DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking

Nathan DeVrio
Carnegie Mellon University
Pittsburgh, PA, USA
ndevrio@cmu.edu

Chris Harrison
Carnegie Mellon University
Pittsburgh, PA, USA
chris.harrison@cs.cmu.edu

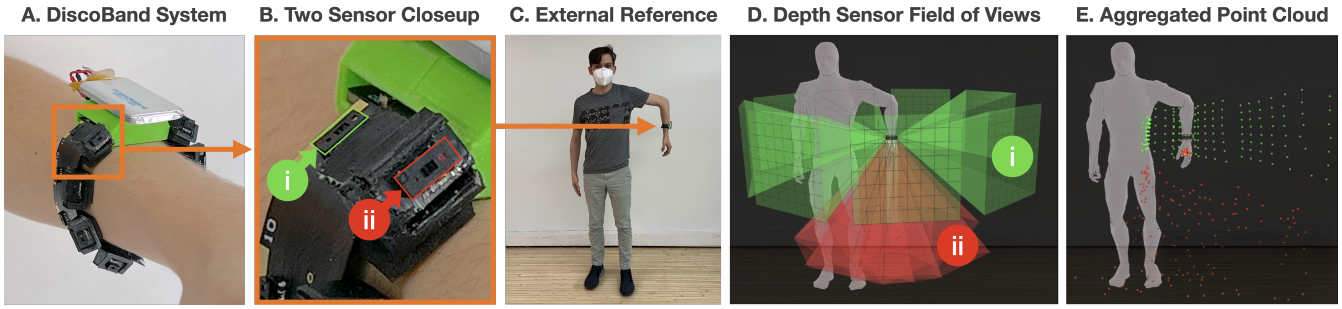

Figure 1: DiscoBand is a smartwatch strap (A-C) featuring sixteen ultra-small, multi-zone depth sensors (closeup of two sensors in B) looking outwards and towards the hands (green and red frustums in D), which capture a fisheye point cloud (E). Many interactive uses are possible, including arm and hand pose tracking.

ABSTRACT

Real-time tracking of a user's hands, arms and environment is valuable in a wide variety of HCI applications, from context awareness to virtual reality. Rather than rely on fixed and external tracking infrastructure, the most flexible and consumer-friendly approaches are mobile, self-contained, and compatible with popular device form factors (e.g., smartwatches). In this vein, we contribute DiscoBand, a thin sensing strap not exceeding 1 cm in thickness. Sensors operating so close to the skin inherently face issues with occlusion. To help overcome this, our strap uses eight distributed depth sensors imaging the hand from different viewpoints, creating a sparse 3D point cloud. An additional eight depth sensors image outwards from the band to track the user's body and surroundings. In addition to evaluating arm and hand pose tracking, we also describe a series of supplemental applications powered by our band's data, including held object recognition and environment mapping.

CCS CONCEPTS

• **Human-centered computing** → **Mobile devices**.

KEYWORDS

Smartwatch; Sensing; Hand gestures; Body pose; Mobile devices; Interaction techniques

UIST '22, October 29-November 2, 2022, Bend, OR, USA

ACM ISBN 978-1-4503-9320-1/22/10.
https://doi.org/10.1145/3526113.3545634

ACM Reference Format:
Nathan DeVrio and Chris Harrison. 2022. DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking. In *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22), October 29-November 2, 2022, Bend, OR, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3526113.3545634

1 INTRODUCTION

Hands are the chief appendage with which humans manipulate the world around them, and for this reason, digitization of the hands for use in interactive computing systems has been sought after for half a century [10, 77]. Applications vary tremendously, including smart environments with hand tracking [5], sign language recognition [57], gesture sensing smartwatches [11, 17, 25, 28, 75], and whole-hand replication in virtual reality for object manipulation [30].

Approaches generally fall into one of three categories. First, we can instrument both the user and the environment, for example with optical [64], acoustic [46], magnetic [45, 47] or other markers worn by the user and sensed by external sensors. Second, it is possible to instrument only the environment, for example with cameras and use computer vision to track a user's body pose [6]. Finally, we can instrument only the user, for example with body-worn IMUs [69], cameras [54], and many other types of active and passive devices (which we review later). The latter category has the significant benefit of being mobile (i.e., self-contained and not limited to a specially-instrumented room), and generally encounters less occlusion from objects in the environment (e.g., furniture) and the user themselves, depending on the instrumentation point.

The wrist is a particularly popular instrumentation point from which to sense the hands for three key reasons. First and foremost, it is a common place to wear jewelry, watches, and other bands. Second, it is a highly practical location to affix a small device to

the body [16]. And third, wrists are proximate to a user's hands, offering the potential for superior data capture. For these same reasons, we too focus on the wrist location. Also similar to prior work, we employ optical sensors to capture the hand pose. However, as we will discuss, many prior systems employing such sensors have had to elevate components centimeters above the skin in order to achieve reliable line-of-sight, which results in form factors less amenable for consumer adoption. Additionally, techniques using cameras tend to elevate privacy concerns (unless otherwise noted, when we discuss "cameras" in regards to related work, we are referring to commonly-used, high-resolution, RGB or infrared cameras).

In response, we set out to create a sensor band that is comparable to a smartwatch strap: thin and self contained. For this, we use low-resolution (8×8 pixels), ultra-small (4.9×2.5×1.6 mm) depth cameras. To mitigate natural occlusion (e.g., fingers blocking line-of-sight to other features), we use eight sensors distributed around the periphery of the wrist (Figure 1, B & D, red frustums). When one view is occluded, the others are generally not, and in this manner they can work together to composite a live 3D point cloud (Figure 1E, red points) and resolve a probable hand pose.

Our band also features eight additional depth sensors facing outwards (i.e., normal to the skin; Figure 1B), which we use to track the arm and upper body (Figures 1, D & E, green frustums and points). Taken together, these sixteen depth cameras capture a fisheye-like 1024 point cloud (Figure 1E), the rays of which are reminiscent of a disco ball reflecting light, and so we dubbed our prototype DiscoBand. The sensors we use have a range of 4 m, allowing us to also capture the proximate environment, opening other application areas discussed later.

Overall, DiscoBand offers a unique combination of features and properties that differentiate it from prior work. First and foremost, our band is thin, and could be plausibly integrated into future smartwatches. Second, our multi-view approach is inherently more robust to occlusion than single-view methods. Third, our low-resolution depth data is more privacy preserving than conventional camera-based wrist systems. Finally, our band's unique design and data opens entirely new capabilities not previously demonstrated with wrist-worn setups, including the ability to estimate user upper body pose, detect held objects, and scan the environment for obstacles and contextual clues. In this paper, we document the implementation of DiscoBand and different example applications we explored, as well as report results from a series of user studies that underscore the potential of our approach.

2 RELATED WORK

We now review three key areas of related work. First, we describe the many disparate application domains that can benefit from hand and arm pose tracking. We then move to a brief survey of hand and arm tracking systems worn on the wrist, but which sense the hands and body using indirect means. We conclude with a discussion of systems closest to our own: wrist-borne systems that use sensors to directly measure the physical configuration of the hands and arms.

2.1 Applications of Hand & Arm Pose Tracking

Gesture and pose tracking of both the hands and arms have been long standing goals in the field of Human-Computer Interaction (HCI). At a basic level, gesture and pose tracking can be used to augment existing devices, such as smartphones, with alternate input channels for interaction. Examples of this include zooming by pinching the index finger and thumb together [70], dismissing notifications with a flick of the wrist [3], and sign language input [57]. Recognition of specific key poses is particularly important in domains such as eating monitoring and medication adherence in healthcare [12, 48, 76], form correction in fitness [52, 60], avatar representation in virtual reality [9, 50], and remote control of robots [65]. Whole-arm pose tracking has also been explored for in-air gestures for handwriting [74] and mapping symbolic body language to emojis [31].

2.2 Wrist-Borne Indirect Sensing

One approach to gesture tracking is through indirect sensing of measurable features that indicate the presence of a pose/gesture without ever imaging the shape of the body itself. Among the many methods in the literature, we were most interested in those that involved sensors worn on the arm or wrist, as this was most similar to our setup.

One of the most popular methods for indirect hand gesture sensing is electromyography (EMG), which measures the electrical activity of muscle tissue. Many systems have implemented this technique because it is relatively non-invasive [29, 32, 39]. Similar techniques that also measure muscle contraction include using air pressure bladders affixed to the forearm [28] and resistive strain sensors on the wrist or back of hand [11, 36].

Another common technique for indirect hand sensing is tomography, which tracks pose configuration by emitting excitation signals into the forearm and measuring how the received signal change based on shifts in the internal composition of the arm. There are many different ways to perform tomography, some involving infrared light [41, 44] or ultrasound [27, 42], while others use pairwise electrical impedance measurements from electrodes surrounding the arm [75]. Less common indirect hand-sensing techniques include acoustic measurements (both active [43, 55] and passive methods [21]), capacitive sensing [49, 63], and measuring skin deformations on the back of the hand [61].

