Carnegie Mellon University

PhD Thesis

Practical Privacy Preserving Ambient Sensing

by
Rushil Khurana

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

CMU-HCII-21-100
July 30, 2021

Supervisor:
Dr. Mayank Goel (Carnegie Mellon University)
Committee:
Dr. Jodi Forlizzi (Carnegie Mellon University)
Dr. Scott Hudson (Carnegie Mellon University)
Dr. Thomas Ploetz (Georgia Institute of Technology)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Human-Computer Interaction

ABSTRACT

Practical Privacy Preserving Ambient Sensing

We have entered a new age of computing where the computer is tied not only to a person's body but may also be present in their environment. The ambient presence of sensors enables unprecedented opportunities to build smart environments that adapt to user needs, tracks activities, enables interactions and assists the user in their daily tasks. Even though ambient sensing exists, its scale until recently was limited by the available hardware and computing power. And even despite some recent advancements in their capabilities, ambient sensing techniques generally tend to be privacy intrusive.

In this dissertation, I identify key challenges of robust ambient sensing *i.e.* the ability to track users and activities via sensors present in the environment. First, there may be multiple users present in the same environment. The need to build reliable novel approaches that detect multiple users and activities from the same sensor stream and identify each user performing those activities presents a unique technical challenge. Additionally, these machine learning powered techniques require a large amount of training data that posits another challenge of data collection and labeling. Lastly, managing the privacy expectations of all users in a shared environment is a socio-technical challenge that influences the design of those approaches.

In my thesis, I focus on two ambient sensors: cameras and mmWave radar. While mmWave radar is inherently a privacy-preserving sensor, the cameras are regarded as highly intrusive. Thus, I first present a mixed-methods approach to understand the privacy preferences of users for cameras being used as sensors in a range of environments. This work highlights how using privacy preserving techniques to sense activities and clearly communicating how it works may instill trust in a user. Next, I discuss three systems I built that tackle the aforementioned challenges of ambient sens-

ing.

1. The ability to sense multiple activities: I showcase a camera-based exercise detection and tracking system that can sense different exercise types and count the number of repetitions for multiple users at the same time.
2. The ability to identify individual users in the same environment: I present a hybrid camera-imu approach that uses motion correspondence from both modalities to identify individual users in a scene.
3. The ability to collect and label data for new sensors: I discuss a novel domain adaptation approach that leverages existing labeled IMU datasets to train mmWave radar sensor for activity recognition.

I have also conducted appropriate evaluations in unconstrained and semi-constrained environments to underscore the practicality of these approaches. Finally, I also outline how all systems tackle a different challenge of ambient sensing and their impact on the privacy of the user.

ACKNOWLEDGEMENTS

I was blessed to have the support and guidance of some incredible people during my journey as a Ph.D. student. I want to express my sincere gratitude to all of you for being with me for the last few years.

Mentors And Collaborators

First, I would like to thank Dr. Mayank Goel for being an incredible mentor, advisor who used his knowledge and expertise to guide me through this journey. He has always been supportive of my ideas, encouraged me to take risks and helped me grow as an independent researcher. I would also like to thank my committee members- Dr. Jodi Forlizzi, Dr. Scott Hudson, and Dr. Thomas Ploetz. They were a valuable resource and provided constructive feedback to help finish my thesis in a timely manner.

My work would not be possible without my incredible collaborators. Julian Ramos, Kareem Bedri and Karan Ahuja have not just been excellent lab mates but were instrumental in helping me push forward my research agenda. Zac Yu, Katelyn Morrison, Sejal Bhalla, Nikola Banovic, Elena Deng and Corentin Dugue are not only talented but a joy to work alongside. I have had some excellent internships with Dr. Kent Lyons, Dr. Steve Hodges, Dr. Jack Yang and Dr. Alanson Sample. They have been excellent mentors and played a vital role in my success.

Family And Friends

I want to thank my family and friends who have always been there and supported me and helped me become who I am today. I would like to thank my parents for their support and care during these years.

I am also thankful to all my friends that made this journey easier for me and filled a part of my life with joy. I want to thank Maria Alejandra

Robles, Anurag Mahajan, Sahil Goyal, Aman Singh, Alexandra To, Cole Gleason, Steven Moore, Michael Madaio, Michal Luria, Franceska Xhakaj, Anurag Maravi, Shreyas Nagare, Sudershan Boovaraghavan, Alex Cabrera, Will Epperson, Joseph Seering, Nathan Hahn, Anhong Guo, Judy Choi, Sauvik Das, Yang Zhang, Kevin El Haddad, Sandeep Acharya, Caio Brito, Vivian Shen, Anouksha Narayan, Steffen Maass, Akanksha Menon, Daksha Yadav, Naman Kohli, Pranav Raj, Aadish Gupta, and Gabriel Reyes. I have been blessed to be surrounded by incredible people who have always been supportive, positive and encouraged me to better myself.

CONTENTS

CHAPTER 1

INTRODUCTION

The famous British writer Arthur C. Clarke articulated three laws regarding the future of technology. His third law [1] states,

"Any sufficiently advanced technology is indistinguishable from magic."

And what makes a magic trick successful is the seamlessness with which it is performed. And as such, the last few decades have seen plethora of research to advance technology to build that seamlessness into everyday lives. From self-driving cars to smartphones, from CCTVs to miniature IoT devices, sensors are ubiquitous. Apart from the newfound ubiquity, powered by state-of-the-art machine learning techniques, these sensors can now transform raw sensor data into usable and intelligent inferences. It makes sensor-laden devices more context-sensitive, adapt to the user's environment, opportunistically capture the scene, and decide what to do with the captured information. To the unfamiliar, this passive computation and sensing is in fact magic.

In my work, I focus on technologies that passively adapt to the user's environment and recognize and track their activities in a seamless manner. There are two prominent philosophies that have guided prior work in this domain. The first uses personal devices that a user may have on their body (*mobile and wearable sensing*), whereas the second uses a device placed in the user's environment (*ambient sensing*).

Here, I make the distinction of using ambient sensors to track singular activities (*e.g.,* using a camera to monitor fall detection in elderly [2]) and tracking activities for multiple users together. The ability of a sensor in the user's environment to track singular activities has been extensively explored in the past. It can be characterized as a building block towards

the next step in practical sensing at scale: tracking multiple activities at the same time. This thesis focuses on solving some of these challenges for ambient sensing to become more practical than it's current state.

Before we dive deeper into the advantages and challenges of ambient sensing, I present a brief overview of prior work in wearable and mobile sensing. I outline its benefits and challenges. It provides a foothold to better understand why ambient sensing has certain advantages over the user-centric wearables approach. Finally, I outline some outstanding challenges of practical ambient sensing and how my thesis contributes in solving them.

1.1 Mobile And Wearable Sensing

Context adaptable applications need to sense rich information with a high fidelity. A common approach to do so is to use the sensors available in a person's own device. This approach has several benefits such as a one-to-one mapping of the device to the user *i.e.,* any technique built to use the personal device for sensing operates under the assumption that all the data is collected for the same primary user. Identification is never an issue, which is in contrast with environmental approaches that need an additional mechanism to determine 'who' in addition to the 'what'. Another big advantage of personal devices is their portability. A smartphone or a smartwatch is present with the user throughout the day, thus moving across different places (environments) throughout the day.

The advent of iPhone in 2007 led to a computing revolution, and within a few years positioned the smartphone as the primary personal device. The wide array of sensors tightly encompassed in a handheld box allows us to sense and recognize a range of activities. From detecting simple activities such as walking, sitting [3] or driving, to more nuanced recognition such as not only that a person is in a vehicle, but also if they are the driver or the passenger [4]. They have also been used extensively in healthcare such as simple step counts, sleep monitoring [5] or to detect Parkinson's disease [6]. In fact, smartphones are so pervasive that they have been used to sense activities at the scale of a crowd to monitor road and traffic conditions[7] and air pollution [8]. Within a span of only few years, the smartphone and the sensing capabilities have co-evolved rapidly.

However, smartphones are not the only personal devices. Recently there

has been a surge in the popularity of wearable devices such as smartwatches and heads-up displays. These on-body devices can be used to detect a myriad of activities such as smart glasses to detect eating [9] or smartwatches for hand-based activities such as typing or brushing teeth [10]. Wrist-worn devices have been particularly popular for healthcare sensing. Historically wearables have been used to track step count and heart rate [11]. But these days, commercial devices dubbed as health and fitness accessories such as Fitbit and the Apple watch are capable of track and collect other kinds of wellness data such as sleep duration and quality. Beyond that, researchers have even used wrist-worn wearables for tracking exercises [12], capture stereotyped movements in children with learning developmental disabilities [13], and tracking a user's smoking habits [14]. Other wearables such as heads up displays have been used as assistive devices for people with Parkinson's [15], and even to improve medical education by providing a first-person view of surgical procedures for medical students [16].

The mere fact that we have been able to leverage personal devices for so many use-cases is astounding. There are numerous benefits, and nobody can deny that personal devices are a powerful source to sense activities. However, they also have certain limitations. The personal devices, especially wearables are limited in their on-body position. It makes it challenging, and sometimes nearly impossible to sense different activities with similar precision. For example, a wrist-worn device is less precise in recognizing activities that do not include the arm that the device is worn on. The restricted position, and the size of personal devices also limits its sensing resolution. Admittedly, different devices are dedicated for different kinds of activities, and excel at those. But even though smartphones may be pervasive, wearables are nowhere near that common. The need for multiple smart devices suited for different roles makes it a costly endeavor for the end user. Even if cost was not an issue, do we really want to instrument every part of the human body with sensors to overcome this challenge of sensing resolution and limited sensing capabilities?

So, until a more unifying approach is developed that uses a single device to sense a wide array of activities- a user must choose between accurately tracking only a small set of activities with a single device, or a person must carry multiple devices on their body to track more activities. This is a limitation of this approach. It points towards a need for a solution that does

not rely on the user and what device they may have, rather is ubiquitous and blends into the environment. Luckily, recent advances in computing such as cheaper and faster hardware, tools for improved computer vision and machine learning have paved a path for practical ambient sensing.

1.2 Ambient Sensing

The challenges borne by mobile and wearable sensors point towards the need for ambient sensing. The second approach for activity recognition relies on using and installing sensors in the user's environment to detect and track their activities. It typically relies on sensors such as cameras, microphones and radars. This approach has several advantages:

1. **Improved Utility:** A sensor placed in the environment is generally more suited to recognize a wide range of human activities. Unlike a personal device, it is not attached to a particular limb and can track all body parts independently.

2. **User Burden:** The individual burden on each user to keep their devices charged, and to remember to carry them along all the time is eliminated. It also requires no instrumentation of the user.

3. **Cost:** It is cost-effective to instrument the environment with one or a handful of sensors than to expect each user to own and carry a personal device.

These advantages make ambient sensing a lucrative approach for activity recognition by opening up the gateway to build smart environments. And, there has been a fair amount of research in this area. The background and related work (Chapter 2) covers the prior work in depth.

We explore ambient sensing from two lenses: privacy and practicality. As we usher into a new era of high computing resources that support ambient sensing, this thesis focuses on the following areas:

1. **Privacy Expectations:** Privacy concerns are a challenge for any kind of sensing approach regardless of modality. However, ambient sensors such as cameras tend to be more invasive. Additionally, the limited control of a user over such sensors that capture a high amount of sensitive information in a foreign environment leads to privacy concerns that need to be understood and addressed.

2. **Reliability:** While wearable and mobile sensors may not suitably detect a wide range of activities, they are able to detect a handful of activities with high reliability and precision. Ambient sensing to track singular activities is able to detect activities with high precision as well. But, the spatio-temporal distinction between different activities of multiple users is still a hard problem.

3. **User Identification:** Modern ambient sensing techniques, especially in shared spaces can passively sense and infer activities, but generally lack the ability to identify individual users in that space. It limits their ability to attribute specific feedback or build interactions that cater to a specific user.

4. **Data Collection to Train a Machine Learning System:** While cameras and microphones provide high sensing resolution, most modern applications use machine learning to draw inferences from the raw data provided by these sensors. The amount of data required to train the system is typically high, thus making it hard to deploy in the real world. Data collection and labeling although a challenge across all approaches; it is slightly easier for mobile and wearable devices where the whole data stream can be attributed with a single label. Ambient sensors such as cameras require a spatial label (*e.g.,* bounding box around a person) and the activity label for its duration; or sensors such as mmWave radar require either user intervention to collect labels or an additional sensor (e.g. cameras) to record the ground truth and generate labels later.

In this thesis, I tackle each of these areas to improve the current state of ambient sensing. My contributions are as follows:

1. I present the results of a mixed-methods user study to understand the privacy perceptions of the user with cameras and various techniques used to sense activities. I discuss how varying levels of privacy preservation offered by different techniques impacts the privacy preference of users.

2. I address the challenge of reliability by developing a novel method to detect activities at scale such as exercises for multiple users using a single camera [17] called GymCam. I use sensing techniques deemed

to be trustworthy by users in the mixed-methods user study. I evaluated this system in an unconstrained environment to demonstrate its robustness and reliability. I also discuss how my approach tackles privacy.

3. Next, I show how a camera-based system such as GymCam may recognize users [18]. A hybrid approach using motion profiles as seen from a camera and a wearable is used to address the challenge of user identification. I discuss how my work empowers users with more control over how they may share their identity with an ambient sensing system thus building a practical approach to privacy preserving user identification.

4. I demonstrate that we can use off-the-shelf smartwatch IMU datasets to train an activity recognition system for mmWave radar sensor with minimally labeled data. I show that despite the lack of extensive datasets for mmWave radar, my domain adaptation approach can be used to build an activity recognition system that can distinguish between 10 activities. My work enables building models for privacy-preserving sensors such as doppler without the data labeling limitation.

CHAPTER 2

BACKGROUND AND RELATED WORK

In this chapter I summarize prior work for the four key challenges of ambient sensing outlined in the introduction. While work in these 4 key challenges has largely been done independently, they are vital pieces to realizing a practical privacy preserving ambient sensing system. Each module (subsection) focuses on one of these challenges and helps ground the reader with a summary to characterize my thesis work appropriately.

First, I summarize prior work that has looked at privacy preferences and concerns with ambient sensors. While my work focuses particularly on camera-based or hybrid systems, these works help lay the groundwork to understand why we should care to build privacy-preserving sensing mechanisms.

With an understanding of the outstanding challenges with privacy with regards to ambient sensing, next I outline some common techniques that have been proposed in the past (both camera and non-camera) to achieve privacy-preserving sensing. This subsection covers the kind of techniques proposed in the past, their limitations and how my presented techniques learn from these prior works.

Similarly, I outline prior work in user identification and point out how there has been very little work in this area to achieve the desired outcome without compromising privacy.

The last subsection examines prior work in using heterogeneous domain adaptation to solve the data collection problem; the last problem before a practical privacy-preserving ambient sensing system can be deployed.

2.1 Privacy Expectations Of Ambient Sensors

In this section, I summarize prior work in understanding privacy expectations when ambient sensors are deployed in an environment. It has been shown that privacy and security practices impact the trustworthiness of systems that leverage ambient sensors [19]. Studies have evaluated a user's privacy perceptions based on the type of data collected [20], the location in which it is collected [21], and who is collecting the data [22].

The studies are usually focused on one environment such as a smart home [23, 24], IoT devices embedded in toys [25], or custom devices with ambient sensors [26] to understand privacy practices employed by the users and their expectations. These works are critical in understanding the privacy perceptions of a user. However, they are generally limited to either one environment or a specific device.

To gain an understanding of the larger context, Emami-Naeini et al. conducted a large-scale survey to understand privacy needs of users with a variety of ambient sensors in different environments [27]. It was one of the first works to capture changes in a sensor's privacy expectations by the user depending on the environment and the data collected. The results of such a study are crucial to understand what ambient sensors may be acceptable in a particular environment. This study was complemented by the work of Apthorpe et al. that captured the privacy perceptions of over 3800 information flows using a wide array of sensors, data types and the conditions in which data is collected [28].

Despite a rich understanding of a user's privacy preferences with different sensors in varying environments, there is a glaring gap in our knowledge. There has been no exploration of how the privacy expectations change based on a user's understanding of the underlying mechanism ultimately used by the sensor. It has been shown that increasing people's awareness about the behavior of a particular sensor can influence their privacy perceptions [29], and sometimes even make them more comfortable [30]. So, a gap that still exists is a better understanding of how the privacy perception of a sensor in different environment would change when the user is aware of the underlying technique being used by the sensor to track their activities.

Next, there have been improvements in multi-object/human detection and tracking in a single video [31], but it comes at the cost of user privacy

as it uses identifying facial features to improve upon the tracking. We look at some existing techniques for user identification as well.

2.2 Ambient Sensing Techniques

Next, we take a look at prior work in ambient sensing. A survey of past techniques is vital to better develop newer methods for ambient sensing. First, I divide the ambient sensing approach into three sub-categories based on the sensor- (1) We take a deeper look at use of cameras, that have primarily been used for environment-centric approaches; and (2) we group non-camera based motion tracking approaches; and (3) techniques that leverage the infrastructure in a space to sense activities. These categories cover the breadth of sensing techniques for activity recognition.

2.2.1 Using Cameras For Activity Recognition

There are several activity recognition systems that focus on detecting singular activities such as fall detection [2, 32], in-home physiotherapy exercises [33, 34] and sleep monitoring [35, 36].

The ideal activity recognition system can robustly learn to detect and track any kind of activity with minimal training. There has been prior research to achieve this lofty goal. Dreher et al. use a depth camera to capture the user's pose and track their hands [37]. They also track objects across the video and construct a scene graph from the object-action relationship as shown in Figure 2.1. They use a graph network classifier to train a system to learn the actions executed by the users with an 86% accuracy. Such a system allows users to demonstrate the activity, and with a couple of hours of training, it can learn to recognize those actions with high precision. Other works have also tried to build actor relation graphs to better understand individual activity for each user in the video, as well as an overall summarization of the group activity [38].

Similarly, Bo et al. use the key idea of preserving temporal information of an action across videos of various lengths[39]. It allows them to capture a video descriptor that captures the important frame features required to recognize actions, but with only a few training parameters. They were able to achieve high accuracy in action recognition using only 8-10

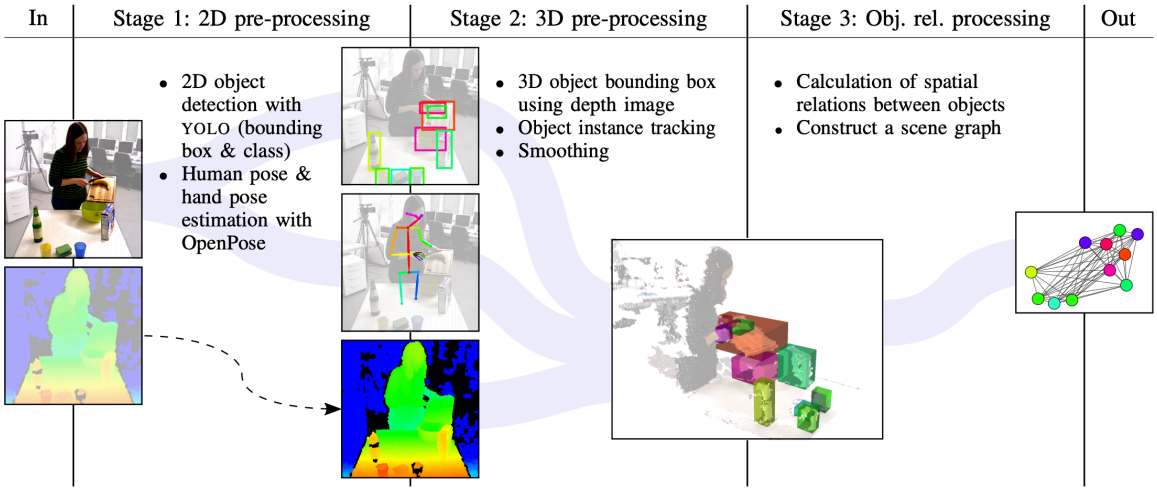


Figure 2.1: System pipeline demonstrating pose tracking with OpenPose, object detection with YOLO [44] in stage 1, tracking the pose and object across frames using depth sensing in stage 2 and building the object-action relationship graph in the final stage.

training samples. It builds and improves upon previous few shot activity recognition systems [40–42].

Others have explored the idea of training activity recognition via demonstration as well, specifically to improve human-robot interactions. Robots in manufacturing industries need to be trained in complex assembly tasks that require them to not only know the objects involved in the process, but also the right sequence of actions to assemble a product. Gu et al. built a portable assembly demonstration system capable of recognizing objects and capturing human motion [43]. These are converted into a symbolic representation. The kinesthetic movements involved in the action together with the symbolic relationship are used to learn the activity demonstrated by the user, to be repeated successfully by the robot.

However, there has been limited exploration of using cameras for activity recognition at scale. One of the key challenges of doing activity recognition at scale is spatiotemporal distinction of one instance of an activity from another.

Kim et al. used depth cameras to obtain pose information and track elderly users in residential homes [45]. They used Hidden Markov Models to detect and recognize activities such as falling down, watching TV, and cooking with an average accuracy of 84.33%. The idea of tracking human poses and building machine learning models to detect activities has been

explored several times in the past [46–49]. But, continued advancements in pose detection techniques [50, 51] and machine learning has led to sustained improvements in pose-based activity recognition techniques.

Caesar [52] uses three cameras to better understand the activities, actions and their interplay for a more complex understanding of the scene. They used seven activities such as approaching another human and handing them stuff or loading the stuff and getting in the car.

The advances in deep learning have led to significant increase in performance for human activity recognition in videos [53], but very few works have extended to multiple users in the same video. Almaadeed et al. used three dimensional convolutional neural network (CNNs) in videos with multiple users to detect activities for each one of them [54]. They detect and track human motion and feed the cropped raw segments of videos for each user into the CNN for activity recognition. While some other works have dealt with multiple users in the same video, they are only capable of classifying and summarizing the video as a whole instead of individual human activities [55].

2.2.2 Non-Camera Based Motion Tracking For Activity Recognition

Akin to the camera, the raw data from 2D lidar can be clustered into human and non-human activities. The motion trajectories obtained from the human activities have been classified into 15 different activities in the kitchen with an accuracy of nearly 99% [56].

A common use case of environmental sensing has been the workplace. Avrahami et al. used a RF radar sensor placed under a surface to recognize work activities in a desk job, convenience store counter, and showrooms [57] as shown in Figure 2.2. They were able to classify common work activities of a cashier such as scanning an item or bagging it with 95% accuracy. Other radar-based sensors such as Doppler have been used to record continuous motions for an individual, which combined with radar cross section and dispersion features are used as inputs to a machine learning model to segment and classify various activities [58].

As stated earlier, an ideal activity recognition system is able to learn different kinds of activities with minimal training. Wu et al. deployed microphones in the environment, and cluster various sounds they hear over

Figure 2.2: The RF radar sensor encased in a 3D printed mount placed under a surface (left). Sensor data as projected into a spherical coordinate system. Image credits: [57]

time [59]. It is a self-supervised model that learns activities over time, and when confident that a set of sounds belong to a singular activity (or cluster), it labels it via a one-shot interaction with the user.

Roy et al. used a combination of ambient sensors and personal devices from users in an environment to recognize complex activities of daily living [60]. They demonstrated that motion sensors mounted on the ceiling combined with data collected from smartphones can be used to effectively recognize activities such as cleaning, cooking or taking medication. Others have used an array of ambient sensors such as motion sensors, temperature sensors, pressure sensors on couches and beds, reed switches on doors and float sensors in bathroom to detect activities of daily living [61–63].

2.2.3 Infrastructure-Mediated Sensing

Another popular approach for privacy-preserving sensing has been leveraging the inherent properties of the infrastructure in the environment to detect, track and recognize various activities.

For example, Wang et al. monitor the change in WiFi signals over time due to reflections from human body and model those changes to detect activities [64]. They are able to distinguish between activities of daily living such as walking, falling and sitting down. WiFi signals are also useful

as a privacy sensitive platform for activity recognition at scale. They can be used to accurately count the number of users present in an environment [65]. And, the variance observed in channel state information from WiFi signals can also be used to distinguish between simple actions such as walking, running and moving hands for multiple users in the same environment [66].

Other work includes using single-point sensing solutions such as microphones [67] and pressure sensors [68, 69]. to track and monitor the water usage of each fixture in the house. Alternatively, researchers have developed specific water activity sensor consisting of a power harvesting circuit and a piezoelectric sensor that can sense similar activities but with less utilization of resources [70]. Similarly pressure sensors mounted on the air filter or the HVAC system typically found in buildings can be used to determine the movement of individuals across different rooms [71].

Furthermore, infrastructure mediated sensing can also be used by creating our own signal and tracking it throughout the house. For example, two modules installed at the extremes of the house that transmit a low frequency signal can be used by receiver tags in different rooms of the house to not only share their location (*e.g.*, object tracking and positioning through the house) but also enable novel interactions [72, 73]. Similarly, another sensor installed on the powerline can use the unique pattern of electric noise generated when devices are in use (*e.g.,* flicking a light switch) to recognize when they are in use [74]. Furthermore, the electromagnetic interference (EMI) signature of these devices is also unique [75] and can be used to build an automated single point sensing event detection system [76].

The characteristic EMI signature generated by these devices not only can be tracked through the aforementioned sensing systems, but can also be used to turn the device into a sensor itself. For example, LightWave monitors the changes in the EMI signature and uses the changes in impedance caused by proximity to a human hand to classify and detect hand gestures [77]. This novel approach allows unmodified light bulbs (infrastructure) to now act as sensors that can recognize human gestures and enable novel interactions. The same principle was also used to demonstrate that the human body can act as an antenna, receive the EM noise present in a user's environment, and can be characterized to detect human gestures on uninstrumented walls across a home [78].

Finally, there has been some exploration into leveraging the physical space and making it interactive too. Wall++ uses patterns of large electrodes painted on the walls of a room using conductive paint to detect user interactions via mutual capacitive sensing, and know which appliances are turned on via airborne EM sensing [79]. This improvement in context awareness allows richer activity recognition at scale within indoor environments. After summarizing the prior work in sensing techniques, I next look at the next challenge of sensing at scale: user identification.

2.3 User Identification

We break down prior research for identifying users into three categories: (1) using facial recognition; (2) using custom hardware; and (3) using video and IMU fusion techniques. We discuss the advantages and shortcomings of each approach.

2.3.1 User Identification By Facial Recognition

As stated earlier, one of the most common biometric based techniques to identify an individual in a group is facial recognition. It has gained prominence with new techniques to improve the quality and robustness of this technology [80, 81]. In fact, with sufficient data, it is robust enough to be used in some countries to identify jaywalkers on a street crossing and fine them [82]. In this section, we look at some use cases where facial recognition has been used in the past. There are several applications of facial recognition ranging from improvements in interactions between human and robots [83] or improving diagnoses by identifying genetic disorders via photos [84]. However, we only focus on prior work that uses facial recognition to determine the identity of the user.

A common use case of such technology has been in the classroom setting. Tang et al. used cameras in an intelligent classroom to detect facial expressions, perform real-time evaluation of student performance and provide feedback to the teacher [85]. Rewari et al. proposed an attendance system to match the faces as seen through the camera with an existing database of faces to automatically mark the presence of an individual [86]. Similarly, Monteiro et al. deployed a facial recognition system in a high school classroom [87]. It compared the faces of students with an image

database collected by the school to mark their attendance and manage access controls to different classrooms.

Similarly, facial recognition is becoming more prevalent to provide access controls to private objects in a public space such as your car [88] or enabling door access [89] in an office. Del Rio et al. examined facial recognition as an approach to build automatic border control gates. Finally, facial recognition has been at the forefront of surveillance systems [90]. In fact, with a growing number of devices with a camera such as drones, prior work has leveraged them to improve the surveillance state [91].

Such systems rely on accurately determining facial landmarks, and thus require high resolution privacy invasive data to be collected. Once, such data is collected- it can be co-opted for use in other systems without a user's explicit consent. This growing lack of privacy is a huge concern for facial recognition-based systems [92].

