatch [40]	Wrist	An upper arm	IMU	Always on	10	9.2cm
BodyTrak	Wrist	Full body	RGB cameras	Always on	13	6.34cm

153 Table 1. Comparison with Other Previous work.

To overcome the constraints of the external non-wearable devices, researchers developed wearable technology with sensors like Internal measuring units (IMU) to reconstruct a 3D body [37, 41]. Mo2cap2 [49] tracks human pose with a cap-wearable fish eye camera pointing downwards. Similarly, xR-EgoPose [43] applies an egocentric fish-eye camera on the VR headset enables 3D avatar motion in VR environment. Other examples like Mono-eye [19] and You2Me [34] wears a single fish-eye chest mounted camera. In contrast to prior systems requiring devices wearing on the body or placing outside, Post-on-the-Go [2] employing iPhone with sensor fusion tracks body posture while users hold the phone in hand. However the above form factor such as chest-mount, hat, or holding smartphone, may not always be available to users in daily activities. Thus, it is important for the future ecosystem of wearable computers to have a variety of form factors such that the users can decides the wearable technology based on the applications or contexts.

Although smartwatches /wristbands are the most popular commodity wearable products on the market, we have not seen any technology that can estimate the full body pose using a single wrist-mounted form factor. Thus, BodyTrak is the first wrist-mounted wearable sensing technology that can track the full body poses. 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 system.

2.3 Wrist-mounted cameras for activity recognition

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 cameras¹. Reserach projets have also explored using wrist-mounted cameras to recognize hand poses[16, 27] or emotional state [36]. However, to the best of our knowledge, we have not seen any prior work that explored using a single wrist-mounted device to estimate the full-body poses. In other words, BodyTrak is the first wearable technology that explores estimating full body poses on a wrist-mounted form factor.

3 THEORY OF OPERATION



Fig. 2. 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, it is impossible to capture the full body images with wrist-mounted cameras due to constraints such as the limited angles of view and occlusion problems. However, 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. In other words, how the body parts were occluded provides rich information on the body pose. We observed that the partial body images/silhouettes

¹<https://www.wristcam.com/>

		Upper Body Movement			
		One Hand Movement		Two Hands Movement	
		Right Hand	Left Hand	Synchronous	Asynchronous
Movement (X)					
	(1) Shoulder Press	(2) Shoulder Press		(3) Lateral Raise	(5) Cross arm
				(4) Front Raise	(6) Boxing
				(9) Sitting	(11) Walking
Movement (O)				(10) Squat	(12) Stair
	(7) Tennis	(8) Kicking			

Fig. 3. Design Space. A matrix of 12 body movements that cover the full range of body movements.

captured from the wrist are unique and different on different body poses, as shown in Fig. 2. Furthermore, recent work has already demonstrated that using partial body images/silhouettes to estimate hand poses[17, 48] and facial expressions [7, 8]. Therefore, we believe that it is possible for AI to learn the 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.

Based on these considerations, we created the 4×2 design spaces (see. Figure 3), allowing BodyTrak to explore various body postures. In each category, we found daily activities from previous 3D full-body reconstruction works [20] (walking, boxing, kicking, sitting) and common daily exercises (e.g., squat, front raise). This design space ensures

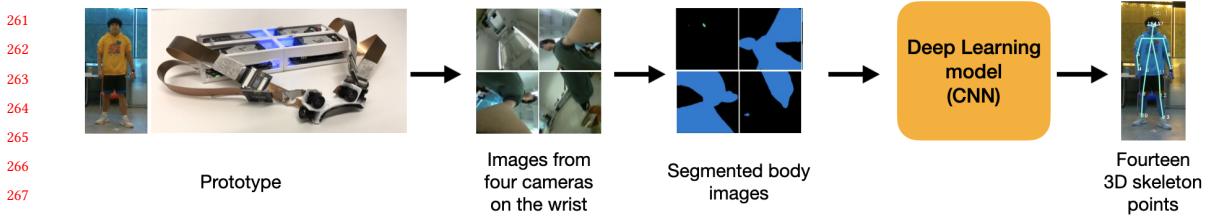


Fig. 4. System Overview of BodyTrak. The prototype is put onto the user. They then perform 13 activities. While the activities are being performed the cameras are capturing images of the body. These images are then segmented and fed into a CNN which outputs an estimated body posture for each activity, using 14 skeleton points.