The last major indirect technique – detecting movement with inertial measurement units (IMUs) – is used for both hand tracking [67, 71] and whole-arm tracking [4, 14, 38, 53]. Advantages of IMUs include their ability to detect gestures beyond just the hand and their ubiquity in many of the smart devices users already own. It is worth mentioning that historically IMUs have been the only major technique that has been used for tracking arm pose using wrist-borne sensors.

2.3 Wrist-Borne Direct Sensing

Most similar to the sensing technique that DiscoBand employs are on-wrist systems that directly image the physical form of the hand or arm they are attempting to track. Within this category there are many different imaging sensors used. In past years, the two most popular methods have been 1) arrays of IR emitters/detectors

Reference systems	Tracking output	Method	High information	Low calibration	Low occlusion	Privacy preserving	Low profile	Low power
Myo [29, 32, 39]	Continuous pose	Electromyography (EMG)	✗	✗	✓	✓	✓	✓
Jung et al. [28]	Discrete gestures	Pressure	✗	✗	✓	✓	✓	✓
SensIR [41]	Discrete gestures	IR tomography	✗	✗	✓	✓	✓	✓
Interferi [27], EchoFlex [42]	Discrete gestures, cont. finger angle	Ultrasound tomography	~	✗	✓	✓	~	✗
Tomo [75]	Discrete gestures	Electrical imp. tomo. (EIT)	✗	✗	✓	✓	✓	✓
GestureWrist [49], CapBand [63]	Discrete gestures	Capacitive	✗	✗	✓	✓	✓	✓
Sugiura et al. [61]	Discrete gestures	Skin deformation	✗	✗	✓	✓	✓	✓
Serendipity [67], Xu et al. [71]	Discrete gestures	Motion/inertial (IMUs)	✗	✓	✓	✓	✓	✓
Beamband [26]	Discrete gestures	Ultrasonic beamforming	✗	✗	✗	✓	✗	✗
WristWhirl [17], RotoWrist [51], ThumbTrak [62]	Discrete gestures, cont. wrist angle	IR rangefinding	✗	~	✓	✓	✓	✓
Arakawa et al. [1], Chen et al. [8], Back-Hand-Pose [68]	Continuous pose	High-res RGB camera	✓	✓	✗	✗	✗	✗
Digits [30], Opisthenar [73]	Continuous pose	High-res IR camera	✓	✓	✗	✗	✗	✓
WatchSense [56]	Cont. pose (of second hand)	High-res depth camera	✓	✓	✗	✗	✗	✗
FingerTrak [25]	Continuous pose	Low-res thermal camera	✗	✗	✓	✓	✓	✗
DiscoBand	Continuous pose	Low-res depth camera	✗	✗	✓	✓	✓	✓*

Figure 2: A high-level overview of wrist-worn, hand tracking methods. We include reference system(s) for each method, which we then evaluate across a range of desirable qualities. Orange tildes are used when the capability is system dependent (and less innate to the method). *Capable of low power operation using the next generation of sensors [59].

positioned around the wrist that capture a single depth value per detector [17, 18] and 2) single cameras mounted above the wrist of many varieties, including RGB [1, 8, 68], depth [56], thermal [72], and infrared [30, 66, 73]. Although not technically a wrist-borne system, ThumbTrak uses IR rangefinders in an ring array to capture finger micro-gestures [62]. Another prominent category of sensing is ultrasonic, ranging from single-sensor measurements [37] to complex beamforming arrays [26].

While often having superior SNR due to direct tracking of the hand, a commonality of the above systems is the need to operate significantly above the surface of the arm to achieve sufficient line of sight to e.g., the fingers [1, 30, 56]. Even still, if the wrist bends away from the sensor, most of these systems will lose tracking due to occlusion [68, 73]. One of the differentiating features of DiscoBand is that it aims to mitigate this occlusion problem by capturing depth maps from many overlapping view points around the wrist. This multi-view approach is underexplored, with the exception of FingerTrak [25], a hand tracking system using four thermal cameras positioned around the wrist. FingerTrak's thermal cameras are low resolution (32×24 pixels), and thus, like DiscoBand, the system more privacy preserving relative to those employing high-resolution cameras. Of course, the two systems are quite different in terms of number of sensors used (and thus views captured; 4 vs. 16) and the type of sensor data captured (temperature vs. distance).

Finally, the most related systems from a sensing standpoint are WristWhirl [17] and RotoWrist [51], which track continuous wrist angle using small infrared range finders (depth sensors). The latter sensors are very similar in style and operation to the sensors we use, but are not multi-zone. In other words, WristWhirl's twelve and RotoWrist's eight sensors provide exactly that many distance measurements, where as DiscoBand's sixteen sensors provide 1024 points in a 3D volume. Moreover, the specific sensor we use has a larger field of view, allowing DiscoBand to capture an aggregated fisheye-like point cloud. These improvements allow us to consider new and interesting use cases, discussed later.

Beyond IMUs, there has been relatively little work on wrist-borne arm tracking. Notable among non-IMU methods is work by Hori et al. [24], which used a wrist-mounted 360° camera to estimate full body pose. With a resolution of 4992×2496, there are obvious privacy implications, and the apparatus itself is approximately 10 cm in height. Even still, there is significant self occlusion, which has been an issue for other worn, single-camera, pose tracking systems. DiscoBand contributes a new method for arm pose tracking with is comparatively practical and compact.

3 DISCOBAND IMPLEMENTATION

DiscoBand is a combination of hardware and software contributions working together to enable a series of use cases, which we subsequently evaluate. In this section, we detail the implementation of these components.

3.1 Hardware

Our early hardware prototypes consisted of a series of linked rigid PCBs (example prototypes shown in Figure 3, top row), but these were awkward to wear and increased the height from the skin, one of the design parameters we sought to minimize. Our final design (Figures 1 and 4) features two flexible PCB "wings" that permit our prototype to be closely wrapped around a user's wrist, emulating a smartwatch strap. Figure 3 (bottom row) offers a mockup of a commercial implementation using a flexible PCB over-molded in silicone.

The most notable component we use is STMicroelectronics' VL53L5CX time-of-flight (ToF) ranging sensor (Figure 1B, green and red highlights). This diminutive surface-mount component is able to capture an 8×8 depth image at 15 FPS, up to a range of 4 m. It uses a 940 nm class 1 laser, which is eye safe. Eight of these sensors face outwards from our band (i.e., normal to the skin), where they can image the wearer's body and environment. Each sensor has a 45° horizontal field of view, so that together, the eight sensors provide a 360° point cloud (Figure 1D, green frustums). A second set of eight sensors faces towards the user's hand (i.e., parallel to

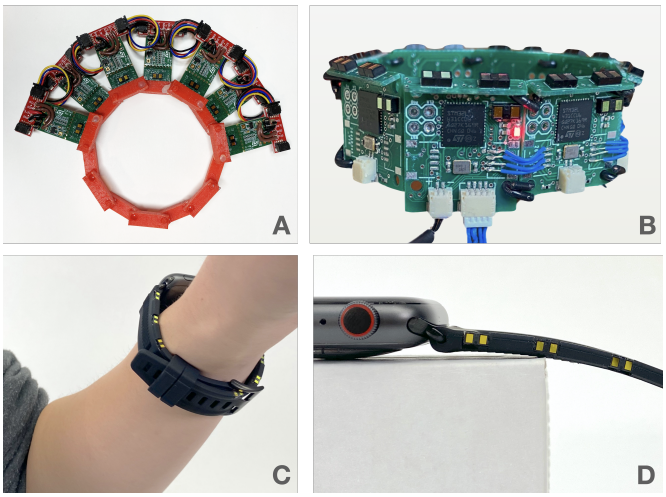

Figure 3: Two early DiscoBand prototypes made from rigid PCBs (A & B). A physical mockup of how DiscoBand could appear in a commercial form factor (C & D).

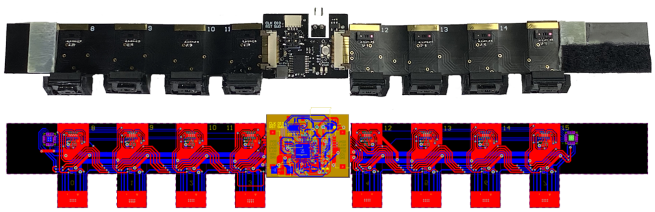

Figure 4: A photo and PCB layout of our final DiscoBand protoype, which can also be seen in Figures 1, 5-8, 13-16.

the skin), sitting on flexible PCB tabs that bend upwards. These sensors have overlapping field of views (Figure 1D, red frustums), but because they operate from different vantage points, they are able to image parts of the hand that might otherwise be occluded in other sensor views.

All sixteen VL53L5CX sensors are controlled by a STM32L431 microcontroller, which sits on a small rigid PCB to which our two flexible PCBs attach. We use all three of the chip's independent I2C buses to maximize sensor throughput. The rigid board also features a BNO055 9-DOF IMU (three axis inertial data, absolute orientation, and magnetometer), 64 kB FRAM, and analog power circuitry for the sensors. Using a modular serial port, our sensor band can run over USB (for power and data) or be battery powered and communicate over Bluetooth. We discuss power consumption in Section 8.