2.3.2 User Identification Using Custom Hardware

Another prominent technique to identify users in a public environment is to embed custom hardware on user's belongings such as name tags or bags. Mokhtari et al. conducted a comprehensive survey of custom hardware-based user identification techniques between 2000 and 2016 [93]. We present some key highlights and summarize the different techniques used in the past.

ID-Match uses RFID tags worn by people and correlates their motion paths with ones observed from a depth camera [94]. Similarly, EyeFi uses a WiFi chipset embedded next to a camera to capture motion traces. The observed motion from the camera is used to improve the angle of arrival estimates of WiFi packets, which are then used to localize users in a shared space [95]. RF-based fingerprinting has also been used for gait-based identification [96–98].

Other approaches include instrumenting different objects in the environment with custom sensors. Hodges and Pollock used RFID sensors placed on objects such as a coffee maker to fingerprint each user's unique coffee brewing pattern to identify them [99]. An accelerometer attached to different objects has also been shown to be capable of identifying users in a similar manner [100]. Yamada et al. used pressure sensors mounted on a chair to distinguish different users based on the differences in their

seating [101].

Lastly, there has been several prior works that have explored the user of smart floors to identify users. Orr and Abowd used a load cell embedded in a floor tile to measure ground reaction force and build user profiles based on their walking pattern [102]. Ubifloor extended this idea and used switch sensors to extract walking patterns across the whole floor and was able to achieve an identification accuracy of 92% for 10 people [103]. Several floor-based identification techniques have followed since then that use different sensors such as microphones [104], accelerometers [105] and piezoelectric sensors [106]. More recently, a geophone sensor has been used to detect and capture footstep-based vibrations used to identify users. Pan et al. were able to identify ten users with an accuracy of 96% with this approach [107].

Despite their ability to robustly identify users, the most obvious challenge with such techniques is instrumentation of the environment. It not only adds cost, but in some cases (such as a smart floor) may require regular upkeep. Techniques that instrument objects in the environment, and not the user with custom hardware also suffer from privacy control issues. In a shared space, it makes it difficult for users to control the user data/profile collected by these sensors.

2.3.3 Video And IMU Fusion Techniques For User Identification

There has been limited exploration of techniques that have taken the approach of fusing video and accelerometer data for user identification. Teixeira et al. used two participants with a network of cameras and an IMU embedded on the belt of one user. Thy used a probabilistic approach to match the locations from camera with the ones predicted from the inertial sensors. This approach yielded 84% accuracy to distinguish this user from the other in experiments where the users were instructed to generate motion trajectories of 4-5 seconds [108].

There have been drastic improvements in the accuracy. CrossMotion uses a depth camera and the onboard IMU on a smartphone to track and identify users in a video [109]. It was able to localize a user within 7cm with a 99% identification accuracy. However, this technique was also limited to a single device, and was not scaled for multiple users in the same en-

vironment. Cabrera-Quiros et al. built upon this work and used a hybrid approach to identify 19 real users in a video [110]. They used an overhead camera, and a badge worn by the users that contained an accelerometer and a proximity detector. The use of the extra proximity detector allowed them to divide the crowd of users into subgroups and reduce the search space for matching the motion traces of the wearable with the camera. They further demonstrated their robustness by creating virtual users from the 19 real motion traces collected during evaluation and accurately identifying 79% of the virtual users.

These techniques either require more sensors (*e.g.,* proximity detector) or are limited in their evaluation. The goal of our work is to make a deployable solution that can leverage existing infrastructure and evaluate it in a semi-constrained environment.

Keep this in mind, there are two systems that are perhaps the closest to our goal. Masullo et al. improved upon prior work and used only 2D information from a camera with an accelerometer to identify 10 users with an average accuracy of 76% [111]. While their overall accuracy is lower than prior works, they were able to do so with a regular camera and IMU and achieve user identification over short clips of only 3 seconds. Finally, Henschel et al. used the orientation measurements as observed from the IMU and the camera to identify 8 users playing soccer with an accuracy of 91% [112].

We have discussed sensing and user identification methods that are largely based on machine learning methods that typically require a large volume of data. One of the most prominent techniques to overcome the issue of lack of labeled data is domain adaptation; a technique in which labeled data from one modality/domain is used to train another. In the next subsection we look at some examples of the same.

2.4 Heterogeneous Domain Adaptation

Most prior work on domain adaptation assumes that data of different domains are of the same dimensionality or are drawn from the same feature space [113–115]. However, this assumption may not hold for many applications. Consequently, recent work has witnessed a rise in heterogeneous domain adaptation (HDA) techniques, which tackle the incongruity of source and target feature spaces by mapping features into a common

and closer subspace [116, 117], or exploiting the correlations between features [118], or directly transforming data from one domain to the other [119, 120]. Although these approaches have shown promising results, they still suffer from challenges. While mapping features into a predefined subspace may lead to the loss of shareable information, feature translators which attempt to synthesize target data that follows source domain distribution (or vice-versa) are domain-specific and often difficult to be constructed in real-world applications. Moreover, most existing HDA methods simply learn multiple binary classifiers by adopting a one-vs-rest strategy to achieve multi-class classification [121–123]. This hinders the full exploration of the underlying structure among multiple classes in the target domain. Lastly, an additional layer of heterogeneity comes into play when the source and target domains belong to different modalities. Cross-modal domain adaptation approaches have succeeded in transferring knowledge between modalities like vision and sound [124], text and vision [125], vision to inertial data [126] and inertial signals to video [116]. Even though these approaches work well, they are limited by their need for paired, synchronous instances in both domains (*e.g.,* [116]). Despite this limitation, there are some key takeaways from these proposed techniques. Most importantly, that it is possible to robustly knowledge transfer between two different modalities/domains by learning a latent shared representation. In fact, some of these works were even able to use lower dimensional modalities such as IMU to knowledge transfer onto a higher dimensional modality such as videos.

Next, we focus specifically on what is perhaps the closest prior work to my thesis project. The following techniques have used different modalities to train the doppler sensor for learning human activities. Vid2Doppler [127] and Cai et al.'s work in RF sensing [128] uses videos, detects and tracks humans in them, reconstruct a 3D mesh and use them to generate a synthetic signal that can be used to train the doppler sensor. In fact, these approaches have been shown to work robustly and accurately without the need for *labeled* paired synchronous data. However, they have a severe limitation. The use and reconstruction of 3D human pose means that the videos that can be used as a source need to have full human body visible without any occlusion. This significantly reduces the size of publicly available labeled datasets that can be used by these two approaches.

Another very successful approach to reduce the data labeling cost for

the doppler sensor uses audio as a source modality to teach the doppler sensor [129]. This approach converts the doppler spectograms into pseudo-audio representations using a generative adversarial network (GAN) and then uses a pre-existing sound classifier to classify activities. This approach is an improvement over other approaches since it is neither limited by signal occlusion issues, nor does it require paired synchronous layer. However, the system still requires a larger amount of initial data to build a model that can convert the doppler spectogram into its pseudo-representation. They used a dataset of 1109 spectograms across six activities where each spectogram with each sample collected over a period of 5 seconds. Again, the data labeling cost for a wide range of activities inhibits the use of this approach.

Based on our learnings from prior work, it is clear that generating synthetic data or pseudo-representations overall performs better as a technique, but also presents severe limitations. Therefore, in our work, we build on the idea of a shared latent feature subspace that shares the knowledge of the source domain while also preserving the target domain characteristics. We achieve the same using a minimal, asynchronously labeled target dataset which is modeled by a multi-objective optimization learning approach that simultaneously constrains the domain confusion and multi-class classification loss, thus overcoming the majority of the challenges outlined in this section.

CHAPTER 3

MY APPROACH

Activity recognition is a crowded research space, and many sensing systems have been proposed to build the ultimate ubiquitous system. In this thesis, I argue that there are limitations to using personal devices such as smartwatches or smartphones for activity recognition. In response, I present a co-evolution of robust algorithms for ambient sensing and strategies to mitigate common challenges that are unique to sensing activities using sensors and devices in the user's environment.

3.1 Key Characteristics Of Practicality

I identified and focused on two key characteristics of a practical sensing system in this thesis. Arguably, there are several facets of what makes a sensing system practical, but based on prior work and drawing from my own experience, I focused on the following key characteristics in my work while solving the ambient sensing challenges outlined in Chapter 1:

1. **Deployability:** Typically, a lot of the sensing systems are tested in the lab in constrained or near ideal conditions. It shields these approaches from challenges that may occur in an in-the-wild deployment. To achieve **practical sensing at scale**, I focus on deployability of my work by rigorous testing in unconstrained and semi-constrained environments.

2. **Unobtrusiveness:** A practical ambient sensing system would require zero (or minimal) interaction from the user. It includes no calibration steps or initialization gestures. The ideal system should automatically detect, recognize and track user activities with minimal input from

the user. All of my work in this thesis require zero calibration or initiation by the user. The only interaction a user may have to do is to set their privacy preference for identification and sharing data.

3.2 Summary Of Explored Approaches

The most obvious question for my work is what if we could eliminate the limitations of mobile and wearable sensors instead of moving to an environment-centric approach? I sought to answer to the same question. I first started to explore if sensing and interactions can be improved by removing some limitations of personal devices. A smartwatch is inherently limited in its functionality by its restrictive position. It leads to issues such as limited one-handed user interactions and user fatigue for sustained interactions [130]. I remove this limitation by building a detachable smartwatch [131]. The extra ability to rotate and hold the device in different positions opens up several new possibilities as shown in Figure 3.1. One example of such an activity is improved fitness tracking, where a detached watchface can be placed on the 'right' body part while performing a specific exercise. This semi-automated approach leads to an increase in performance and reliable de-

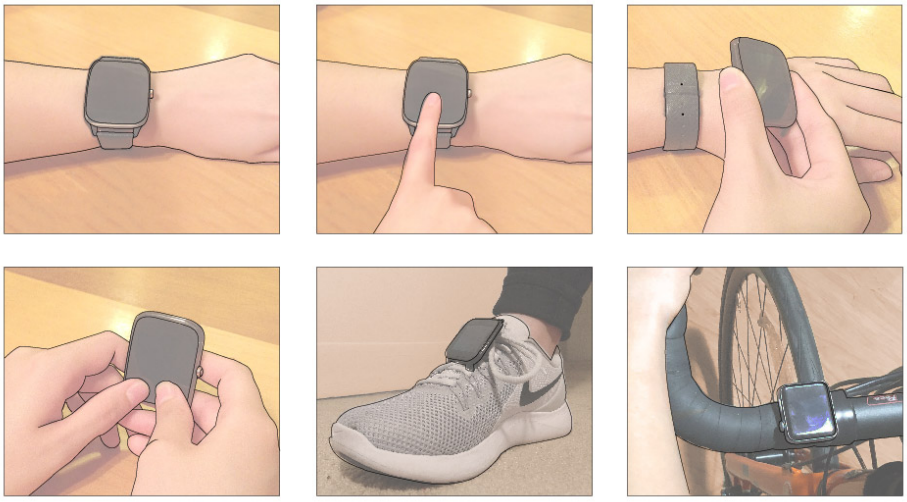


Figure 3.1: (Top Row) In its default state, the watch on wrist can only be used by one hand. If the watchface is then detached: (Bottom Row) Three different uses of a detachable watch are shown; (1) mobility: ability to use the watch in both hands; (2) heterogeneity: the ability to morph into a better sensor like attaching it to the shoe for better gait analysis, and (3) docking on bike for navigation.

tection of the exercise, but it also introduces a new challenge: user burden. The need to put the detached watchface in different locations is cumbersome and another extra step a user has to perform before every exercise. **It makes it obtrusive to use.** There is a trade-off between reliability and user burden in personal device approaches.

I move to ambient sensing and using the same use case of fitness tracking, I demonstrate that ambient sensing approaches can be used to have both high reliability and zero user burden. I present Gymcam [17], a passive exercise tracking and recognition system that uses a single camera to track multiple users at the same time(Figure 3.2). It leverages the insight that almost all repetitive motion in a gym represents some form of exercise. Such motions can be readily captured by a camera, despite heavy occlusion, and used to segment and recognize various simultaneous exercises. I evaluated Gymcam in the university gym over a period of 5 days without any intervention for a truly unconstrained testing. The system also requires no feedback from the user and can automatically detect up to 18 exercises with an accuracy of 85%.

Even though I have demonstrated high reliability in capturing and tracking the exercise, attributing it to a specific person in a room of 100 users is still an outstanding challenge. Next, I worked on MotionID [18] to explore the interplay between environmental sensors and personal devices of users in a space to address user attribution in a privacy-preserving man-

Figure 3.2: Multiple users doing various exercises being tracked using a single camera using Gymcam.

ner. Currently, we are able to identify each user individually in a group of 2-8 people by observing and correlating (just) their motion from a camera and a smartwatch with over 95% accuracy. MotionID was tested across three different activities (volleyball, dancing and poster sessions) with minimal instructions from the researcher in an unconstrained environment across several sessions recorded over multiple days. Despite only subtle differences between users' motion for the same activity, our approach is able to capture these small differences for robust user identification. The user only needs to turn on tracking when they want to be tracked akin to common features such as location sharing on a smartphone.

These systems demonstrate robust algorithms for ambient sensing in an unobtrusive manner. The high performance of both systems also shows resilience to change in environmental conditions.

While both Gymcam and MotionID rely on cameras, I sought out efficient methods to train a large machine learning system without incurring significant labeling cost. Due to its popularity, videos have an extensive library of datasets but the larger problem of training an ambient sensor still persists. Therefore, I shifted my focus to explore how would one train a new ambient sensor that does not share the same luxury of extensive datasets. One such sensor is the doppler radar. It's ability to measure motion in the environment makes it a suitable privacy preserving (by default) alternative to the use of cameras in Gymcam and MotionID. The same functionality can potentially be replicated using a doppler sensor without the need of a information-rich camera. Thus, to solve the problem of data labeling, I showcase IMU2Doppler- an approach to build activity recognition models for the doppler sensor using labeled data from IMU.

I also briefly touch upon the privacy-centric approach taken to develop the sensing algorithms. As stated earlier, privacy is a big concern with ambient sensors such as cameras and microphones. Microphone based products such as Amazon Alexa have improved their privacy using wake words. A system that listens, processes the sounds but does not keep any information until a specific wake word is used. An analogous method for our camera-based systems is a hard problem. It requires the camera system to reliably detect a specific gesture. A subtle and discrete gesture will be hard to detect, whereas a loud gesture may not be socially appropriate. In an effort to still preserve privacy, we incorporated it in our design. Both Gymcam and MotionID use motion profiles of users using optical flow and

pose information respectively. A sensor to capture motion information [1] can replace the camera without any loss of performance. We also provide the user with more control over who tracks them using MotionID. If the data sharing is turned off, the system cannot determine the identity of the user. To better understand how these privacy preserving mechanisms would impact a user's comfort, we first conduct a large scale study to measure the impact of the knowledge of an underlying sensing mechanism on a user's preference.

[1] https://www.racedayquads.com/products/matek-3901-l0x-optical-flow-lidar-sensor

CHAPTER 4

UNDERSTANDING THE IMPACT OF DIFFERENT SENSING TECHNIQUES ON A USER'S PRIVACY PREFERENCES

4.1 Introduction

There has been a meteoric rise in sensors to infer context, recognize human activities and build smarter environments. Faster computation, network resources and smaller form-factors have led to rapid deployment of sensors to build smart environments. Despite a vast literature that explores the privacy preferences of users especially with personal devices [132–134], the privacy expectations and impact of different kinds of sensors in a smart environment (shared or otherwise) have not been investigated as deeply. One of the more information-rich yet privacy divisive sensor is a camera. Cameras traditionally used for security and surveillance, now have been shown to be capable of recognizing context [135] and track a wide range of activities [17, 136, 137]. However, this information-rich sensing capability of a camera is also perceived as a privacy nightmare [138, 139]. While prior work has established the relationship between cameras, data type collected, data sharing practices etc. in different environments [27, 140], one critical piece that has currently not been studied is the impact of the 'knowledge of the underlying activity recognition mechanism' on a user's privacy preferences. Prior work has discussed how privacy concerns in sensing and interaction technologies need to be understood and discussed in detail [141]. And, there have been improvements in sensing and interaction technologies since then too. Newer sensing algorithms employ techniques such as optical flow [142] or pose detection [143] to track user ac-

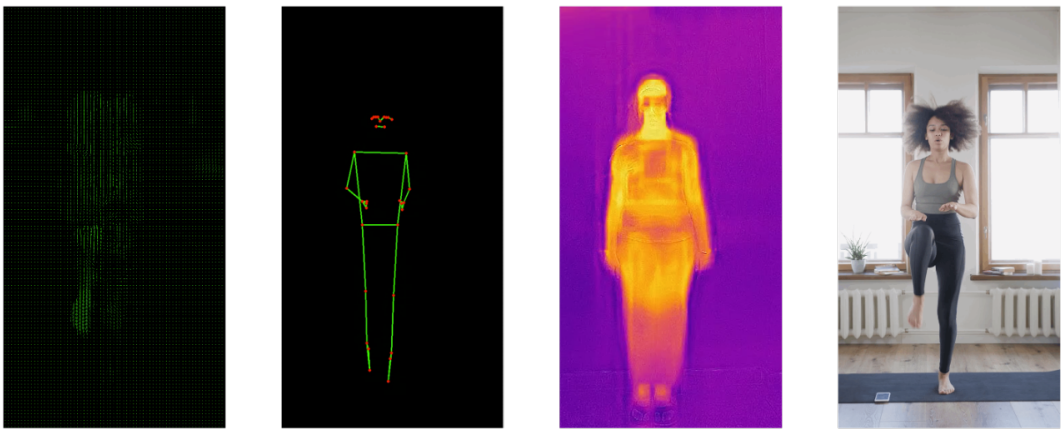


Figure 4.1: Figure showing four different camera sensing techniques: (A) optical flow; (B) pose detection; (C) thermal imaging; and (4) RGB imaging

tivities [17, 144, 145] that obfuscate personally identifying information by default as opposed to prior 'blurring' techniques that don't preserve privacy [146].

In this work, I examine how three camera-based techniques (both software and hardware) that can obfuscate personally identifying information influence the privacy perception of the user. The three techniques used in our work (Figure 4.1) are described below:

1. **optical flow:** This software-based technique allows a camera to track the motion of moving objects in a scene. It tracks a pre-defined number of pixels across the video and tracks movements of objects as they shift between the pixels of the video. It allows a camera to recognize activities based on just the motion without using any identifying information from the scene (*e.g.,* exercises in a gym [17]).

2. **pose detection:** This software-based technique allows a camera to track human bodies by approximating the position of different body parts and reconstructing a body pose (or skeleton figure) from them. OpenPose provides a more detailed explanation of how pose detection works [143].

3. **thermal imaging:** This hardware based technique uses a specialized thermal camera that is capable of recording images using infrared radiation in the range of 1,000nm - 14000nm. Broadly, higher temper-

ature objects emit higher radiation and thus, this imaging technique can be used to capture the variance in temperature across a room.

We employ a mixed-methods approach. We first conduct qualitative interviews with 10 participants to better understand their existing relationship (if any) with cameras in their house. We probed the participants on their current privacy preferences, utility of the sensors and factors that may influence their perception. Next, we demonstrated that three camera-based obfuscation techniques to the participants and recorded their feedback towards the use of those approaches with regards to privacy and utility.

We analyze the semi-structured interviews using thematic analysis [147]. The key results from this first study are:

1. There was an improvement in trust with camera based sensing when optical flow and pose tracking techniques were used. Thermal imaging received mixed response with participant viewing that option as only a marginal improvement.

2. There were two contrasting approaches when a subject purchases a camera; the first prioritizes functionality over privacy whereas the other group takes a privacy-first approach. This trade-off is an important dimension to understand how different sensing techniques will impact privacy preferences.

3. Users had varying levels of comfort with cameras in different environments. A substantial number of participants were okay (and some even expected) to be recorded in public spaces but all participants were privacy conscious in their home. This result corroborates prior work in this area.

Next, we use our results from study one and relevant prior work [23, 27, 28] to design an online vignette study (conducted on mTurk with 633 participants) to identify how the knowledge of the underlying sensing and obfuscation mechanism may influence the privacy perception of uses. Our survey demonstrates that if a user is made aware that a camera is using a privacy preserving sensing mechanism (optical flow, pose detection), then it significantly enhances the trust and comfort of a user with the device regardless of the location.

Our work establishes a new factor that influences the privacy preferences of users, and one that is increasingly becoming more important with

the advent of ambient sensing techniques. With the increase in number of sensors in shared spaces, our work could prove crucial to building trust between the user and the environment, facilitating machine learning powered services and improved user experience.

4.2 Mixed Methods Approach

We took a mixed-methods approach to explore how the knowledge of an underlying sensing mechanism may influence the user's privacy perception of cameras. In our approach, we first conducted a qualitative study to unearth user attitudes towards cameras and validate our hypothesis that knowledge of different sensing techniques does in fact influence a user's privacy perception. The qualitative interview focused mainly on understanding the current privacy attitudes towards cameras as sensors, user practices surrounding this sensors and their comfort with a public camera in a shared space.

Based on our findings, we developed a large scale vignette study to determine the relationship between factors listed in Table 4.1.

4.3 Study 1: Qualitative Interviews

4.3.1 Pilot

We first developed a list of around ten questions to capture the privacy preferences of users with cameras as sensors in their house and shared spaces such as gyms or malls. We conducted the pilot with two participants where one participant was currently using cameras for home security purposes and the other did not employ cameras as sensors in their house. Based on their responses, we refined our questionnaire to include probes about practices users may follow to maintain their privacy. For example, the participant with the camera exclaimed that despite knowing that their video feed was encrypted, they would unplug the camera when they were at home as an additional precautionary measure. Based on the pilot interviews, we also added another question to ask the participants about perceived utility of the three new camera-based techniques shown to them. While a technique might be more privacy-preserving and per-

ceived as safer, it might not be suitable for the specific tasks that the participants were interested in tracking.

4.3.2 Study Procedure

We conducted 10 semi-structured interviews to elicit how participants currently use cameras as sensors in their house. We probed them on the expected utility, their privacy concerns, preferences and current practices. We also probed them on their privacy expectations in shared environments (*e.g.,* gym and mall) where cameras may be used to track certain activities. Next, we showed the participants three camera-based techniques that used for activity recognition but may obfuscate personally identifying information. These techniques are: (1) optical flow; (2) pose detection and (3) thermal imaging. We showed an example video of a person performing activities in the house and the output video if recorded using the aforementioned techniques (Figure 4.1). All techniques were shown together before we asked the participants any questions about their privacy preferences. This parallel prototyping method [148] has been shown to produce better results and feedback from participants.

The interviews were conducted remotely using audio-video conferencing tools. Each interview lasted 40 minutes on average.

4.3.3 Participants

We used snowball sampling to recruit interviewees. We tried to balance participants in age, gender, and their technical expertise. The participants' [5 male, 5 female] age ranged from 22 - 40 (avg. = 29.3, std. dev = 5.94). We conducted an initial screen to assess the participants' security attitudes. SA6 is a self-report survey shown to accurately capture security attitudes [149]. The SA6 score of our participants ranged from 1.3 - 4 (avg. = 3.08, stdev. = 0.81). This is slightly lower than the US average sample of SA6 scores but our range covers participants with varying security attitudes.

4.3.4 Data Collection And Analysis

All interviews were audio recorded and transcribed by the authors. We performed open coding on the transcribed interviews to identify major

themes in the data. We used thematic analysis [147] and identified three major themes that inform not only how our participants perceived the privacy-preserving activity recognition methods, but also their attitudes towards a camera in different environments.

4.3.5 Findings

Based on the thematic analysis from the data collected in this initial interview study, we present three key findings. First, we report our findings on the additional comfort enabled by the knowledge of the three sensing system techniques presented in this study. Next, we discuss how our participants evaluated privacy risk against the utility of the camera for their respective needs. We also discuss the current practices employed by our participants to maintain their privacy comfort. And finally, we discuss the mixed response attitude of our participants towards privacy concerns with cameras and how it varies across different environments.

Impact of Sensing Technique on Privacy Comfort:

3 out of 10 participants were familiar with all of the presented sensing mechanisms shown int he study whereas an additional 3 were familiar with a subset (most commonly, thermal imaging). When asked about their comfort with different techniques, all participants exclaimed that pose detection and optical flow would make them feel more comfortable compared to thermal imaging and the standard information-rich RGB video feed. Thermal imaging received a mixed response from the participants. 4 out of 10 participants stated that while they thought thermal imaging was more privacy preserving than a standard rgb feed, they did not feel that it offered enough 'protection' that they would consider it a significant improvement over the standard video feed. Whereas all 10 participants agreed that the other two mechanisms were a significant improvement in terms of privacy over the standard feed and thermal imaging.

For example, one participant [M, 36] stated, *Yeah, for me, I am most comfortable with the slide 4 [pose detection]. And my second most comfortable would be slide two [optical flow]. [...]. Slide three [thermal imaging] is the same as slide one [raw video feed], because it does capture as much data; because it capture[s] the background, [and] objects as well.*

To further explore this dimension, we also probed the participants about what kind of activities (cooking, if fans/light were left on, exercising, watching kids and elderly etc.) they would wish to track around the house and how useful would such a mechanism would be for some of these activities. Despite not have the technical expertise, the participants were able to grasp the underlying working of each technique from the videos presented to them and were able to differentiate why one technique would provide more utility over another. For example, one participant [M, 35] correctly identifies that while a motion based technique like optical flow would be suitable for tracking if the fan was on, it will not be able to track if the lights were left on. Another participant [M, 27]stated that they would want the full rgb video feed for home security purposes in case they want to identify an intruder, but would like to have the option of switching modes to a privacy-preserving technique when they are at home and use the camera to track home activities. The trade-off between utility and privacy is hard to navigate but there is supporting evidence from our work that the knowledge of the underlying sensing mechanism can improve the privacy comfort of cameras in a house.

Even though comparing perceived utility and privacy is not the focus of this work, it helps us further expand how two similar cameras with differently programmed sensing techniques may be perceived. If a more privacy preserving sensing mechanism (e.g. on-camera pose detection) is able to provide the same utility for specific use cases, it may become a viable option for some. In fact, one user [F, 24] stated the following about using pose detection for exercise tracking (a utility they cared about): *So, I will feel comfortable putting it in a corner where I exercise and not worry about it being turned on all the time.* which is in contrast to the current behavior of our participants.

In fact, 4 out of 10 participants agreed that they would be okay keeping the video on if a more privacy preserving mechanism was being used by the camera.

Privacy vs Utility:

5 out of 10 participants in our study currently owned a camera that they use primarily for home security and as baby monitors. We asked these participants about their decision making process when they purchased

the camera. 3 out of 5 participants reported that they evaluated the utility of the camera service video quality, data storage, easy to use accompanying app etc.) and then looked for common security measures such as encrypted video feeds. The other 2 participants reported taking a privacy-first approach where they sought out tech reviews, tech specs pertaining to security and privacy. Their rationale was that most products catering to a specific need (*e.g.,* baby monitoring) would provide similar functionality. Hence they prioritized sorting their preferences based on the security first. This brings out an interesting dynamic between privacy versus utility. While one group optimizes for utility and is satisfied by an acceptable level of privacy measures; the other group maximizes for privacy and is comfortable with an acceptable level of utility.

Despite this early difference between the purchase-making behavior, all 5 participants reported using additional measures to improve their privacy comfort in the home. All 5 participants reported that they either physically unplug/turn off the camera when they are at home. A common reason cited by the participants was that they don't want to the camera to inadvertently record compromising videos and/or the utility of the camera as a home security system is minimal if the owner is at home.