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 PoseTrak, 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.

5.1 Prototype

5.1.1 Hardware. Our hardware prototype consists of three main components, the wristband with four RGB cameras, a depth camera for tracking groundtruth of 3D body poses, and a PC for data processing. The 3D printed wristband houses four miniature RGB cameras (i.e. b006605 RGB Arducam) with a dimension of 60mm × 11.5mm × 9.5mm and a fixed FOV of 160°. Each camera 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 weighing in at 350g, which are held in the user's hand during the study. 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 camera² 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.

²<https://www.intelrealsense.com/depth-camera-d435/>

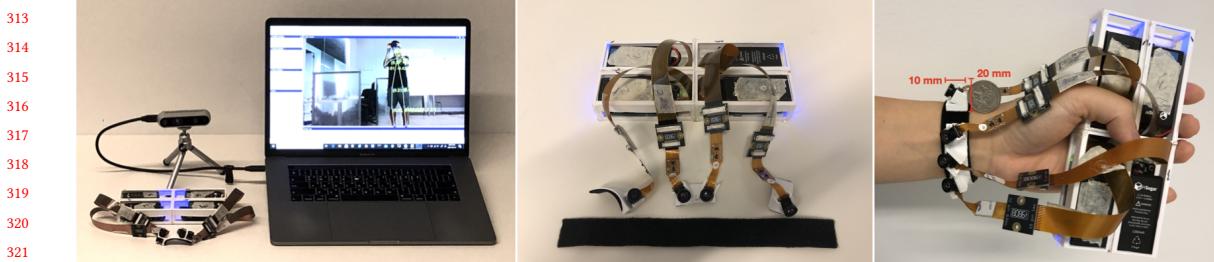


Fig. 5. Implementation and Hardwear Setting. The full prototype consists of the wristband, the Intel RealSense Depth Camera, and a PC. When setting up the prototype for data collection, the cameras are mounted onto the velcro strap which is then wrapped around the participant's wrist. The participant then holds the 3D box which holds the raspberry pi boards and the accompanying power sources.

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 [31, 33]. 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. 2. After applying several body segmentation techniques such as PixelLib³, Pose2Seg⁴, and CDCL [28], we decided to use a well-known pre-trained model named FCN-ResNet101 [31]⁵ 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 6.

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 results to train and test the performance on