3.2 Firmware

Our firmware handles three main responsibilities: interfacing with the depth sensors, packaging data frames, and transmitting them from the watch to a computer for processing. We note that in the future, all processing could occur on the watch itself, as the latest generation of smartwatches contain very capable processors

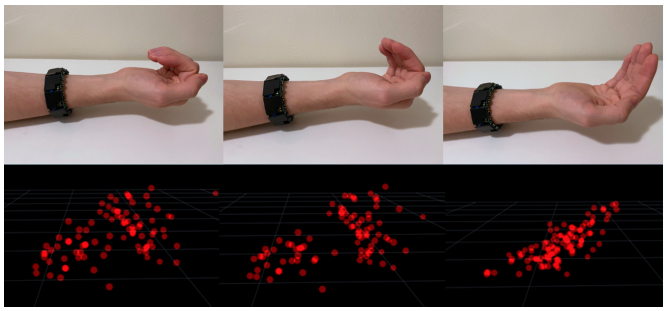

Figure 5: DiscoBand's depth sensors capture different views of the hand, which can be composited into a unified 3D point cloud. In this example sequence, the user uncurls their fingers, which is apparent in the point cloud data.

and machine learning hardware accelerators. To maximize our framerate, we utilize all three I2C busses on the STM32L431 microcontroller, and also use direct memory access (DMA) to free the processor for other parallel tasks. One bus is devoted to each wing's array of hand-facing depth sensors, while the third bus communicates with all of the outward-facing sensors. We found that the VL53L5CX sensors do not significantly interfere with one another even when operating simultaneously, and we can further select sensor pairs that are on opposite sides of the body to further reduce any interference. This strategy allows us to overlap as many as three sensor read requests at once (two hand-facing sensors and one outward-facing sensor). Rather than wait for all sensors to be read before transmitting a single data frame, we send sensor values piecemeal as they become available, which helps to reduce latency. As we have half as many sensors on our two hand-facing sensor buses, we can receive all hand-facing depth maps at 7.3 FPS. For our single bus of eight outward-facing sensors, our band runs at 3.7 FPS, and this is also the frequency at which we send 9-DOF IMU data. This data is transmitted over serial via USB or Bluetooth. We note that the depth sensors we use can go up to 15 Hz, and that our prototype's slower framerate is primarily a consequence of using a single low-cost microcontroller and sending data back to a computer. If a superior processor was used, enabling on-device compute, both of these bottlenecks could be eliminated.

3.3 Compositing Multi-View Point Clouds

Although our flexible PCBs can deform, the relative geometry of the sixteen depth sensors is largely constrained. If we take into account a user's wrist circumference (which could be entered once during a setup wizard), we can make even tighter estimates. We use this geometry to composite all sensor data into a unified point cloud, as shown in Figure 5. We can orient this point cloud in three ways, useful for different applications. First, and most simple, is to keep the point cloud aligned to the wrist's coordinate system (essentially the band's first person view). Second, we can rotate it to align with the world using the BNO055's gravity and magnetic north vectors. Finally, we can align it to the wearer's body, using our upper body tracking described later. At this stage, we also perform basic signal

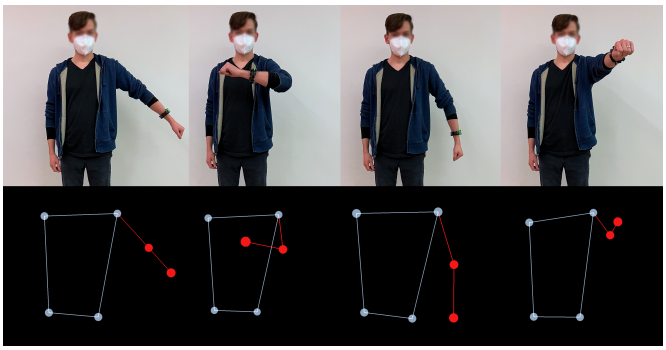

Figure 6: Four example arm poses (top row) alongside real-time output of DiscoBand's arm pose pipeline (bottom row).

filtering to reject outlier or intermittent noisy depth values, which improves the robustness of our downstream example applications.

At present, this multi-view compositing process occurs on a laptop using a custom application we wrote in Python. This app also offers a real-time visualization of the unified point cloud, with the ability to toggle on and off individual sensors, which was helpful for debugging and exploring possible use cases. However, we note that our compositing process is not computationally expensive, and could occur on the smartwatch in the future, as could the machine learning components described in the next section.

4 EXAMPLE USES & IMPLEMENTATIONS

Our primary motivation for creating DiscoBand was to enable arm and hand pose tracking. Our band's design reflects this, with two rings of depth sensors explicitly targeted for these applications. For these primary use cases, we built functional, real-time implementations, which we evaluate in a later user study. However, during our development process (and particularly after we were able to visualize the multi-view point clouds generated by our band), many additional uses came to light. These application areas, explored but not rigorously implemented, are described in Section 7.

4.1 Arm Pose Tracking

Tracking a user's arm relative to their upper body has applications in context sensing (e.g., activity detection), fitness (e.g., rep counting), rehabilitation (e.g., range of motion tracking), and other domains. As discussed in Related Work, the most common way this data is captured today is with external sensors, such as cameras placed in a room. There are comparatively few self-contained, worn systems that track the arm and upper body in the literature, with the most popular method being worn IMUs requiring per-worn-session calibration.

For this application, DiscoBand uses data from its eight outward-facing depth sensors. The raw point cloud data is geometrically complex, and thus we sought an efficient way to reduce dimensionality. For this, we cluster the point cloud using DBSCAN [15] (min cluster size 16, max inter-point distance of 75 mm) and identify the closest large cluster (generally the user's torso). We then compute the phi, theta, and distance values as machine learning features.

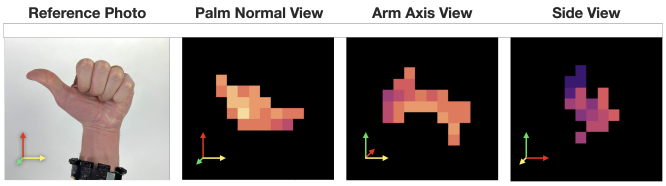

Figure 7: Left: photo of example hand pose. Right three: the three virtual view depth maps our pipeline creates as features for hand pose estimation.

We also include the absolute orientation of the wrist (three Euler angles) provided by our band's IMU.

To capture ground truth arm pose for training, we use an RGB webcam and BlazePose [2] via MediaPipe Pose [20], which provides 3D estimates for 33 body keypoints. Of these, we save six keypoints: the left wrist, left elbow, left shoulder, right shoulder, left hip and right hip. In cases where a user wears a smartwatch on their right hand, we would capture right wrist and right elbow instead. We use SciPy's ExtraTreesRegressor (default parameters) to predict the aforementioned six upper body points, with the midpoint between the two hip joints as the root node. Example tracking is shown in Figure 6 and our Video Figure.

4.2 Hand Pose Tracking

More widely explored in the literature is hand pose tracking, which has immediate applications in VR input (e.g., grasping objects), free-space interactions (e.g., gestural control of IoT devices), and activity detection (e.g., carpal tunnel mitigation). We note that data capture inevitably suffers from heavy hand self-occlusion, and when rendered as a point cloud (examples in Figure 5), the data more resembles that of a 3D convex hull. Nonetheless, the external geometry that is captured is indicative of many different hand poses.

As with arm pose tracking, we featurize the point cloud to provide a lower-dimensional representation that is both descriptive and stable. More specifically, we create three synthetic 14×14 depth maps looking at the hand from different virtual viewpoints: looking

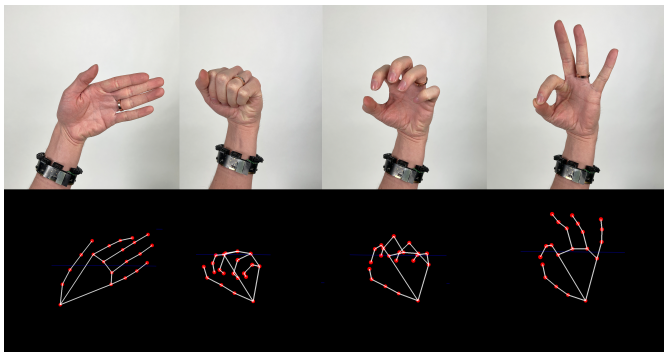

Figure 8: Four stills of example hand poses (top row) alongside renders of the 21 hand pose joints predicted by our model (bottom row).