One user said [F, 29], *"No I physically turn it off when I'm at home, and whenever it's on it's always connected to the internet, because that's how the app interacts with the camera. [The] reason is, there might be some hacking, and someone might gain unauthorized access to the feed, that's the main reason I keep it turned off. Yeah, just to avoid any unauthorized access to when I'm actually at home. That's the main point."*

Another user exclaimed [M, 35], *Yeah, I don't want the camera in my house to see me walking around naked all the time, that sort of thing.*

Despite each participant owning a camera that met their own bar for privacy and security standards, we saw practices such as these that instill an extra sense of comfort in users. This lack of trust is a key finding from our interviews. However, in some cases this impacts the utility of the camera as well. Two participants reported that sometimes they would forget to turn the camera back on before they left the house, making the camera obsolete for security purposes.

I have to remember to turn it on, which is like, 10% of the time actually remember. [M, 27]

This trade-off between privacy and utility is an interesting dimension

that needs to be studied in more detail. However, it is clear that privacy preferences are impacted by the knowledge of the sensing mechanism; and for certain use cases, the knowledge that a camera system is programmed to use a certain technique can factor into the decision-making process for a user's purchase.

Variance of Privacy Preferences Across Environments

Unsurprisingly, it was evident that there was a lot of variance in people's attitude towards cameras depending on the location. 6 out of 10 participants shared that they assume they are always being recorded in public **and** if they are in a public space such as a mall, they do not care if there is a video camera that is recording them. However, upon further probing 4 out of those 6 participants stated that they would not want to be recorded in places such as gyms or workplaces without knowing what the data is being used for. This result is corroborated by prior work in this area [150]. When we further probed these participants about why the location makes a difference, each participant had a different response/reasoning. For example, while one participant [F, 30] was comfortable being recorded while at work; they did not wish the gym to have access to their exercise videos in fear of them being misused for marketing or otherwise. In stark contrast, another participant [M, 24] did not care about them being tracked at the gym but did not want their supervisor to track their every move at work.

The remaining 4 participants said that while they would ideally prefer not to be recorded, the limited control over a public space leaves them helpless. All 10 participants shared that they would at least like to know if a camera is actively recording them; and only 4 participants cared about knowing the purpose of the data collection.

We also probed the participants on their thoughts about receiving smartphone notifications about the presence of a camera if they enter a new space. In fact, 2 participants brought up this idea even before the interviewer could probe them. All participants reacted positively to the idea and 2 participants stated that they would like to receive this notification only the first time they enter a new space whereas the other 8 participants stated that they would like to receive such a notification the first time and every other time there is a 'significant' update that impacts their privacy or consent in any form.

While most participants were comfortable with a camera in a public space, the most important takeaway from these findings is the variance seen in how people react to cameras in different environment. This study also reveals that the underlying sensing mechanisms can help mitigate privacy concerns in different scenarios. For example, knowing that the optical flow technique is only being used to track movement to control lights and/or possible accidents in a workplace; it might mitigate privacy concerns of an employer tracking the employees every single move.

4.4 Study 2: Large Scale Vignette Study

To further confirm our hypothesis at a larger scale, we conducted a within-subjects vignette study on Amazon MTurk with 633 participants. We showed each participant 5 different vignettes with different data collection scenarios. We varied the factors listed in Table 4.1 in the vignettes. Each factor was chosen based on our data from qualitative interviews to determine how an individual's privacy preferences would be impacted by the knowledge of the sensing algorithm; and which algorithm would be more suitable for varying locations.

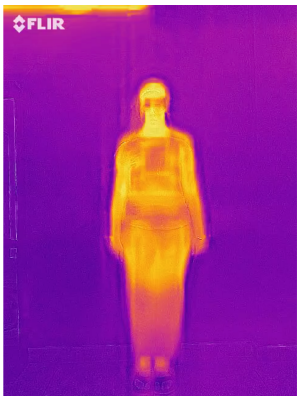
After accepting the MTurk HIT, the participants were directed to a survey link where they were shown the vignettes. The vignettes were structured consistently and followed the same order of factors each time starting with the location and ending with an example view (except the no algorithm scenario). An example vignette is presented below. The variables are shown in bold.

There is a camera in your **workplace**. It **knows** your identity. When you enter your *workplace*, you receive a notification on your smartphone:

*A camera in this space uses **thermal imaging** to track users. An example view of what the camera sees is shown here:*

From all possible combinations of factors in a scenario, we created 8 subsets of 5 scenarios where each subset contained all levels of a factor at least once. Each participant viewed one subset chosen randomly (with even distribution). The ordering of scenarios was also randomized.

For each scenario, the participants were asked if they were comfortable in the scenario (5-point likert scale). We also asked the participants if they would allow the data collection in the described scenario and how



Factors	Levels
Location	(1) own house, (2) friend's house, (3) workplace, (4) mall
Algorithm	(1) raw image (regular rgb camera), (2) thermal imaging, (3) optical flow, (4) pose detection
identity known	(1) knows identity, (2) does not know identity
notification type	(1) notification about presence of sensor with algorithm and image, (2) notification about presence of sensor without algorithm mentioned.

Table 4.1: Table showing all factors and their corresponding levels used in the vignette study

useful was the notification. We also probed the participants on additional questions such as what kind of data would they be comfortable sharing and why. At the end of all scenarios, the participants were asked summary questions. We questioned the participants on how difficult was to understand the camera tracking technique (sensing algorithm) from the notification. We also asked them about what information would they seek from a notification like this. Lastly, we collected demographic information from the participants.

4.5 Results

Our survey was completed by 644 participants. We removed the answers of 11 participants because they took less than 5 minutes to finish the survey whereas the average completion time was 14 minutes. Our results are

	OWN HOUSE	FRIEND'S HOUSE	WORKPLACE	MALL
EXTREMELY UNCOMFORTABLE	14.06%	17.22%	19.91%	16.43%
SOMEWHAT UNCOMFORTABLE	13.90%	25.75%	25.75%	21.48%
NEITHER UNCOMFORTABLE NOR COMFORTABLE	11.70%	13.58%	12.32%	17.06%
SOMEWHAT COMFORTABLE	31.59%	29.55%	29.70%	31.45%
EXTREMELY COMFORTABLE	28.75%	13.90%	12.32%	13.58%

	RAW IMAGE	THERMAL IMAGE	OPTICAL FLOW	POSE DETECTION
EXTREMELY UNCOMFORTABLE	25.27%	16.59%	12.64%	13.11%
SOMEWHAT UNCOMFORTABLE	28.60%	19.27%	20.06%	18.96%
NEITHER UNCOMFORTABLE NOR COMFORTABLE	10.27%	16.27%	13.27%	14.85%
SOMEWHAT COMFORTABLE	22.27%	31.75%	34.12%	34.12%
EXTREMELY COMFORTABLE	13.59%	16.12%	19.91%	18.96%

Figure 4.2: Figure showing participants comfort level across different levels of locations and different levels of sensing algorithms

computed based on the remaining 633 responses. Participants were required to be from United States, have an approval rate greater than 98%, and a minimum of 500 approved HITs. The participants were given $2 for their time. According to our demographics, our survey was taken by 352 males, 269 females, 4 non-binary and 8 unspecified. The median age of our participants was 26 (stdev = 11.68).

In our survey, after describing a scenario (vignette)- we asked the participants how comfortable do they feel in a given scenario. Figure 4.2 shows the general distribution of participants' comfort level across location and sensing algorithm. Unsurprisingly, participants were most comfortable with having a camera in their own home (not factoring the sensing technique). This can be attributed to the control they exert over the device if they own it and the environment it is placed. While participants in our qualitative were more open to being tracked in a mall, as opposed to a workplace; our survey participants rated them almost equally when the sensing technique is not factored in.

Next, our results from the qualitative interviews were substantiated in the vignette study with optical flow and pose detection being the most

	NO ALGORITHM SPECIFIED	RAW IMAGE	THERMAL IMAGE	OPTICAL FLOW	POSE DETECTION
OWN HOUSE	50.59	52.2%	57.24%	66.25%	66.23%
FRIEND'S HOUSE	38.4%	25.9%	44.65%	52.98%	59.24%
WORKPLACE	29.68%	28.02%	48.98%	45.91%	45.18%
MALL	39.49%	37.74%	40.13%	51.2%	50.31%

Figure 4.3: Figure showing participants' comfort level with different sensing techniques in different locations. The value represents percentage of people that chose 4 or 5 on the comfort 5-point likert scale for each of those conditions.

preferred method followed by thermal imaging and the regular rgb camera placing last on the comfort scale (Figure 4.2). This means regardless of the scenario, our participants felt more comfortable in the privacy-preserving sensing techniques that do not capture as much information compared to a regular raw image/video.

Now, we explore the relationship between the various factors in our study. Figure 4.3 outlines the comfort level of participants with different sensing techniques across different locations. We calculated the percentage of participants that chose 4 (somewhat comfortable) and 5(extremely comfortable) on the 5-point likert scale for scenarios that had an interaction between the values presented in the figure (location x algorithm). We can see that all techniques were comfortable for participants when the device was present at their own house. While optical flow and pose detection ranked much higher, even the full rgb image capture was comfortable for at least 50% of the respondents. We attribute this to the privacy control that a user has over the device, its policies and the environment. With the exception of workplace where thermal imaging was the preferred choice (by a small margin), optical flow and pose detection were consistently the most valued options.

We can confirm from these results that knowing the underlying sensing mechanism indeed influences how comfortable a user would be with a device in a particular environment. We also confirm this by comparing it to the condition where the algorithm was not specified. We use a Wilcoxon Rank Sum Test to measure significant differences (if any) between the two conditions. The results are reported in Table 4.2. As evidenced from the

	Raw Imaging / No Algo	Thermal Imaging / No Algo	Optical Flow / No Algo	Pose Detection / No Algo
Own house	p=0.81	p=0.41	**p=0.001**	**p=0.013**
Friend's house	**p=0.007**	p=0.18	**p=0.006**	**p=0.002**
Workplace	p=0.83	**p=0.0002**	**p=0.001**	**p=0.003**
Mall	p=0.32	p=0.33	**p=0.016**	**p=0.014**

Table 4.2: Table showing results of wilcoxon sign-ranked test when a notification with algorithm is compared with the condition where the notification does not contain information about the algorithm.

table, optical flow and pose detection significantly influence the privacy perception of individuals in every location. Combined with the prior results (Figure 4.3 we can infer that the user's comfort level in the camera improves when they are informed about these two specific sensing mechanisms. Interestingly, the notification reduced the trust of users in a camera (rgb condition) when the device is at a friend's house. A possible explanation of this result could be that while users are aware that they are most likely being recorded in public (also supported by our earlier interviews), they may not expect to be recorded at a friend's house. Making users aware of the possibility that a friend may have a device that records them when they are there may reduce their trust in the scenario. However, this possible explanation needs further examination.

We did not find a significant impact of the identity variable in any of the conditions. There was an average increase of 6% in usefulness of the notifications when they contained the description of the algorithm, however the results were not statistically significant. When asked in the summary questions, 76.4% of the participants stated that they were able to understand the camera tracking technique (sensing algorithm) from the notification easily (extremely easy or somewhat easy on the likert scale).

4.6 Discussion & Conclusion

Our mixed methods approach verifies that the knowledge of a sensing algorithm influence a user's privacy perception and may help build trust in certain scenarios. We conducted qualitative interviews with 10 participants and probed them on their current practices and preferences with cameras

in their house and other locations. We also showed them several different sensing techniques and captured their feedback on how it influences their privacy. Finally, we probed them on other factors that impact their privacy preferences.

Some of the factors that came up in our interviews have been studied extensively in prior work [20, 21, 27, 151]. We chose not to include these factors in our follow-up vignette study for two reasons. First, we wanted to ensure lower number of factors in our vignettes to truly capture the relationship between the camera, location and the sensing algorithm. And secondly, the relationship between other factors (except the sensing algorithm) is well-established. We acknowledge that understanding the relationship and interaction effects of these factors with the sensing algorithm is important and would be an important future work.

Our follow-up vignette study with 633 participants provides evidence that knowing about privacy preserving sensing techniques (optical flow, pose detection) can instill trust in the users. Both techniques significantly increased trust in users regardless of the location and scenario.

As the prevalence of IoT and devices increases, we must continue to improve privacy awareness. Our work demonstrates that making users aware of some of the inner workings of a sensing system can massively improve the trust of users in certain scenarios. Our work also used minimal notifications to establish a relationship between the factors studied in this paper. We demonstrate that sensing algorithm is one key part of a privacy-awareness notification in a shared space and is a first step in designing appropriate privacy notifications for ambient sensing.

CHAPTER 5

GYMCAM: DETECTING, RECOGNIZING AND TRACKING SIMULTANEOUS EXERCISES IN UNCONSTRAINED SCENES

5.1 Introduction

Regular physical workout improves well-being and reduces the risk of obesity, diabetes, and hypertension [152–154]. To maintain overall health and build strength, the Centers for Disease Control and Prevention (CDC) recommends adults to strength train at least twice a week[1]. However, despite the benefits of regular exercise, most people struggle to maintain steady progress. This failure is often attributed to lack of motivation and feedback [155–157].

One way to tackle lack of motivation is through gamification and tracking [158]. The ability to view personalized data enhances awareness and enables reflection of exercise regimens [159]. However, capturing and tracking a regimen is challenging. Manual tracking is most accurate, but this is tedious for end users. Thus, numerous commercial and academic efforts have focused on automatically tracking and quantifying physical activity, the most pervasive being step count captured by a worn device (*e.g.,* FitBit, Apple Watch, [160, 161]). Nowadays, consumer devices can track some cardio and strength-training exercises using special applications[2,3]. These ap-

[1]Centers for Disease Control and Prevention. Physical activity recommendations for adults: cdc.gov/physicalactivity/everyone/guidelines/adults.html

[2]Gymaholic: http://www.gymaholic.me

[3]Gymatic: https://itunes.apple.com/us/app/vimofit-auto-exercise-tracker/

Figure 5.1: GymCam uses a camera to track exercises. ***(Top)*** Optical flow tracking motion trajectories of various points in the gym. Green showcases points classified as exercises and red showcases non-exercise points. ***(Bottom Left)*** Individual exercise points are clustered based on similarity to combine points belonging to the same exercise. ***(Bottom Right)*** For each exercise (cluster) GymCam infers the type of exercise and calculates the repetition count.

plications generally rely on a wearable's inertial measurement unit (IMU) to monitor *e.g.,* arm motion as users perform different exercises. Such techniques can be robust for some specific exercises, but fail for many others due to sensor placement where there is limited signal. For example, Maurer *et al.* found that detecting ascending motion such as climbing stairs is more accurate when the IMU is attached to the bag than when attached to a person's shirt [162]. Similarly, data from a smartwatch is inadequate for exercises involving other parts of the body (*e.g.,* leg presses). An alternative is to instrument the exercise machine rather than the user, but that is too intrusive and also makes free-weight and body weight exercises harder to track. This presents a need for a method to robustly identify and track a wide range of exercises that a user might perform, while maintaining the seamlessness offered by wearable devices.

To this end, we present *GymCam* (Figure 5.1), a vision-based system that uses off-the-shelf cameras to automate exercise tracking and provide high-fidelity analytics, such as repetition count, without any user or environment specific training or intervention. Instead of requiring each user in the gym to wear a sensor on their body, *GymCam* is an external single-point sensing solution, *i.e.,* a single camera placed in a gym can track **all** people and exercises simultaneously. One camera-based approach would be to track body motions to detect user pose [163, 164]. However, these techniques are error-prone due to significant occlusion in gym settings (*e.g.,* Figure 5.2). Thus, instead of attempting to accurately estimate body keypoints (*i.e.,* skeletons), GymCam leverages the insight that *almost all repetitive motion in a gym represents some form of exercise*. Such motions can be readily captured by a camera, despite heavy occlusion, and used to segment and recognize various simultaneous exercises. We also found that it is extraordinarily rare for two separate people to exercise at the *exact* same rate and time, allowing for robust segmentation even when users are adjacent.

To develop and evaluate our machine learning algorithms, we collected data in our university's gym for five days. In total, we recorded 42 hours of video and annotated 597 different exercises. We did not record the number of gym users because our protocol required immediate anonymization of the data (*i.e.,* faces blurred). Users of the gym were informed that a research team was recording video, but there was no other interaction with participants, minimizing observer effects (*e.g.,* intentional or unintentional

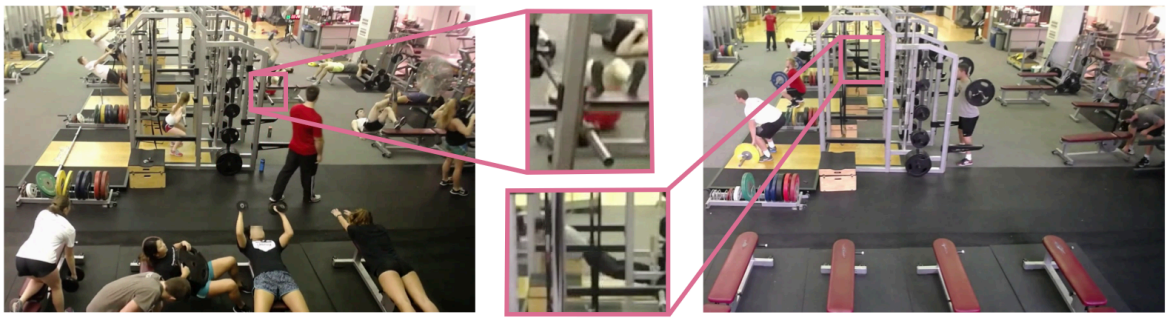


Figure 5.2: In gym settings, user pose can be challenging to determine due to significant occlusion.

changes to their routine). We note this problem often affects research studies where users are aware they are part of an exercise tracking research study, and the evaluation setting is constrained [12]. We believe this paper presents the first truly unconstrained evaluation of exercise tracking.

The overall process of GymCam is as follows:

1. Detect all exercise activities in the scene (acc. = 99.6%), then
2. Disambiguate between simultaneous exercises (acc. = 84.6%), then
3. Estimate repetition counts ($\pm$ 1.7 counts)
4. Recognize common exercise types (acc. = 93.6% for 5 most common exercise types).

5.2 Theory Of Operation

We now discuss the underlying premise behind GymCam that allows it to: (1) detect motion, (2) cluster motions into separate exercises, and (3) identify and track individual exercises.

GymCam leverages the insight that almost all repetitive motion in a gym represents some form of exercise. Even if a camera cannot see an entire person, it is still often able to see a small part of the body exhibiting repetitive motion, and can track that body part, linking it to an exercise later. However, when multiple users are exercising and potentially overlap in a video, it can be hard for camera-based systems to delineate the exact boundaries between the exercises – an issue worn sensors do not have to

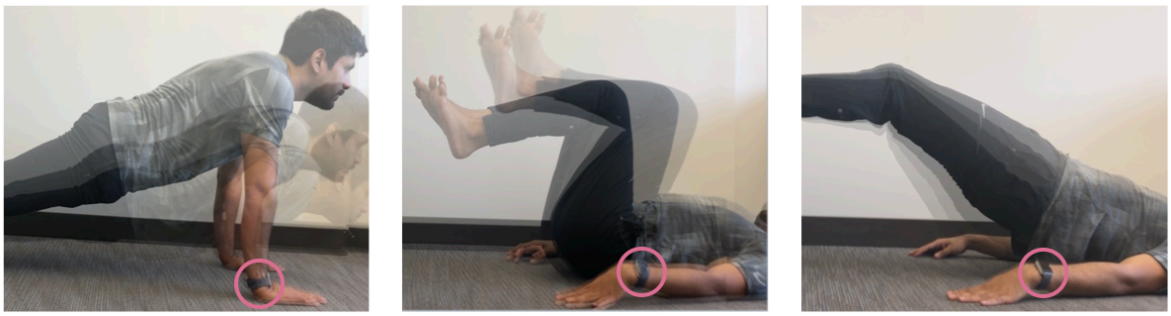


Figure 5.3: A wristworn IMU (circled in the photos) is not ideally positioned to monitor many exercises.

handle. Fortunately, we found it is extremely rare for two users to perform their exercises at exactly same time, speed, and phase (Figure 5.4). Thus, by calculating features that capture these dimensions, GymCam is able to differentiate between simultaneous exercises without any supervised training data.

Apart from distinguishing different users, there are other challenges when relying solely on repetitive motion tracking. Foremost, periodicity can be exhibited by a user's gait or warm up before starting an exercise. Secondly, when placed in an unconstrained environment, users tends to be less deliberate with between-exercise moments (*e.g.* fidgeting, stretching, walking). These interludes can be quite periodic, and thus indistinct from exercises. Moreover, in the unconstrained environment of a gym, users may challenge themselves (*e.g.* lift challenging weights). Morris *et al.* [12] observed that "*self-similarity [or periodicity] may break down in intensive strength-training scenarios. For this reason, more validation of intensive weightlifting is important future work.*" We believe that the only viable approach to solve the problem of variations in exercise and noisy human behavior, is to collect extensive training data in the user's actual workout environment without significant observer effect.

5.3 Data Collection

We collected data in the Carnegie Mellon University's varsity gym over a five-day period.

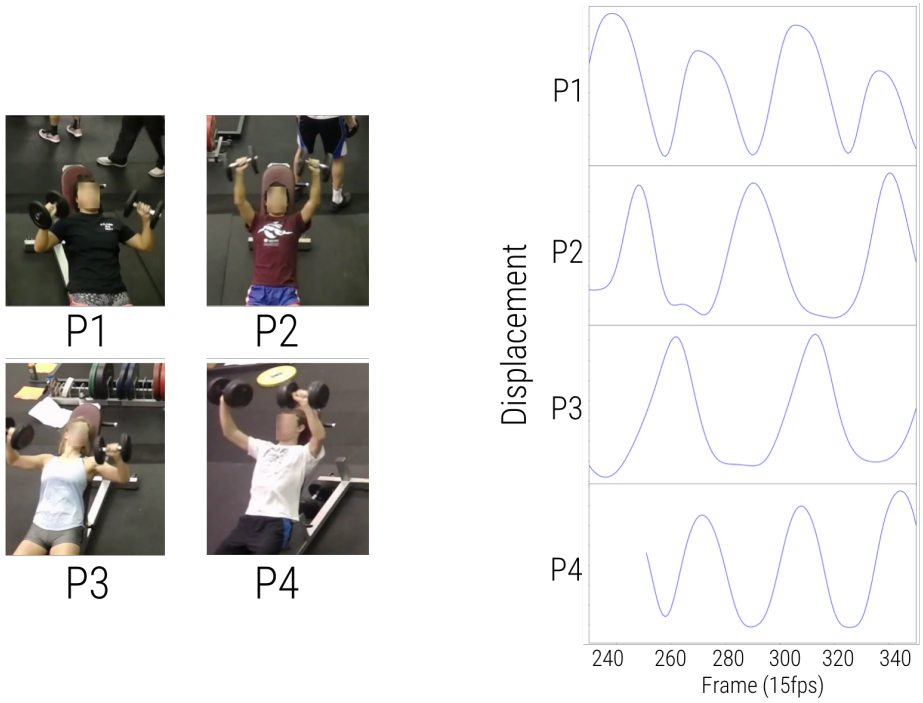


Figure 5.4: Four users performing same exercise at different phase and frequency. *y*-axis shows displacement in terms of pixels.

5.3.1 Participants And Protocol

To ensure a wide, unobstructed view, we placed one camera on a wall at a height of approximately 4 meters. This placement was also inconspicuous, aiming to minimize observer effects (*e.g.*, users altering their warm-up or stretching routine, lifting usual weights). The university's Institutional Review Board and Department of Athletics officials agreed that as long as videos were immediately anonymized, we did not need signed consent from participants. Nonetheless, gym users were informed that a research team was recording anonymized videos and any questions, comments or objections should be raised to the gym staff (though none did). Thus, gym users were given no instructions regarding exercises, repetitions, breaks, etc., and is as close to unconstrained data collection as practically possible.

We used a Logitech C922 camera at a resolution of1920×1080 to record 15 frames per second (fps) video. We used a state-of-the-art face detection algorithm [165] to blur the faces of gym users and anonymize the videos. After dropping periods when the gym was empty, we had 42 hours of data spanning 5 days. We hand annotated 15 hours of this video, which contained 597 exercise instances.

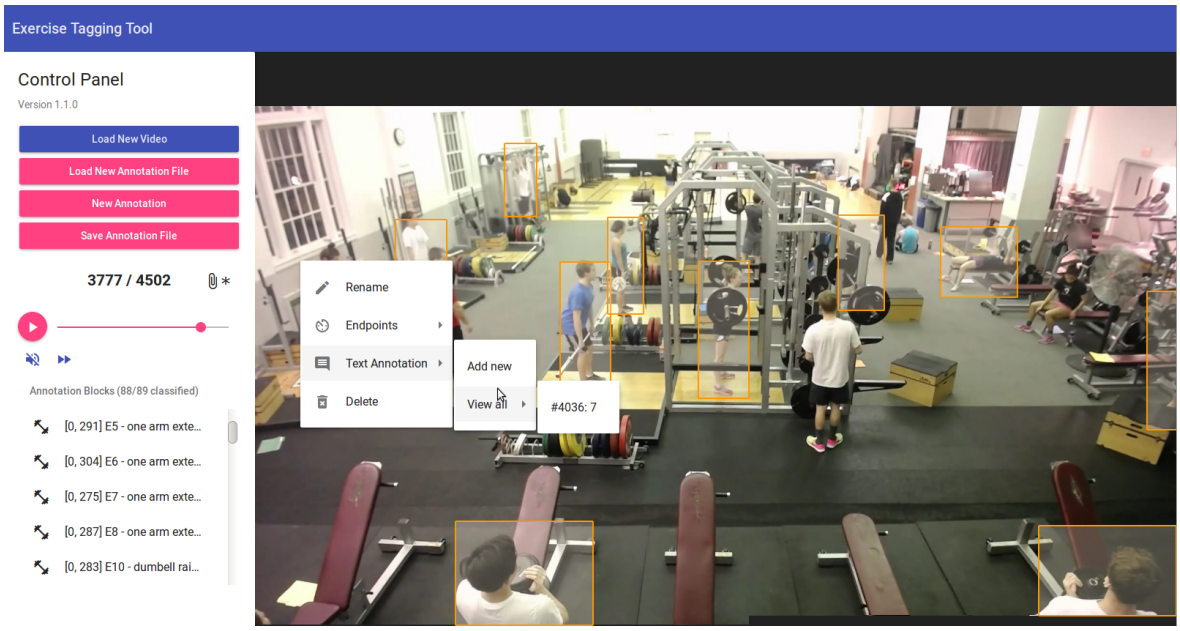


Figure 5.5: Custom video annotation tool. Annotators drew the bounding boxes, and marked the start and end time for each exercise. They used the text annotation option to add exercise type and repetition count.

5.3.2 Labeling

To annotate our captured videos, we tested several tools, but found that it was hard to efficiently annotate multiple exercises in the same frame while simultaneously recording their location, the repetition count and the type of exercise performed. In response, we built a custom annotation software (see Figure 5.5).

This software allows an annotator to load a video and use a mouse to draw bounding boxes around an exercise. The annotator can then edit the start and end frame for the annotation to correspond to the start and end of the exercise. The software also provides basic functions such as play/pause and fast forward at different rates. When the annotator indicates an end frame for a bounding box (*i.e.* an exercise), the software requests the repetition count. The annotator also has the ability to add text annotations to each bounding box, recording any notes of interest. We open-sourced our tool[4], and researchers can easily customize it for their specific use.

We recruited and trained two student annotators. They were instructed to not count any exercise with fewer than 3 repetitions. For exercises such

[4] https://github.com/zacyu/exercise-annotation-tool

as running on a treadmill, elliptical or other cardio machines, the annotators were simply asked to label ”cardio” when asked for a repetition count. The annotators did not annotate ill-defined periods as exercises, but well-defined warm ups were labeled appropriately.

To allow multiple annotators to work simultaneously, we split each recorded video into 5 minute segments and the annotators processed these fragments in batches. If any exercises got split across two video segments, we counted them as two different exercises. This would never occur in a practical scenario since the input would be a continuous video stream. However, it was a procedural decision for us to ensure efficient parallelization of effort. The annotation software recorded all annotations as a JSON file. These files could be reloaded, along with their corresponding videos, to make any post-hoc changes to the annotations if needed. After all exercise start times, end times, and repetition counts were annotated, a single annotator labeled the exercise types. As there are several variants of each exercise, and different individuals may call one exercise by different names, this process ensured quality and consistent labels.

