Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition

Karan Ahuja
Carnegie Mellon University
Pittsburgh, PA, USA
kahuja@cs.cmu.edu

Yue Jiang
Max Planck Institute for Informatics
Saarbrücken, Germany
yuejiang@mpi-inf.mpg.de

Mayank Goel
Carnegie Mellon University
Pittsburgh, PA, USA
mayank@cs.cmu.edu

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

ABSTRACT

Millimeter wave (mmWave) Doppler radar is a new and promising sensing approach for human activity recognition, offering signal richness approaching that of microphones and cameras, but without many of the privacy-invading downsides. However, unlike audio and computer vision approaches that can draw from huge libraries of videos for training deep learning models, Doppler radar has no existing large datasets, holding back this otherwise promising sensing modality. In response, we set out to create a software pipeline that converts videos of human activities into realistic, synthetic Doppler radar data. We show how this cross-domain translation can be successful through a series of experimental results. Overall, we believe our approach is an important stepping stone towards significantly reducing the burden of training such as human sensing systems, and could help bootstrap uses in human-computer interaction.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Interaction techniques*; Gestural input.

KEYWORDS

Human activity recognition, HAR, Doppler sensing, Datasets, Cross domain translation, Privacy-preserving sensing

ACM Reference Format:
Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 2021. Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3411764.3445138

1 INTRODUCTION

Future smart homes and offices will need to be able to sense the activities of their occupants in order to intelligently adapt to the environment and respond to their users' needs. An incredible variety of technical approaches for recognizing the user's actions has been considered over many decades of research (see [73] for a survey). While tagging every object in the environment with sensors can be used to infer activity [88, 89], this approach is expensive, hard to maintain, and often visually obtrusive. Therefore, the trend has been towards centralized sensing, either a worn device (*e.g.*, smartwatches) or with *e.g.*, microphones or cameras operating in an environment [52, 59]. While more practical, these high-fidelity sensors also raise significant privacy concerns. Indeed, many users are wary of microphones and cameras recording them in their homes, especially after recent data leaks [25]. For this reason, there is renewed interest in identifying and exploring sensing modalities that are inherently more privacy preserving, yet sufficiently rich to enable fine-grained activity recognition.

In this work, we explore one such sensor: the millimeter wave (mmWave) Doppler radar. Owing to their extensive use in security and automobile applications, the price of these sensors has fallen dramatically, to even just a few dollars for basic units (*e.g.*, RCWL-0516, HB100 and LV002 Doppler sensors). More sophisticated frequency-modulated continuous wave (FMCW) sensors cost around $30 USD [26]. Both types of radar sensors are solid state and small enough to be integrated into consumer devices, such as smart speakers and smartphones [66]. These radar sensors emit a known RF signal, and any motion in the scene (either from users or objects) causes reflected signals to be Doppler-shifted, which can then be used to create a 1-D Doppler plot. In the case of FMCW sensors, a 2-D plot of range *vs.* the Doppler shift of signals can be produced. Although some biomechanical attributes are expressed in the Doppler signal (*e.g.*, limb gait while walking), this has only been shown to recognize people from a small set of users [77], and not from the population at large. Indeed, it would seem hard to be embarrassed by leaked Doppler data, in contrast to a video or audio recording that can easily reveal identity and capture sensitive content [13, 74].

That said, Doppler radar faces a significant challenge: bootstrapping machine learning classifiers. Unlike audio and computer vision approaches that can draw from huge libraries of videos to train machine learning models, Doppler radar has no existing large datasets. All prior Doppler sensing work we could find in the literature had to collect their own bespoke training data for recognition tasks. The scale of data appears to be so limited that the full potential of techniques like deep learning techniques remains to be seen.

In this paper, we propose a unique software pipeline that allows unstructured videos to be transformed into synthetic Doppler radar data that can then be used for training. This process opens up an unparalleled volume of training data for Doppler sensors, closing an important gap and elevating the feasibility of Doppler sensing for activity recognition. Results from our user study show that training a model using our proof-of-concept synthetic data output (81.4%

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445138

accuracy) is roughly comparable in accuracy to training with native sensor data (90.2% accuracy) – a loss of around 8.8% in accuracy in a 12-class activity recognition task. If we augment our large synthetic dataset with just a few minutes of user data captured with an in-situ sensor, accuracy jumps to 95.9%, suggesting a mixed approach could be successful while minimizing user burden.

2 RELATED WORK

We first briefly review related work on activity recognition powered many different sensing modalities, and then more specifically focus on approaches that leverage Doppler shifts induced by human motion.

2.1 Human Activity Recognition

Over the years, researchers have studied human activity recognition from different perspectives and used numerous sensing and machine learning modalities; including microphones [59, 65], pedometers [30], IMUs [3, 15] and optical sensors [52, 103] to name a few. Lara *et al.* [60] and Ke *et al.* [51] provide a comprehensive survey of wearable and video-based human activity recognition techniques, respectively.

With the advent of deep learning algorithms, the applicability of camera- and microphone-based methods for human activity recognition have significantly increased. SoundNet [8] uses unlabelled videos to learn sound representations of various activities. Ubicoustics [59, 97] utilized audio from sound effect libraries and focused on domestic activities. Overall, sound-based approaches to activity recognition are promising, but the main challenges are privacy, environmental noise, and limited sensing range due to sound attenuation. Cameras are also powerful, and able to capture certain activities that are not possible with microphones. Researchers have explored different motion and temporal feature representations of videos learned by 3D Convolutional Neural Networks (CNNs) [14, 46, 96]. In addition to CNNs, long short-term memory (LSTM) models are popular, taking advantage of dependencies across video frames [27, 64, 80, 101]. Instead of using raw visual information, researchers have also explored the idea of relying on high-level semantic representations for human activity. For example, using body pose [9, 28, 50, 92, 98, 109], motion of semantic keypoints [24], and joint representations [29, 44, 68, 79, 85]. To facilitate comparison, we provide a summary of other externally sensed activity recognition systems in Table 1. Finally, we reiterate that cameras and microphones provide high signal fidelity and richness, but carry increased privacy concerns [13, 74].