³<https://github.com/ayoolaolafewa/PixelLib>

⁴<https://github.com/erezposner/Pose2Seg>

⁵https://pytorch.org/hub/pytorch_vision_fcn_resnet101/

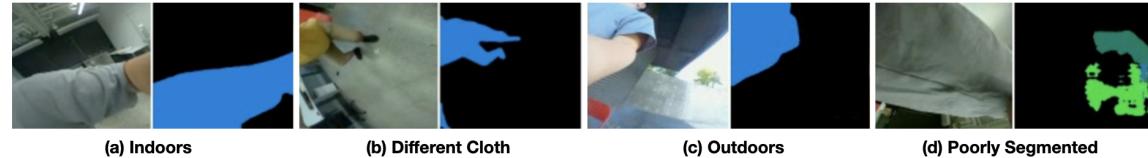
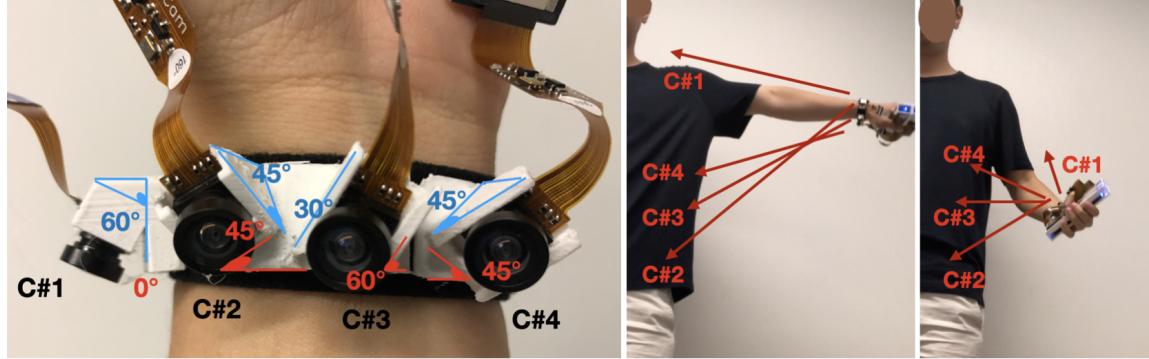


Fig. 6. Body Segmentation. The images provided ensure that we can visually distinguish between properly and poorly segmented images. The blue in the segmented images represent the user's body and the black indicates the background. In image (b) the user is performing a different posture and wearing different clothing. In the outdoor environment (c), segmented images are displayed when the user is performing the activity in an outdoor setting.

365 BodyTrak. Therefore, if the body segmentation in the images improves, the performance of BodyTrak can potentially
 366 be even better. What we present in this paper is just a baseline.
 367



382 Fig. 7. Camera Setting. In this figure is the camera arrangement, along with their angular measurements. The x plane runs horizontally
 383 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
 384 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
 385 demonstrates how the wristband is held and the directions in which the cameras point.
 386

387 **5.2.2 Camera Setting: Position Arrangement.** An optimized camera setting (number, position and orientation) is the key
 388 for this system to accurately estimate full body posture. By experimenting with our research team, we aimed to find the
 389 best camera setting to capture maximum information on body poses.
 390

391 After conducting a pilot study on segmentation, we determined the camera settings using the following criteria.

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

396 Based on these criteria, we investigated the camera arrangement considering 1). the range of arm movement such as
 397 holding up or folding the arm. 2). the natural rotation of the wrist reaching up to 150° [39] and 3) our cameras' FoV is
 398 160°. As a result, the first camera (See. Fig. 7 C#1) is empirically placed on the side of the arm at 60° perpendicular to
 399 the arm. This ensures that we can capture more body information, including head pose, when users stretch their arms
 400 without folding. The other three cameras (see. Fig. 7 C#2, C#3, and C#4) are positioned side by side on the inner part of
 401 the arm considering the rotation of the wrist when the arm is folded. The positions and arrangement were empirically
 402 decided based on the preliminary experiments on researchers. Here, we titled and rotated the cameras as seen in Fig. 7
 403 (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
 404 each camera pointing more towards the body. In addition, all three cameras excepting C#1 were rotated from 45° to 60°,
 405 which was helpful to capture the torso with lower/upper body for proper segmentation. In Section 7.1, we will discuss
 406 the impact of camera settings on the performance.
 407

412 **5.3 Ground Truth Acquisition**

413 We used the depth camera (Intel RealSense) and associated body Skeleton tracking application to capture the groundtruth
 414 of body poses, including 3D positions of 18 body joints on eyes, ears, nose, arms, torso, and legs Fig.8(a). We exclude
 415