5.4 Algorithm

The goal of our work is to detect, identify, and track exercises, including when people are only partially visible. In fact, the real test of our approach is when the user is *barely* visible, but the camera can merely see a weight or a handlebar moving. Thus, GymCam starts by identifying all movements, and classifying them as repetitive or not. There could be several movements in a video that belong to the same exercise (*e.g.,* movement of different limbs, weights, and handlebar), so we combine similar repetitive exercise movements into exercise clusters. Next, for all motion trajectories in each identified cluster, we derive a combined trajectory to recognize the exercise type and estimate repetition count for that exercise (cluster). We will now describe our pipeline (Figure 5.6) in detail.

5.4.1 Detecting Exercise Trajectories

To detect movement, we start by extracting optical flow trajectories from our video. We initially investigated OpenCV’s implementation of Lucas-Kanade sparse optical flow [166]. However, the algorithm failed to track

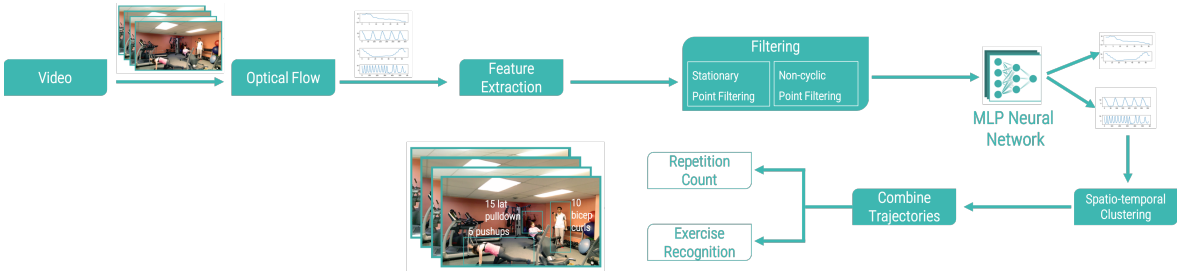


Figure 5.6: GymCam system architecture.

large, sudden movements and we switched to Wang *et al.*'s [167] dense optical trajectory extraction method to process all motion captured by the camera. For every video frame, the algorithm generates new keypoints, which are tracked continuously across frames to produce a motion trajectory. We found a keypoint max lifespan of 11 seconds was ideal for capturing several exercise repetitions, while also managing the processing time needed to track thousands of points in a video stream.

These motion trajectories are then converted into features and passed to a classifier. To limit the number of data points, we trim motion trajectories by removing stationary points (*i.e.*, any keypoint that moved less than 4-pixels between frames). We then normalize motion trajectories by their maximum translation and calculate a feature vector over an (empirically determined) sliding window of five seconds, with a stride of one second. Our feature vector consists of 27 features, a subset of which have also been used in prior work (see [12, 168]).

- **Frequency-based features:** Our working principle is that exercises are more periodic than non-exercises. We use frequency-based features to encode this property:
  - **Number of zero crossings:** We calculate the number of zero crossings of the keypoint motion trajectory, only in the x-axis, and only in the y-axis.
  - **Variance in zero crossings:** Exercises will be more periodic and have a lower variance in zero crossings than non-exercises.
  - **Dominant Frequency:** The dominant frequency of the signal calculated by frequency transformation.
  - **Autocorrelation:** Autocorrelation characterizes the periodicity of a signal.

- **Maximum autocorrelation peak:** Higher value indicates higher periodicity.
- **Frequency via autocorrelation:** The dominant frequency of signal determined via autocorrelation.
- **Number of autocorrelation peaks:** Unusually high number of peaks indicate noisy signals, which are more likely to be non-exercises.
- **Number of prominent peaks:** Represents the number of peaks higher than their neighboring peaks by a threshold (25%). A greater number of prominent peaks indicates higher periodicity.
- **Number of weak peaks:** Similarly, we calculate the number of peaks smaller than their neighboring peaks by a threshold (25%). A greater number of weak peaks represents noisy and less periodic motion.
- **Height of first autocorrelation peak after first zero crossing.** The height of the first peak after a zerolag provides an estimate of the signal's periodicity.

· **Non-frequency-based features:** Apart from the frequency-based features, we also calculate some non-frequency based features:

- **RMS:** The root-mean-square amplitude of the signal.
- **Span:** The span of the motion helps to characterize the intensity of the motion. We use overall span, and span in both *x*- and *y*-axes as features.
- **Displacement Vector:** Displacement helps us distinguish between exercises and other periodic motions such as walking. Non-exercise motions (such as walking) often have a higher displacement than exercise motions. We use the coefficients of the overall displacement vector, and displacement in both *x*- and *y*-axes, for a total of 9 features.
- **Decay:** Decay signifies the loss of intensity over time, a characteristic of exercise motions. We fit a line to the observed trajectory and use its coefficients as features.

We filter motion trajectories to bias our classifier to minimize false positives, at the cost of lower precision. This is because when an person is

exercising, not all body parts may be involved in the motion. For example, legs do not move during a bicep curl, so a keypoint on a person's leg may be inside the bounding box created by an annotator, but would not be periodic. Similarly, improper form may cause a point to move while performing an exercise. Thus, not every motion trajectory inside an "exercise" bounding box is indicative of actual exercise motion. To protect the classifier from inaccurate training data, we filter motion trajectories with aggressive thresholds on frequency-based features. By filtering, we only provide the strongest and most representative examples of exercise trajectories to train our classifier. However, we do not perform any such filtering while validating the algorithm.

We use a multilayer perceptron (Scikit-Learn implementation with default hyperparameters) to classify every 5 second window segment of each keypoint trajectory as an exercise or not. The neural network optimizes the log-loss function using stochastic gradient descent. To smooth the output, we take a majority vote of three consecutive classifications and assign that as the output for each of those three classifications. Finally, we combine all consecutive *positive* classifications to construct a motion trajectory that was predicted as an exercise.

5.4.2 Clustering Points For Each Exercise

Exercises are often captured by many keypoint motion trajectories. Thus, our next step is to cluster keypoint motion trajectories into exercise groups. We perform clustering in two steps: (1) use spatio-temporal distribution of motion trajectories, and (2) use phase-differences between motion trajectories (Figure 5.4).

Given an exercise, the motion trajectories of its encapsulating keypoints will likely be close to one another in space *and* time. For space, we bootstrap the clustering algorithm by drawing bounding boxes next to each workout machine and station. Note, this only needs to happen once at the start of the system deployment (assuming machine and stations do not move). These boxes are non-overlapping and are representative of the exercise areas of the gym. Figure 5.7 shows these bounding boxes and also a distribution of exercises in our dataset.

Apart from spatial distribution, we also investigated the temporal separation between exercises. The exercise keypoints that overlap temporally

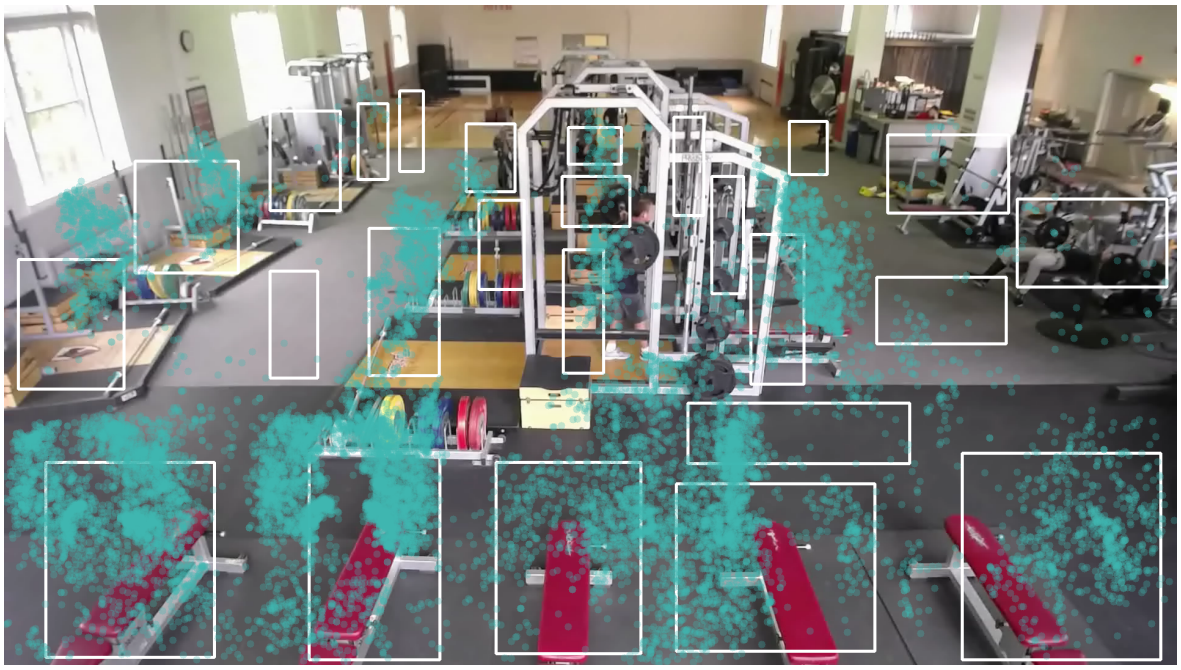


Figure 5.7: An image of the gym with a distribution of all the exercise motions observed across all videos. The white boxes are the manually-drawn boxes to aid in clustering.

as well as spatially are assigned to the same cluster. However, there is still a chance that exercises that are close to one another and occur together will be wrongly combined. To seperate such clusters, we also use phase information. For each cluster, we compute a phase-based similarity score between each trajectory-pair. For a pair of points that are not temporally co-located, the similarity is set to zero, and for others, the similarity is equal to the phase difference. We then threshold the phase difference (15 degrees) to assign a binary similarity score. In the end, we have a complete $N \times N$ adjacency matrix, where N denotes the number of motion trajectory points classified as an exercise. Given such a matrix, we calculate all connected graphs. Each graph denotes one exercise cluster associated with the nearest bounding box.

At the end of clustering, we combine the trajectories of all keypoints within a cluster to create a representative, average trajectory for further analysis (Figure 5.8). More specifically, we take the average of all the points within the cluster, accounting for the duration of each point, and smoothing it with a Hann window (size=1 second). This trajectory is used in our next process: exercise recognition and repetition count.

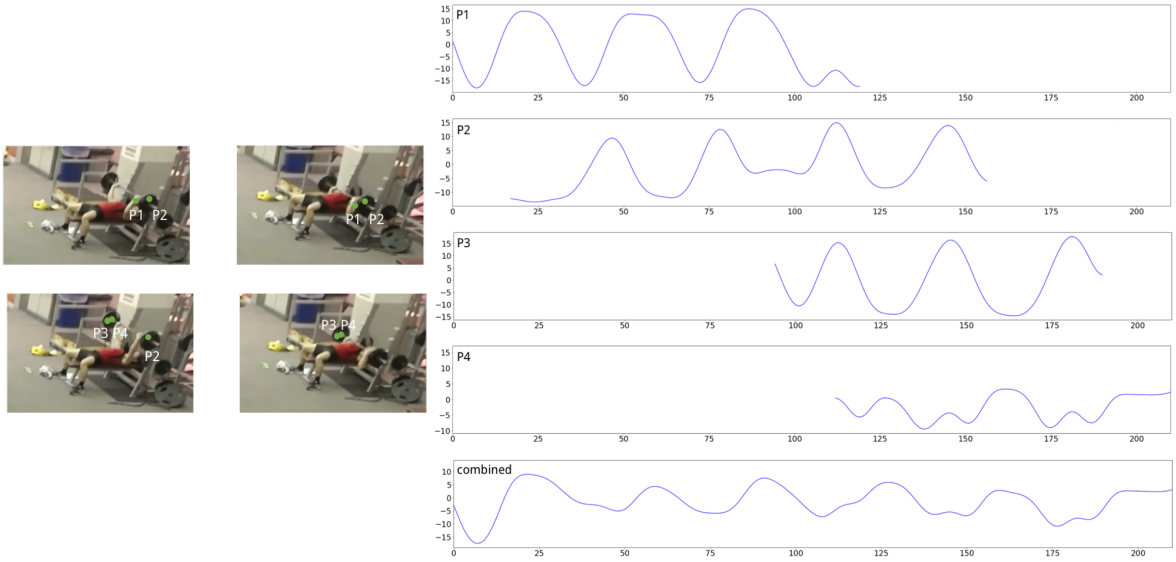


Figure 5.8: Individual and combined trajectories for an exercise. The *x*-axis is frame numbers (15 fps).

5.4.3 Repetition Count

Once a representative, average trajectory for each cluster is obtained from the previous step, we calculate the repetition count. To objectively disambiguate actual exercises from warm ups, we disregard any exercises that have less than five repetitions in the ground truth annotations. We train a multilayer perceptron regressor (Scikit-Learn; default hyperparameters) that uses our frequency-based features for each combined trajectory (as detailed in section 5.4.1), and outputs an estimate for the repetition count.

5.4.4 Exercise Recognition

Similar to repetition count, we leverage the cluster-average trajectory to infer the exercise type. We first quantize the trajectory into fixed-length segments as input to our classifier. We then run a sliding window (length five seconds, stride of one second) over this motion trajectory. Each window is passed to a multi-layer perceptron classifier (Scikit-Learn; default hyperparameters) to predict the exercise label, and we take a probability-based majority vote over all windows in the trajectory.

5.5 Results

5.5.1 Detecting Exercise Trajectories

We first report the results for distinguishing keypoint motion trajectories as *exercise* or *non-exercise*. For this, we performed a leave-one-day-out-cross-validation, which yielded a per-day, mean cross-validation exercise detection accuracy of 99.86%, with a mean false positive rate of 0.001% and precision of 23%. Again, we optimized our algorithm to reduce false positives at the expense of precision.

5.5.2 Clustering Points For Each Exercise

There are 597 distinct exercises in our ground truth annotated data. GymCam was able to accurately track 84.6% of these exercises. It also had a false positive rate of 13.5%, with most errors due to miscellaneous cyclic non-exercise motion such as warm-ups, rocking while seated, and walking.

5.5.3 Repetition Count

Repetition count accuracy helps in objective assessment of the time overlap between a predicted cluster and its corresponding ground truth match. We used 5-minute folds for cross-validation and achieved an accuracy of ± 1.7 for counting repetitions with a standard deviation of 2.64.

5.5.4 Exercise Recognition

As discussed previously, our data was collected in an uncontrolled environment where participants were not instructed to perform a specific set of exercises, and so the distribution of exercise types was not uniform. Participants performed numerous atypical exercises and curating a balanced training set of conventional exercises from our data was challenging. We identified 18 common gym exercises and annotated their instances in our dataset (Table 5.1). We decided to disregard warm-up exercises because the annotator labeled many different exercises as "warm-up". The remaining 17 exercise types were classified with an accuracy of 80.6% with cross-validation across 5-minutes folds (Confusion Matrix: Figure 5.9). The five most frequently performed exercise types constituted roughly 69% of our data. We noticed that a lack of training data caused the less frequently

17 exercises	Arm extension	D. Flies	Benchpress	Squats	D. raises	Tricep extension	Deadlift	Pullup	Pushup	Lying D. flies	D. Benchpress	Bicep curl	Lat pulldown	D. raises	Shoulder press	D. press	Cardio
Arm extension	77	0	0	2	0	0	4	0	0	0	18	0	0	0	0	0	0
D. Flies	13	13	13	13	13	0	0	0	0	0	13	0	0	25	0	0	0
Benchpress	0	0	88	3	0	0	0	0	0	0	3	0	0	5	0	0	0
Squats	0	0	1	97	0	0	0	1	1	0	0	0	0	1	0	0	0
D. raises	0	0	0	0	0	0	0	0	0	50	50	0	0	0	0	0	0
Tricep extension	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
Deadlift	0	0	1	0	0	0	96	0	0	0	0	0	0	0	3	0	0
Pullup	0	0	0	13	0	0	33	21	13	0	0	0	0	0	21	0	0
Pushup	0	0	0	70	0	0	10	0	20	0	0	0	0	0	0	0	0
Lying D. flies	50	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0
D. Benchpress	10	0	2	0	2	0	0	0	0	0	84	0	0	2	0	0	0
Bicep curl	0	0	0	33	0	0	67	0	0	0	0	0	0	0	0	0	0
Lat pulldown	0	0	0	0	0	0	50	0	0	0	0	0	25	0	25	0	0
D. raises	0	0	15	0	0	0	5	0	0	0	10	0	0	65	5	0	0
Shoulder press	0	0	0	0	0	0	57	3	0	0	0	0	0	0	40	0	0
D. press	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
Cardio	0	0	0	11	0	0	0	0	0	0	0	0	0	0	11	0	78

Figure 5.9: Confusion matrix for exercise identification across 17 exercises.

5 exercises	Arm extension	D. Benchpress	Squats	Deadlift	Benchpress
Arm extension	77	16	4	4	0
D. Benchpress	14	84	0	0	2
Squats	0	0	99	0	1
Deadlift	0	0	0	99	1
Benchpress	0	3	5	0	92

Figure 5.10: Confusion matrix for exercise identification across 5 exercises.

Table 5.1: Count of different exercise types

Exercise Type	Count
Squats	126
Deadlift	124
Benchpress	55
Arm Extension	56
Dumbell Bench-press	51
Shoulder Press	27
Pullups	24
Dumbell raises	18
Pushups	10
Cardio	9
Lat Pulldown	8
Dumbell Flies	8
Lying Dumbell Flies	4
Bicep Curl	3
Dumbell Raises	2
Tricep Extension	2
Dumbell Press	1
~~Warmup~~	~~69~~

seen exercises to be misclassified. Thus, if we only focus on the most frequent exercises, GymCam recognition accuracy increases to 93.6% (Confusion Matrix: Figure 5.10). Figure 5.11 shows the average identification accuracy as the number of recognized exercise types increase. This result indicates that our approach has the potential to differentiate between exercises based on our feature set, but a larger annotated dataset is needed.

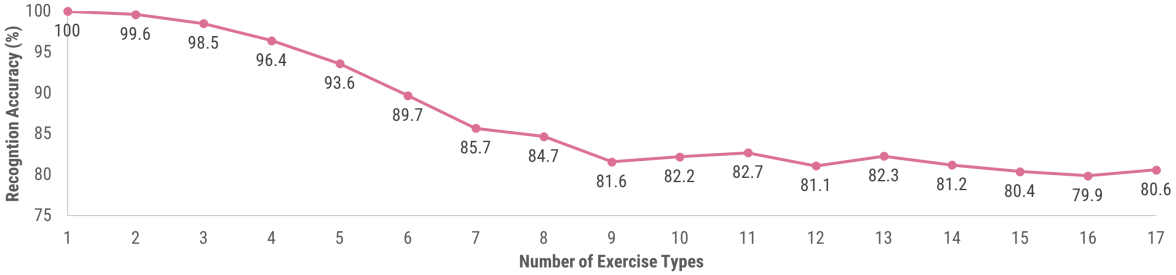


Figure 5.11: Plot showing the accuracy of exercise recognition *vs.* number of exercise types

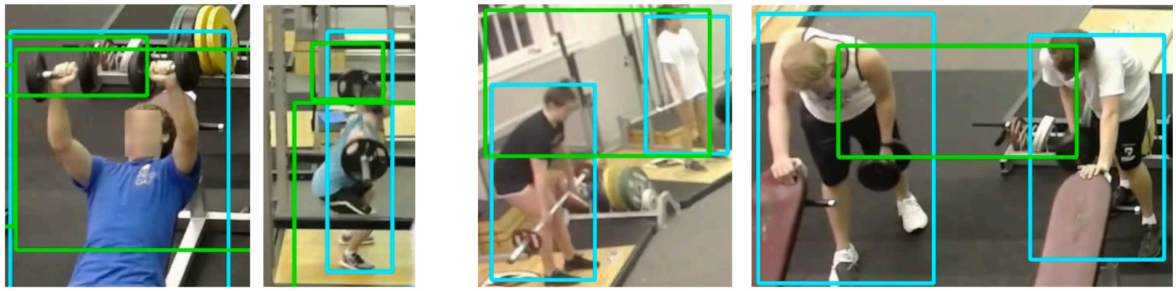


Figure 5.12: The blue boxes represent ground truth, and the green boxes represent predicted exercises (clusters) in each image. **Left enclosed in red:** Examples of exercises where one exercise gets broken into two separate clusters. **Right enclosed in green:** Examples of exercises where two exercises get combined into a single cluster.

5.6 Discussion And Limitations

GymCam provides an end-to-end pipeline to detect, track, and recognize multiple people and exercises in real-world settings, overcoming issues such as noisy data and visual occlusion. Based on our observations and experiences from building the system, we now discuss limitations to our approach. Besides completely missing an exercise, there are two types of major failure modes when an exercise is not recognized properly: (1) two exercises get clustered as a single exercise; and (2) one exercise gets split into two separate exercises as shown in Figure 5.12.

5.6.1 Reliance On Motion Differences For Clustering

When two or more individuals are exercising close to each other, and exhibit similar motion features such as phase and frequency, the individual exercise motion keypoints for each exercise may get combined into a single cluster. For example, it may occur during a group workout, when many people are roughly synchronized. Such cases are unavoidable, and should be expected to occur in uncontrolled environments. A potential solution is to augment GymCam with depth or body pose information to improve spatial clustering of nearby keypoints. Additionally, higher framerate cameras could offer finer-grained phase information (*i.e.,* at 15 fps, participants synchronized to within 70 ms are indistinguishable).

Secondly, fatigue or improper exercise form may cause a person to show high variance between repetitions of the same exercise. Such irregular

movements affects GymCam's performance as the algorithm relies on phase-based similarity of repetitions of the same exercise. In cases with irregular movements, the similarity between trajectories decreases, which may introduce clustering errors and cause one exercise to be incorrectly broken into separate clusters.

5.6.2 Tracking Irregular Motions

The backbone of our algorithm is effective capture of motion trajectories across many keypoints. One of the most popular approaches to capture motion trajectories is Lucas-Kanade Optical Flow. While highly regarded and versatile, in our dataset, the algorithm failed to track exercise motion reliably. After investigating many failure cases, we realized that the algorithm failed to **continuously** track a keypoint if a person made sudden big movement, and instead initialized a new keypoint (not necessarily in the same location). Trajectories obtained from such keypoints do not contain sufficient information to classify repetitive and non-repetitive motions. To solve this problem, we used a variant of optical flow that is more resilient to irregular movements [167] and allows GymCam to classify individual motion trajectories as exercise/non-exercise with high accuracy (acc. = 99.6%). It highlights two potential points of failure in our approach: (1) choosing relevant keypoints to obtain motion trajectories in a frame; and (2) successfully capturing motion trajectories for the duration of the exercise.

5.6.3 User Identification

GymCam detects, recognizes, and tracks the exercise, but does not identify the user. Correlating the information between two sensors could be used to identify users. For example, Amazon Inc.'s Go Stores[5] combine information from users' phones, in-store cameras, and on-the-shelf sensors to track shoppers and their purchases. Similarly, practical deployment of GymCam could combine information from either a camera-based identification system or correlate data from a user's wearable device [18] or use some form of manual identification step by the user (*e.g.,* using RFIDs [169] or QR codes [170] next to each workout station). It is arguable that relying only on pose-tracking might help in detecting the exercises *as well as*

[5]http://www.wired.co.uk/article/amazon-go-seattle-uk-store-how-does-work

identifying the users. However, our pilot experiments showed that the current state-of-the-art pose tracking approaches were unable to handle the occlusion challenges.

5.6.4 Viewpoint Invariance

An ideal camera-based system would be viewpoint invariant and not require calibration for every camera position. Considering GymCam uses some spatial information, it is not entirely viewpoint invariant. For **exercise recognition**, we use a simple bootstrapping and each area where users are likely to exercise is annotated. In a regular gym, where machines do not change positions, this annotation will be a one-time process. To detect **whether a user is exercising**, we only use time- and frequency-based features that do not change with position. Thus, it can be viewpoint invariant but we have not evaluated it formally.

5.6.5 Privacy

Using cameras enables accurate exercise tracking that is not limited to certain kinds of motion, but of course also raises privacy concerns. We acknowledge that the end-to-end system in our system required capturing the raw video, but our prior work in measuring user's privacy preferences has shown that if optical flow can be computed on the chip, that instills a sense of trust in the users. What we showed in this work is that with this processed signal, GymCam can detect exercises, but sensitive user information is not easily recoverable. Indeed, with on-camera compute power, this could be the only data transmitted from the device, or perhaps the entire classification pipeline could be run locally.

5.6.6 Unconstrained Evaluation Environment

The primary insight of our algorithm is the periodicity of repetitions. However, as Morris *et al.* point out [12], periodicity is especially hampered during strength-training scenarios. When lifting challenging weights, for example, users often become fatigued and lose pace. Such issues are uncommon in cases where users participate in a user study, as the primary goal is not to challenge participants physically. In contrast, we developed and

evaluated GymCam in a truly unconstrained environment, offering greater ecological validity and a more strenuous test. In addition to between-exercise movements and warm-ups, GymCam had to face an extra noise source: moving objects. For example, some participants in our dataset performed TRX (Total Body Resistance Exercise) workouts using ropes. In many cases, the suspended ropes kept oscillating after the exercise ended. Considering GymCam does not detect body pose and treats all repetitive motions equally, these rope oscillations resulted in a few falsely-recognized exercises.

5.7 Conclusion

With the surge in quantified self and health-focused wearable devices, the interest in exercise tracking is also rising. While tracking cardio exercises is relatively easy, tracking repetitive strength training exercise is still an outstanding problem. Most popular solutions involve using motion sensor-equipped wearables, but these devices are inadequate for capturing a wide range of exercises, especially ones involving other limbs. In this paper, we presented *GymCam*, a system that leverages a single camera to track a multitude of simultaneous exercises. GymCam relies on tracking motion and assumes that most repetitive motion in a gym are exercises in progress. To develop and evaluate our machine-learning algorithms, we collected data in CMU's varsity gym for five days. We segmented *all* concurrently occurring exercises from other activities in the video with an accuracy of 84.6%; recognized the type of exercise (acc.=92.6%) and counted the number of repetitions ($\pm$ 1.7 counts). GymCam advances the field of real-time exercise tracking by filling some crucial gaps, such as tracking whole body motion, handling occlusion, and enabling single-point sensing for a multitude of users.

CHAPTER 6

MOTIONID: A HYBRID CAMERA-WEARABLE APPROACH TO IDENTIFY USERS IN A GROUP

6.1 Introduction

Recent improvements in computer vision and machine learning have led to a surge in applications and products that can passively sense different activities, enable personalized service, provide feedback to the user and enhance interactions. More specifically, cameras have enabled sensing at scale such as smart hospitals [171], smart offices [172] or even smart gyms [17]. Applications that work at scale with tens and hundreds of users at the same time require a method to identify each user in the scene. There are two common strategies to identify users:

1. biometric-based identification such as facial recognition [82, 173].
2. using custom hardware such a RFID tags [94] or smart insoles [174].

Facial recognition is economic, accurate to a reasonable degree but requires highly privacy invasive data from users. On the other hand, custom hardware based solutions do not infringe on the user's privacy but are not economical or easy to deploy. Therefore, there is a need for an identity recognition system that is cheap, easy to deploy and privacy sensitive, while maintaining a similar level of utility and usefulness as other techniques.