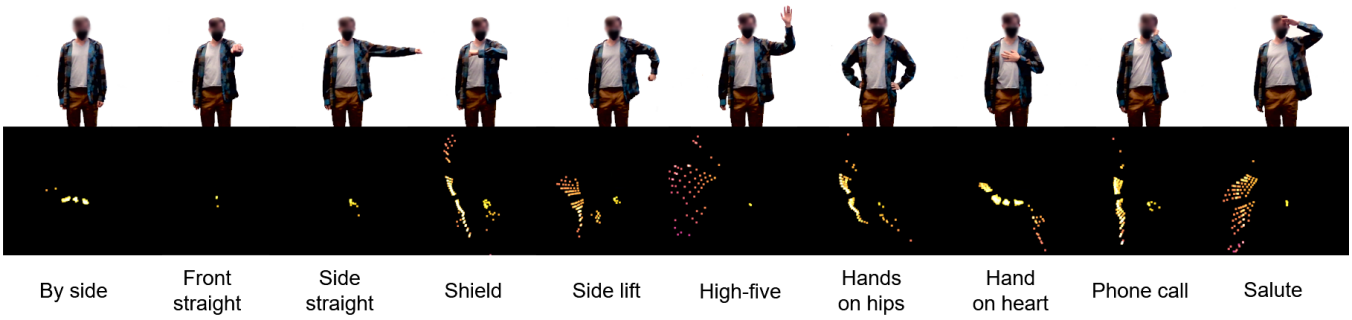

Figure 9: The ten terminal arm poses used in our study. Top row: Reference pose images. Bottom row: point cloud of body as seen by the eight outward-facing sensors.

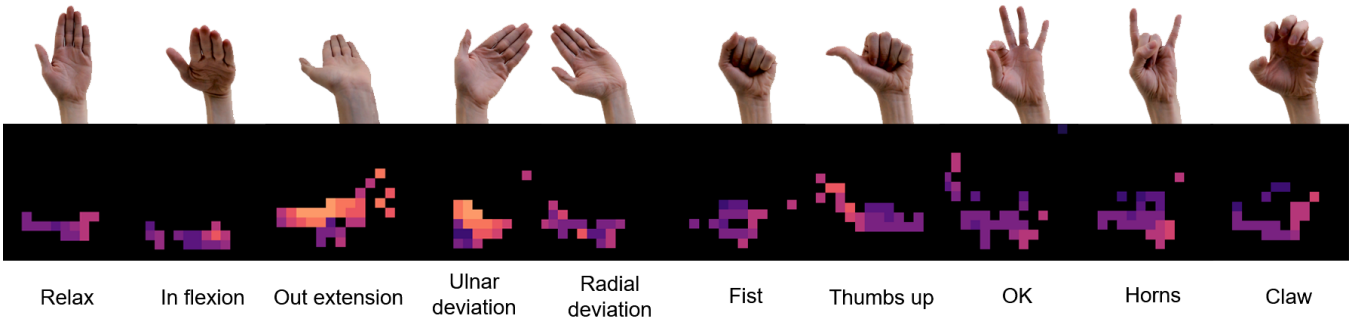

Figure 10: The ten terminal hand poses used in our study. Top row: Reference pose images. Bottom row: palm-normal-view synthetic depth maps (see Section 4.2 and Figure 7) of the hand as captured by the eight hand-facing sensors.

down the palm normal, along the arm's axis, and a side view of the hand (pinky finger closest to virtual camera). An example of these three views for a thumbs-up pose can be seen in Figure 7.

For ground truth data capture, we use a webcam and MediaPipe Hands [19, 40], which provides 3D coordinates for 21 hand keypoints. In pilot testing, we found this software offered better tracking accuracy and stability than a Leap Motion Controller, which is often used in such experiments. Our model (SciPy ExtraTreesRegressor, default parameters) predicts MediaPipe's 21 hand keypoints. Example hand pose tracking can be seen in Figure 8.

5 EVALUATION PROCEDURE

To evaluate the efficacy of our multi-view depth sensing approach, we recruited ten participants (mean age 25, all right-handed) for a 60 minute study, which paid $20 in compensation. The study was conducted in a standard office space with large pieces of furniture nearby and next to a set of large windows letting in outside light. After a brief orientation, participants were fitted with DiscoBand, proceeding once they felt comfortable. We then recorded six body measurements: shoulder width, arm length, wrist diameter, palm width, band to wrist crease distance, and band to middle finger tip distance. These values are passed to our machine learning models as user descriptors, and also used to normalize our data in analysis.

Our procedure was divided into two parts, starting with an arm pose study. Lacking a common set of arm poses from the literature to work with, we devised our own set of ten poses, exemplifying a variety of arm movements. In designing our arm pose set, we aimed to capture 1) a variety of joint movements (shoulder/elbow) and 2) joint orientations, as well as poses that 3) provided varied spatial endpoints (i.e., above/below the shoulders, as well as in-front/to-the-side of the body). Importantly, we were not just considering the poses as static endpoints, but also considering variation in the dynamic movements between poses. Pictures of our final pose set can be seen in Figure 9.

These poses were visually requested using a computer monitor. Participants were instructed to slowly move their body to match and then hold that pose for a few seconds, after which the monitor showed the next arm pose to be performed. One round of data collection consisted of all ten arm poses in a random order. Ten rounds of data were collected in this manner, which formed one session of data. DiscoBand was then removed and participants were given a few minutes break. The band was then re-worn and a second session of data was collected.

Importantly, we did not just record data at the terminal arm poses, but rather continuously throughout the experiment. This provided significantly more pose variety (and tracking challenge) than a defined pose set. Consider, for instance, all the intermediate pose states between the salute pose and hands-on-hips. In essence, each round of data capture can be thought of not as ten random pose trials, but rather as nine paired pose transitions (90 possible

pairwise combinations). For ground truth 3D body pose, we used a webcam (1 m away) and BlazePose [2] running on MediaPipe Pose [20], which provides 33 body keypoints at 30 FPS. Each time our band transmits a frame of data, it is saved alongside the most recent BlazePose output. Across our ten participants, this data collection procedure yielded 50,000 body pose instances.

Next, our study moved to hand poses. This procedure was similar to arm pose, this time with a ten-gesture hand pose set drawn from the literature [26, 28, 75] (Figure 10). As before, a computer monitor requested the hand pose, and participants slowly moved their hands to match, holding the pose for a few seconds, before continuing to the next trial. One round of data collection consisted of all ten hand poses requested in a random order, and ten rounds were collected per participant to complete one session. As before, two sessions of data collection were completed, with a break in between when the band was removed. To capture ground truth hand pose, we placed a webcam 25 cm below the participant's hands, looking upwards. We use MediaPipe Hands [19, 40] to estimate the 3D hand pose (21 keypoints) which, like before, is recorded in each frame of data. Across all ten participants, this collection procedure yielded 100,000 hand pose instances.

Similar to arm pose, we recorded the continuous motions between hand poses and not just the terminal poses. With this process, we were able to collect data and eventually track hand movements down to the granularity of individual fingers. Tracking of individual finger movements specifically occurred during the evaluation when performing transitions such as *Fist* → *Thumbs up* and *Relax* → *OK* and can also be seen in our Video Figure.

6 RESULTS & DISCUSSION

For our evaluation, we were interested in investigating DiscoBand's performance in predicting arm and hand pose when moving continuously between poses in our gesture sets. In addition, we also wished to quantify performance stability for a single user across multiple worn sessions, and across multiple users. To accomplish this, we compared our regression model results against the BlazePose ground truth in three separate analyses of within-session, cross-session, and cross-user performance.

6.1 Within-Session Performance

We conducted an investigation of within-user continuous joint error to simulate the performance of the system when it is calibrated to each user when first worn. We train/test our model using a modified k-fold cross validation procedure ($k = 5$) whereby we divide the data in the two session into 5 folds each, and then select one fold from each session to concatenate into a test data set and concatenate the remaining 8 folds into a training set. Of note is that all folds were divided at gesture transitions and data was not shuffled, ensuring there was never training and testing on data from the same instance of a performed gesture. We averaged the results from all 25 fold combinations to estimate within-session performance for a single user, and repeated this process for all users. Combining the per-user results, we then calculated mean per-joint position error (MPJPE) across all participants, which is shown along with standard error bars in Figure 12 for arm pose tracking and Figure 11 for hand pose tracking (dark blue bars).

For within-session arm tracking, we found a MPJPE of 8.83 cm (SD=2.83 cm) for the arm joints (wrist and elbow) and 5.88 cm (SD=2.75 cm) for all upper-body points. The maximum joint error was 11.05 cm for the wrist, which is expected given that it has the greatest range of motion and is furthest from the root node. When viewing arm pose prediction outputs, the most common observation was that the model would get the general arm gesture correct, and any error often arose from the fine pitch of the elbow or shoulder joint bends.

For within-session hand tracking, we found a MPJPE of 11.69 mm (SD=2.37 mm) and a maximum joint error of 21.84 mm (tip of the middle finger). Joint error decreases for each finger joint moving inwards from the distal end, as one would expect. Error was lowest at the middle finger metacarpal joint, rather than the wrist joint, because when the hand pitches MediaPipe returns keypoints rotated about a root close to that joint rather than the wrist, giving it the smallest range of motion.