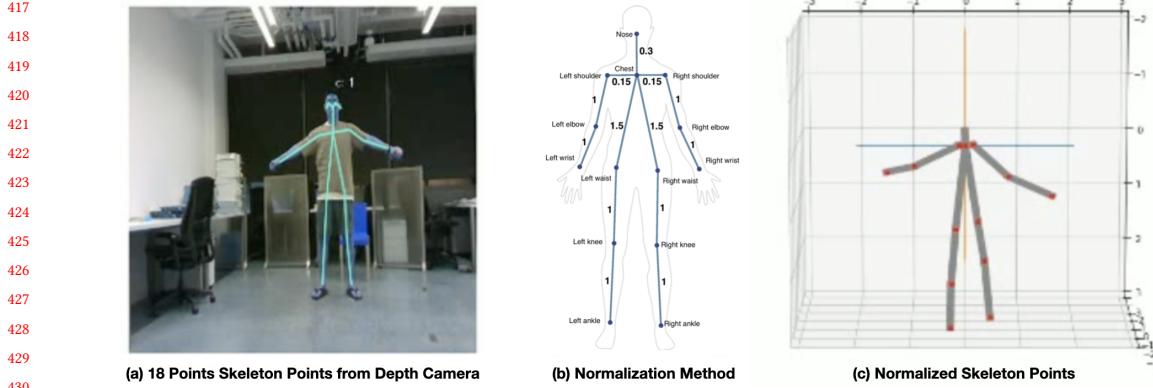


Fig. 8. Ground Truth Acquisition. Image (a) represents the ground truth that is displayed when using the RealSense depth camera. In image, (b) we depict the normalization values for each body part. In figure (c) the image displays the normalized skeleton after it is passed through the deep learning pipeline.

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 skeleton information before feeding it into a deep learning model. As shown in Fig. 8(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. 8(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,

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

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 [15]. 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. 9. We then conduct late fusion by concatenating the

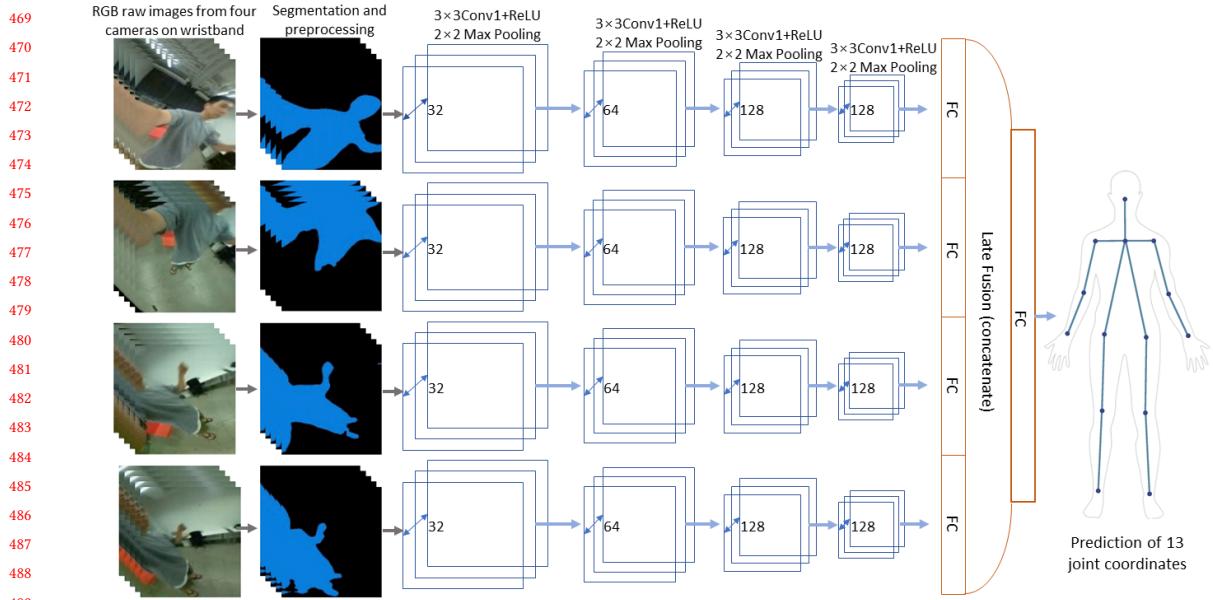


Fig. 9. Deep Learning Model Structure

output of fully connected layer in each branch and add final fully connected layer to obtain the estimation of 14 joint coordinates.