In this paper, we present MotionID- a hybrid computer vision and inertial sensor system that leverages the similarity in motion as observed from

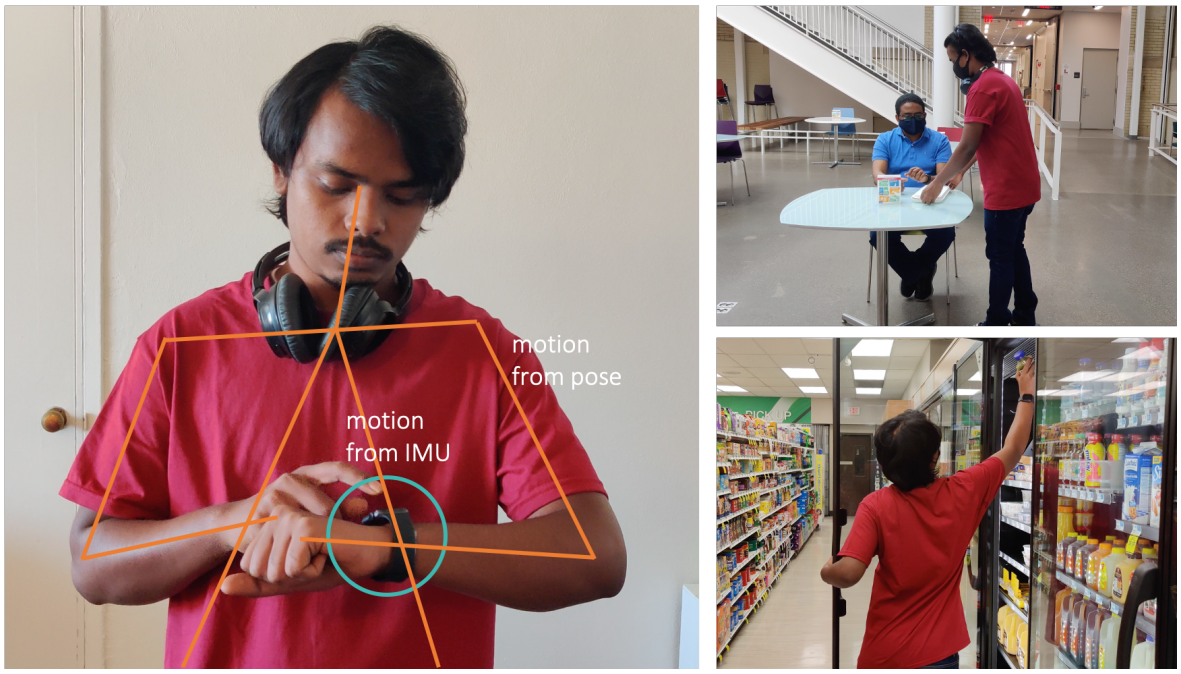


Figure 6.1: Figure showing the intuitive approach of our system to use motion from body pose and motion from IMU to identify users and enable applications in restaurants and grocery stores.

the two modalities. It uses a regular RGB camera to obtain the motion profile of a person using pose detection [50]. It also uses the IMU present in a smartwatch worn by the user to capture the motion of their arm, and provide the system with the identity of the user. We measure the correlation between the two motion profiles to calculate a measure of similarity, which is used to attribute the identity obtained from the smartwatch to one of the poses as seen through the camera. MotionID does not rely on any privacy invasive features, and a smartwatch reliant approach allows an opt-in ecosystem to be built around this technology. Additionally, as discussed in Chapter 4, open pose instills a sense of trust in the users. While we used cameras in our setup, an on-device pose detection chip can be used to preserve user privacy while identifying users in a group.

To develop and validate our approach, we collected data from three different activities (playing a sport, social gathering, and dancing), each with a group of 2, 4 and 8 people for a total of 9 conditions. These activities are representative of some common use-cases of sensing at scale applications. Additionally, we chose these specific activities to validate our approach across: (1) degree of motion observed in an activity; and (2) similarity of motion across people at the same time.

Our results demonstrate that on an average we are able to identify users in a group of 2 with 100% accuracy regardless of the activity. We were able to identify users with an average accuracy of 90% in a group of 4, and 95% in a group of 8.

Finally, we discuss the implications and use-cases of our work in retail for cashier-less shopping, smart offices, and smart gyms as shown in Figure 1.

6.2 Theory Of Operation

In this section we outline the working principle behind our approach. We discuss some intuitive approaches that we tried, and postulate why they did not work, and how they informed our proposed algorithm.

The most common approach is to measure the correlation between the acceleration measured from both the pose tracking in the video and the IMU from the smartwatch. First, we attempted to correlate acceleration signals. We calculated the acceleration from the pose information in the video, and used varying window sizes for finding correlation. In each video, we calculated the correlation between each pair of pose and watch. This technique seemed to work well for smaller groups. We were able to achieve 100% accuracy in groups of 2 with small window sizes. However, for larger groups of 4 and 8 individuals, the results did not look promising. From our results, there were two noticeable challenges-

1. Activities with a low degree of motion such as poster presentations were performing worse. We attribute it to the error caused by differences in how the acceleration is calculated in both modalities. The error amplifies with a higher number of comparisons, especially when the motions are small. Even one incorrectly matched pair propagates the error for other poses as well. For example,

   correlation{P1, W1} = 0.7

   correlation{P4, W1} = 0.72

   where P(n) refers to a pose in the video and W(n) refers to the data from the watch/person in the scene. Here, The actual match should have been between P1 and W1, which objectively has high correlation value. However, due to the accumulated error in acceleration calculation for the tiny differences between the movements of P1 and P4,

P4 ends up matching with W1 with a higher correlation value, forcing P1 to find the next most suitable pair.

2. If a person would leave the video frame, the missing data in acceleration causes issues with matching the right pair. This was apparent in high degree of motion activities such as playing volleyball. Whenever a person would leave the video frame to fetch the ball, they would run, bend quickly to pick the ball, and sometimes even kick it. These are all prominent events being captured by the IMU, but not by the video- leading the falsely correlated pairs. This problem was further exacerbated by imperfect tracking of poses. If the pose is not detected accurately for some frames, it would degrade the acceleration measurements from the poses.

To reduce noise, we then attempted to extract relevant features to describe the accelerometer signal from both video and IMU data. For varying lengths of windows, we extracted features that would describe the shape of a motion trace- max value, min value, standard deviation for both X and Y axis, RMS of the combined XY signal and so on. It was able to reduce the error for high motion data. The loss of data was "smoothed" over by extracting only relevant features from the entire signal. However, this reductive approach further amplified the error in small degree of motion activities such as meeting.

Next, we sought out more informative signals. We used the yaw and pitch values obtained from the Apple Watch [1] and used them as a proxy for motion traces in X and Y axis, similar to the ones obtained from poses in the video. This signal was less noisy compared to the acceleration. The correlation of these signals yielded similar results for people in groups of 2, and worked well for high motion activities like volleyball in groups of 4 as well. However, low motion activities were still getting confused with an increase in number of participants. We attribute this to yaw and pitch being a proxy for movement in the 2D coordinate space, and their calculation for smaller movements is more prone to noise. In a small window size, this can lead to a lot of variance in matching pairs of poses and watches. We confirmed this hypothesis by using large window sizes ($>$ 2 mins) to reduce the effect of these noise prone movements. All activities in a group of 4 were recognized with high accuracy in this scenario.

[1] sensorlog.berndthomas.net/

Using such a large window size may be suitable for some activities, but would make it unusable in a multitude of scenarios. However, we do have informative signals that are able to help us correlate poses with their respective user watches. To overcome the issue of error prone low movement moments, instead of looking in a specific window, we changed our algorithm to search "loud and highly distinguishable" moments. We started looking for specific moments in the pose signal where it change in signal is above a certain threshold within a small duration of time. The intuitive idea is to look for moments that would be highly distinguishable in the pose data, and look for matches in the IMU data by finding the correlation for that part of the signal. This technique works well in most cases. We were able to identify users in groups of 2 and 4 with high accuracy regardless of activity. In groups of 8, high motion activities were recognized with high accuracy, but the low motion activity was only able to identify half the users correctly. Even though the technique itself was sound and robust, the lack of highly distinguishable moments made it harder to distinguish between signals.

We have outlined these techniques, and where they failed to not only underscore the difficulty of this task, but also to provide insights into what approaches may work in different circumstances. For example, in a scenario where there is no expectation of a user leaving the frame, or the activities are deemed to be high motion- then some of the aforementioned approaches will also be suitable for use. With those assumptions, it may even work in a smaller window size compared to our proposed approach.

One key observation we made with our prior approaches was that different approaches were able to match different pairs of poses and watches. While all of them had some problems, they worked quite well for different kinds of challenges. Armed with this insight, we use several different signals in our proposed approach. We also take the PCA of combined signals, and extract the same features that serves as a proxy for only retaining the 'interesting' moments in a signal. Besides more input signals, we want to overcome the issue of requiring a long signal ($>$ 2 mins) for accurate results as seen in one of the earlier approaches. To do so, we also extract features to measure similarity of signals in both frequency and time domain, and combine it with correlation values to determine similarity between pair of signals. **We seek a trade-off between accuracy and the amount of time required by the system to identify users in a group.**

6.3 Data Collection

The key parameter in this work is motion. We selected three activities based on two different dimensions:

1. **degree of motion:** the expected size of movement in an activity
2. **expected synchronized motion:** how likely is it that two or more users would perform the same motion at the same time

Based on these criteria, we chose three different activities:

1. **Volleyball:** Volleyball is a team sport where the expected movement size is big. The players run towards a ball, try to place it in a particular position or hit it hard enough to make it on the other side. However, there is very little expected synchronization. Even if two people are performing actions at the same time, they are likely to be different (defense vs offense; or hitting the ball from opposite ends of the net).
2. **Poster Sessions:** The second activity we chose was poster sessions. In this activity, the primary task is for a person to talk about their poster. The expected movements involve them moving their hands and gesturing, which would be small in size. But a little synchronization can be expected. It is possible for the presenter and one of the audience members to perform similar gestures when talking or pointing to a specific part of the poster.
3. **Dancing:** The last activity in our dataset is dancing. The movement size in a dance is big. And, we specifically looked for dancing groups that perform a piece in synchronization. This activity is a true test of our system's reliability.

For each activity, and each group size of 2 and 4 individuals, we recorded five sessions of data. For poster sessions and volleyball, we randomly selected groups from a participant pool of 18 individuals. For dancing, we recruited a large dance group, and randomly shuffled groups for each session.

For a group size of 8, we again recorded five sessions of volleyball. However, we recorded one long 18-minute poster session. Groups of 2 and 4 capture the smaller natural interactions that happen at one poster. In the

larger 8-person group condition, we wanted to capture the whole activity of a poster session where the audience is free to move from one poster to another. Lastly, due to covid19, we were unable to record any data for a group of 8 people dancing together.

We collected the data over several days and in different locations. We used a single iPhone SE (1st generation) to capture and record RGB videos at a resolution of 1080p and 30fps. All participants also wore an Apple Watch Series 4.

6.3.1 Extracting Motion Information

Smartwatch: The SensorLog application [2] available on the Apple app store was used to record IMU data in the background. In addition to the user acceleration, it also logs the yaw, pitch and roll values. The data was initially recorded at 50Hz but was later down sampled to 30Hz.

Video: The videos were recorded at 30fps. We developed a custom tool that uses Deep Sort [175] on top of OpenPose [50] to track body poses across the whole video. We manually fixed any tracking errors generated from our custom tool. We logged the position of all keypoints available via OpenPose but we only used the LWrist (left wrist) in our setup. All participants wore the apple watches on their left arm, so we used only the motion information captured by LWrist in OpenPose. Hereon, we refer to the position of LWrist as pose information, unless explicitly stated otherwise.

6.4 Algorithm

The goal of our work is to identify users in a group in a privacy preserving manner. To this end, we leaned on an approach that uses both the camera and the smartwatch for identification. The reliance on smartwatch to identify users allows them to gain more control over their privacy, and how and when their data is shared. Similar to prior work, our approach also relies on finding similarities in the motion as observed from the video and the smartwatch (IMU) tied to a user's wrist.

[2] http://sensorlog.berndthomas.net/

6.4.1 Feature Computation

As stated in Theory of Operation, we built on aspects of prior work and our failed approaches. We developed a machine learning algorithm that relies on features that describe the shape of each signal from both modalities. It also takes input features that themselves are a measure of similarity between two signals (from different modalities). Altogether we have a total of 576 features. First, we describe the different signals that we use to compute these features.

The list of the signals is:

1. **accX-v:** the X-axis accelerometer signal from the video.
2. **accX-w:** the X-axis accelerometer signal from the smartwatch.
3. **accY-v:** the Y-axis accelerometer signal from the video. Both accX-v and accY-v calculated from the pose information obtained from the video.
4. **accY-w:** the Y-axis accelerometer signal from the smartwatch.
5. **accMag-v:** the magnitude of the accelerometer signal at each sample *i.e.* sqrt(a_x^2 + a_y^2).
6. **accMag-w:** the magnitude of the accelerometer signal at each sample *i.e.* sqrt(a_x^2 + a_y^2)
7. **accPC1-v:** the project of the X and Y axes of the video acceleration signal onto its first principal component.
8. **accPC1-w:** the project of the X and Y axes of the watch acceleration signal onto its first principal component.
9. **poseX:** the X-axis value of pose information obtained from OpenPose for the left wrist.
10. **pitch:** movement around the pitch axis recorded on the smartwatch.
11. **poseY:** the Y-axis value of pose information obtained from OpenPose for the left wrist.
12. **yaw:** movement around the yaw axis recorded on the smartwatch.

13. **poseMag:** the magnitude of the pose signal at each sample *i.e.* sqrt($pose_x^2$ + $pose_y^2$)
14. **movMag:** the magnitude of the combined pitch and yaw signal at each sample *i.e.* sqrt($pitch_i^2$ + yaw_i^2).
15. **posePC1:** the project of the X and Y axes of the pose signal onto its first principal component.
16. **movPC1:** the project of the pitch and yaw axes from the watch onto its first principal component.

We calculate the following features for each of the 16 signals.

1. **statistical features:** We calculate mean, standard deviation, variance, kurtosis and skew for a total of 5 statistical features.
2. **RMS:** the root-mean squared amplitude of the signal.
3. **prominent peaks:** we calculated the total number and location of two most prominent peaks in the signal for a total of 3 features. The intuition here is that the location of prominent peaks in signals from both modalities should occur close to each other.
4. **power spectrum features:** we calculate the magnitude and mean of the power spectrum in 10 bands spaced equally between 0.1-15Hz for a total of 20 features.

We calculated the following features for every signal pair (listed in order). For example (accX-v, accX-w) is one pair and (poseX, pitch) is another.

1. **Pearson's correlation:** it is a measure of how closely associated are the two input signals.
2. **similarity in time:** the two signals are multiplied and added in place to generate a single similarity score.
3. **similarity in the 90th percentile shifted time domain:** multiply the fast Fourier transform of each signal, and then take an inverse fast Fourier transform of the resulting signal. Then we only retain the 90th percentile of the signal to retain the most interesting parts. We then calculate the sum, mean and max of this signal for a total of 3 features. This feature set is similar to calculating correlation between two signals shifted in time domain.

4. **similarity in frequency:** we take the Fourier transform of the two signals, which are then multiplied and added in place. We calculate the mean and max for a total of 2 features.

5. **similarity in the 90th percentile shifted frequency domain:** We first multiply both the signals and take the fast Fourier transform of the resulting signal. Then we only retain the 90th percentile of this signal. We calculate the sum, mean and max of this signal for a total of 3 features.

6. **similarity in signal variance:** both signals are squared, normalized and then a measure of similarity in signal variance is obtained by subtracting one from the other.

7. **circular correlation:** we calculate the circular correlation between the two signals and take the sum, mean and max of the resulting correlation values for a total of 3 features.

A total of 29 individual features calculated from each signal results in a total of 464 signal. And 14 features calculated for 8 pairs of signals adds another 112 features for a total of 576 features used in our system.

6.4.2 Matching Users

We take a stringent approach to testing our system. No data from the condition being tested (even a different session) was included in the training for that activity. So, if we were testing the group of 8 participants playing volleyball, our machine learning system was only trained on users playing volleyball in groups of 2 and 4. And some activities were recorded in different locations. For example, the group of 2 dancing were recorded in a different room than the group of 4 subjects dancing. Similarly, the poster session of 2 users was recorded in a room different than the one used for poster session with 4 and 8 individuals. For volleyball, even though the same court was used- the data was recorded over several days, without a fixed point for the camera setup. The researchers used an approximately similar viewpoint each time. Not using the data from the condition being tested strengthens our results and demonstrates location and viewpoint invariance.

We trained a Random Forest Classifier with the default parameters to classify if two signals were similar or not. We use the probability score from the output of the classifier to rank potential candidates and find the best match for each pose in the system. If two poses have the same watch as their highest ranked candidate, the pose-watch pair with the higher probability score is matched together, and the other pose looks for its next best match.

For our evaluation that retains prior history of classification, we use a normalized running score for probability scores added over all windows.

6.5 Results

We look at the results in the size of the group. Identifying users in a smaller group size regardless of the activity is easier than a larger one. So, we first look at the simplest case of two users in the same environment.

6.5.1 Group Of 2

We used a window size of 45s to calculate features and regardless of the activity poster session, volleyball or dancing- on an average across 5 sessions, we were able to identify 100% of the users within the first 45s. All sessions of all activities of this group size were recorded for 3 minutes each. When we retain the history, we were able to continue identify 100% of the users with 90s, 135s and 180s of video duration as well.

6.5.2 Group Of 4

We again used a window size of 45s to calculate features for each activity.

For **volleyball**, on an average across 5 sessions, we were able to identify 100% of the users within the first 45s. When we retain the history, we were able to continue identify 100% of the users within 90s, 135s and 180s of video duration as well as shown in Figure 6.2.

For **meeting**, on an average across 5 sessions, we were able to identify 90% of the users within the first 45s. When we retain the history, we were able to improve and identify 100% of the users with 90s. For both 135s and 180s duration of video, we continued to identify 100% of the users as shown in Figure 6.2.

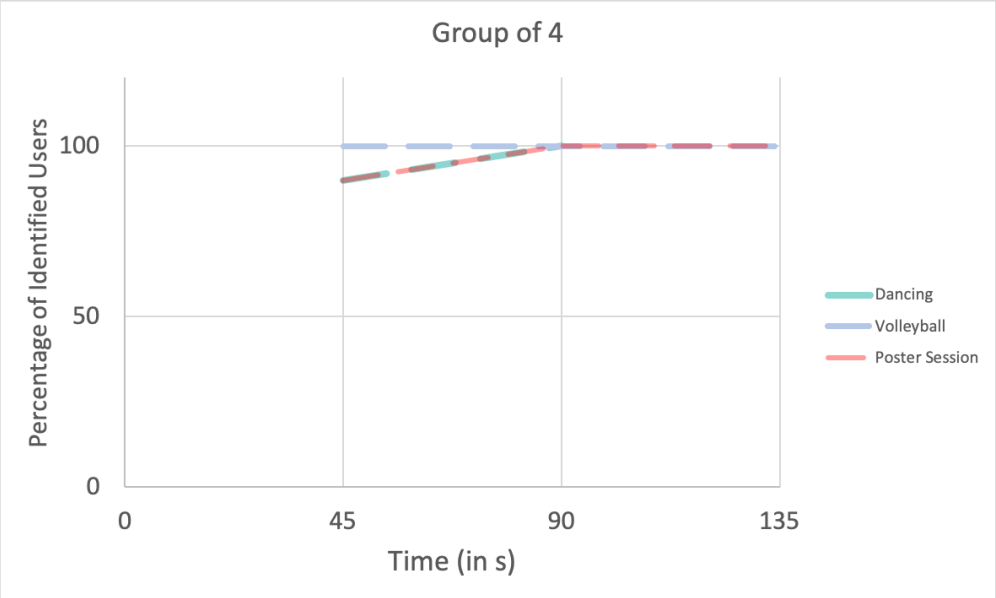


Figure 6.2: Figure showing the percentage of accurately identified users in a group of 4 for three activities.

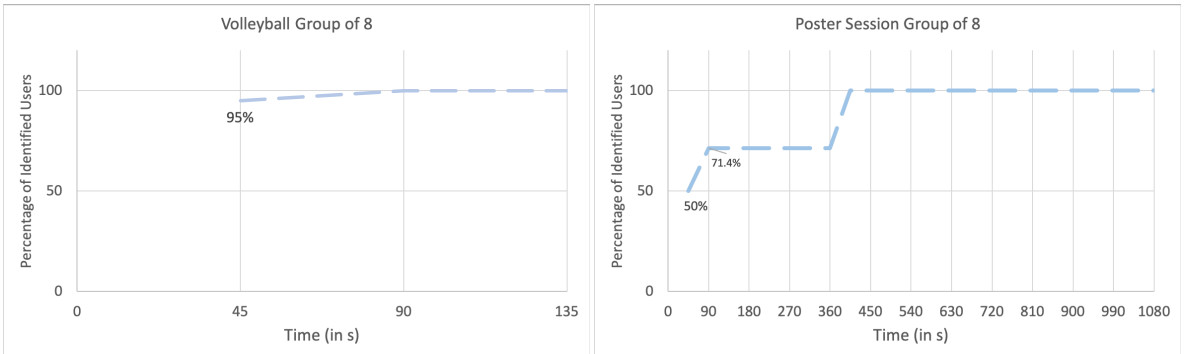


Figure 6.3: Figure showing the percentage of accurately identified users in a group of 8 for volleyball (top) and poster session (bottom).

For **dancing** with a group size of 4, we only recorded sessions of 1 minutes and 30 seconds. This was done keeping in mind the challenge of 4 dancers being in sync for a longer duration and the fatigue it may cause. On an average across 5 sessions, we were able to identify 90% of the users within the first 45s. When we retain the history, we were able to improve and identify 100% of the users with 90s of data as shown in Figure 6.2.

6.5.3 Group Of 8

Using the same window size, for **volleyball**, on an average across 5 sessions, we were able to identify 95% of all users within the first 45s. When

Figure 6.4: Example application of cashier-less checkout. A person walks into a store, is recognized with their motion of reaching for the bottle on the top shelf, and walks out of the store as the purchase automatically gets charged to their account.

we retain the history, we were able to improve and identify 100% of the users with 90s of data as shown in Figure 6.3.

For **meeting**, we only recorded on long poster session to capture the entire event as if it would happen in the real world. It also allows us to do a longer analysis. We initially used 8 individuals, but the tracking for one person did not work, so we took out their data leaving a total of 7 individuals. For the first 45 seconds, only 6 individuals were tracked and we were able to identify 50% of the users. However, after 90s we were able to identify 71.4% (or 5 out of 7 users) using our approach. We are able to continuously identify 5 out of 7 users, while retaining the confusion between the other two for 6 minutes and 45 seconds. After that, we are able to identify 100% of the users at all times. Figure 6.3 shows the percentage of users identified over time.

As stated earlier, due to covid19, we were unable to collect any data for a group of 8 subjects in the dancing activity.

6.5.4 Example Applications

The ability to identify users in a group in a privacy preserving manner enables several applications. We outline a few of them in this section.

Automatic user identification can be used for cashier-less checkouts in different stores. Most stores have an app and just a software update would be required to allow user data to be shared with a camera-based system in physical stores. An example is shown in Figure 6.4 where a person is able to pick up an item, their motion is captured- used to recognize their identity from their smartwatch and the item is directly charged to their ac-

Figure 6.5: Example application of improved service in a cafe. A person orders food from their smartphone app. Their motion in the moment is used to recognize the user and tag the order with their location. The server is then able to bring the food directly to the table without the need to ask for who order that particular dish.

count. The user simply walks out of the store without having to go through a cumbersome payment process with a cashier or self-checkout machine.

Similarly, a lot of bars and cafes allow users to order food and drinks from an app, and the food is delivered to their table. In the current system, the server typically shouts the name of each customer for identification purposes. However, this process can be automated where the identity of the user can be shared when they order the food. A little tracking icon tied to the order number can show up on a map on the server's iPad. An example scenario is shown in Figure 6.5. A similar concept can be applied to help centers such as the Apple store, where one employee typically describes your outfit and shares it with the other employees via an iPad app. This cumbersome method of tracking people can be replaced by automatic identification and tracking of users in real time within a store.

And perhaps the most obvious use of automatic identification would be to use it for access control in a building. A person sharing their identity in their work building could be tracked and identified to automatically grant access (or restrict access) to different parts of the building.

6.6 Discussion And Conclusion

The goal of our work is to identify users in a group in a privacy preserving manner. We use the motion traces as observed from the camera and the smartwatch to find similarities and identify users from those signals. The

fact that smartwatch data is a required part of the system enables a user with more control over how and when their data is shared. Keeping our approach in mind, user identification can now be built as a feature similar to sharing location. When a user walks into an environment, where they would like to be recognized, they can simply turn on sharing and their IMU data would be shared with the camera system in place. For example, a user walks into a gym that provides a service to track their workout. A user that wants to leverage this service can turn on their identity sharing just for their gym and reap the benefits of the automatic workout logging service. However, if the same user does not feel comfortable sharing their identity in a mall, they can simply turn sharing off. A simple push of a button prevents the ambient sensing system alone to recognize users.

In this work, we have demonstrated that we are able to identify users while they perform different kind of activities with varying degree of motions and sync in the user's actions. We were able to identify users in a group of 2 with 100% accuracy regardless of results in only 45 seconds. In a group of 4 individuals, we were able to recognize 90% of the users in both the poster session, and dancing scenario and 100% of the users when they are playing volleyball. Similarly, for group of 8 individuals, we are able to identify 96% of the users in only 45 seconds and we are able to identify 5 out of 7 individuals in a poster session after only 90 seconds of activity. Our results demonstrate that for activities that last longer than a couple of minutes, a hybrid approach of smartwatches and camera can be both reliable and privacy preserving in identifying the users.

CHAPTER 7

IMU2DOPPLER: CROSS-MODAL DOMAIN ADAPTATION FOR DOPPLER-BASED ACTIVITY RECOGNITION USING IMU DATA

7.1 Introduction

Researchers and developers often rely on sensors in smartphones [176, 177], smartwatches [178, 179], cameras [17, 180, 181], and even microphones [182, 183] to infer context, recognize user activities, and adapt to the user's needs. Recently, we have seen many activity recognition systems that rely on Doppler Effect-based mmWave radars to measure activity movements [184–186]. An advantage of a mmWave radar is its ability to characterize fine-grained motion. It has the ability to capture micro-motion dynamics of subtle activities (*e.g.,* hand activities such as brushing, eating etc.) captured via the *micro-Doppler Effect* [187]. A mmWave radar-based activity recognition system also offers a higher degree of privacy preservation compared to other popular ambient sensors such as cameras or microphones.

These activity recognition and sensing systems are typically built using machine learning models that need labeled, *in-situ* sensor data for training. These training labels typically rely on manual annotation or user intervention to segment and label specific activities performed by users. This process introduces time and resource constraints that impedes our ability to quickly deploy and use new doppler sensors. Moreover, the ground truth collection (cameras, user intervention etc.) tends to be intrusive and may be unsuitable in scenarios with elevated privacy constraints. This **data**

collection and labeling cost is one of the biggest challenges of building any new activity recognition system. These systems need to work well out-of-the-box with no or very-little in-situ calibration. Ideally, the machine learning models would not need any *in-situ* training, and ultimately facilitate easier deployability.

One method to overcome the challenge of data labeling cost is automated domain adaptation. Such approaches rely on successful knowledge transfer from labeled data collected in one domain and use it to assist the training of a model in a target domain with no (or limited) data of its own. Here, one popular approach has been to use videos as the source domain [127, 128]. Videos provide a rich source of information with a considerable feature space. Moreover, the extensive library of labeled video datasets make it an attractive choice for a source domain. However, using videos as the source domain requires the full body of a human to be visible in the source videos. The approach cannot handle occlusion or partial capture of the body. This limitation significantly reduces the available video datasets that can be reliably used for domain adaptation.

In this chapter, we present *IMU2Doppler* and evaluate the use of off-the-shelf inertial measurement units (IMU) datasets as the source domain to build an activity recognition model for the mmWave radar sensor. IMU data does not share the same limitations as videos. The uninhibited signal that captures the motion performed as a part of the activity is rotation and environment invariant which makes it a good candidate as a source. Additionally, IMU retains some of the advantages of video datasets *i.e.,* prior works in activity recognition have extensively collected IMU data for a gamut of activities and made it publicly available.