6.2 Across-Session Performance

In addition to testing performance in single worn sessions, we also wished to test reproducibility, namely the ability to have stable performance each time the user puts on the band without the need for retraining. To test across-session performance, we evaluated our model with a simple leave-one-session-out cross validation, whereby we first train on data from session one and test on session two, and vice versa, averaging the results to obtain performance for a single user. This process was repeated for all users and all per-users results were combined to calculate MPJPE as shown in Figure 12 for arm pose tracking and Figure 11 for hand pose tracking (orange bars).

For cross-session arm tracking, we found a MPJPE of 12.16 cm (SD=4.32 cm) for the arm joints (wrist and elbow) and 7.40 cm (SD=3.28 cm) for all upper-body points. For cross-session hand tracking, we found a MPJPE of 17.87 mm (SD=2.89 mm). In both studies, but especially for hand tracking, there was a noticeable increase in error when testing cross-session performance. This is almost certainly due to the band varying slightly in worn position across sessions, and given the model is only trained on data from one session (i.e., one location), it cannot extrapolate to other locations. It seems likely that more varied training data would be of great benefit. For arm pose tracking, the effect is less severe as surfaces are both farther away and more coarse. Finally, we note that cross-session robustness is often a challenge for worn sensor systems, and DiscoBand is no exception. As one point of reference, FingerTrak [25] saw a 127% increase in error when moving from within-session to cross-session testing (vs. DiscoBand's 53% increase in error).

6.3 Cross-User Performance

Beyond testing performance for individual users with custom models, we also wished to evaluate the ability of our system to be trained once (i.e., model flashed at the factory) and work for all users. To investigate this, we performed a leave-one-user-out cross validation, where data from nine participants was used for training and the tenth for testing. This process was repeated for all possible combinations of participants and the results were averaged to calculate

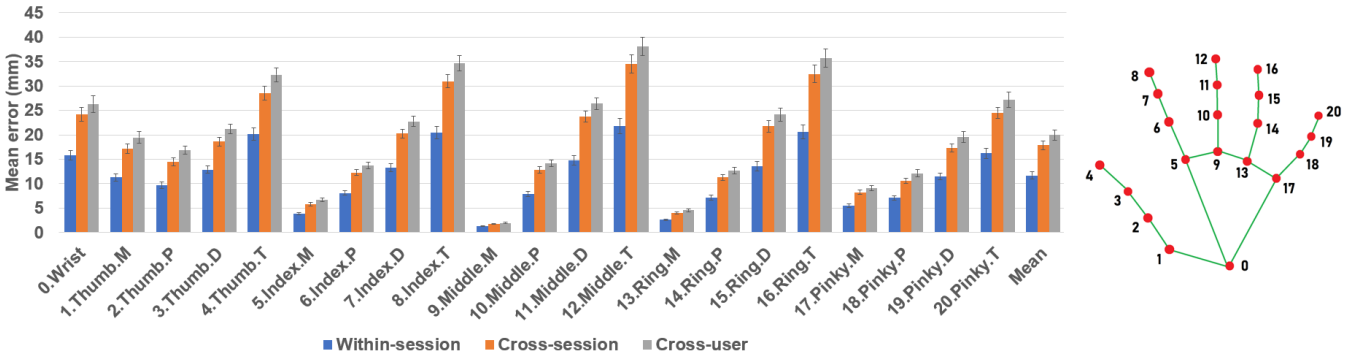

Figure 11: Hand pose study per-joint position error averaged across all participants for 21 hand keypoints. Our lowest mean per-joint position error was 11.7 mm for within-session.

MPJPE. The results are shown in Figure 12 for arm pose tracking and Figure 11 for hand pose tracking (grey bars).

For cross-user arm tracking, we found a MPJPE of 12.42 cm (SD=2.97) for the arm joints (wrist and elbow), and 8.24 cm (SD=2.96) for all upper-body points. For cross-session hand tracking, we found a MPJPE of 19.98 mm (SD=3.29). We were surprised to find that for the wrist joint, the most important for arm pose tracking, there was almost no increase in error moving from cross-session to cross-user. This is in contrast to other approaches (e.g., EIT, tomography) which often see significant drops in performance across multiple users [27, 75].

7 FUTURE USE CASES

As discussed in Section 4, our DiscoBand implementation focused on two core use cases: arm and hand pose tracking. However, over the course of more than a year of development, we identified several other interesting use cases. Although only brief explorations with basic implementations, we believe they convey how DiscoBand is an enabling technology with many interesting uses.

7.1 Bimanual Activities

While prior work has shown that smartwatches can be used to detect activities performed with the hand (see e.g., [33–35]), they are inherently limited by the fact that users wear one smartwatch on one arm. Worse still, the most common location to wear a smartwatch is on the non-dominant hand; if an activity is performed with the other (dominant) hand, it is likely impossible to detect.

In the case of bimanual activities — such as cutting food on a plate, steering a car, typing on a keyboard, riding a bicycle, tying one's shoes, or performing jumping jacks — DiscoBand can provide useful data. These examples are illustrated in Figure 13 where we can see DiscoBand readily captures the other arm using its outward-facing depth sensors. The resolution is crude at present, permitting only arm detection and angle estimation. However, forthcoming advances in sensor resolution will greatly enable this use case. We speculate that virtual IMU data could be synthesized by tracking the opposing arm mass, such that a single smartwatch has inertial data for both hands, which would be a huge boon to many context sensing applications.

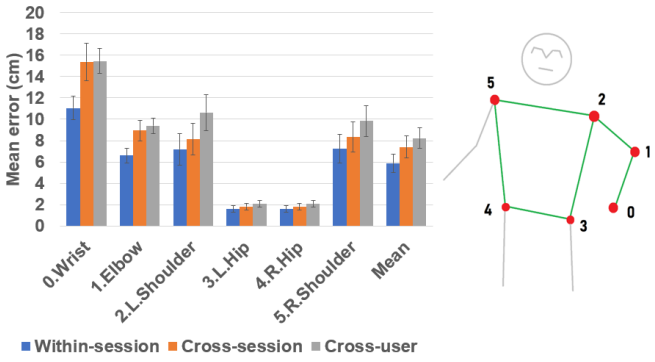

Figure 12: Arm pose study MPJPE across all participants for six upper body keypoints. Within-session accuracy was the strongest, with a MPJPE of 8.8 cm for the arm joints (keypoints 0 & 1) and 5.9 cm for all upper-body points.

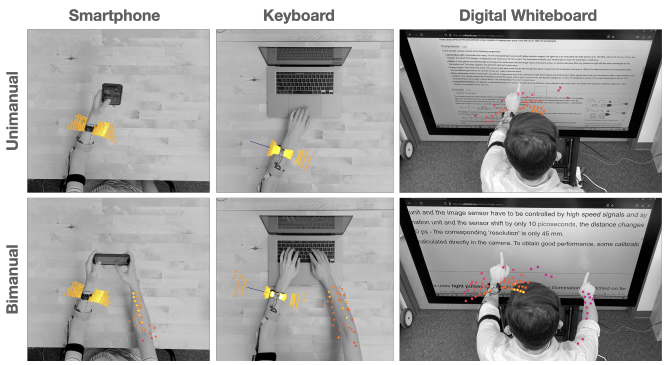

Figure 13: In these paired examples, we can see how DiscoBand is able to image the other (uninstrumented) arm. For illustration, the real point clouds have been rotated to match a greyscale reference photo.

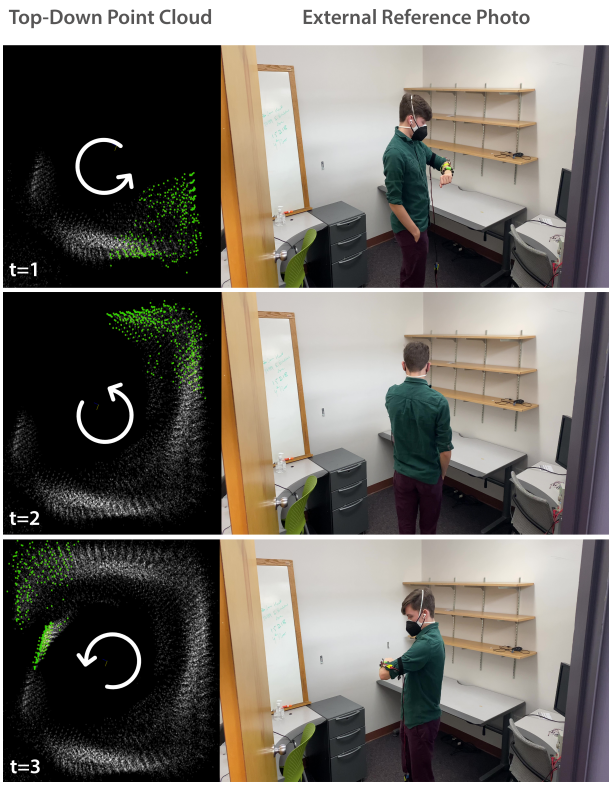

Figure 14: As the user turns, a higher-resolution, multi-frame point cloud of the office is built up. Note the ajar door to the left is captured.