	Training Data	No. Classes	Accuracy
TSN [94]	RGB Video	20	94.2%
TTDD_FV [93]	RGB Video	20	90.3%
LTC [90]	RGB Video	20	91.7%
KVMF [108]	RGB Video	20	93.1%
VideoDarwin [36]	RGB Video	51	63.7%
MPR [71]	RGB Video	51	65.5%
Zhao *et al.* [106]	RGB-D video	12	89.1%
Ubicoustics [59]	Audio	30	82.1%
Liang *et al.* [65]	Audio	15	83.6%
Fu *et al.* [37]	Doppler Ultrasound	3	92.0%
Radhar [83]	Doppler Radar	5	94.7%
Erol *et al.* [32]	Doppler Radar	7	92.8%
Kim *et al.* [53]	Doppler Radar	8	82.6%
Our Approach	Doppler Radar	12	95.9%
Our Approach	Synthetic Doppler	12	81.4%

Table 1: Activity recognition systems that make use of external sensors (i.e., not worn).

2.2 Doppler-Based Sensing

Energy waves undergo Doppler shift when reflecting off moving objects. These waves can be sound [76, 77], radio frequencies (RF) [22, 83], visible light [6, 43], or even gravitational waves [10]. Given the focus on practical and deployable systems, the HCI community typically relies on microphone- and RF-based Doppler sensing. The ubiquity of microphones makes them extremely popular for sensing Doppler shifts (most often in ultrasonic frequency ranges so as to not interfere with human hearing). Using sound-based approaches, researchers have enabled large in-air gestures [41, 76], fine-grained hand gestures [35, 95], multi-device interactions [7, 17], and activity recognition [37].

In recent years, RF-based Doppler sensors have become significantly cheaper and more accessible. RF systems also tend to offer superior range than ultrasonic Doppler techniques, and can sometimes operate through walls. Prior work has explored through-the-wall person detection [23, 76, 87], human gesture recognition [22, 40, 63], respiratory monitoring [62], and signs of life detection [18]. Closer to this work are papers investigating RF-based Doppler sensing for activity recognition. Chen *et al.* proposed an in-home Wi-Fi signal-based activity recognition framework using passive micro-Doppler signatures [22]. Using deep learning, Chen *et al.* monitored daily activities and detected falling accidents [19–21]. Similarly, Singh *et al.* used a sparse point cloud from a mmWave Radar sensor for recognizing five different human activities [83]. A commonality in this prior work is the need for in-situ training data to develop machine learning models, which are specific to the use domain and collection environment. Given this data is manually collected, the volume of training data used in these systems is comparatively small compared to audio- and video-derived datasets.

2.3 Synthesizing Doppler Data

Using data sources such as videos, motion capture, and animated 3D models, prior work has synthesized training data for IMU [45, 57], audio [8, 107], depth camera [75] and human pose [91] powered-systems. The idea to specifically synthesize Doppler data to mitigate training data issues is not new either. In particular, Lin *et al.* explored using MoCap data to synthesize Doppler data for walking and running with some success [67]. Unfortunately, MoCap data is generally sparse (often a dozen or so key joints), so researchers have also generated synthetic Doppler data using point clouds captured from depth-cameras [31, 33, 61]. In both MoCap and depth camera cases, we found datasets to be much smaller than video sources and missing many commonplace activities. The fact is, people capture and upload video data freely, but do not go out and capture depth-camera datasets for research use. Perhaps most similar to our work is [32], which captured seed data using an actual Doppler radar

	Type	Wave	Climb Staircase	Walk	Squat	Run	Lunge	Jump Rope	Jumping Jack	Jump	Cycle	Clean	Clap
CMU [1]	MoCap	–	–	~300	–	~60	–	–	–	~110	–	–	–
SFU [2]	MoCap	–	–	~20	–	~5	–	–	–	~10	–	–	–
RHA [58]	RGB Video	~100	–	~100	–	~100	–	–	–	–	–	–	~100
UCF101 [86]	RGBVideo	–	–	–	~115	–	~130	~145	~125	–	~135	~110	–
HMDB [56]	RGB Video	–	~50	~155	–	–	–	–	–	–	~105	–	–
YouTube8M [4]	RGB Video	–	–	317	2422	7628	–	–	–	4692	31080	206	–
STAIR [100]	RGB Video	–	~900	~900	–	~900	–	–	–	~900	–	–	–
ActivityNet [12]	RGB Video	–	–	–	–	–	–	–	–	–	–	~65	–
RadHAR [83]	RF Doppler	–	–	~50	~50	–	–	–	~40	~35	–	–	–
Gambi [38]	RF Doppler	–	–	231	–	–	–	–	–	–	–	–	–
NTU [79]	Depth Cam	~1000	–	~1000	1000	~1000	–	–	–	~1000	–	–	~1000
MHAD [72]	Depth Cam	50	–	–	–	–	–	–	25	25	–	–	25
UTD [16]	Depth Cam	–	–	–	–	–	–	–	25	25	–	–	25
YouTube	RGB Video	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 2: We selected 12 activities from [4, 79, 86] to gauge data availability. This table shows multiple datasets, crossing four different categories of data (depth cameras, motion capture, Doppler radar and video). Counts are number of video snippets for that class. We use ✓ to denote classes with more data than could be practically utilized (i.e., functionally unlimited).

sensor and then generated synthetic Doppler data using Generative Adversarial Networks (GANs). This approach is complementary to ours, and could expand the volume of training data even further. To summarize, our approach to synthetic Doppler data creation is unique in that the inputs are videos, allowing researchers to tap into a near-limitless amount of training data.

3 POSSIBLE TRAINING DATA SOURCES

The decision to use video as the input into our synthesis pipeline was not a forgone conclusion. At the very start of the research, our preference was to use datasets that required less dramatic transformation. We now briefly describe the four main data categories we considered, and their relative pros and cons that led us to pursue a video-based approach. As a benchmark, we picked 12 activities (listed in Table 2) drawn from [4, 79, 86] as a sort of feasibility litmus test.