We demonstrate that IMU2Dopper can map the doppler data (input to the untrained ML model) to a latent feature representation of the pre-trained IMU model. In addition to this representation of the IMU model, we use minimally labeled (akin to a calibration step) doppler data to classify 10 activities of daily living. This novel approach allows us to recognize these activities with an accuracy of **70% with only 15 seconds of labeled data** from the mmWave radar sensor. We acknowledge that this is not the performance we should expect from a real world system. However, IMU2Doppler provides an out-of-the-box model that can benefit from a quick personalization and calibration step. Our contribution lies in facilitating rapid development of a 'good enough' base model that can then be

used with other techniques such as active learning [188] or meta-labeling [189] that personalize to the user's environment and improve over time without a need for significant data labeling.

In this work, we also demonstrate that we can combine multiple IMU datasets recorded in completely different environments with different users as a unified source of training data. Typically, using multiple sources is a significant challenge for domain adaptation due to the domain shift that exists across the sources. Zhao *et al.* have summarized the numerous challenges of multi-source domain adaptation [190]. However, we show that our approach is resilient to such issues. In fact, when we combine the training data from two IMU datasets, IMU2Doppler demonstrated a small increase of 1% in recognition performance. From a practical perspective, it means that not only any publicly shared IMU data can be used, but a user who wishes to record a completely new activity may incur a one-time-cost, use an app on their smartwatch to collect IMU data for that activity, and personalize the machine learning model. Moreover, if the user chooses to share their data of this new activity, other users can leverage it to train their doppler sensor without incurring the same time and resource penalty.

In summary, our contributions are as follows:

1. An activity recognition system for 10 different activities using mmWave radar. Prior work has shown the use of mmWave radar to capture gross movements. We include and expand the set of activities to include subtle activities such as brushing teeth, folding laundry *etc*.

2. A novel multi-class heterogeneous domain adaptation approach that learns a feature mapping between inertial sensors worn on a user's wrist and a mmWave radar sensor placed in the environment. It means that our approach is viewpoint and translation invariant.

3. The domain adaptation approach uses off-the-shelf IMU dataset as the source domain. It means that the source data was not only collected on different users as the target domain, but also at a different time. We also show that we can use muliple datasets and combine them as a single source to achieve the same results.

7.2 Algorithm

IMU2Doppler is a transfer learning-assisted ambient sensing system that uses mmWave radar sensors to detect and distinguish between a set of activities of daily living with minimal labeled data. To account for the lack of labeled radar data, we implement a multi-objective optimisation technique that uses domain adaptation. It uses a neural network pre-trained on inertial measurement data from multiple datasets specifically curated for the task of activity recognition. Below, we describe our sensing principle and algorithm in detail.

7.2.1 Sensing Principle

Millimeter-wave (mmWave) radar sensors transmit pulses of electromagnetic energy and receive reflections when obstructed by rigid targets in the environment. By exploiting the Doppler Effect, it is possible to measure certain motion characteristics of the target like its relative velocity, angle of arrival and distance to the radar system. While the Doppler effect arises from the bulk motion of the target, micro-motion dynamics of the target or its structure such as vibration, rotation, tumbling and coning motions induce the *micro-Doppler Effect* [187]. For instance, in case of a moving person, the arms and legs act as independent elements in motion [191]. Since the intensity of the micro-Doppler effect is dependent on the velocity and direction of the motion, individual movements of the target with discernible motion characteristics produce distinct micro-Doppler signatures, which can be used for human activity recognition [192, 193].

We collect the synthetic aperture radar (SAR) data from the doppler sensor and used the azimuth-range-doppler algorithm to parse the continuous data. As shown in Figure 7.1, the images corresponding to different activities represent distinct patterns. These patterns can be modeled and recognised by appropriate learning algorithms, as described in the following sections.

7.2.2 Knowledge Transfer

Machine learning algorithms show exceptional predictive power in a range of human activity recognition (HAR) tasks [17, 194–196] but require an abun-

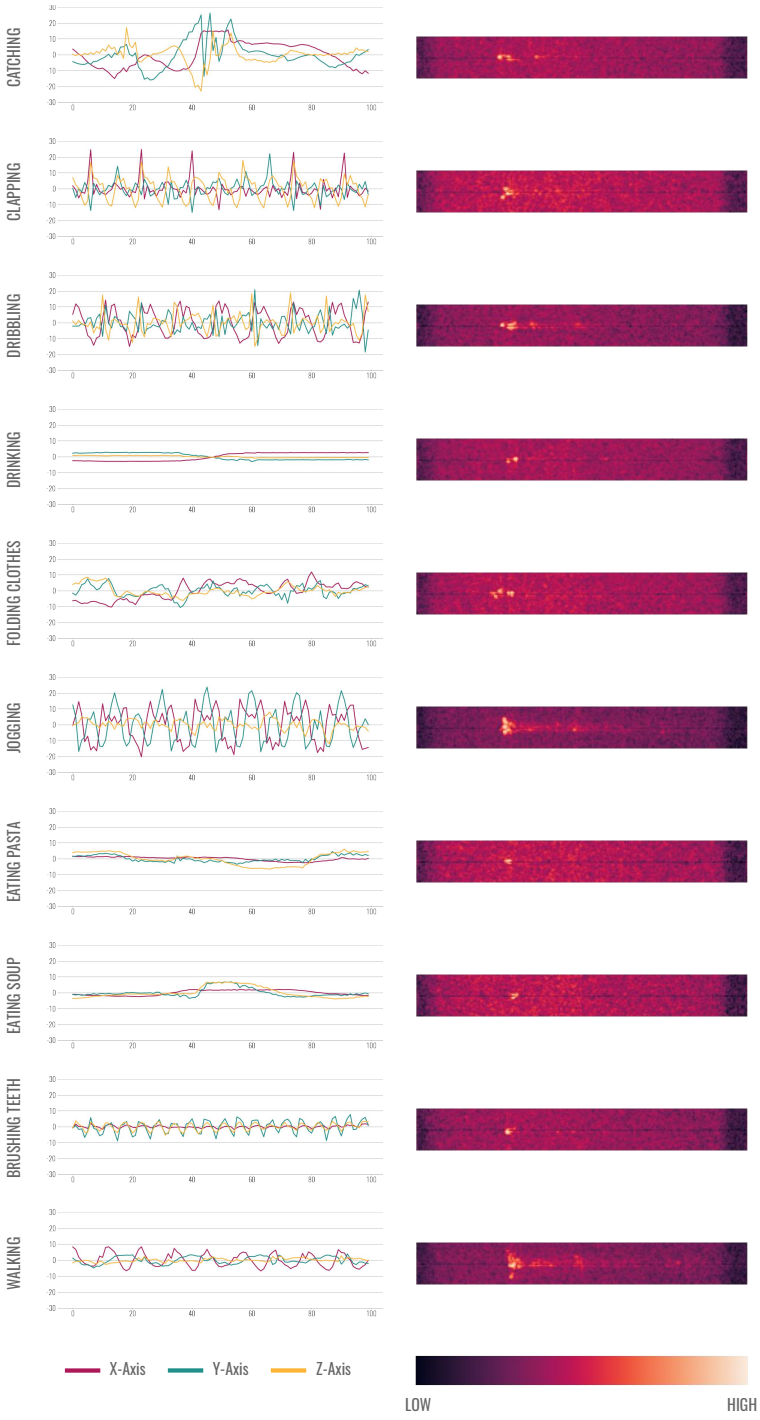


Figure 7.1: Image showing corresponding doppler and IMU signals for various activities

dance of annotated data. Although such labeled data exists for a number of sensing modalities, the newfound promise of doppler radar sensing is limited by the lack of a sufficiently large labeled dataset. To solve this problem, we use transfer learning, specifically domain adaptation, wherein we can leverage neural networks trained on a sufficiently large dataset of a different but related modality (source domain) to accelerate the learning of micro-doppler signatures (target domain).

The accelerometer data captured by a wearable inertial measurement units (IMU) characterizes similar motion characteristics as that of a doppler sensor. It captures an environment and position invariant snapshot of the motion of human movement. We postulate that this characteristic of IMU makes it a suitable candidate for source domain. Besides heterogeneous domain adaptation across two different modalities, we also use off-the-shelf datasets to demonstrate that the same events do not need to be recorded synchronously for knowledge transfer across different modalities.

For knowledge transfer, we propose a supervised, cross-modal domain adaptation approach that maps the input of the untrained doppler model to the shared latent feature representation of the pre-trained IMU model. Further, to preserve the information about the target domain or doppler data, we adopt multi-task learning to simultaneously minimize the domain discrepancy (between the latent representation of the two modalities) along with the classification loss (between the predicted and actual target label). This multi-objective optimisation ensures that the underlying structure of the target data is retained even in the latent feature subspace. The rationale for our approach is rooted in the observation that task-specific and domain-invariant semantic features can be better associated with the higher layers (close to the output side) of a network [197]. This observation allows us to choose different neural architectures that are best suited for each modality, with the constraint of having identical fully connected layers that are responsible for producing the shared latent feature representation.

7.2.3 IMU: Data Processing And Neural Architecture

We use the ”WISDM Smartphone and Smartwatch Activity and Biometrics Dataset”[198] for training an activity recognition classifier with inertial data. The dataset was collected from the accelerometer and gyroscope

sensors on both the smartwatch and smartphone of a total of 51 users. It consists of 18 unique activities, ranging from basic ambulation like walking and jogging to other activities of daily living like eating and drinking. For the purpose of our work, we chose a subset of 10 activities for evaluation, as listed in Figure 7.1. We chose these activities based on their suitability for detection with doppler sensor. For example we did not include activities such as kicking a soccer ball or two other different eating related activities (sandwich, chips). We also excluded activities that do not include any motion such as sitting.

The four streams of data, namely phone accelerometer, phone gyroscope, watch accelerometer and watch gyroscope, are each recorded at a sampling rate of 20 Hz. For our purposes, we consider only the smartwatch accelerometer data. We segment the raw watch acceleration data using a sliding window of size 5 secs and an overlap of 2.5 secs. The extracted tri-axial frames are reshaped into 3-channel windows with a length of 100 samples (input size: 100×3) that are ready for classification.

IMU Model Selection

To assess the discriminability of the activities in the source domain, we evaluate the performance of a set of deep neural networks including 1D CNN (5 Convolutional Units, each consisting of a Convolutional layer, a Batch Normalisation layer and a Max Pooling layer), LSTM (2 LSTM layers, Units: [128, 256]) and Bidirectional-LSTM (1 Bi-LSTM layer, Units: 256). One fully-connected layer (Units: 128) and the final output layer were added at the end of each model. The networks were trained from scratch with Adam optimizer (Learning Rate: 0.01) coupled with a learning rate decay of 0.1 (to check for the saturation of validation loss). We use the Categorical Crossentropy loss function to optimise the outputs of the final layer, which uses the Softmax activation to classify activities. We used Keras [199] and Python to implement and train these models.

We followed a subject-independent scheme for evaluation and split the dataset into 5 folds of 10 subjects each. Each train-test split resulted in approximately 28.4K training instances and 7.8K test instances. Table 7.1 provides the classification performance of all the models along with their total number of trainable parameters. We found the bidirectional LSTM classifier to produce the best accuracy owing to the superiority of Recurrent

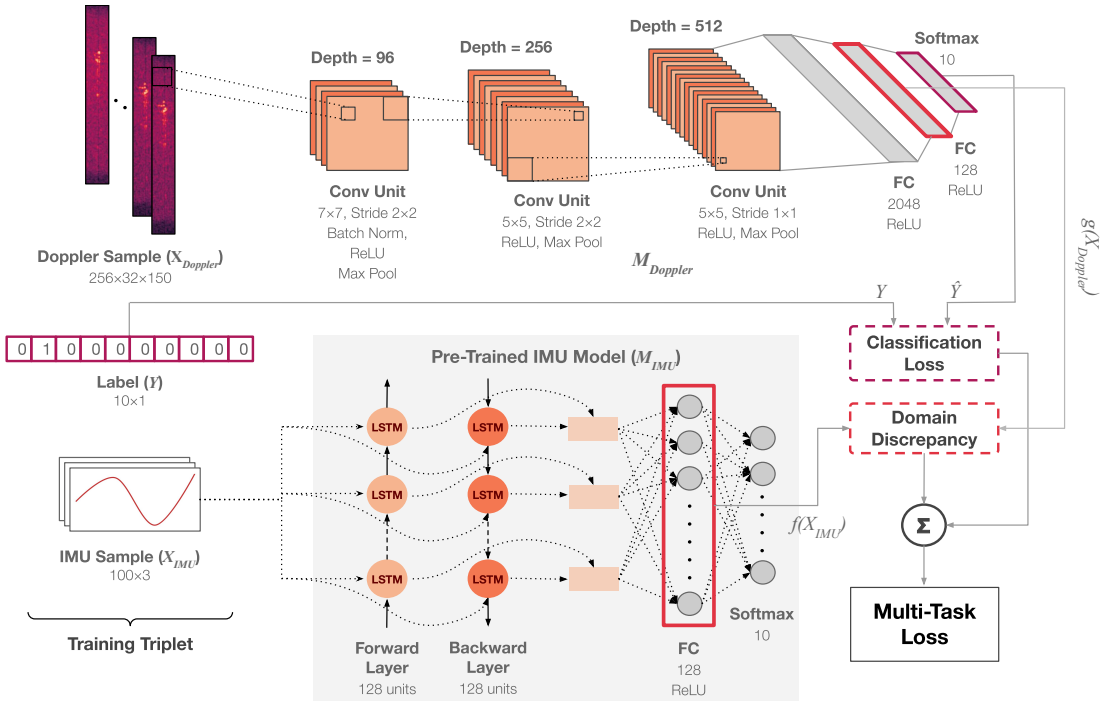


Figure 7.2: Training schematic for cross-modal domain adaptation with IMU data as the souce domain and doppler data as the target domain.

Neural Networks (RNNs) like LSTMs and Bi-LSTMs in modeling long-range temporal dependencies. Moreover, since a bidirectional-LSTM layer consists of two LSTM layers that operate on the original and reversed copy of the data in parallel, it preserves the information from both the future and the past, thus outperforming an LSTM. Hence, for all further experiments, we use the Bi-LSTM as the neural architecture for the source domain.

Table 7.1: Classification results of different models on a subset of WISDM Dataset (10 activities)

Model	Trainable Parameters	Accuracy ± SD
1D CNN	453,002	79.16 ± 3.35
LSTM	496,010	80.67 ± 2.92
Bi-LSTM	169,354	**83.34** ± **4.23**

7.2.4 End-To-End Learning Algorithm

The pipeline begins with training a Bi-LSTM classifier, M_{IMU}, on the entire IMU dataset. M_{IMU} acts as the pre-trained model used for domain adaptation. The goal is to train a new model, $M_{Doppler}$, for classifying the limited la-

beled doppler data. Each doppler sample, $X_{Doppler}$, is paired with a random IMU sample, X_{IMU}, having the same activity label Y, to form training triplets of the form $(X_{IMU}, X_{Doppler}, Y)$. To extract a latent feature representation, we consider the output of the second-to-last fully connected layer, which has an identical configuration in both M_{IMU} and $M_{Doppler}$. Let the sequential transformation of all the layers before the pre-final layer be denoted by $f(.)$ and $g(.)$ for M_{IMU} and $M_{Doppler}$ respectively. The learning objective of $M_{Doppler}$ is to optimise the weighted sum of the mean-squared error between the latent representations, i.e. $|g(X_{Doppler}) - f(X_{IMU})|^2$, and the categorical cross-entropy loss between the predicted softmax values, $\hat{Y}$, and actual label, Y. Mathematically, we define our objective function $\mathcal{L}$ as follows:

$$\mathcal{L}(X_{IMU}, X_{Doppler}, Y) = \alpha \times |g(X_{Doppler}) - f(X_{IMU})|^2 + \beta \times -\sum_i Y^{(i)} log \hat{Y}^{(i)}$$

Here, $Y^{(i)}$ and $\hat{Y}^{(i)}$ denote the value of the i^{th} class in the actual and predicted one-hot encoded labels respectively. Adam optimiser (Learning Rate: 0.001), coupled with a learning rate decay of 0.1 (to check for validation loss saturation) is used to optimise $\mathcal{L}$. Empirically, we found the value of $\alpha = 1.3$ and $\beta = 0.7$ to produce the best performing classifier. To further accelerate the learning, we initialise the final layer of $M_{Doppler}$ with weights of the corresponding layer in the pre-trained M_{IMU}. In this way, the knowledge in M_{IMU}, in the form of learned parameter values and input-output mapping, is effectively transferred to $M_{Doppler}$. The entire training schematic is visualised in Figure 7.2.

7.2.5 Doppler: Data Processing And Neural Architecture

We apply the azimuth-range-doppler algorithm on our collected doppler dataset to get rolling spectrograms consisting of 256 frequency bins and 32 time steps (representing nearly 0.01s of data). We construct sequences of these spectrograms by using a sliding window of 5s with a step size of 1s. When stacked together, each window consists of 150 (5s $\times$ 30 Hz) frames of 256 $\times$ 32 spectrograms (final input size: 150 $\times$ 256 $\times$ 32).

To compare the results of our domain adaptation approach and determine the best neural architecture for the target domain, i.e. doppler data, we trained and evaluated different models chosen from state-of-the-art deep learning architectures that are generally adapted in a wide range of

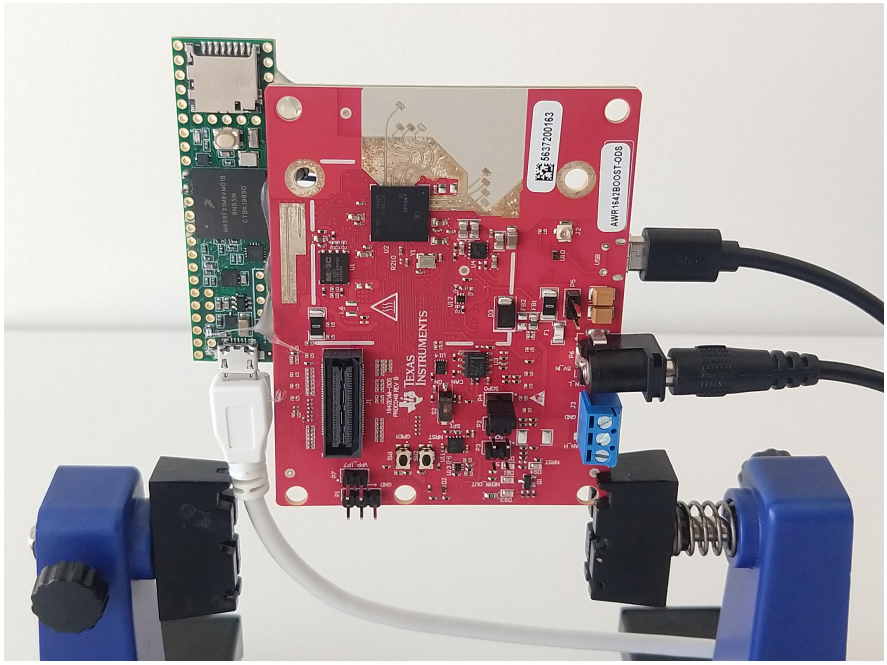


Figure 7.3: We used TI's AWR1642 Radar sensor for our data collection.

applications. We compared the performance of a 3-layer 2D CNN, a 5-layer 2D CNN, a CNN-LSTM and a CNN-Bi LSTM, on our processed dataset of spectrogram sequences. Standalone LSTMs and Bi-LSTMs can't be considered since the dataset comprises sequences of 2D images, which need to be condensed into sequences of 1D vectors before they can be processed by an LSTM layer. For this purpose, we added a CNN encoder before the first LSTM layer in order to extract the spatial information from each image while modeling the temporal dependencies of a sequence. Thus, both CNN-LSTM and CNN-Bi LSTM consist of a 3-layer time distributed 2D-CNN, followed by 2 LSTM layers (Units: [128, 256]) and 2 Bi-LSTM layers (Units: [128, 256]) respectively. On the other hand, the 5-layer and 3-layer 2D-CNNs just consist of 5 and 3 Convolutional Units (convolution layer and a max pooling layer) respectively. Following the same structure as the IMU model, each model is connected to a 128-unit fully-connected layer, followed by the output layer with Softmax activation for classification. As shown in Section 7.4.1, the 3-layer 2D-CNN proved to be the optimal neural architecture for modeling our dataset.

7.3 Data Collection

7.3.1 Participants And Apparatus

We collected data from 9 participants (6 males, 3 females), ranging in age from 20-32 (Mean: 26.3, SD: 3.8). The data was collected in a lab space roughly 5.2 x 6.5 x 2.8 m. We used TI's AWR1642 doppler radar sensor (Figure 7.3) to record SAR data at a sampling rate of 30 Hz. The sensor was placed at a distance of approximately 2m from the participants. All activities were recorded using a laptop and ground truth was collected with an accompanying video camera.

All activities except clapping, jogging and walking required additional apparatus. The food was packaged in the same takeout containers for each participant. All participants used manual toothbrushes from the same brand. We provided the participants with a tennis ball and a basketball for the catching and dribbling tasks. We did not control the apparatus for drinking and folding clothes. The participants were given different sized cups out of convenience; and the participants typically folded their own clothes (e.g. jackets).

7.3.2 Experimental Design And Protocol

Before beginning the data collection, the researcher introduced the protocol and briefed the participants about the activities. We conducted an extensive within-subject study in which we recorded 1500s of data (10 activities $\times$ 150 seconds) from each participant. For each activity, we divided the recording into 10 sessions of 15s each.

7.4 Results

In this section, we evaluate the performance of the proposed cross-modal domain adaptation approach via extensive experiments on our collected dataset comprising 10 different subjects.

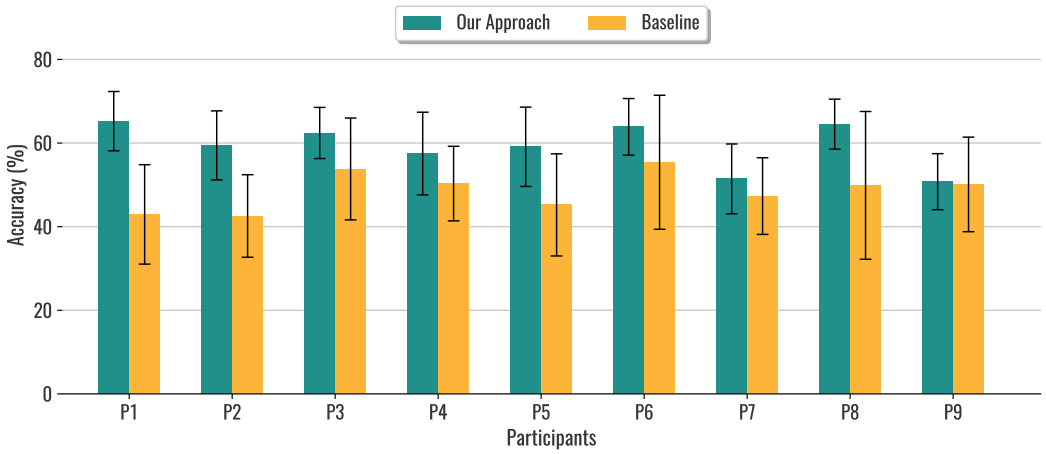
7.4.1 Doppler Model Selection/Baselines

We consider a subset of the doppler data, consisting of three subjects, for determining an appropriate baseline. We built three per-user classifiers, one for each subject, for each type of model using a leave-n-sessions-out

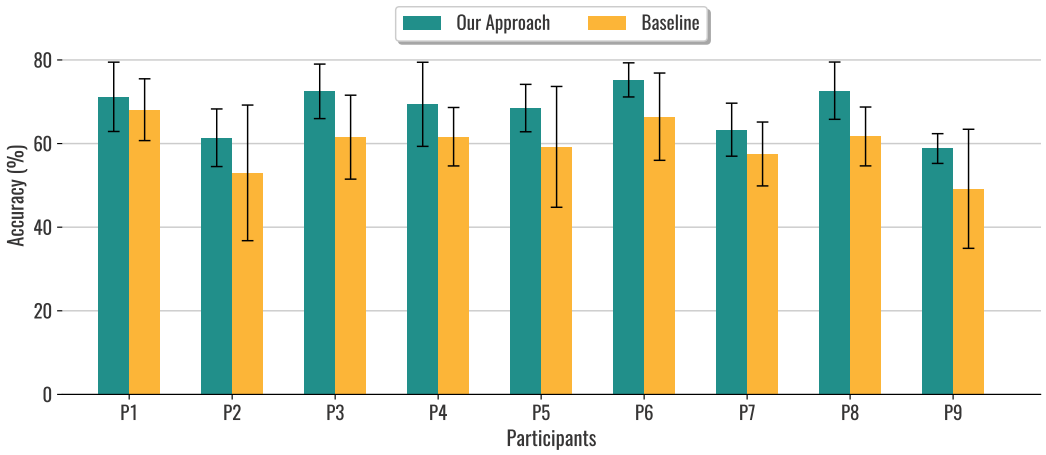
Table 7.2: Baseline Results for different models across 10 activities and 3 subjects

Model	Input Size	Trainable Parameters	Accuracy $\pm$ SD
3-Layer 2D CNN	$N \times 256 \times 32 \times 150$	11,154,922	**76.55 $\pm$ 5.25**
5-Layer 2D CNN	$N \times 256 \times 32 \times 150$	15,874,538	68.12 $\pm$ 0.01
CNN-LSTM	$N \times 150 \times 256 \times 32 \times 1$	4,765,504	52.57 $\pm$ 3.01
CNN-Bi LSTM	$N \times 150 \times 256 \times 32 \times 1$	9,698,624	53.94 $\pm$ 22.53

scheme (n = 9). Here, in each fold, we train the model on one session of data (training set), calibrate the hyperparameters on another session (validation set) and evaluate the performance on the remaining eight sessions (test set). Thus, for this experiment, we obtain a total of 110 training instances (10 activities $\times$ 1 session $\times$ 11 instances/session; each session is 15s long per activity), 110 validation instances and 880 test instances. Table 7.2 summarizes the classification accuracies of all models, the number of associated trainable parameters and the required shape of the input data. Despite the limited training data, the 3-layer CNN produces an accuracy of 76.55%, thereby outperforming the rest. In fact, we can generalise that for our task, CNNs perform significantly better than the CRNNs (Convolutional-Recurrent Neural Networks; CNN-LSTM and CNN-Bi LSTM). The superiority of CNNs can be attributed to the way in which the input data is modeled by the two architectures. While both use convolutional layers as feature extractors, the CNNs interpret the entire sequence as a multi-channel image (stacked spectrograms), unlike the CRNNs, which treat each frame of the sequence individually before fusing the extracted features and passing them through an LSTM. The former allows a more comprehensive representation of the sequence by systematically organizing the temporal information as spatial neighbours. The latter, on the other hand, diminishes the local intra-frame temporal dependencies. Lastly, additional convolutional layers in the 5 layer network make the model unnecessarily complex for a small dataset, thus leading to overfitting. Therefore, we find a 3-layer CNN to be the most suitable for effectively modeling doppler spectrograms. We deploy our domain adaptation approach on the same to compare against the best-performing baseline.



(a) (5s, 5s)



(b) (10s, 5s)

Figure 7.4: Per-user comparison of our approach and the baseline under two labeled data distributions consisting of 5s and 10s of training data per class respectively, combined with 5s of validation data per class. The error bar indicates the variation (Standard Deviation) across different folds.