7.2 Ad Hoc Touch Tracking

Our band's hand-facing depth sensors not only capture the hand, but also objects and surfaces in front of the hand. These surfaces could be opportunistically appropriated for ad hoc touch input [22]. To explore this opportunity, we created a functional application that can detect touches and basic 2D gestures on ad hoc surfaces. To detect everyday planar surfaces (uninstrumented walls, tables, countertops, etc.), we run Optimal RANSAC [23] on the point cloud (ignoring points further than 50 cm away). This process identifies and fits a plane to the point cloud, if one exists. For "click" detection, we extend a ray from the band and test for collision with the fitted plane at the tip of the index finger (a distance we measure per user, but which could be captured once in a setup wizard). We note a more advanced version of our software would detect a pointing gesture using our hand pose pipeline, but we did not combine these two processes. Finally, once the user has "clicked" a surface, our band tracks in-plane 2D gestures using its IMU (see Video Figure).

7.3 Held Objects

Our eight hand-facing depth sensors are not only capable of imaging hand pose, but also objects held in the hand. Figure 15 includes example point clouds for a coffee cup, water bottle, notebook, duffle bag, grocery bag, umbrella and door handle. We threshold depth values further than 1 m, as these are unlikely to be an object held in the hand.

The low-resolution nature of our point clouds will naturally lead to confusion among objects if only geometry is leveraged for classification. However, machine learning could utilize two other important pieces of information. First is distance the object appears from the wrist, and the second is absolute hand orientation. For instance, an umbrella is not only distinctive for its large curved surface, but also the fact it operates 1 meter away and above the user. Arm pose data, which our system excels at, could also prove helpful in disambiguating some generic object geometries, such as a user bringing a cylinder-like object to their mouth (i.e., a bottled beverage).

7.4 Environments

While held-object and bimanual activity detection operate in the near field, we can also use DiscoBand to image the far field for environment detection and scanning. As noted previously, our current prototype can sense surfaces up to 4 m away.

Although a single frame of depth data does not typically allow for fine-grained geometric details to be resolved, some environments are distinctive. For example, a car interior is typified by a small enclosed volume, in which the hands operate in front of the body. Further, the presence of a vertical surface to the immediate left of a user would suggest they are in the driver seat (in countries with left-hand drive), which could be a useful contextual clue for a smart assistant (e.g., managing cognitive load). Figure 16 offers an example sequence of single-frame depth data, where the geometry of a tabletop and then laptop become visible. Even more interesting is when multiple point clouds are combined over time to build up a fine-grained 3D model of the environment. As a simple demo, we used our arm pose tracking (which provides 6-DOF wrist orientation data) to layer multiple point clouds. Figure 14 offers an example scene. Due to the nearly 360-degree nature of our sensing frustums, it is also possible to perform scans like this from a variety of arm poses, not just the one used in Figure 14. We note that more advanced multi-frame registration methods (e.g., SLAM [13]) would provide even better output, an implementation we leave to future work. Finally, we also speculate that a detailed model of the ground could be captured simply by the arms naturally swinging during locomotion, which could have assistive uses for those with impaired vision.

7.5 Body Scanning

Finally, DiscoBand is also capable of opportunistically scanning the user's body. The arms naturally sway in front and behind the user as they move about, providing multiple view points of the body. Like environment scanning, the point clouds can be correlated and overlaid to build up a 3D scan of the torso. Such a feature could be used to detect clothing, such as jackets, or potentially tracked over time for fitness and health applications.

8 LIMITATIONS

We believe our approach and proof-of-concept implementation demonstrates significant promise, but there are nonetheless important limitations worth discussing. Chief among these is the

Figure 15: Example point clouds of various hand-held objects as captured by DiscoBand's eight hand-facing depth sensors. For illustration, the real point clouds have been rotated to match a greyscale reference photo.

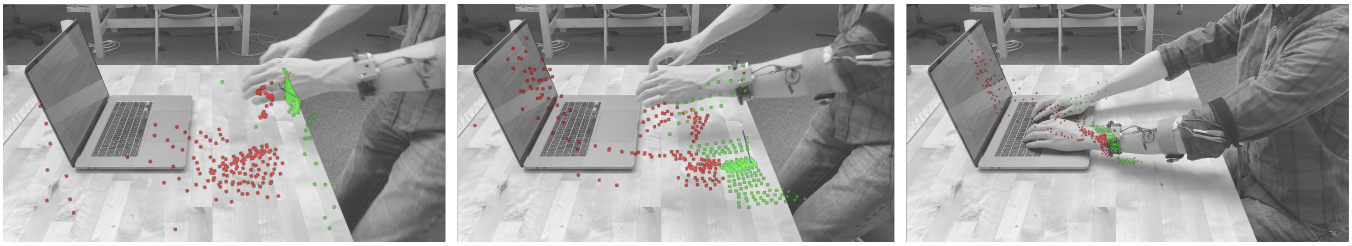

Figure 16: Example single-frame point clouds as the user approaches a laptop to begin work. Green and red point clouds are from outward- and hand-facing depth sensors, respectively. In the right two examples, the hand-facing sensors capture the geometry of the keyboard and screen of the laptop.

outstanding challenge of hand pose estimation in the face of significant hand self-occlusion. This issue is inherent to all wrist-borne direct sensing systems, and severity of the effect only increases as the sensors operate ever-closer to the skin. DiscoBand's multi-view approach partially mitigates this issue, and superior depth sensor resolution and machine learning models could yield further accuracy gains. Further, a multi-modal sensing approach, combining multi-view depth sensing with non-line-of-sight methods such as EIT [75] and EMG [29, 32], could prove successful as they are robust in orthogonal ways [7].

Like almost all worn sensing systems, our band achieved its best hand pose accuracy when trained and tested using within-worn-session data (MPJPE of 11.69 mm). This result is not surprising, but it is also unrealistic, as consumers cannot be expected to re-calibrate their device each time it is worn. When we look at cross-session hand pose accuracy (i.e., where the band is removed and later re-worn), MPJPE increased to 17.87 mm, which is a more realistic appraisal of our band's pose accuracy. Even more challenging for the pose model is when it is only trained on data from other people, but never on its own user. This simulates "out-of-the-box" accuracy. Like most worn systems, across-user accuracy was DiscoBand's worst performing train/test condition. However, we were surprised to see that it was competitive with our cross-session performance. We suspect this is because the model was trained on more data (nine users instead of one), and hypothesize that across-user accuracy could improve further with a larger corpus.

Our lab study also has some limitations, First, it was conducted exclusively indoors, as the one-off hardware was tethered and reasonably fragile. Although we did not formally test our system outdoors, it is expected that IR interference from direct sunlight would impact performance (though we note that time-of-flight depth sensors can operate in direct sunlight, unlike structured light approaches). Second, with regards to the biometric calibration data we collected from our participants at the beginning of the study, we would hope to eliminate this in the future. Instead of manual input of body data, it may be possible to conduct a one-time, automatic calibration where hand and arm length could be determined by performing known poses.

We also note that our wrist band is very much a proof-of-concept implementation, and several engineering challenges would have to be overcome in a consumer version. One issue is cost – our band cost roughly $200 to build (as a one-off prototype). While each VL53L5CX sensor only costs around $6 USD in large volumes, we utilize sixteen of these sensors that drives up the bill of materials. Perhaps in very large volumes, economies of scale could reduce the price. Alternatively, future iterations of the sensor could increase the field of view and sensor resolution, such that that the number of sensors could be halved.

Power consumption is also a significant challenge, especially in a wrist-worn device where battery size is constrained. At full duty cycle, our present band consumes 3.6 W (which we note is lower than comparable systems such as FingerTrak [25]). For this reason, we completed much of our development and testing with the band tethered to a power source (runtime was 3.7 hours when using a 4 Ah LiPo battery). To make our approach compatible with mobile device power budgets, the sensing would have to be made intermittent, sampling the scene only occasionally. It could wake and run at full speed opportunistically, perhaps triggered by an application or IMU data (similar to raise-to-wake functionality in contemporary smartwatches). Fortunately, the time-of-flight depth

sensors we employ are seeing continuous improvements. Indeed, the next generation of the sensors has already been announced by STMicroelectronics [58, 59] and will consume 50% less power than the components we used (while also offering higher resolution).

Finally, in this paper, we described seven example uses for DiscoBand: arm pose, hand pose, bimanual tracking, ad hoc touch tracking, held object detection, environment scanning, and body scanning. We believe this long list highlights the generalizability of our approach and the long tail of applications it could enable. However, our explorations were not exhaustive, and we hope others will join us in applying this general approach to new application domains, especially those that were not previously possible in a mobile context.