Doppler Radar Datasets - We started by surveying projects that used radar sensors for activity recognition and cataloged how they sourced their training data. For example, RadHAR [83] manually collected data for five activities (walking, jumping, jumping jacks, squats and boxing) while Gambi *et al.* [38] collected 231 sequences of 29 people walking at different speeds. In [32], the authors manually collected 1356 sequences of 14 people performing 8 actions at different angles. In all cases, these manually-created datasets were very small in volume and limited in their classes. Most importantly, every minute of recorded data took at least one minute of researcher time.

Depth Camera Datasets - Next we considered RGBD and depth-camera video datasets (captured with sensors such as the Microsoft Kinect). We surmised the 3D point cloud of segmented users could be converted into synthetic Doppler [31, 61], which could then be used for training. Unlike with Doppler radar, large datasets exist for research use, which was encouraging. We surveyed 13 such datasets, but found them to be missing several of our sample activities (see Table 2), and thus several datasets (in different formats) would have to be combined.

MoCap Datasets - We then looked at 3D motion capture (MoCap) datasets, digitized by professional optical tracking systems. These are very spatially accurate, but only provide a sparse 3D model of users – often just 17 key joints – and thus can only provide a very coarse synthetic Doppler signal [67]. Any Doppler-shifted reflections between e.g., the wrist and elbow must be interpolated. Additionally, of the 13 datasets we surveyed, many activities were missing, as noted in Table 2.

Video Datasets - Given they contain no innate 3D data, we were initially skeptical of videos as a data source. Cameras are generally very high resolution in the plane orthogonal to the camera axis, but are largely intensive along their Z axis. However, the incredible wealth of video content – with more than 500 hours of video uploaded every minute to just YouTube alone – it soon became the clear winner. Beyond unstructured sources like Youtube, there exist scores of excellent and very large video repositories that cover all of our poses (Table 2). We were likewise encouraged by recent advances in computer vision that enabled 3D pose and even 3D meshes of users to be extracted from videos, offering the building blocks to explore synthesizing Doppler data.

Motivated by these findings, we set out to create a software pipeline that converts videos into realistic, but synthetic Doppler radar data. If achievable, it would offer an unparalleled volume of training data for this emerging sensing modality, closing an important gap and elevating the feasibility of Doppler sensing for activity recognition.

4 IMPLEMENTATION

We now describe in detail the iterative steps of our software pipeline, illustrated in Figure 1.

4.1 Mesh fitting

Doppler sensors measure the radial velocity of reflective surfaces in a scene. To replicate this signal, we require equivalent 3D data of a user's body against a static background. Fortunately, computer vision has made enormous strides in fitting a 3D mesh to a person's image [42, 55, 70, 99]. Hence, as a first step, we compute the position of all vertices of the human body by fitting a mesh to it. For this, we use VIBE [55], which estimates the mesh via an adversarial learning framework for human pose estimation. Given an input video, the VIBE model estimates the human pose and outputs a human pose mesh for each frame. We track vertices across frames and also smooth their positions to increase stability.

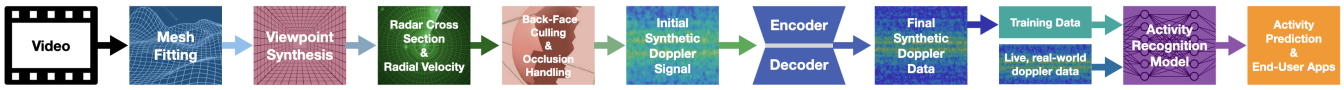

Figure 1: Overview of our software pipeline for synthetic Doppler data generation for training activity recognition models.

4.2 Viewpoint Synthesis

Once we have the 3D mesh of a user, we can place a virtual camera in the scene to synthesize any view. Indeed, we place nine synthetic cameras in a spherical coordinate system around the user mesh to simulate different viewpoints (see Figure 2 and Video Figure). Specifically, we include a "head on" view, along with views 45° to the left and right, as well as 45° up and down (forming a 3×3 polar coordinate grid). This simple manipulation, essentially a form of training data augmentation, has a multiplicative effect on every second of input video. Moreover, it helps to make the later machine learning model more view-invariant.

4.3 Radar Cross-Section & Radial Velocity

Given a viewpoint and a mesh of a user, we can compute the radar cross-section of every vertex with respect to a virtual Doppler sensor. We do this by taking the user mesh returned by VIBE [55] (which uses SMPL mesh [69]) and calculate each vertex's surface area and normal. To compute radial velocity, we look back at each vertex's movement history (previous frames), again with respect to a virtual Doppler sensor (Figure 3, center). We also further augment our training data by slightly varying the framerate of the input video (e.g., by assuming consecutive frames are not 1/30th of a second apart, but rather 1/29th or any other value) to produce realistic variations of the same activity being performed at different speeds.

4.4 Vertex Visibility & Occlusion

At this point in the pipeline, we have each vertex's contribution to the synthetic doppler signal, assuming all were visible. Of course, there are vertices on the reverse side of the user, and also vertices that are occluded by other body parts (e.g., arms crossing the torso). Since these would not contribute to an RF Doppler signal, they must be filtered. We first perform back-face culling [104] for each viewpoint, and then calculate if a vertex is occluded by another. Only vertices with line-of-sight to the virtual Doppler sensor (Figure 3, right) are passed to the next step of our pipeline.

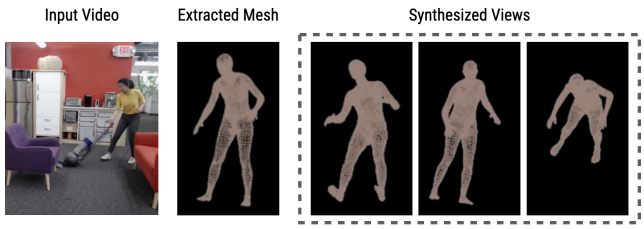

Figure 2: An input video is first processed to extract a 3D mesh. We can then simulate different virtual viewpoints (here you can see 3 of our 9 synthesized views). This both increases our training data volume and improves robustness in real world conditions.

4.5 Synthesizing Initial Whole-Body Doppler

A useful and popular visualization of Doppler sensor data is a radial velocity profile at a given instant in time. To mimic this, we create a 32-bin histogram of the radial velocities of the visible user vertices in the velocity range of our real world sensor (in this case, -2 to 2 m/s). By stacking such signatures over time, we create a sliding, Doppler-time plot (see Figure 4, second row). The envelope of this initial simulated Doppler signal roughly follows that of actual Doppler data, but this can be further improved as we will explain.