7.4.2 Domain Adaptation Results

Finally, we evaluate the performance of our proposed approach under varying conditions. We primarily vary the amount of labeled data used in the learning procedure with the objective of navigating the tradeoff between minimizing the amount of annotated data learned by the model and increasing the resultant performance. Starting from 10 seconds of data per class, we examine different sizes of annotated data up to 30 seconds. With increments of 5s, we obtain a total of 5 durations: 10s, 15s (or 14s), 20s, 25s, and 30s. Each of these durations entail different combinations of training and validation size (see Table 7.3), represented by (T, V), where T denotes the training size per activity (in seconds) and V denotes the validation size per activity (in seconds). For instance, a model can be exposed to 20s of labeled data in two ways: (15s, 5s) or (10s, 10s). Further, if an entire session is not consumed in the training or validation set, we discard the remainder to prevent information leakage. This ensures that no two windows belonging to the same session are present in two different sets.

We adopt a leave-n-sessions-out scheme to train the models, where *n* represents the number of sessions that are not a part of the training set. *n* can take different values depending on the size of the training set. For example, if we consider a training size that is greater than the session size (15s), say 20s, we will require 2 sessions for training. This leaves us with 1 session for validation and 7 sessions for the test set (*n* = 8), thus leading to a total of ${}^{10}C_8$ combinations of train-validation-test sets. On the other hand, with a training size of 10s, one session would suffice for training (*n* = 9) and we'll obtain ${}^{10}C_9$ train-validation-test sets. Each set was trained for a maximum of 500 epochs with Adam optimizer (Learning Rate: 0.01) coupled with a learning rate decay of 0.1 and early stopping on the validation set with a patience of 100. The average of the results across ${}^{10}C_n$ runs is reported for each (T, V) configuration.

Augmentation: Different (T, V) configurations imply varying numbers of training and validation instances. After following the window segmentation procedure described in Section 7.2.5, we obtain 1, 3, 6, 11, 12, and 17 instances from a total of 5s, 7s, 10s, 15s, 20s, and 25s of data respectively. To account for the limited training samples in case of (5s, 5s) and (7s, 7s), we augment the doppler dataset by pairing each doppler sample with three IMU samples in order to get three training triplets, $(X_{Doppler}, X_{IMU\_a}, Y)$,

Table 7.3: Classification accuracy of our proposed domain adaptation approach and baseline for different amounts of training and validation data, averaged across 10 subjects.

Amount of Labeled Data	Configuration (Training size, Validation size)	Accuracy of Our Approach (in %)	Accuracy of Baseline (in %)
10s	(5s, 5s)	59.36	48.61
15s	(7s, 7s)	70.00	64.16
	(10s, 5s)	68.18	59.81
20s	(10s, 10s)	72.68	68.01
	(15s, 5s)	71.80	66.67
25s	(15s, 10s)	74.58	70.03
	(20s, 5s)	74.46	69.28
30s	(15s, 15s)	75.68	72.55
	(20s, 10s)	77.15	73.18

$(X_{Doppler}, X_{IMU\_b}, Y)$, and $(X_{Doppler}, X_{IMU\_c}, Y)$. As a result, we have 3 and 9 instances per activity for (5s, 5s) and (7s, 7s) respectively.

Table 7.3 summarizes the domain adaptation results for different (T, V) configurations and their corresponding baseline results. It shows convincing evidence that micro-doppler based human activity classifiers can learn from the knowledge of a pre-trained IMU model. All configurations show a jump of at least 3% from the baseline, with maximum difference observed in the lower training and validation sizes. The (5s, 5s), (7s, 7s) and (10s, 5s) configurations show an improvement of approximately 10%, 6% and 9% over the baseline respectively. A deeper look at the classifiability of individual activities for the (10s, 5s) configuration (see Figure 7.5) shows that our model confuses between classes like eating soup, eating pasta and drinking. These activities can be broadly represented by a similar hand-to-mouth gesture, thus showing limited distinction in the source domain as well. In the course of knowledge transfer, these inherently similar motion characteristics transfer from the source to the target domain and consequently translate into the observed confusion. Nevertheless, this result is quite encouraging as the difference in recognition performance of our proposed approach from the baseline is considerable in spite of the heterogeneity of the source and target domains. As anticipated, the classification accuracy increases with an increase in the amount of annotated data used for training. The classifier trained on 20s and validated on 10s of data per activity produces the best average accuracy of 77.15% (max participant accuracy: 85.47%), highlighting the role of proportionately distributing our minimally labeled data into training and validation.

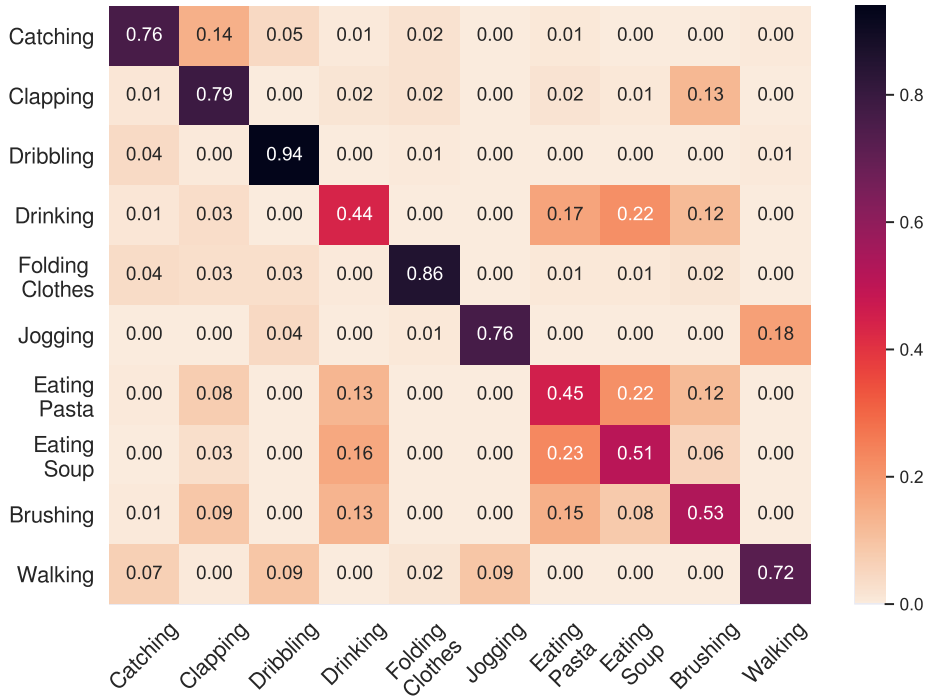


Figure 7.5: Confusion matrix for a 3-layer 2D CNN trained with 10s of training and 5s of validation data. The figure represents the combined results of all participants.

The results of these experiments substantially support the feasibility and effectiveness of the proposed supervised domain adaptation approach. Demonstrated over a range of locomotion and other complex daily activities, domain transformed micro-doppler representations are seen to better capture motion information in comparison with the original micro-doppler spectrograms.

7.4.3 Domain Adaptation Using Multiple Datasets Combined As A Single Source

Although our approach relies on the transfer of higher-level domain-invariant features, we verify the same by distilling information from multiple datasets combined into a single source domain. We constructed a new IMU dataset by replacing the data of two activities in our current dataset, namely *walking* and *jogging*, by corresponding samples drawn from the Wearable Activity Recognition Dataset (WARD) [200]. WARD consists of sequences of 13 human actions (including walking and jogging) collected from 20 participants by a network of 5 sensors placed at different body positions (in-

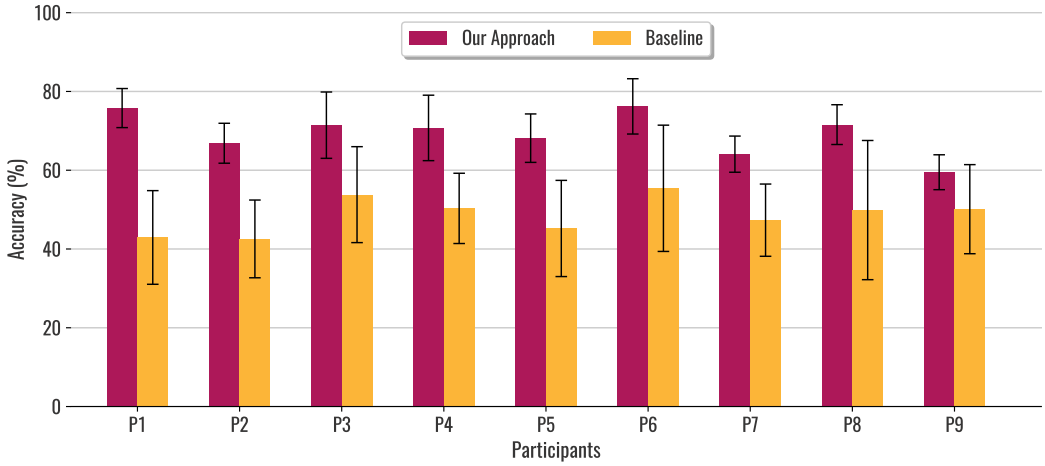


Figure 7.6: Per-user comparison of baseline and proposed domain adaptation approach with a heterogeneous source domain comprising of two datasets. The training and validation data contain 15s of labeled data per class (10s, 5s). The error bar indicates the variation (Standard Deviation) across different folds.

cluding the wrist), each carrying a triaxial accelerometer and a biaxial gyroscope. In order to maintain class balance in the proposed dataset, we randomly selected 20 participants from our current dataset before combining it with the wrist accelerometer data from WARD. Due to the inconsistency in the sensor specifications, participants and other external factors, we normalised the mixed dataset as a whole to account for the incompatible data distributions across activities from the two datasets.

We trained a Bidirectional LSTM, which proved to be the best classifier for IMU data, from scratch for the mixed dataset. Adopting the same training procedure as followed in Section 7.2.3, we achieved an accuracy of 80.36%, which is at par with the results for a homogeneous IMU dataset. Using this model for extracting learned latent feature representations, we evaluated the performance of our domain adaptation approach with a training and validation size of 10s and 5s, respectively. With an average **per-user accuracy of 69.37%** (see Figure 7.6), the results not only indicate invariance to heterogeneity in the source domain but also show a marginal increase in comparison to the previous results together with a high accuracy/classifiability for the classes belonging to the minority dataset (walking, jogging). Thus, the results of this experiment highlight the potential of leveraging the sizeable collection of IMU datasets that cover a multitude of human actions, to build a more comprehensive human activity classifier

for doppler data.

7.5 Limitations And Discussion

In this section, we discuss some key limitations of our work and reflect on how it may impact the usability and deployability of our approach. We also discuss how our work may contribute to future research directions.

7.5.1 Classifier Accuracy For Real World Use

Our work demonstrates success in domain adaptation and is able to outperform the baseline consistently. However, even with 30 seconds of labeled data, our approach is able to classify these 10 activities with an accuracy of 77.15% (compared to 73.18% with baseline). We acknowledge that this is not sufficient for a system to be deployed in the real world. However, we believe that our approach can be used in combination with other strategies such as meta-labeling [189] and active learning [188] designed to improve the classifier accuracy and robustness over time. Such approaches typically require a 'good enough' base model that can be used to make initial, out-of-the-box predictions. However, building that base model is also not easy without significant labeled data. Our approach can assist in rapidly building these base models to facilitate these techniques that can learn and improve over time without introducing significant data labeling cost.

7.5.2 Limited Activities In Source Domain

Our work shows that we can use existing off-the-shelf IMU datasets to train a mmWave radar sensor. While our approach is robust, the multi-task learning method can only be leveraged to train the mmWave radar with activities that are distinctively recognizable in the source domain. Our approach only helps augment the training process for the activities that IMU can reliably characterize. Our solution is not a catch-all and despite this limitation, a vast body of prior IMU work means that our approach can be a catalyst to improve deployability of mmWave radar sensor for activity recognition. In fact, our work can potentially leverage prior work to convert the extensive video datasets into virtual IMU streams [201] and then

use those virtual IMU streams to train the doppler sensor to recognize a wide gamut of activities.

7.5.3 Controlled Environment For User Study

Despite promising results, one key limitation of our work is that the study was conducted in a controlled environment. The users were free to perform the actions/activities as they normally would but they were recorded in a largely static environment. There were no other motions except the primary user in the field of view of the mmWave radar sensor. The source domain (IMU) is impervious to this challenge, but the doppler sensor captures a wide range of motions in the environment. This limitation needs to be overcome before our work can be deployed widely. Fortunately, newer doppler radar sensors are bundled with person tracking algorithms [1] that can be leveraged to sample the doppler from the primary user. Secondly, our work can still be used in scenarios where only a single user with (mostly) static background would be expected. For example, a small office, single-owned apartment or a home gym.

7.6 Conclusion

A fundamental challenge of scaling up any machine-learning system, especially activity recognition systems has been collecting and labeling the data required to train a model. Every few years there is a new sensor in the market that shows promise either due to the signal it is able to capture or advancements in the software and compute capabilities (*e.g.,* surge of computer vision in recent years). In this paper, we tackle the challenge of data collection and labeling with the new and promising mmWave radar sensor. We showcase that we can use existing IMU datasets to learn a latent feature representation that can be used by the mmWave radar sensor to classify between 10 activities with minimal data labeling of its own data (10 seconds).

Our approach not only demonstrates successful heterogeneous domain adaptation, but importantly also works with off-the-shelf datasets. From a real world perspective, it means that not only existing IMU datasets can be used to train the mmWave radar sensor, we can catalogue and label

[1] https://www.ti.com/tool/TIDEP-01000

a library of activities recorded using IMU-ladden smartwatches which can be then be used to train sensors such as the mmWave radar. This is an improvement over simply collecting and labeling doppler data because: (1) mmWave radar sensors are not widely adopted or used which makes it hard to do a large data collection; and (2) it is harder to collect the ground truth required for doppler data as it would potentially require cameras (and video coders) or dedicated user time in front of the doppler for direct labeling. On the other hand, smartwatches are popular with a large user base and they have the capability to passively sense, record and label activities with minimal user disruption.

CHAPTER 8

CONCLUSION

The thesis started with a lofty goal of tackling four different challenges that inhibit a practical privacy-preserving system.

First I conducted a large scale study to understand how different camera based sensing systems impact a user's privacy preferences. I demonstrate that just the mere fact that a user is aware of how a camera-based system processes the data being collected instills more trust in the system. I also show variance in trust depending on the technique being used with optical flow and pose detection being considered the most trustworthy.

I build upon my results and use these privacy preserving sensing techniques to solve the next two big challenges: (1) building robust and accurate activity recognition techniques using ambient sensors; and (2) user identification in a shared space.

In GymCam [17], I built a vision-based system that uses off-the-shelf cameras to automate exercise tracking and provide high-fidelity analytics, such as repetition count, without any user or environment-specific training or intervention. To develop and evaluate our machine learning algorithms, we collected data in our university's gym for five days. It was a first of its kind unconstrained evaluation of a fitness tracking system. In our dataset, there were several instances where 25 users were tracked at the same time using GymCam. It is a practical sensing system that is resilient to challenges seen in a real gym. In our dataset, there were several instances where multiple users were tracked at the same time using GymCam. A challenge with applications that sense or track activities at scale such as GymCam is how to identify each user in the space. The ability to identify users in a shared space is crucial to enable applications and products that can passively sense different activities, offer personalized service, and pro-

vide feedback to the user. In MotionID [18], I built a user identification approach that uses a lightweight machine learning model to couple motion profiles from a regular RGB camera and a smartwatch worn by the user. This reliance of this hybrid approach on the smartwatch makes user identification a feature similar to location sharing that can be enabled/disabled by the user depending on their privacy preferences in any environment. I evaluated my approach in different group sizes (2, 4, and 8) across three different activities (poster session, playing sports, and coordinated dancing). Besides collecting data in a natural setting, these activities allowed me to evaluate MotionID with a gamut of motions ranging from tiny movements in poster sessions to large displacements in team sports. To further validate its robustness, I also evaluated MotionID in coordinated dancing where the observed difference between user movements is miniscule.

The last grand challenge outlined at the start of this thesis was the need for labeled data to build a robust sensing system such as the ones described above. In my thesis work, I tackle the challenge of data collection and labeling with the new and promising mmWave radar sensor. I showcase that we can use existing IMU datasets to learn a latent feature representation that can be used by the mmWave radar sensor to classify between 10 activities with minimal data labeling of its own data (10 seconds). My approach not only demonstrates successful heterogeneous domain adaptation, but importantly also works with off-the-shelf datasets.

In summary, the key to my research has been making sensing and interaction practical. I will further my agenda to make deployable AI and sensing platforms, particularly on edge devices. My work is inherently interdisciplinary and offers opportunities to collaborate with other machine learning experts, industrial designers, privacy gurus and other domain-specific specialists. Here are some future directions that spawn from my thesis work.

8.1 Future Work

In the future, I plan to build upon my research and expand it to other application domains. I have previously explored ambient sensing opportunities in fitness and sports, but I am eager to solve high-impact problems in other domains and shared spaces such as airports, factories, and evens entire neighborhoods. Simultaneously, I want to expand sensing for the

individual to domains such as accessibility, privacy and education where individual agency and ownership is of paramount importance. Here I outline some research avenues that I am excited to explore:

8.1.1 Improving Privacy Control Of Sensing Systems:

Privacy management is a huge concern with sensing systems, especially in shared user spaces. Ambient sensors in such environments like cameras tend to be privacy invasive. My thesis work addresses some of these concerns by using privacy-by-design techniques for featurizing the raw image data and providing user control over who gets access to their identity. However, these do not completely mitigate all privacy concerns a user may have. The limited control of a user over such sensors that capture a high amount of sensitive information in a foreign environment leads to privacy concerns. These ambient systems have the same privacy policy for everyone in the shared space. I want to explore methods to enable each user with individualistic control over their privacy. For example, if Alice does not want to be tracked through the mall, they should be able to opt out but if Bob wants to be tracked to leverage personalized recommendations, they should also be able to do so in the same shared space.

8.1.2 Data Labeling Techniques For Building Deployable Systems:

IMU2Doppler has shown promise of using domain adaptation to teach ambient sensors how to recognize activities with minimally labeled data. We can take this work a step further and address some of the limitations of the current work. One avenue would be to examine a range of ambient sensors that can learn from a translation invariant IMU dataset. Another research avenue that I wish to pursue in this domain is to explore other new approaches such as active learning that can leverage IMU2Doppler and then be used to deploy a practical system. Efforts in this space have largely been concentrated on specific adaptations due to the high degree of complexity; however, if one were to work towards the grand challenge of making domain adaptation more accessible to a developer, that may not only propel this space forward, but also lead to curbing data labeling as a challenge in machine learning.

8.1.3 Practical & Robust Evaluation Systems

A key component of this thesis was implementing unconstrained evaluation protocols to determine the robustness and practicality of our work. While these protocols were carefully designed to capture realistic data, besides measuring accuracy/precision, there is no feedback loop built into the system to tell the developer that the data is not representative. In fact, there are even no metrics that capture this underspecificity of data. This area is in its nascent stages, however determining the right metrics to capture the underspecificity and providing useful feedback to the developer for data collection has the potential to impact how we build machine learning systems in the future.

BIBLIOGRAPHY

[1] A. C. Clarke. *Profiles of the Future, rev. ed*, (1973).

[2] K. I. Withanage, I. Lee, R. Brinkworth, S. Mackintosh, and D. Thewlis, *Fall recovery subactivity recognition with rgb-d cameras*, IEEE transactions on industrial informatics **12**, 2312–2320 (2016).

[3] C. A. Ronao and S.-B. Cho, *Human activity recognition with smartphone sensors using deep learning neural networks*, Expert systems with applications **59**, 235–244 (2016).

[4] R. Khurana and M. Goel. *Eyes on the Road: Detecting Phone Usage by Drivers Using On-Device Cameras*. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11, (2020).

[5] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell. *Unobtrusive sleep monitoring using smartphones*. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 145–152. IEEE, (2013).

[6] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, *Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study*, Parkinsonism & related disorders **21**, 650–653 (2015).

[7] P. Mohan, V. N. Padmanabhan, and R. Ramjee. *Nericell: rich monitoring of road and traffic conditions using mobile smartphones*. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 323–336, (2008).

[8] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, *Participatory*

air pollution monitoring using smartphones, Mobile Sensing **1**, 1–5 (2012).

[9] A. Bedri, D. Li, R. Khurana, K. Bhuwalka, and M. Goel. *FitByte: Automatic Diet Monitoring in Unconstrained Situations Using Multimodal Sensing on Eyeglasses*. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, (2020).

[10] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber. *Smartwatch-based activity recognition: A machine learning approach*. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 426–429. IEEE, (2016).

[11] R. Khurana, *The past, the present, and the future of fitness tracking*, XRDS: Crossroads, The ACM Magazine for Students **25**, 30–33 (2019).

[12] D. Morris, T. S. Saponas, A. Guillory, and I. Kelner. *RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises*. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3225–3234. ACM, (2014).

[13] Y. Lee and M. Song, *Using a smartwatch to detect stereotyped movements in children with developmental disabilities*, IEEE Access **5**, 5506–5514 (2017).

[14] M. Shoaib, H. Scholten, P. J. Havinga, and O. D. Incel. *A hierarchical lazy smoking detection algorithm using smartwatch sensors*. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6. IEEE, (2016).

[15] R. McNaney, J. Vines, D. Roggen, M. Balaam, P. Zhang, I. Poliakov, and P. Olivier. *Exploring the acceptability of google glass as an everyday assistive device for people with parkinson's*. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 2551–2554, (2014).

[16] W. Glauser, *Doctors among early adopters of Google Glass*, Canadian Medical Association. Journal **185**, 1385 (2013).

[17] R. Khurana, K. Ahuja, Z. Yu, J. Mankoff, C. Harrison, and M. Goel, *GymCam: Detecting, recognizing and tracking simultaneous exercises*

in unconstrained scenes, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**, 1–17 (2018).

[18] R. Khurana, C. Dugue, Z. Yu, J. Ramos, and M. Goel, *MotionID: Using Camera and Smartwatch Motions to Recognize Users in a Group*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**, 1–17 (2020).

[19] I. D. Addo, S. I. Ahamed, S. S. Yau, and A. Buduru. *A reference architecture for improving security and privacy in internet of things applications*. In *2014 IEEE International conference on mobile services*, pages 108–115. IEEE, (2014).

[20] H. Lee and A. Kobsa. *Understanding user privacy in Internet of Things environments*. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pages 407–412. IEEE, (2016).

[21] H. Lee and A. Kobsa. *Privacy preference modeling and prediction in a simulated campuswide IoT environment*. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 276–285. IEEE, (2017).

[22] S. Lederer, J. Mankoff, and A. K. Dey. *Who wants to know what when? privacy preference determinants in ubiquitous computing*. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 724–725, (2003).

[23] S. Zheng, N. Apthorpe, M. Chetty, and N. Feamster, *User perceptions of smart home IoT privacy*, Proceedings of the ACM on Human-Computer Interaction **2**, 1–20 (2018).

[24] E. K. Choe, S. Consolvo, J. Jung, B. Harrison, and J. A. Kientz. *Living in a glass house: a survey of private moments in the home*. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 41–44, (2011).

[25] E. McReynolds, S. Hubbard, T. Lau, A. Saraf, M. Cakmak, and F. Roesner. *Toys that listen: A study of parents, children, and internet-connected toys*. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5197–5207, (2017).

[26] P. Worthy, B. Matthews, and S. Viller. *Trust me: doubts and concerns living with the Internet of Things*. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pages 427–434, (2016).

[27] P. E. Naeini, S. Bhagavatula, H. Habib, M. Degeling, L. Bauer, L. F. Cranor, and N. Sadeh. *Privacy expectations and preferences in an IoT world*. In *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*, pages 399–412, (2017).

[28] N. Apthorpe, Y. Shvartzshnaider, A. Mathur, D. Reisman, and N. Feamster, *Discovering smart home internet of things privacy norms using contextual integrity*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**, 1–23 (2018).

[29] N. Sadeh, J. Hong, L. Cranor, I. Fette, P. Kelley, M. Prabaker, and J. Rao, *Understanding and capturing people's privacy policies in a mobile social networking application*, Personal and Ubiquitous Computing **13**, 401–412 (2009).

[30] J. Y. Tsai, P. Kelley, P. Drielsma, L. F. Cranor, J. Hong, and N. Sadeh. *Who's viewed you? The impact of feedback in a mobile location-sharing application*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2003–2012, (2009).

[31] G. Khan, Z. Tariq, and M. U. G. Khan. *Multi-Person Tracking Based on Faster R-CNN and Deep Appearance Features*. In *Visual Object Tracking in the Deep Neural Networks Era*. IntechOpen, (2019).

[32] A. Nunez-Marcos, G. Azkune, and I. Arganda-Carreras, *Vision-based fall detection with convolutional neural networks*, Wireless communications and mobile computing **2017** (2017).

[33] I. Ar and Y. S. Akgul, *A computerized recognition system for the home-based physiotherapy exercises using an RGBD camera*, IEEE Transactions on Neural Systems and Rehabilitation Engineering **22**, 1160–1171 (2014).

[34] D. Antón, A. Goñi, A. Illarramendi, et al., *Exercise recognition for Kinect-based telerehabilitation*, Methods Inf Med **54** (2015).

[35] W.-H. Liao and C.-M. Yang. *Video-based activity and movement pattern analysis in overnight sleep studies*. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, (2008).

[36] D. Falie and M. Ichim. *Sleep monitoring and sleep apnea event detection using a 3D camera*. In *2010 8th International Conference on Communications*, pages 177–180. IEEE, (2010).

[37] C. R. Dreher, M. Wächter, and T. Asfour, *Learning Object-Action Relations from Bimanual Human Demonstration Using Graph Networks*, IEEE Robotics and Automation Letters **5**, 187–194 (2019).

[38] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu. *Learning actor relation graphs for group activity recognition*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, (2019).

[39] Y. Bo, Y. Lu, and W. He. *Few-Shot Learning of Video Action Recognition Only Based on Video Contents*. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 595–604, (2020).

[40] S. Kumar Dwivedi, V. Gupta, R. Mitra, S. Ahmed, and A. Jain. *ProtoGAN: Towards Few Shot Learning for Action Recognition*. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, (2019).

[41] C. Careaga, B. Hutchinson, N. Hodas, and L. Phillips. *Metric-Based Few-Shot Learning for Video Action Recognition*. (2019).

[42] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. Torr, and P. Koniusz. *Few-shot Action Recognition via Improved Attention with Self-supervision*. (2020).

[43] Y. Gu, W. Sheng, C. Crick, and Y. Ou, *Automated assembly skill acquisition and implementation through human demonstration*, Robotics and Autonomous Systems **99**, 1–16 (2018).

[44] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. *You only look once: Unified, real-time object detection*. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, (2016).

[45] K. Kim, A. Jalal, and M. Mahmood, *Vision-Based Human Activity Recognition System Using Depth Silhouettes: A Smart Home System for Monitoring the Residents*, Journal of Electrical Engineering & Technology **14**, 2567–2573 (2019).

[46] L. Liu, L. Shao, X. Zhen, and X. Li, *Learning discriminative key poses for action recognition*, IEEE transactions on cybernetics **43**, 1860–1870 (2013).

[47] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, *Silhouette-based human action recognition using sequences of key poses*, Pattern Recognition Letters **34**, 1799–1807 (2013).

[48] R. Vemulapalli, F. Arrate, and R. Chellappa. *Human action recognition by representing 3d skeletons as points in a lie group.* In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, (2014).

[49] A. Jalal, S. Kamal, and D. Kim, *A Depth Video-based Human Detection and Activity Recognition using Multi-features and Embedded Hidden Markov Models for Health Care Monitoring Systems.*, International Journal of Interactive Multimedia & Artificial Intelligence **4** (2017).

[50] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. *Realtime multi-person 2d pose estimation using part affinity fields.* In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, (2017).

[51] J. Li, W. Su, and Z. Wang. *Simple Pose: Rethinking and Improving a Bottom-up Approach for Multi-Person Pose Estimation.* (2019).

[52] X. Liu, P. Ghosh, O. Ulutan, B. Manjunath, K. Chan, and R. Govindan. *Caesar: cross-camera complex activity recognition.* In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, pages 232–244, (2019).

[53] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. *Large-scale video classification with convolutional neural networks.* In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, (2014).