5.4.2 Model Training. Using the Adam loss function, our model is trained to predict 42 parameters (i.e., 14 joint points \times 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 the left wrist.

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. fig3). 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.

521 6.1.3 *Data Collection.* We have four sessions for data collection. In each session, participants were asked to perform 12
 522 activities. We intentionally asked the participants to slowly perform these activities such that the depth camera could
 523 capture the ground truth reliably. In the first session, the participants repeated each activity ten times. In the later three
 524 sessions, the participants repeated each activity three times. All activities were randomly ordered.
 525

526 The first session evaluated the system performance within the same session where the device was not remounted.
 527 In the second session, we asked the participant to remount the wristband to evaluate how the performance would be
 528 impacted after remounting. In the third session, the participants were asked to wear a different shirt to evaluate how
 529 would BodyTrak perform if the cloth was different. In the fourth session, the participants performed the activities in an
 530 outdoor setting to investigate how our system operates in outdoor environments where lighting and background vary.
 531

533 6.2 Result

534 In this section, we reported the performance of BodyTrak in different settings, including without remounting, after
 535 remounting, after changing the cloth, and outdoor environments. We used the images captured from all four cameras in
 536 the following experiments.
 537

538 6.2.1 *Evaluation Matrix.* We calculated the performance, i.e., accuracy, by taking the Euclidean distance between the
 539 predicted 3D joints and the true value of the 3D joints from the ground truth and averaging over the sequence. This
 540 distance is then scaled to centimeters (cm) based on each participant's arm (Mean = 22.78 cm, SD = 1.4 cm). We report
 541 an average mean error over all 14 body joints.
 542

Motion Type	Avg	Nose	Ch	RS	RE	RWr	LS	LE	LWr	RWa	RK	RA	LWa	LK	LA
Average	6.34	1.95	0.00	1.12	9.44	14.97	1.18	9.16	13.52	5.36	6.46	6.87	5.65	6.41	6.71
Standing	4.21	1.37	0.00	0.74	7.00	10.12	0.91	5.43	6.18	4.12	4.05	4.49	4.04	4.44	4.55
Right Shoulder Press	4.03	1.36	0.00	0.78	9.51	12.91	0.84	6.12	7.11	4.17	4.33	4.47	3.21	4.16	4.11
Left Shoulder Press	4.26	1.60	0.00	0.89	5.61	8.12	0.83	7.01	10.45	3.88	4.43	4.44	3.85	4.28	4.38
Lateral Raise	4.59	1.65	0.00	0.83	7.06	11.17	1.05	7.64	11.50	4.15	4.71	4.86	4.09	4.57	4.96
Front Raise	4.76	1.57	0.00	0.98	7.20	12.80	1.03	8.08	12.39	4.01	4.75	5.01	4.21	5.11	5.02
Cross Arm	4.34	1.36	0.00	0.81	6.95	14.53	0.90	6.85	11.47	3.66	4.38	4.61	4.24	4.54	4.78
Boxing	5.26	1.76	0.00	1.20	11.29	15.41	1.07	10.31	12.66	4.85	5.39	5.64	5.08	5.01	5.47
Tennis	6.28	1.77	0.00	1.11	10.04	14.68	1.03	8.80	13.84	5.20	6.51	6.61	5.62	6.58	7.10
Kicking	6.76	1.80	0.00	1.04	7.19	12.55	1.05	8.19	12.94	5.63	7.84	9.59	6.14	6.67	7.53
Sitting	4.74	1.69	0.00	0.86	5.29	7.41	0.76	5.23	6.87	3.90	5.72	5.33	4.63	5.55	5.13
Squat	6.12	1.76	0.00	1.02	9.18	18.70	1.10	7.63	15.16	4.90	6.08	6.45	5.68	7.03	6.65
Walking	5.94	1.70	0.00	1.02	8.06	12.25	1.02	8.01	11.37	4.89	5.78	6.41	5.62	6.26	7.15
Stepping a Stair	7.19	1.86	0.00	1.03	8.59	13.43	1.24	9.20	13.35	5.30	7.56	7.91	6.03	7.46	7.75