4.6 Encoder-Decoder for Domain Translation

The aforementioned synthetic Doppler-time plot is very coarse, as it is heavily quantized and even small mesh fitting errors can produce big jumps in radial velocity. Additionally, it does not contain any of the characteristic noise or non-linearities found in real RF Doppler sensors. However, we found that the correspondence between synthetic and real-world Doppler can greatly be improved by making use of an encoder-decoder model. To train our encoder-decoder model, we choose a U-Net architecture [78], which contains a convolutional and deconvolutional block with an embedding layer of size 128 in between. Each block has 16, 32 and 64 2D filters respectively with a kernel size of 3×3. A Leaky ReLU activation function and a batch normalization are applied to each convolutional layer. We take corresponding pairs of real world Doppler and synthetic Doppler to train our model for 1000 epochs with a root mean square error loss and Adam optimizer [54] with a learning rate of 0.001.

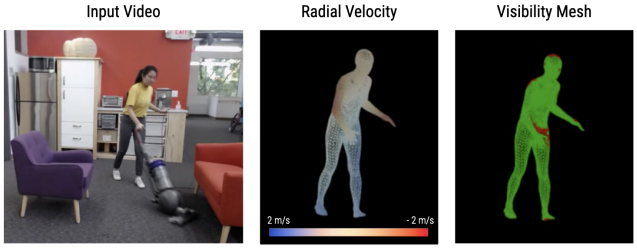

Figure 3: Left: video of user vacuuming. Center: extracted user mesh color-coded by radial velocity. Right: mesh color-coded by visibility - green for visible and red for occluded points. In this example, for illustration, the virtual Doppler sensor is located just below the virtual camera, and so a sensor "shadow" from the arm is cast onto the torso.

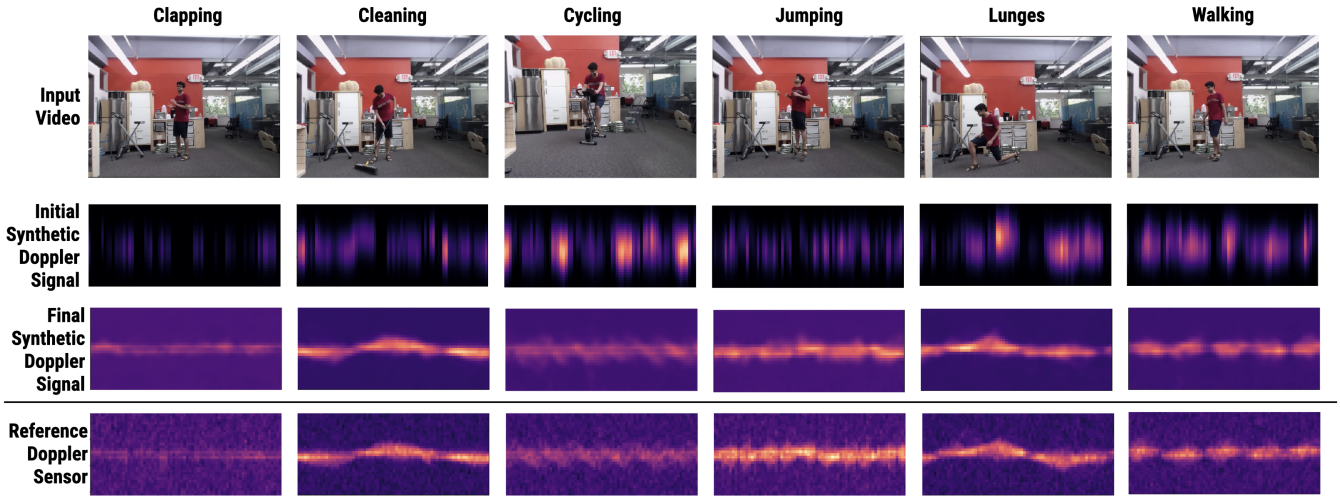

Figure 4: Top row shows a user performing different activities. Below each activity is our initial and final synthetic Doppler signal. The bottom row shows the corresponding signal captured by an actual Doppler radar sensor (positioned next to the camera that filmed the user). Note our synthetic pipeline produces comparable signal (see also Video Figure).

4.7 Final Whole-Body Doppler Signal

The final output of our pipeline is a synthetic Doppler-time plot generated by our encoder-decoder model. This plot represents radial velocity from -2 m/s to 2 m/s (Y-axis, 32 bins) and 3.0 seconds of data (X-axis, 72 bins). As can be seen in Figure 4 (third row), this synthetic signal has a strong correspondence to real-world RF doppler data (bottom row), despite only utilizing 2D video (see also Video Figure). It is this signal that we use to train our deep learning model for activity recognition, described next.

4.8 Activity Recognition

As a proof-of-concept Doppler sensor, we used a Texas Instruments AWR1642 mmWave radar board [47], which costs around $30 USD [26] when purchasing just the sensor (Figure 6). Doppler data is streamed to a MacBook Pro laptop (3.1GHz dual-core i5) over USB at 31 FPS. Our machine learning model is a VGG-16-based convolutional neural network [82] that ingests a real-world Doppler-time plot (32x72) and outputs one of 12 activity classes. We train the model with a categorical cross-entropy loss [105] and Adam optimizer [54] with a learning rate of 0.001 for 1000 epochs. On our laptop, this model takes 47 ms of compute, allowing it to run in realtime. Example activities recognized by our model are shown in Figure 5.

Figure 5: Live classification of three example activities by our activity recognition classifier with confidence scores overlaid over real-time Doppler signals.

5 OPEN SOURCE MODEL AND DATA

To enable other researchers and practitioners to build upon our system and study results, we have made our synthetic Doppler data and real world Doppler data available at https://github.com/FIGLAB/Vid2Doppler.