[54] N. Almaadeed, O. Elharrouss, S. Al-Maadeed, A. Bouridane, and A. Beghdadi. *A novel approach for robust multi human action detection and recognition based on 3-dimentional convolutional neural networks*. (2019).

[55] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, *Summarization of user-generated sports video by using deep action recognition features*, IEEE Transactions on Multimedia **20**, 2000–2011 (2018).

[56] F. Luo, S. Poslad, and E. Bodanese. *Temporal convolutional networks for multi-person activity recognition using a 2D LIDAR*. IEEE, (2020).

[57] D. Avrahami, M. Patel, Y. Yamaura, S. Kratz, and M. Cooper, *Unobtrusive Activity Recognition and Position Estimation for Work Surfaces Using RF-Radar Sensing*, ACM Transactions on Interactive Intelligent Systems (TiiS) **10**, 1–28 (2019).

[58] C. Ding, H. Hong, Y. Zou, H. Chu, X. Zhu, F. Fioranelli, J. Le Kernec, and C. Li, *Continuous human motion recognition with a dynamic range-Doppler trajectory method based on FMCW radar*, IEEE Transactions on Geoscience and Remote Sensing **57**, 6821–6831 (2019).

[59] J. Wu, C. Harrison, J. P. Bigham, and G. Laput. *Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition*. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, (2020).

[60] N. Roy, A. Misra, and D. Cook, *Ambient and smartphone sensor assisted ADL recognition in multi-inhabitant smart environments*, Journal of ambient intelligence and humanized computing **7**, 1–19 (2016).

[61] D. Singh, E. Merdivan, I. Psychoula, J. Kropf, S. Hanke, M. Geist, and A. Holzinger. *Human activity recognition using recurrent neural networks*. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 267–274. Springer, (2017).

[62] A. De Paola, P. Ferraro, S. Gaglio, and G. L. Re. *Context-awareness for multi-sensor data fusion in smart environments*. In *Conference*

of the Italian Association for Artificial Intelligence, pages 377–391. Springer, (2016).

[63] L. Lu, C. Qing-Ling, and Z. Yi-Ju, *Activity recognition in smart homes*, Multimedia Tools and Applications **76**, 24203–24220 (2017).

[64] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, *Device-free human activity recognition using commercial WiFi devices*, IEEE Journal on Selected Areas in Communications **35**, 1118–1131 (2017).

[65] S. Liu, Y. Zhao, F. Xue, B. Chen, and X. Chen. *DeepCount: Crowd counting with WiFi via deep learning*. (2019).

[66] S. Arshad, C. Feng, R. Yu, and Y. Liu. *Leveraging transfer learning in multiple human activity recognition using WiFi signal*. In *2019 IEEE 20th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, pages 1–10. IEEE, (2019).

[67] J. Fogarty, C. Au, and S. E. Hudson. *Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition*. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pages 91–100, (2006).

[68] J. E. Froehlich, E. Larson, T. Campbell, C. Haggerty, J. Fogarty, and S. N. Patel. *HydroSense: infrastructure-mediated single-point sensing of whole-home water activity*. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 235–244, (2009).

[69] J. Froehlich, E. Larson, E. Saba, T. Campbell, L. Atlas, J. Fogarty, and S. Patel. *A longitudinal study of pressure sensing to infer real-world water usage events in the home*. In *International conference on pervasive computing*, pages 50–69. Springer, (2011).

[70] T. Campbell, E. Larson, G. Cohn, J. Froehlich, R. Alcaide, and S. N. Patel. *WATTR: A method for self-powered wireless sensing of water activity in the home*. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 169–172, (2010).

[71] S. N. Patel, M. S. Reynolds, and G. D. Abowd. *Detecting human movement by differential air pressure sensing in HVAC system ductwork: An exploration in infrastructure mediated sensing*. In *International Conference on Pervasive Computing*, pages 1–18. Springer, (2008).

[72] S. N. Patel, K. N. Truong, and G. D. Abowd. *Powerline positioning: A practical sub-room-level indoor location system for domestic use*. In *International Conference on Ubiquitous Computing*, pages 441–458. Springer, (2006).

[73] G. Cohn, E. Stuntebeck, J. Pandey, B. Otis, G. D. Abowd, and S. N. Patel. *SNUPI: sensor nodes utilizing powerline infrastructure*. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 159–168, (2010).

[74] S. N. Patel, T. Robertson, J. A. Kientz, M. S. Reynolds, and G. D. Abowd. *At the flick of a switch: Detecting and classifying unique electrical events on the residential power line (nominated for the best paper award)*. In *International Conference on Ubiquitous Computing*, pages 271–288. Springer, (2007).

[75] M. Enev, S. Gupta, T. Kohno, and S. N. Patel. *Televisions, video privacy, and powerline electromagnetic interference*. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 537–550, (2011).

[76] S. Gupta, M. S. Reynolds, and S. N. Patel. *ElectriSense: single-point sensing using EMI for electrical event detection and classification in the home*. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 139–148, (2010).

[77] S. Gupta, K.-Y. Chen, M. S. Reynolds, and S. N. Patel. *LightWave: using compact fluorescent lights as sensors*. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 65–74, (2011).

[78] G. Cohn, D. Morris, S. N. Patel, and D. S. Tan. *Your noise is my command: sensing gestures using the body as an antenna*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 791–800, (2011).

[79] Y. Zhang, C. Yang, S. E. Hudson, C. Harrison, and A. Sample. *Wall++ Room-Scale Interactive and Context-Aware Sensing*. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–15, (2018).

[80] D. Tao, Y. Guo, Y. Li, and X. Gao, *Tensor rank preserving discriminant analysis for facial recognition*, IEEE transactions on image processing **27**, 325–334 (2017).

[81] B. Nguyen and B. De Baets, *Kernel distance metric learning using pairwise constraints for person re-identification*, IEEE Transactions on Image Processing **28**, 589–600 (2018).

[82] L. Tao, *Jaywalkers under surveillance in Shenzhen soon to be punished via text messages*, South China Morning Post **27** (2018).

[83] E. Rodríguez, C. Gutiérrez, C. Ochoa, F. Trávez, L. Escobar, and D. Loza. *Construction of a Computer Vision Test Platform: VISART for Facial Recognition in Social Robotics*. In *International Conference on Applied Technologies*, pages 637–651. Springer, (2019).

[84] A. C. Hurst, *Facial recognition software in clinical dysmorphology*, Current opinion in pediatrics **30**, 701–706 (2018).

[85] J. Tang, X. Zhou, and J. Zheng. *Design of Intelligent classroom facial recognition based on Deep Learning*. In *Journal of Physics: Conference Series*, volume 1168, page 022043. IOP Publishing, (2019).

[86] S. Rewari, A. Shaha, and S. Gunasekharan, *Facial Recognition Based Attendance System*, Journal of Image Processing & Pattern Recognition Progress **3**, 43–49 (2016).

[87] C. Monteiro, E. Ogasawara, L. Gonçalves, and J. R. de Toledo Quadros. *Control and Security System for Classroom Access Based on Facial Recognition*. In *2018 XLIV Latin American Computer Conference (CLEI)*, pages 654–661. IEEE, (2018).

[88] K. J. Bhojane and S. Thorat, *A review of Face Recognition Based Car Ignition and Security System*, International Research Journal of Engineering and Technology **5**, 532–53 (2018).

[89] A. Patel and A. Verma, *Iot based facial recognition door access control home security system*, International Journal of Computer Applications **172**, 11–17 (2017).

[90] V. A. Kumar, V. A. Kumar, S. Malathi, K. Vengatesan, and M. Ramakrishnan, *Facial Recognition System for Suspect Identification Using*

a Surveillance Camera, Pattern Recognition and Image Analysis **28**, 410–420 (2018).

[91] N. H. Motlagh, M. Bagaa, and T. Taleb, *UAV-based IoT platform: A crowd surveillance use case*, IEEE Communications Magazine **55**, 128–134 (2017).

[92] S. Naker and D. Greenbaum, *Now you see me: Now you still do: Facial recognition technology and the growing lack of privacy*, BUJ Sci. & Tech. L. **23**, 88 (2017).

[93] G. Mokhtari, N. Bashi, Q. Zhang, and G. Nourbakhsh. *Non-wearable human identification sensors for smart home environment: a review*. Emerald Publishing Limited, (2018).

[94] H. Li, P. Zhang, S. Al Moubayed, S. N. Patel, and A. P. Sample. *Id-match: A hybrid computer vision and rfid system for recognizing individuals in groups*. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4933–4944, (2016).

[95] S. Fang, T. Islam, S. Munir, and S. Nirjon. *EyeFi: Fast Human Identification Through Vision and WiFi-based Trajectory Matching*. In *IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, (2020).

[96] E. R. Schafermeyer, E. A. Wan, S. Samin, N. Zentzis, N. Preiser, J. Condon, J. Folsom, and P. G. Jacobs. *Multi-resident identification using device-free IR and RF fingerprinting*. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5481–5484. IEEE, (2015).

[97] Y. Zeng, P. H. Pathak, and P. Mohapatra. *WiWho: wifi-based person identification in smart spaces*. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 1–12. IEEE, (2016).

[98] F. Hong, X. Wang, Y. Yang, Y. Zong, Y. Zhang, and Z. Guo. *WFID: Passive device-free human identification using WiFi signal*. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 47–56, (2016).

[99] M. R. Hodges and M. E. Pollack. *An 'object-use fingerprint': The use of electronic sensors for human identification*. In *International Conference on Ubiquitous Computing*, pages 289–303. Springer, (2007).

[100] T. Terada, R. Watanabe, and M. Tsukamoto. *A user recognition method using accelerometer for electric appliances*. In *2013 16th International Conference on Network-Based Information Systems*, pages 350–355. IEEE, (2013).

[101] M. Yamada, M. Kudo, H. Nonaka, and J. Toyama. *Hipprint person identification and behavior analys*. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 533–536. IEEE, (2006).

[102] R. J. Orr and G. D. Abowd. *The smart floor: A mechanism for natural user identification and tracking*. In *CHI'00 extended abstracts on Human factors in computing systems*, pages 275–276, (2000).

[103] J. Yun, *User identification using gait patterns on UbiFloorII*, Sensors **11**, 2611–2639 (2011).

[104] R. L. de Carvalho and P. F. F. Rosa. *Identification system for smart homes using footstep sounds*. In *2010 IEEE International Symposium on Industrial Electronics*, pages 1639–1644. IEEE, (2010).

[105] A. Mostayed, S. Kim, M. M. G. Mazumder, and S. J. Park. *Foot step based person identification using histogram similarity and wavelet decomposition*. In *2008 International Conference on Information Security and Assurance (isa 2008)*, pages 307–311. IEEE, (2008).

[106] R. Vera-Rodríguez, J. S. Mason, J. Fiérrez, and J. Ortega-García, *Analysis of spatial domain information for footstep recognition*, IET computer vision **5**, 380–388 (2011).

[107] S. Pan, T. Yu, M. Mirshekari, J. Fagert, A. Bonde, O. J. Mengshoel, H. Y. Noh, and P. Zhang, *Footprintid: Indoor pedestrian identification through ambient structural vibration sensing*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**, 1–31 (2017).

[108] T. Teixeira, D. Jung, and A. Savvides. *Tasking networked cctv cameras and mobile phones to identify and localize multiple people*. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 213–222, (2010).

[109] A. D. Wilson and H. Benko. *Crossmotion: fusing device and image motion for user identification, tracking and device association*. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 216–223, (2014).

[110] L. Cabrera-Quiros and H. Hung. *Who is where? Matching people in video to wearable acceleration during crowded mingling events*. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 267–271, (2016).

[111] A. Masullo, T. Burghardt, D. Damen, T. Perrett, and M. Mirmehdi. *Who Goes There? Exploiting Silhouettes and Wearable Signals for Subject Identification in Multi-Person Environments*. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, (2019).

[112] R. Henschel, T. von Marcard, and B. Rosenhahn. *Simultaneous identification and tracking of multiple people using video and IMUs*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, (2019).

[113] Y. Zhu, Y. Chen, Z. Lu, S. Pan, G.-R. Xue, Y. Yu, and Q. Yang, *Heterogeneous Transfer Learning for Image Classification*, AAAI **25** (2011).

[114] S. J. Pan and Q. Yang, *A Survey on Transfer Learning*, IEEE Trans. Knowl. Data Eng. **22**, 1345–1359 (2010).

[115] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, *A theory of learning from different domains*, Mach Learn **79**, 151–175 (2010).

[116] T. Xing, S. S. Sandha, B. Balaji, S. Chakraborty, and M. Srivastava. *Enabling Edge Devices that Learn from Each Other: Cross Modal Training for Activity Recognition*. In *Proceedings of the 1st International Workshop on Edge Systems, Analytics and Networking*, pages 37–42, Munich Germany, (2018). ACM.

[117] B. Tan, Y. Zhang, S. J. Pan, and Q. Yang. *Distant Domain Transfer Learning*. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 2604–2610. AAAI Press, (2017). event-place: San Francisco, California, USA.

[118] J. Blitzer, R. McDonald, and F. Pereira. *Domain Adaptation with Structural Correspondence Learning*. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, USA, (2006). Association for Computational Linguistics. event-place: Sydney, Australia.

[119] M. Harel and S. Mannor. *Learning from Multiple Outlooks*. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 401–408, Madison, WI, USA, (2011). Omnipress. event-place: Bellevue, Washington, USA.

[120] Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang. *Learning Cross-Domain Landmarks for Heterogeneous Domain Adaptation*. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5081–5090, Las Vegas, NV, USA, (2016). IEEE.

[121] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. *Translated Learning: Transfer Learning across Different Feature Spaces*. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, pages 353–360, Red Hook, NY, USA, (2008). Curran Associates Inc. event-place: Vancouver, British Columbia, Canada.

[122] L. Duan, D. Xu, and I. W. Tsang. *Learning with Augmented Features for Heterogeneous Domain Adaptation*. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 667–674, Madison, WI, USA, (2012). Omnipress. event-place: Edinburgh, Scotland.

[123] C. Wang and S. Mahadevan. *Heterogeneous Domain Adaptation Using Manifold Alignment*. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1541–1546. AAAI Press, (2011). event-place: Barcelona, Catalonia, Spain.

[124] Y. Aytar, C. Vondrick, and A. Torralba. *SoundNet: Learning Sound Representations from Unlabeled Video*. In D. Lee, M. Sugiyama,

U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., (2016).

[125] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba, *Cross-Modal Scene Networks*, IEEE Trans. Pattern Anal. Mach. Intell. **40**, 2303–2314 (2018).

[126] V. Radu and M. Henne, *Vision2sensor: Knowledge transfer across sensing modalities for human activity recognition*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **3**, 1–21 (2019).

[127] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison. *Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition*. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, (2021). Association for Computing Machinery.

[128] H. Cai, B. Korany, C. R. Karanam, and Y. Mostofi, *Teaching RF to Sense without RF Training Measurements*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **4**, 1–22 (2020).

[129] K. T. Tran, L. D. Griffin, K. Chetty, and S. Vishwakarma. *Transfer Learning from Audio Deep Learning Models for Micro-Doppler Activity Recognition*. In *2020 IEEE International Radar Conference (RADAR)*, pages 584–589, Washington, DC, USA, (2020). IEEE.

[130] R. Khurana, N. Banovic, and K. Lyons. *In only 3 minutes: perceived exertion limits of smartwatch use*. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 208–211, (2018).

[131] R. Khurana, M. Goel, and K. Lyons, *Detachable Smartwatch: More Than A Wearable*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **3**, 1–14 (2019).

[132] F. Shih, I. Liccardi, and D. Weitzner. *Privacy tipping points in smartphones privacy preferences*. In *Proceedings of the 33rd Annual ACM*

Conference on Human Factors in Computing Systems, pages 807–816, (2015).

[133] B. Liu, J. Lin, and N. Sadeh. *Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help?* In *Proceedings of the 23rd international conference on World wide web*, pages 201–212, (2014).

[134] J. Lin, B. Liu, N. Sadeh, and J. I. Hong. *Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings*. In *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*, pages 199–212, (2014).

[135] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. *You only look once: Unified, real-time object detection*. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, (2016).

[136] A. Jalal, Y. Kim, S. Kamal, A. Farooq, and D. Kim. *Human daily activity recognition with joints plus body features representation using Kinect sensor*. In *2015 International Conference on Informatics, Electronics & Vision (ICIEV)*, pages 1–6. IEEE, (2015).

[137] C. Zhang and Y. Tian, *RGB-D camera-based daily living activity recognition*, Journal of computer vision and image processing **2**, 12 (2012).

[138] A. Doyle, R. Lippert, and D. Lyon, *Eyes everywhere: The global growth of camera surveillance*, Routledge (2013).

[139] K. Albrecht and L. Mcintyre. *Privacy nightmare: When baby monitors go bad [opinion]*. volume 34, pages 14–19. IEEE, (2015).

[140] S. Zhang, Y. Feng, L. Bauer, L. F. Cranor, A. Das, and N. Sadeh. *"Did you know this camera tracks your mood?": Understanding Privacy Expectations and Preferences in the Age of Video Analytics*. volume 2021, pages 282–304. Sciendo, (2021).

[141] L. Palen and P. Dourish. *Unpacking" privacy" for a networked world*. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 129–136, (2003).

[142] B. K. Horn and B. G. Schunck. *Determining optical flow*. volume 17, pages 185–203. Elsevier, (1981).

[143] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. *OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields*. volume 43, pages 172–186. IEEE, (2019).

[144] A. Piergiovanni and M. S. Ryoo. *Fine-grained activity recognition in baseball videos*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1740–1748, (2018).

[145] F. M. Noori, B. Wallace, M. Z. Uddin, and J. Torresen. *A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network*. In *Scandinavian conference on image analysis*, pages 299–310. Springer, (2019).

[146] C. Neustaedter, S. Greenberg, and M. Boyle, *Blur filtration fails to preserve privacy for home-based video conferencing*, ACM Transactions on Computer-Human Interaction (TOCHI) **13**, 1–36 (2006).

[147] V. Braun and V. Clarke, *Using thematic analysis in psychology*, Qualitative research in psychology **3**, 77–101 (2006).

[148] S. P. Dow, A. Glassco, J. Kass, M. Schwarz, D. L. Schwartz, and S. R. Klemmer. *Parallel prototyping leads to better design results, more divergence, and increased self-efficacy*. volume 17, pages 1–24. ACM New York, NY, USA, (2010).

[149] C. Faklaris, L. A. Dabbish, and J. I. Hong. *A self-report measure of end-user security attitudes (SA-6)*. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*, (2019).

[150] A. T.-Y. Chen, M. Biglari-Abhari, I. Kevin, and K. Wang. *Context is King: Privacy Perceptions of Camera-based Surveillance*. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, (2018).

[151] D. Barua, J. Kay, and C. Paris. *Viewing and controlling personal sensor data: what do users want?* In *International Conference on Persuasive Technology*, pages 15–26. Springer, (2013).

[152] I. Janssen and A. G. LeBlanc, *Systematic review of the health benefits of physical activity and fitness in school-aged children and youth*, International journal of behavioral nutrition and physical activity **7**, 40 (2010).

[153] S. R. Colberg, R. J. Sigal, J. E. Yardley, M. C. Riddell, D. W. Dunstan, P. C. Dempsey, E. S. Horton, K. Castorino, and D. F. Tate, *Physical activity/exercise and diabetes: a position statement of the American Diabetes Association*, Diabetes Care **39**, 2065–2079 (2016).

[154] F. R. Roque, A. M. Briones, A. B. García-Redondo, M. Galán, S. Martínez-Revelles, M. S. Avendaño, V. Cachofeiro, T. Fernandes, D. V. Vassallo, E. M. Oliveira, et al., *Aerobic exercise reduces oxidative stress and improves vascular changes of small mesenteric and coronary arteries in hypertension*, British journal of pharmacology **168**, 686–703 (2013).

[155] J. Kruger, H. M. Blanck, and C. Gillespie, *Dietary and physical activity behaviors among adults successful at weight loss maintenance*, International Journal of Behavioral Nutrition and Physical Activity **3**, 17 (2006).

[156] G. Heath, E. H. Howze, E. B. Kahn, and L. T. Ramsey, *Increasing physical activity. A report on recommendations of the Task Force on Community Preventive Services*, MMWR Recomm Rep **50**, 1–14 (2001).

[157] M. Standage, J. L. Duda, and N. Ntoumanis, *A model of contextual motivation in physical education: Using constructs from self-determination and achievement goal theories to predict physical activity intentions.*, Journal of educational psychology **95**, 97 (2003).

[158] D. M. Bravata, C. Smith-Spangler, V. Sundaram, A. L. Gienger, N. Lin, R. Lewis, C. D. Stave, I. Olkin, and J. R. Sirard, *Using pedometers to increase physical activity and improve health: a systematic review*, Jama **298**, 2296–2304 (2007).

[159] R. O. Nelson and S. C. Hayes, *Theoretical explanations for reactivity in self-monitoring*, Behavior Modification **5**, 3–14 (1981).

[160] C. Seeger, A. Buchmann, and K. Van Laerhoven. *myHealthAssistant: a phone-based body sensor network that captures the wearer's exercises throughout the day.* In *Proceedings of the 6th International*

Conference on Body Area Networks, pages 1–7. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), (2011).

[161] C. Seeger, A. Buchmann, and K. Van Laerhoven. *Adaptive gym exercise counting for myHealthAssistant*. In *Proceedings of the 6th International Conference on Body Area Networks*, pages 126–127. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), (2011).

[162] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. *Activity recognition and monitoring using multiple sensors on different body positions*. In *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*, pages 4–pp. IEEE, (2006).

[163] K. Rector, C. L. Bennett, and J. A. Kientz. *Eyes-free yoga: an exergame using depth cameras for blind & low vision exercise*. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, page 12. ACM, (2013).

[164] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. *Convolutional pose machines*. In *CVPR*, (2016).

[165] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, *Joint face detection and alignment using multitask cascaded convolutional networks*, IEEE Signal Processing Letters **23**, 1499–1503 (2016).

[166] J.-Y. Bouguet, *Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm*, Intel Corporation **5**, 4 (2001).

[167] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, *Dense trajectories and motion boundary descriptors for action recognition*, International journal of computer vision **103**, 60–79 (2013).

[168] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, T. Prioleau, M. Y. Beh, M. Goel, T. Starner, and G. Abowd, *EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**, 37 (2017).

[169] R. Chaudhri, J. Lester, and G. Borriello. *An RFID based system for monitoring free weight exercises*. In *SenSys*, (2008).

[170] S. R. Watterson, D. Watterson, and M. D. Watterson. *Systems and Methods to Generate a Customized Workout Routine*, (2013). US Patent App. 13/754,361.

[171] A. Haque, M. Guo, A. Alahi, S. Yeung, Z. Luo, A. Rege, J. Jopling, L. Downing, W. Beninati, A. Singh, et al. *Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance*. (2017).

[172] B. D. Carolis and S. Ferilli, *Learning Daily Routines in Smart Office Environments*, State of the Art in AI Applied to Ambient Intelligence **298**, 122 (2017).

[173] O. Henniger, N. Damer, and A. Braun. *Opportunities for biometric technologies in smart environments*. In *European Conference on Ambient Intelligence*, pages 175–182. Springer, (2017).

[174] S. E. R. Poluan and Y.-A. Chen. *Using Smart Insoles and RGB Camera for Identifying Stationary Human Targets*. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–6. IEEE, (2019).

[175] N. Wojke, A. Bewley, and D. Paulus. *Simple online and realtime tracking with a deep association metric*. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, (2017).

[176] R. Khurana and M. Goel. *Eyes on the Road: Detecting Phone Usage by Drivers Using On-Device Cameras*. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11, (2020).

[177] G. Dogan, I. Cay, S. S. Ertas, Ş. R. Keskin, N. Alotaibi, and E. Sahin. *Where are you? Human activity recognition with smartphone sensor data*. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, pages 301–304, (2020).

[178] C. Zhang, J. Yang, C. Southern, T. E. Starner, and G. D. Abowd. *WatchOut: extending interactions on a smartwatch with inertial sensing*. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, pages 136–143, (2016).

[179] G. Reyes, J. Wu, N. Juneja, M. Goldshtein, W. K. Edwards, G. D. Abowd, and T. Starner, *Synchrowatch: One-handed synchronous smartwatch gestures using correlation and magnetic sensing*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**, 1–26 (2018).

[180] A. Bedri, D. Li, R. Khurana, K. Bhuwalka, and M. Goel. *Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses*. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, (2020).

[181] P. Voigt, M. Budde, E. Pescara, M. Fujimoto, K. Yasumoto, and M. Beigl. *Feasibility of human activity recognition using wearable depth cameras*. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 92–95, (2018).

[182] C. Zhang, Q. Xue, A. Waghmare, S. Jain, Y. Pu, S. Hersek, K. Lyons, K. A. Cunefare, O. T. Inan, and G. D. Abowd, *Soundtrak: Continuous 3d tracking of a finger using active acoustics*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**, 1–25 (2017).

[183] L. Sicong, Z. Zimu, D. Junzhao, S. Longfei, J. Han, and X. Wang, *Ubiear: Bringing location-independent sound awareness to the hard-of-hearing people with smartphones*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**, 1–21 (2017).

[184] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava. *RadHAR: Human Activity Recognition from Point Clouds Generated through a Millimeter-wave Radar*. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems - mmNets'19*, pages 51–56, Los Cabos, Mexico, (2019). ACM Press.

[185] Y. Lin, J. Le Kernec, S. Yang, F. Fioranelli, O. Romain, and Z. Zhao, *Human Activity Classification With Radar: Optimization and Noise*

Robustness With Iterative Convolutional Neural Networks Followed With Random Forests, IEEE Sensors J. **18**, 9669–9681 (2018).

[186] T. Stadelmayer, M. Stadelmayer, A. Santra, R. Weigel, and F. Lurz. *Human Activity Classification Using mm-Wave FMCW Radar by Improved Representation Learning*. In *Proceedings of the 4th ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, pages 1–6, London United Kingdom, (2020). ACM.

[187] V. Chen, Fayin Li, Shen-Shyang Ho, and H. Wechsler, *Micro-doppler effect in radar: phenomenon, model, and simulation study*, IEEE Trans. Aerosp. Electron. Syst. **42**, 2–21 (2006).

[188] B. Settles. *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences, (2009).

[189] M. L. De Prado. *Advances in financial machine learning*. John Wiley & Sons, (2018).

[190] S. Zhao, B. Li, P. Xu, and K. Keutzer. *Multi-source domain adaptation in the deep learning era: A systematic survey*. (2020).

[191] L. Senigagliesi, G. Ciattaglia, A. De Santis, and E. Gambi, *People Walking Classification Using Automotive Radar*, Electronics **9**, 588 (2020).

[192] B. Cagliyan and S. Z. Gurbuz, *Micro-Doppler-Based Human Activity Classification Using the Mote-Scale BumbleBee Radar*, IEEE Geosci. Remote Sensing Lett. **12**, 2135–2139 (2015).

[193] R. Zhang and S. Cao, *Real-Time Human Motion Behavior Detection via CNN Using mmWave Radar*, IEEE Sens. Lett. **3**, 1–4 (2019).

[194] O. D. Lara and M. A. Labrador, *A Survey on Human Activity Recognition using Wearable Sensors*, IEEE Commun. Surv. Tutorials **15**, 1192–1209 (2013).

[195] S.-R. Ke, H. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, *A Review on Video-Based Human Activity Recognition*, Computers **2**, 88–131 (2013).

[196] D. Liang and E. Thomaz, *Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online*

Videos, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**, 1–18 (2019).

[197] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition*. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–647–I–655. JMLR.org, (2014). event-place: Beijing, China.

[198] G. M. Weiss, K. Yoneda, and T. Hayajneh, *Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living*, IEEE Access **7**, 133190–133202 (2019).

[199] F. Chollet et al. *Keras*, (2015).

[200] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, *Distributed recognition of human actions using wearable motion sensor networks*, Journal of Ambient Intelligence and Smart Environments **1**, 103–115 (2009).

[201] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Ploetz, *IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition*, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **4**, 1–29 (2020).