562 Table 2. Results of Within Session Evaluation (cm). R: Right, L: Left; Ch: Chest, S: Shoulder, E: Elbow, Wr: Wrist, Wa: Waist, K: knee,
 563 A: Ankle.

564
 565
 566
 567 6.2.2 *Within Session Evaluation.* At first we conducted a within-session evaluation on BodyTrak using the data from
 568 the first session as the training and testing data. Specifically, we used the first 8 instances of the 12 activities (based on
 569 chronicle order) as the training data, and the last 2 instances as the testing data. Please note that the collected data was
 570 not shuffled when splitting the training and testing set. The results show that BodyTrak achieved an average of 6.34 cm
 571

(SD = 0.61 cm), indicating that the body silhouettes captured using wrist-mounted cameras are informative to estimate body pose as shown in Fig. 1.

Similar to other works [19, 34], 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 of 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 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 9.30 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. Although the overall accuracy was worse than within-session performance, it indicates the potential of applying BodyTrak in real-world settings.

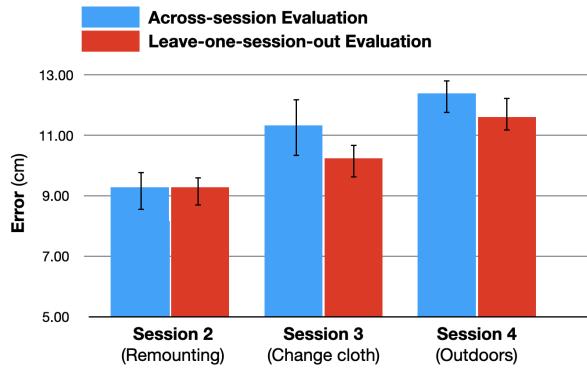


Fig. 10. Comparison on Performance between Across-session and Leave-one-session-out. In this figure the blue bars indicate the error when data from the Across-session evaluation was included. The red bars indicate the error from Leave-one-session-out evaluation. As demonstrated in this figure by including more data (i.e. Leave-one-session-out evaluation), our performance improved.

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 repeated four times, so that each session was used as 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. Compared to the second experiment, leave-one-session-out improved the performance for around 1 cm in each session(See. Fig 10). It indicates that if a 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.

Evaluation Type	Across-session			Leave-one-session-out		
Testing Session	Session 2	Session 3	Session 4	Session 2	Session 3	Session 4
Avg.	9.30	11.16	12.33	9.30	10.12	11.33
Nose	2.62	2.96	3.70	2.78	2.96	3.28
Chest	0.00	0.00	0.00	0.00	0.02	0.00
Right Shoulder	1.39	1.66	2.21	1.57	1.75	2.17
Right Elbow	13.05	15.93	16.93	13.24	14.20	15.89
Right Wrist	23.03	27.03	28.94	22.61	24.03	26.94
Left Shoulder	1.64	1.75	2.23	1.75	1.77	2.08
Left Elbow	12.59	15.18	16.88	12.57	13.79	15.19
Left Wrist	20.00	24.67	27.37	18.83	20.84	23.85
Right Waist	7.63	8.35	9.99	8.00	8.29	9.33
Right Knee	9.64	11.26	12.25	9.50	10.41	10.92
Right Ankle	10.48	12.73	14.29	10.25	11.55	13.01
Left Waist	8.57	10.72	11.99	9.07	9.60	11.26
Left Knee	9.29	11.43	12.41	9.56	10.68	11.82
Left Ankle	10.26	12.63	13.37	10.43	11.82	12.88