6 TRAINING DATA

As a proof-of-concept class set, we used the same 12 activities used to survey data sources earlier in the paper, which were drawn from the literature [4, 79, 86]. To train our activity recognition model, we aggregated 10.4 hours of video data to serve as the input for our synthetic Doppler data generation (expanded roughly ten fold via data augmentation). Most of the video datasets that we use (8.4 hrs of the 10.4 hrs) are structured ("RGB Video" datasets in Table 2), i.e., have activity labels associated with them. Similar to [57], we also mined data from unstructured sources (e.g., YouTube) using queries related to our activity set and then filtered them manually. In the future, as the sophistication of vision-based Human-Activity Recognition [11] modules improve, we could rely on them for automatic labeling.

To train our encoder-decoder model, we had two participants perform the 12 activities in a different room than our later study (a living room of dimension 7.2 × 5.6 × 3.6 m). Each user provided one hour of data, varying their angle to the sensor. As the encoder-decoder model works with unsupervised data, collection required

no labeling and motions that were not part of the activity dataset were also captured when transitioning from one activity to another.

7 USER STUDY

We now describe the physical arrangement of our study, and then walk through a series of specific experiments that we used to elucidate the feasibility of our approach.

7.1 Apparatus & Location

For this study, we used the same Texas Instruments AWR1642 RF Doppler sensor [47] and MacBook Pro laptop as described in the previous section. We mounted the Doppler sensor to a tripod alongside a Logitech HD webcam to capture footage (Figure 6). We cleared a small space in our lab where users could safely perform activities we requested.

7.2 Procedure

We recruited 10 participants (8 male, 2 female) with an average age of 25.3 years. For each participant, we captured two sessions of data back-to-back in a lab space roughly 12.8 × 6.5 × 3.8 m. Within each session, participants were asked to perform 12 activities (enumerated in Table 2) at 3 different angles with respect to our sensor tripod (0°, 45° and -45°), resulting in a total of 36 trials per session per participant. Thus, in total we collected: 10 participants × 2 sessions × 3 angles × 12 activities = 720 total trials (roughly 3.4 hours of data).

For each trial, we collected synchronized Doppler and video data. Note that video data was collected as a reference to benchmark the accuracy of our generated synthetic Doppler signal vs. the real world Doppler signal. Importantly, reference videos were never used to generate any synthetic data for training. As described previously, our training dataset consisted of 10.4 hours of video data curated from various external data sources, which we processed into synthetic Doppler data.

7.3 Results

We designed our study procedure in order to analyze and isolate different factors that affect performance. First we discuss the quality of the generated synthetic Doppler data. We then describe a series of varying train/test configurations to assess activity recognition accuracy.

7.3.1 Quality of Synthetic vs. Real-World Doppler Signal. To test the efficacy of our synthetic Doppler data generation pipeline, we make use of the videos we captured in tandem with real-world Doppler data. Specifically, we run participant videos through our pipeline and compare the synthetic Doppler output to the real-world Doppler sensor signal, finding a Mean Absolute Error (MAE) of 0.09 (SD = 0.03) for the normalized (between 0 and 1) amplitude of Doppler shift.

7.3.2 Recognition Accuracy: Only Synthetic Training Data. We evaluated the performance of our model trained *only* on synthetic data generated from our external video dataset (detailed in Section 3) and tested using participants' real-world Doppler signals. In this configuration, our model achieved an accuracy of 81.4% (chance is 8.34%) across the 12 activities and all participants. Figure 7 (left) provides the confusion matrix. Note that in this train/test configuration, we do not train our model on *any* real-world Doppler data (i.e., only synthetic Doppler data). This result can be thought of as "out-of-the-box" accuracy, without any calibration to the local environment or user.

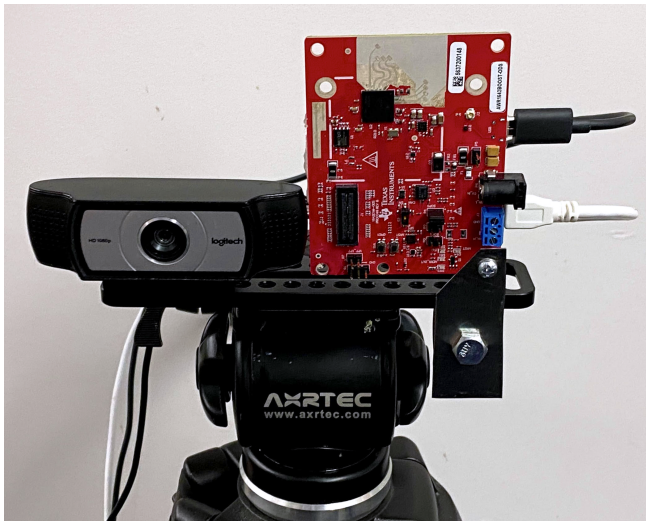

Figure 6: For data capture, we use a Texas Instruments AWR1642 FMCW Radar Sensor (red board) and Logitech HD webcam.

7.3.3 Recognition Accuracy: Only Real-World Training Data. To better contextualize our model accuracies, we wished to train a model using real-world Doppler data and then test it on real-world Doppler data. As already mentioned several times, there are not good existing datasets to run such an analysis, so instead we had to use our own study data. Specifically, we performed a leave-one-user-out cross validation. In this process, we train on real-world Doppler data from nine of our participants and test on a tenth (all combinations, results averaged). These models achieved an average accuracy of 90.2% (SD = 4.8%). Unsurprisingly, training on data captured using the actual sensor in the same location outperforms our model trained only on synthetic data, though the difference is only 8.8%, which we view as a strong result for our proof-of-concept pipeline. At a high level, we believe it shows that a purely synthetic training data pipeline can be competitive with training procedures that rely on in-situ captured data.