9 CONCLUSION

We have presented our work on DiscoBand, a novel sensing wristband utilizing a distributed array of ultra-small depth cameras to digitize a user's hand and upper body pose. Our depth sensors also image the environment around a user, enabling additional uses cases. Our approach allows for a uniquely thin form factor (<1 cm), which could permit integration into future worm consumer electronics. Further, our low-resolution point clouds are much more privacy preserving than comparable camera-based systems. In our user study, we found that DiscoBand can track the six upper body keypoints with a mean joint positional error of 5.88 cm. For hand pose, we found a mean joint positional error of 1.70 cm. Overall, we believe DiscoBand is a powerful enabling technology, and we describe five additional application areas enabled by our approach that we hope to more fully explore in the future.

REFERENCES

[1] Riku Arakawa, Azumi Maekawa, Zendai Kashino, and Masahiko Inami. 2020. Hand with Sensing Sphere: Body-Centered Spatial Interactions with a Hand-Worn Spherical Camera. In *Symposium on Spatial User Interaction*. ACM, Virtual Event Canada, 1–10. https://doi.org/10.1145/3385959.3418450

[2] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. https://doi.org/10.48550/ARXIV.2006.10204

[3] Yannick Bernaerts, Jo Vermeulen, Matthias Druwé, Johannes Schöning, and Sebastiaan Steensels. 2014. The Office Smartwatch – Development and Design of a Smartwatch App to Digitally Augment Interactions in an Office Environment. (2014), 4.

[4] Ali Bigdelou, Loren Schwarz, and Nassir Navab. 2012. An Adaptive Solution for Intra-Operative Gesture-Based Human-Machine Interaction. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (Lisbon, Portugal) *(IUI '12)*. Association for Computing Machinery, New York, NY, USA, 75–84. https://doi.org/10.1145/2166966.2166981

[5] Richard A. Bolt. 1980. Put-that-there: Voice and gesture at the graphics interface. *ACM SIGGRAPH Computer Graphics* 14, 3 (July 1980), 262–270. https://doi.org/10.1145/965105.807503

[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. *arXiv:1812.08008 [cs]* (May 2019). http://arxiv.org/abs/1812.08008 arXiv: 1812.08008.

[7] Maria Claudia F Castro. 2015. Selection of suitable hand gestures for reliable myoelectric human computer interface. (2015), 11.

[8] Feiyu Chen, Jia Deng, Zhibo Pang, Majid Baghaei Nejad, Huayong Yang, and Geng Yang. 2018. Finger Angle-Based Hand Gesture Recognition for Smart Infrastructure Using Wearable Wrist-Worn Camera. *Applied Sciences* 8, 3 (2018). https://doi.org/10.3390/app8030369

[9] Pascal Chiu, Kazuki Takashima, Kazuyuki Fujita, and Yoshifumi Kitamura. 2019. Pursuit Sensing: Extending Hand Tracking Space in Mobile VR Applications. (2019), 5.

[10] T DeFanti and DJ Sandin. 1977. Sayre Glove Final Project Report. *US NEA R60-34-163 Final Project Report* (1977).

[11] Artem Dementyev and Joseph A Paradiso. 2014. WristFlex: low-power gesture input with wrist-worn pressure sensors. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 161–166.

[12] Yujie Dong, Adam Hoover, Jenna Scisco, and Eric Muth. 2015. A New Method for Measuring Meal Intake in Humans via Automated Wrist Motion Tracking. (2015), 25.

[13] Hugh Durrant-Whyte and Tim Bailey. 2006. Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms. (2006), 9.

[14] M. El-Gohary and J. McNames. 2012. Shoulder and Elbow Joint Angle Tracking With Inertial Sensors. *IEEE Transactions on Biomedical Engineering* 59, 9 (Sept. 2012), 2635–2641. https://doi.org/10.1109/TBME.2012.2208750 Conference Name: IEEE Transactions on Biomedical Engineering.

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon) *(KDD'96)*. AAAI Press, 226–231.

[16] F. Gemperle, C. Kasabach, J. Stivoric, M. Bauer, and R. Martin. 1998. Design for wearability. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*. 116–122. https://doi.org/10.1109/ISWC.1998.729537

[17] Jun Gong, Xing-Dong Yang, and Pourang Irani. 2016. WristWhirl: One-Handed Continuous Smartwatch Input Using Wrist Gestures. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan, 2016) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 861–872. https://doi.org/10.1145/2984511.2984563

[18] Jun Gong, Yang Zhang, Xia Zhou, and Xing-Dong Yang. 2017. Pyro: Thumb-Tip Gesture Recognition Using Pyroelectric Infrared Sensing. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 553–563. https://doi.org/10.1145/3126594.3126615

[19] Google. 2020. MediaPipe Hands. https://google.github.io/mediapipe/solutions/hands.html

[20] Google. 2020. MediaPipe Pose. https://google.github.io/mediapipe/solutions/pose.html

[21] Teng Han, Khalad Hasan, Keisuke Nakamura, Randy Gomez, and Pourang Irani. 2017. SoundCraft: Enabling Spatial Interactions on Smartwatches using Hand Generated Acoustics. (2017), 13.

[22] Chris Harrison, Hrvoje Benko, and Andrew D. Wilson. 2011. OmniTouch: wearable multitouch interaction everywhere. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*. ACM Press, Santa Barbara, California, USA, 441. https://doi.org/10.1145/2047196.2047255

[23] Anders Hast, Johan Nysjö, and Andrea Marchetti. 2013. Optimal RANSAC - Towards a Repeatable Algorithm for Finding the Optimal Set. *Journal of WSCG* 21 (2013), 10.

[24] 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)*. 1304–1308. https://doi.org/10.1109/ICIP42928.2021.9506043

[25] 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 (June 2020), 71:1–71:24. https://doi.org/10.1145/3397306

[26] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–10. https://doi.org/10.1145/3290605.3300245

[27] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300506

[28] P. Jung, G. Lim, S. Kim, and K. Kong. 2015. A Wearable Gesture Recognition Device for Detecting Muscular Activities Based on Air-Pressure Sensors. *IEEE Transactions on Industrial Informatics* 11, 2 (April 2015), 485–494. https://doi.org/10.1109/TII.2015.2405413 Conference Name: IEEE Transactions on Industrial Informatics.

[29] Frederic Kerber, Michael Puhl, and Antonio Krüger. 2017. User-Independent Real-Time Hand Gesture Recognition Based on Surface Electromyography. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 36, 7 pages. https://doi.org/10.1145/3098279.3098553

[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*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/2380116.2380139

[31] Jung In Koh, Josh Cherian, Paul Taele, and Tracy Hammond. 2019. Developing a Hand Gesture Recognition System for Mapping Symbolic Hand Gestures to Analogous Emojis in Computer-Mediated Communication. *ACM Trans. Interact. Intell. Syst.* 9, 1, Article 6 (mar 2019), 35 pages. https://doi.org/10.1145/3297277

[32] Thalmic Labs. 2016. Unveiling the Final Design of the Myo™ Armband. https://medium.com/thalmic/unveiling-the-final-design-of-the-myo-armband-10576c0ae95b

[33] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, Berlin Germany, 213–224. https://doi.org/10.1145/3242587.3242609

[34] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. (2019), 13.

[35] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, Tokyo Japan, 321–333. https://doi.org/10.1145/2984511.2984582

[36] Jhe-Wei Lin, Chiuan Wang, Yi Yao Huang, Kuan-Ting Chou, Hsuan-Yu Chen, Wei-Luan Tseng, and Mike Y. Chen. 2015. BackHand: Sensing Hand Gestures via Back of the Hand. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 557–564. https://doi.org/10.1145/2807442.2807462

[37] Shu-Yang Lin, Chao-Huai Su, Kai-Yin Cheng, Rong-Hao Liang, Tzu-Hao Kuo, and Bing-Yu Chen. 2011. Pub - point upon body: exploring eyes-free interaction and methods on an arm. In *Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST '11)*. Association for Computing Machinery, New York, NY, USA, 481–488. https://doi.org/10.1145/2047196.2047259

[38] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time Arm Skeleton Tracking and Gesture Inference Tolerant to Missing Wearable Sensors. (2019), 13.