Table 3. Results of User Study under different conditions

7 DISCUSSIONS

7.1 The impact of cameras settings

In our evaluation, we used four cameras to estimate fully 3D body poses. However, 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 within-session experiment and across-session experiments. The results (detailed in Table ??) 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 in within-session experiment, compared to 6.3 cm using all four cameras. Our interpretation is that camera 1 with a wide view angle, may covers in average more information on the upper body and lower body. It shows that, it is possible to estimate the full body pose using just one camera, which is significantly more practical than using four cameras. Because as we have discussed previously, some of the existing commodity smartwatches already has a built-in camera. By slightly changing the position and orientation, these built-in camera can potentially be repurposed to track full body poses.

7.2 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 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 using the session 2,3,4 as the testing data respectively. This process 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, 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 user-independent model in the future, especially if a 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 [19]. 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.

Evaluation Type		Within Session	Accross Session		
Setting	Camera	Session 1	Session 2	Session 3	Session 4
One Camera	c#1	6.90	11.74	12.92	13.49
	c#2	6.89	13.37	13.50	14.53
	c#3	7.25	12.91	13.41	15.11
	c#4	7.41	14.05	14.42	14.42
Two Cameras	c#1, c#2	6.69	12.25	13.03	13.80
	c#1, c#3	6.82	10.68	11.98	14.27
	c#1, c#4	6.99	12.65	13.03	14.27
	c#2, c#3	6.79	10.84	13.05	13.05
	c#2,c#4	6.98	11.49	13.48	13.48
Three Cameras	c#3,c#4	7.25	13.90	14.69	14.69
	c#1,c#2,c#3	6.46	11.79	12.80	13.64
	c#1,c#2,c#4	6.51	10.51	12.12	12.94
	c#1,c#3,c#4	6.63	10.02	11.59	12.71
Four Cameras	c#2,c#3,c#4	6.63	13.17	13.71	14.10
	c#1,c#2,c#3,c#4	6.34	9.30	11.16	12.33

Table 4. The Impact of Camera Settings

7.3 Real World Application

As the development of commercial smartwatches with cameras such as WristCam⁶, 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.

7.3.1 Areas of Application. By using 3D body reconstruction in a smart watch we can get important information about the behavior we exhibit in daily activities. Namely, this technology can improve individual fitness performance analysis by examining the energy we expend in our workout routines [5]. Posture is also important in a variety of sports such as swimming [46], running[42], and soccer [24]. The incorporation of 3D body pose estimation in a smart watch contributes to greater mobility to track these types of activities, helping people in the process of improving their athletic performances. Alongside workouts, the development of a smartwatch that can estimate our body posture can inform us about our posture while performing activities in the environment we are regularly exposed to. Such information can guide us to decide what shoes to wear for a specific walking route, or how much weight we can hold onto while performing blue collar work until our posture begins to worsen. To build on this point, the future

⁶<https://www.wristcam.com/>

of smartwatches and 3D body estimation reflects a future of precision in personalized health informatics. Moreover, 3D body estimation is making its way into commercial industries with products like Vicon [32], which can monitor 3D body posture and provide bio mechanic analyses, and the WristCam [47] which enables users to video record using either of the two cameras. 3D Posture estimation in smart watches may also potentially help us care for elderly individuals in emergency situations by recognizing uncommon body postures [35]. The ubiquity of smartwatches presents a tremendous opportunity for companies to use 3D body posture technology to develop the next generation of wrist-mounted device, which is novel, personalized experiences for users.

7.3.2 *Hardware Considerations.* 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 camera module in video mode measured by PowerJive USB Power Meter [3]. 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, we recorded the images and then offline processed all data on a workstation in our study. However, moving forward, the implementation of our system can be altered by considering by applying one of 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.