7.3.4 Recognition Accuracy: Synthetic + Real-World Training Data. It is also possible for models to leverage both synthetic and real-world Doppler data for training. This could offer the best of both worlds: a large corpus of videos for creating an even larger synthetic Doppler dataset, as well a smaller real-world dataset captured in-situ that is inherently better tuned to the local environment and physical sensor. To explore this, we again ran a leave-one-user-out cross validation. This time, we trained models using all synthetic Doppler data and real-world data from nine participants, testing on a tenth holdout user (all combinations, results averaged). In this scenario, our model achieves an accuracy of 93.4% (SD = 5.4%),

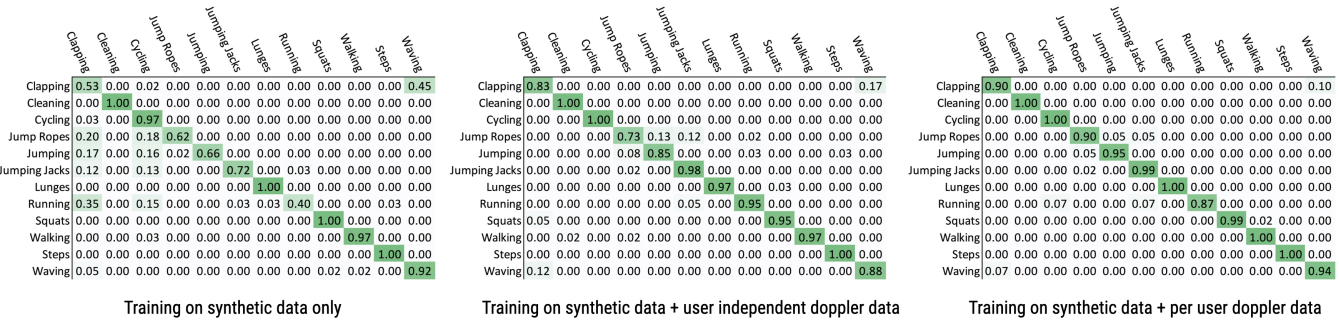

Figure 7: Confusion matrix across different train/test conditions.

a boost of 3.2% over using only real-world data for training. See confusion matrix in Figure 7, center.

7.3.5 Recognition Accuracy: Per-User Training. In all of the previous experiments, the model is never exposed to training data from the user it is testing. However, it is not uncommon for sensing systems to collect some training data from users, often in the guise of a calibration during setup (e.g., voice transcription systems that are trained using a large general corpus, but also ask the user to speak some phrases to calibrate). To simulate this, we performed a leave-one-session-out cross validation. Specifically, we train a model using all of our synthetic Doppler data (i.e., a large general corpus) and then add one round of a participant's real-world Doppler data, testing on the holdout round (both round combinations, for all participants, results averaged). This achieves the best accuracy of all of our tests: 95.9% (SD = 0.3%). The confusion matrix can be found in Figure 7, right.

7.3.6 Comparison to Prior Work. The accuracy of our approach compares favorably to prior work. RadHAR makes use of point clouds generated from mmWave radar (the same sensor as ours) and achieves an accuracy of 90.5% across 5 activities employing a deep learning model. [53] uses SVM's trained on Doppler radar to recognize 7 activities with a per-user model accuracy of 92.8% and a cross-user accuracy 91.9%. The synthetic Doppler data approach in [32] achieves an accuracy of of 82.6% across 8 activities on one participant. In contrast, our per-user model on 12 activities achieves an accuracy of 95.9%. However, it is to be noted that our goal is not to make a better framework for sensing activities via Doppler data, but rather to create a framework for synthesizing Doppler data for training a myriad of different activity recognizers. That said, higher accuracy is a nice side-effect of leveraging synthetically-created training data derived from video sources. An overview of accuracies can be found in Table 1, though we emphasize these systems are tested on different datasets and have different applications.

8 DISCUSSION

In general, as the richness of a sensing modality increases, so does the range of activities it can sense. Unfortunately, privacy implications tend to also grow in lockstep, to the point where many people do not want such sensors in their homes. We can use these two abstract axes to formulate a design space (Figure 8). We propose that most sensing approaches lie along the diagonal. In the lower-left are low-richness sensors that rarely, if ever, reveal sensitive information (e.g., temperature, barometric, magnetic, and proximity sensors). In the upper-right are high-richness, highly-invasive sensors, such as microphones and cameras. Using the latter sensors, researchers have demonstrated high accuracy activity recognition [5, 34, 59, 103], but many consumers remain skeptical and even early adopters are wary of leaked recordings and data [25, 39]. We believe that Doppler radar is among a handful of sensing techniques that starts to pull away from the diagonal trend, offering good signal richness and good privacy preservation. For this reason, the modality deserves attention, and we hope that our pipeline can serve as a tool to bootstrap such exploration efforts.

Prior research utilizing cameras and microphones has already demonstrated many applications of user activity recognition, ranging from exercise tracking, health and wellness monitoring, life logging, to context-aware assistants. RF Doppler radar and our training approach could not only enable similar applications, but do so in a more privacy-preserving manner that could help to realize the vision of ubiquitous sensing. Additionally, mmWave radar sensors are sufficiently low cost and compact to allow for integration into almost all consumer electronics. Indeed, commercial smartphones have already shipped with Doppler radar sensors, such as the Google Pixel 4 [84, 95].

9 LIMITATIONS AND FUTURE WORK

There are several key technical limitations that will need to be overcome before consumer use and widespread adoption. First, our model was trained and tested with static background scenes. Our model would fail in cases where there is motion in the background or the sensor itself was in motion (i.e., the environment would create Doppler-shifted reflections in addition to a user). To overcome this in the future, it may be possible to add random Doppler signals or even simulate the physics of moving objects and walls as part as part of the synthetic data generation pipeline. Another limitation of our current system is the inability to handle multiple simultaneous users. Our current approach sums all the Doppler profiles across distances to create a distance-invariant Doppler-time plot. However, some Doppler radar systems can segment and track multiple people if they are far enough apart [48], so this may be overcome in the future. Lastly, our current machine learning approach makes use of convolutions on the Doppler-time plot. However, alternate feature