[39] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D Hand Pose Tracking Using EMG Wearables. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1471–1482. https://doi.org/10.1145/3442381.3449890

[40] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. https://doi.org/10.48550/ARXIV.1906.08172

[41] Jess McIntosh, Asier Marzo, and Mike Fraser. 2017. SensIR: Detecting Hand Gestures with a Wearable Bracelet Using Infrared Transmission and Reflection. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 593–597. https://doi.org/10.1145/3126594.3126604

[42] Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. 2017. EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1923–1934. https://doi.org/10.1145/3025453.3025807

[43] Adiyan Mujibiya, Xiang Cao, Desney S. Tan, Dan Morris, Shwetak N. Patel, and Jun Rekimoto. 2013. The Sound of Touch: On-Body Touch and Gesture Sensing Based on Transdermal Ultrasound Propagation. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces* (St. Andrews, Scotland, United Kingdom) *(ITS '13)*. Association for Computing Machinery, New York, NY, USA, 189–198. https://doi.org/10.1145/2512349.2512821

[44] Santiago Ortega-Avila, Bogdana Rakova, Sajid Sadi, and Pranav Mistry. 2015. Non-invasive optical detection of hand gestures. In *Proceedings of the 6th Augmented Human International Conference (AH '15)*. Association for Computing Machinery, New York, NY, USA, 179–180. https://doi.org/10.1145/2735711.2735801

[45] Polhemus. 2020. Polhemus motion tracking. https://polhemus.com

[46] Nissanka B. Priyantha, Anit Chakraborty, and Hari Balakrishnan. 2000. The Cricket Location-Support System. In *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking* (Boston, Massachusetts, USA) *(MobiCom '00)*. Association for Computing Machinery, New York, NY, USA, 32–43. https://doi.org/10.1145/345910.345917

[47] F. H. Raab, E. B. Blood, T. O. Steiner, and H. R. Jones. 1979. Magnetic Position and Orientation Tracking System. *IEEE Trans. Aerospace Electron. Systems* AES-15, 5 (Sept. 1979), 709–718. https://doi.org/10.1109/TAES.1979.308860 Conference Name: IEEE Transactions on Aerospace and Electronic Systems.

[48] Blaine Reeder. 2016. Health at hand: A systematic review of smart watch uses for health and wellness. *Journal of Biomedical Informatics* (2016), 8.

[49] J. Rekimoto. 2001. GestureWrist and GesturePad: unobtrusive wearable interaction devices. In *Proceedings Fifth International Symposium on Wearable Computers*. 21–27. https://doi.org/10.1109/ISWC.2001.962092

[50] K. Martin Sagayam and D. Jude Hemanth. 2017. Hand posture and gesture recognition techniques for virtual reality applications: a survey. *Virtual Reality* 21, 2 (June 2017), 91–107. https://doi.org/10.1007/s10055-016-0301-0

[51] Farshid Salemi Parizi, Wolf Kienzle, Eric Whitmire, Aakar Gupta, and Hrvoje Benko. 2021. RotoWrist: Continuous Infrared Wrist Angle Tracking using a Wristband. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. ACM, Osaka Japan, 1–11. https://doi.org/10.1145/3489849.3489886

[52] Matthias Seuter, Alexandra Pollock, Gernot Bauer, and Christian Kray. 2020. Recognizing Running Movement Changes with Quaternions on a Sports Watch. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 151 (dec 2020), 18 pages. https://doi.org/10.1145/3432197

[53] 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 (MobiSys '16)*. Association for Computing Machinery, New York, NY, USA, 85–96. https://doi.org/10.1145/2906388.2906407

[54] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K. Hodgins. 2011. Motion Capture from Body-Mounted Cameras. In *ACM SIGGRAPH 2011 Papers* (Vancouver, British Columbia, Canada) *(SIGGRAPH '11)*. Association for Computing Machinery, New York, NY, USA, Article 31, 10 pages. https://doi.org/10.1145/1964921.1964926

[55] Nabeel Siddiqui and Rosa H. M. Chan. 2020. Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist. *PLOS ONE* 15, 1 (01 2020), 1–12. https://doi.org/10.1371/journal.pone.0227039

[56] Srinath Sridhar, Anders Markussen, Antti Oulasvirta, Christian Theobalt, and Sebastian Boring. 2017. WatchSense: On- and Above-Skin Input Sensing through a Wearable Depth Sensor. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3891–3902. https://doi.org/10.1145/3025453.3026005

[57] T. Starner, J. Auxier, D. Ashbrook, and M. Gandy. 2000. The gesture pendant: a self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In *Digest of Papers. Fourth International Symposium on Wearable Computers*. 87–94. https://doi.org/10.1109/ISWC.2000.888469

[58] STMicroelectronics. 2020. STMicroelectronics Introduces World's First All-in-One, Multi-Zone, Direct Time-of-Flight Module. https://newsroom.st.com/media-center/press-item.html/p4281.html

[59] STMicroelectronics. 2022. 2nd-generation multi-zone direct Time-of-Flight sensor from STMicroelectronics uses less energy and can range 2X as far as existing products. https://www.globenewswire.com/news-release/2022/06/09/2459735/0/en/2nd-generation-multi-zone-direct-Time-of-Flight-sensor-from-STMicroelectronics-uses-less-energy-and-can-range-2X-as-far-as-existing-products.html

[60] David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. MM-Fit: Multimodal Deep Learning for Automatic Exercise Logging across Sensing Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 168:1–168:22. https://doi.org/10.1145/3432701

[61] Y. Sugiura, F. Nakamura, W. Kawai, T. Kikuchi, and M. Sugimoto. 2017. Behind the palm: Hand gesture recognition through measuring skin deformation on back of hand by using optical sensors. In *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. 1082–1087. https://doi.org/10.23919/SICE.2017.8105457

[62] Wei Sun, Franklin Mingzhe Li, Congshu Huang, Zhenyu Lei, Benjamin Steeper, Songyun Tao, Feng Tian, and Cheng Zhang. 2021. ThumbTrak: Recognizing Micro-finger Poses Using a Ring with Proximity Sensing. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. ACM, Toulouse & Virtual France, 1–9. https://doi.org/10.1145/3447526.3472060

[63] Hoang Truong, Shuo Zhang, Ufuk Muncuk, Phuc Nguyen, Nam Bui, Anh Nguyen, Qin Lv, Kaushik Chowdhury, Thang Dinh, and Tam Vu. 2018. CapBand: Battery-Free Successive Capacitance Sensing Wristband for Hand Gesture Recognition. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems* (Shenzhen, China) *(SenSys '18)*. Association for Computing Machinery, New York, NY, USA, 54–67. https://doi.org/10.1145/3274783.3274854

[64] Vicon. 2020. Vicon | Award Winning Motion Capture Systems. https://www.vicon.com

[65] Richard Voyles, Jaewook Bae, and Roy Godzdanker. 2008. The gestural joystick and the efficacy of the path tortuosity metric for human/robot interaction. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems (PerMIS '08)*. Association for Computing Machinery, New York, NY, USA, 91–97. https://doi.org/10.1145/1774674.1774689

[66] Cheng-Yao Wang, Min-Chieh Hsiu, Po-Tsung Chiu, Chiao-Hui Chang, Liwei Chan, Bing-Yu Chen, and Mike Y. Chen. 2015. PalmGesture: Using Palms as Gesture Interfaces for Eyes-free Input. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. Association for Computing Machinery, New York, NY, USA, 217–226. https://doi.org/10.1145/2785830.2785885

[67] Hongyi Wen, Julian Ramos Rojas, and Anind K. Dey. 2016. Serendipity: Finger Gesture Recognition using an Off-the-Shelf Smartwatch. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3847–3851. https://doi.org/10.1145/2858036.2858466

[68] 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 (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1147–1160. https://doi.org/10.1145/3379337.3415897

[69] Xsens. 2020. Xsens Motion Capture. https://www.xsens.com/motion-capture

[70] Chao Xu, Parth H. Pathak, and Prasant Mohapatra. 2015. Finger-Writing with Smartwatch: A Case for Finger and Hand Gesture Recognition Using Smartwatch. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications* (Santa Fe, New Mexico, USA) *(HotMobile '15)*. Association for Computing Machinery, New York, NY, USA, 9–14. https://doi.org/10.1145/2699343.2699350

[71] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E Hudson, Charlie Maalouf, Jax Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 496, 19 pages. https://doi.org/10.1145/3491102.3501904

[72] Yuki Yamato, Yutaro Suzuki, Kodai Sekimori, Buntarou Shizuki, and Shin Takahashi. 2020. Hand Gesture Interaction with a Low-Resolution Infrared Image Sensor on an Inner Wrist. (2020), 5.

[73] Hui-Shyong Yeo, Erwin Wu, Juyoung Lee, Aaron Quigley, and Hideki Koike. 2019. Opisthenar: Hand Poses and Finger Tapping Recognition by Observing Back of Hand Using Embedded Wrist Camera. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 963–971. https://doi.org/10.1145/3332165.3347867

[74] Yafeng Yin, Lei Xie, Tao Gu, Yijia Lu, and Sanglu Lu. 2019. AirContour: Building Contour-Based Model for In-Air Writing Gesture Recognition. *ACM Trans. Sen. Netw.* 15, 4, Article 44 (oct 2019), 25 pages. https://doi.org/10.1145/3343855

[75] Yang Zhang and Chris Harrison. 2015. Tomo: Wearable, Low-Cost Electrical Impedance Tomography for Hand Gesture Recognition. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, Charlotte NC USA, 167–173. https://doi.org/10.1145/2807442.2807480

[76] Hongyang Zhao. 2018. MobiGesture_ Mobility-aware hand gesture recognition for healthcare. *Smart Health* (2018), 15.

[77] Thomas G Zimmerman, Jaron Lanier, Chuck Blanchard, Steve Bryson, and Young Harvill. 1987. A HAND GESTURE INTERFACE DEVICE. (1987), 4.