7.4 Privacy Issue

A wearable camera will always raise concerns about privacy. In BodyTrak, as the cameras mostly point towards the user's body and do not directly capture the user's face, the images are arguably less sensitive. However, the cameras can still occasionally capture sensitive information of the user or the people around. 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.

7.5 Limitations and future work

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

First, our training data set is relatively small. If 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.

Lastly, the ground truth acquired by RealSense depth camera was overall stable. However, it still includes errors and noise occasionally. If we have a more accurate and reliable ground truth, the performance can potentially be improved too. Besides, using depth camera as the ground truth acquisition method also limited the activities that the participants can perform. For instance, the participants can not actually walk or run in large distance. Otherwise the depth camera would not capture the full body images, which lead to inaccurate estimation of body poses.

8 CONCLUSION

This paper presents BodyTrak, the first AI-powered wristband that can continuously estimated the full body poses including the 3D positions of 14 body joints on legs, head, arms and torso. It utilizes four RGB cameras wearing on the wrist to capture the body silhouettes, which are learned by a customized deep learning model to infer the full body poses. A user study with 11 participants demonstrated that it estimate the full body poses, with an average of 6.34 cm error over 14 body joints using four cameras and 6.89 cm using only one camera. Based on the study results, we discuss the opportunities and challenges associated with deploying our system and implementing it in real-world scenarios.

REFERENCES

- [1] 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.
- [2] 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.
- [3] Amazon. [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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] Intel Corporation. 2021. RealSense. In <https://www.intelrealsense.com/>.
- [10] Microsoft Corporation. 2021. Microsoft Kinect.. In <https://en.wikipedia.org/wiki/Kinect>.
- [11] 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.
- [12] Amit Das, Ivan Tashev, and Shoab Mohammed. 2017. Ultrasound based gesture recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 406–410.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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>

- [17] 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.
- [18] 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*.
- [19] 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.
- [20] 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.
- [21] NaturalPoint Inc. 2021. OptiTrack. In <http://optitrack.com>.
- [22] Northern Digital Inc. 2021. trakSTAR. In <https://www.ndigital.com/msci/products/drivebay-trakstar/>.
- [23] PhaseSpace Inc. 2021. PhaseSpace. In <https://phasespace.com/>.
- [24] Alen Kapidžić, Tarik Huremović, and Alija Biberović. 2014. Kinematic analysis of the instep kick in youth soccer players. *Journal of Human Kinetics* 42 (2014), 81.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] ALT LLC. 2021. Antilatency. In <https://antilatency.com/>.
- [31] 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.
- [32] Vicon Motion Systems Ltd. 2021. Vicon. In <https://vicon.com/>.
- [33] 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.
- [34] 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.
- [35] Chankyu Park, Jaehong Kim, and Ho-Jin Choi. 2012. A watch-type human activity detector for the aged care. In *2012 14th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 648–652.
- [36] 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.
- [37] 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).
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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>
- [42] 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.
- [43] 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.
- [44] Vive. 2021. HTC VIVE.. In <https://www.vive.com/>.
- [45] 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.

- 885 [46] Zhelong Wang, Jiaxin Wang, Hongyu Zhao, Sen Qiu, Jie Li, Fengshan Gao, and Xin Shi. 2019. Using wearable sensors to capture posture of the
886 human lumbar spine in competitive swimming. *IEEE Transactions on Human-Machine Systems* 49, 2 (2019), 194–205.
887 [47] WristCam. 2021. WristCam. In <https://www.wristcam.com/product>.
888 [48] 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
889 Wrist-worn Camera via Dorsum Deformation Network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*.
890 1147–1160.
891 [49] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. 2019. Mo 2 cap 2:
892 Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics* 25, 5 (2019),
893 2093–2101.
894 [50] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human
895 pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936