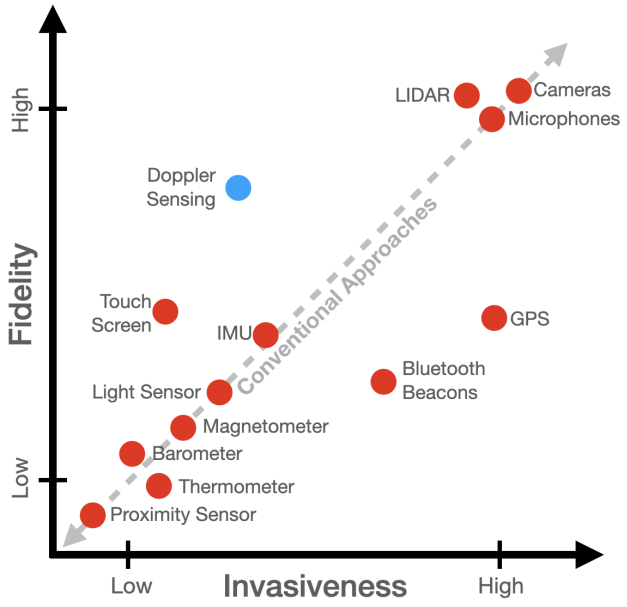

Figure 8: A design space plotting sensing fidelity vs. invasiveness. Most sensors lie along the diagonal, where invasiveness increases as fidelity increases. Ideally, we want sensors that offer higher fidelity without commensurate privacy trade-offs. We believe Doppler radar sensing is one approach that begins to pull away from the traditional axis.

representations that treat the Doppler-shift histograms as a time series and make use of RNN-based architectures [81, 102] could help take the model away from fixed window lengths. Furthermore, apart from a UNet architecture, conditional adversarial losses [49] can also be explored where the encoder-decoder and discriminator (activity classification) are combined in a single step.

Finally, We would like to acknowledge that while RF Doppler radar and our approach is privacy-preserving in comparison to cameras and microphones, the logging of activity data in itself can have significant privacy implications. This is a long standing HCI research topic, and as Doppler radar sensors become more pervasive, they too will need special scrutiny given their unique pros and cons.

10 CONCLUSION

Activity recognition enables a plethora of applications, demonstrated in much prior research. With the high-fidelity, yet privacy-preserving sensing afforded by Doppler radar sensors, the ubiquity of this modality is held back by the lack of available training data. In our paper, we aim to mitigate this important issue by created a software pipeline that takes videos of users performing activities and outputs realistic, synthetic Doppler data. This can then be used to train activity recognition models. As a proof of concept, we created a model using our pipeline that can detect 12 exemplary activities at accuracies competitive with prior systems that collected in-situ data with physical Doppler sensors.

ACKNOWLEDGMENTS

We especially thank NVIDIA Corporation for the donation of a Titan V GPU and Yang Zhang for his hardware expertise. We are also grateful to Cathy Fang for appearing in our video as an example user. Finally, we thank our reviewers for their invaluable feedback and comments.

REFERENCES

[1] 2004. CMU MoCap. http://mocap.cs.cmu.edu/.
[2] 2004. SFU MoCap. https://mocap.cs.sfu.ca/.
[3] Nimsiri Abhayasinghe and Iain Murray. 2014. Human activity recognition using thigh angle derived from single thigh mounted imu data. In *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 111–115.
[4] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).
[5] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical classroom sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.
[6] H-E Albrecht, Nils Damaschke, Michael Borys, and Cameron Tropea. 2013. *Laser Doppler and phase Doppler measurement techniques*. Springer Science & Business Media.
[7] Md Tanvir Islam Aumi, Sidhant Gupta, Mayank Goel, Eric Larson, and Shwetak Patel. 2013. DopLink: using the doppler effect for multi-device interaction. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 583–586.
[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*. 892–900.
[9] Fabien Baradel, Christian Wolf, and Julien Mille. 2018. Human activity recognition with pose-driven attention to rgb.
[10] Barry C Barish and Rainer Weiss. 1999. LIGO and the detection of gravitational waves. *Physics Today* 52 (1999), 44–50.
[11] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. 2020. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications* 79, 41 (2020), 30509–30555.
[12] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.
[13] Kelly E Caine, Arthur D Fisk, and Wendy A Rogers. 2006. Benefits and privacy concerns of a home equipped with a visual sensing system: A perspective from older adults. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. SAGE Publications Sage CA: Los Angeles, CA, 180–184.
[14] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
[15] Pierluigi Casale, Oriol Pujol, and Petia Radeva. 2011. Human activity recognition from accelerometer data using a wearable device. In *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 289–296.
[16] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*. IEEE, 168–172.
[17] Ke-Yu Chen, Daniel Ashbrook, Mayank Goel, Sung-Hyuck Lee, and Shwetak Patel. 2014. AirLink: sharing files between multiple devices using in-air gestures. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 565–569.
[18] Qingchao Chen, Kevin Chetty, Karl Woodbridge, and Bo Tan. 2016. Signs of life detection using wireless passive radar. In *2016 IEEE Radar Conference (RadarConf)*. IEEE, 1–5.
[19] Qingchao Chen, Yang Liu, Francesco Fioranelli, Matthiew Ritchie, and Kevin Chetty. 2019. Eliminate Aspect Angle Variations for Human Activity Recognition using Unsupervised Deep Adaptation Network. In *2019 IEEE Radar Conference (RadarConf)*. IEEE, 1–6.
[20] Qingchao Chen, Yang Liu, Bo Tan, Karl Woodbridge, and Kevin Chetty. 2020. Respiration and Activity Detection Based on Passive Radio Sensing in Home Environments. *IEEE Access* 8 (2020), 12426–12437.
[21] Qingchao Chen, Matthew Ritchie, Yang Liu, Kevin Chetty, and Karl Woodbridge. 2017. Joint fall and aspect angle recognition using fine-grained micro-Doppler classification. In *2017 IEEE Radar Conference (RadarConf)*. IEEE, 0912–0916.

[22] Qingchao Chen, Bo Tan, Kevin Chetty, and Karl Woodbridge. 2016. Activity recognition based on micro-Doppler signature with in-home Wi-Fi. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 1–6.

[23] Kevin Chetty, Graeme E Smith, and Karl Woodbridge. 2011. Through-the-wall sensing of personnel using passive bistatic wifi radar at standoff distances. *IEEE Transactions on Geoscience and Remote Sensing* 50, 4 (2011), 1218–1226.

[24] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. 2018. Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7024–7033.

[25] CNBC. 2020. Google Data Leaks. Retrieved 2020 from https://www.cnbc.com/2019/07/11/google-admits-leaked-private-voice-conversations.html

[26] Digikey. 2020. DigiKey Doppler Sensor TI. Retrieved 2020 from https://www.digikey.com/product-detail/en/texas-instruments/AWR1642ABIGABLRQ1/296-49116-2-ND/9169365

[27] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[28] Wenbin Du, Yali Wang, and Yu Qiao. 2017. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 3725–3734.

[29] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1110–1118.

[30] Miikka Ermes, Juha Parkka, and Luc Cluitmans. 2008. Advancing from offline to online activity recognition with wearable sensors. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 4451–4454.

[31] Barıs Erol and Sevgi Zubeyde Gurbuz. 2015. A kinect-based human micro-doppler simulator. *IEEE Aerospace and Electronic Systems Magazine* 30, 5 (2015), 6–17.

[32] Baris Erol, Sevgi Z Gurbuz, and Moeness G Amin. 2019. GAN-based synthetic radar micro-Doppler augmentations for improved human activity recognition. In *2019 IEEE Radar Conference (RadarConf)*. IEEE, 1–5.

[33] Barış Erol, Cesur Karabacak, Sevgi Zübeyde Gürbüz, and Ali Cafer Gürbüz. 2014. Simulation of human micro-Doppler signatures with Kinect sensor. In *2014 IEEE Radar Conference*. IEEE, 0863–0868.

[34] Facebook. 2020. Facebook. Retrieved 2020 from https://portal.facebook.com/

[35] Tenglong Fan, Chao Ma, Zhitao Gu, Qinyi Lv, Jialong Chen, Dexin Ye, Jiangtao Huangfu, Yongzhi Sun, Changzhi Li, and Lixin Ran. 2016. Wireless hand gesture recognition based on continuous-wave Doppler radar sensors. *IEEE Transactions on Microwave Theory and Techniques* 64, 11 (2016), 4012–4020.

[36] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. 2015. Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5378–5387.

[37] Biying Fu, Florian Kirchbuchner, Arjan Kuijper, Andreas Braun, and Dinesh Vaithyalingam Gangatharan. 2018. Fitness activity recognition on smartphones using Doppler measurements. In *Informatics*, Vol. 5. Multidisciplinary Digital Publishing Institute, 24.

[38] Ennio Gambi, Gianluca Ciattaglia, Adelmo De Santis, and Linda Senigagliesi. 2020. Millimeter wave radar data of people walking. *Data in brief* 31 (2020), 105996.

[39] Clint Gibler, Jonathan Crussell, Jeremy Erickson, and Hao Chen. 2012. AndroidLeaks: automatically detecting potential privacy leaks in android applications on a large scale. In *International Conference on Trust and Trustworthy Computing*. Springer, 291–307.

[40] Mayank Goel, Chen Zhao, Ruth Vinisha, and Shwetak N Patel. 2015. Tongue-in-cheek: Using wireless signals to enable non-intrusive and flexible facial gestures detection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 255–258.

[41] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1911–1914.

[42] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. 2020. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5052–5063.

[43] Eugene Hecht. 2002. *Optics, 5e*. Pearson Education India.

[44] Yonghong Hou, Zhaoyang Li, Pichao Wang, and Wanqing Li. 2016. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 3 (2016), 807–811.

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

[46] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. 2019. Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 254–263.

[47] Texas Instruments. 2020. Texas Instruments mmWave. Retrieved 2020 from https://www.ti.com/product/AWR1642

[48] Texas Instruments. 2020. Texas Instruments Person Tracking. Retrieved 2020 from https://www.ti.com/tool/TIDEP-01000

[49] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[50] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. 2013. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*. 3192–3199.

[51] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A review on video-based human activity recognition. *Computers* 2, 2 (2013), 88–131.

[52] Rushil Khurana, Karan Ahuja, Zac Yu, Jennifer Mankoff, Chris Harrison, and Mayank Goel. 2018. GymCam: Detecting, recognizing and tracking simultaneous exercises in unconstrained scenes. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–17.

[53] Youngwook Kim and Hao Ling. 2009. Human activity classification based on micro-Doppler signatures using a support vector machine. *IEEE Transactions on Geoscience and Remote Sensing* 47, 5 (2009), 1328–1337.

[54] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[55] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5253–5263.

[56] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*. IEEE, 2556–2563.

[57] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *arXiv preprint arXiv:2006.05675* (2020).

[58] Ivan Laptev and Tony Lindeberg. 2004. Velocity adaptation of space-time interest points. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 1. IEEE, 52–56.

[59] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 213–224.

[60] Oscar D Lara and Miguel A Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials* 15, 3 (2012), 1192–1209.

[61] Jiayi Li, Aman Shrestha, Julien Le Kernec, and Francesco Fioranelli. 2019. From Kinect skeleton data to hand gesture recognition with radar. *The Journal of Engineering* 2019, 20 (2019), 6914–6919.

[62] Wenda Li, Bo Tan, and Robert J Piechocki. 2016. Non-contact breathing detection using passive radar. In *2016 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

[63] Wenda Li, Bo Tan, Yangdi Xu, and Robert J Piechocki. 2018. Log-likelihood clustering-enabled passive RF sensing for residential activity recognition. *IEEE Sensors Journal* 18, 13 (2018), 5413–5421.

[64] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. 2018. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding* 166 (2018), 41–50.

[65] Dawei Liang and Edison Thomaz. 2019. Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–18.

[66] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–19.

[67] Yier Lin and Julien Le Kernec. 2017. Performance analysis of classification algorithms for activity recognition using micro-Doppler feature. In *2017 13th International Conference on Computational Intelligence and Security (CIS)*. IEEE, 480–483.

[68] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*. Springer, 816–833.

[69] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[70] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2020. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *ACM Transactions on Graphics (TOG)* 39, 4

(2020), 82–1.

[71] Bingbing Ni, Pierre Moulin, Xiaokang Yang, and Shuicheng Yan. 2015. Motion part regularization: Improving action recognition via trajectory selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3698–3706.

[72] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 53–60.

[73] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. 2013. Context aware computing for the internet of things: A survey. *IEEE communications surveys & tutorials* 16, 1 (2013), 414–454.

[74] Joseph Phelps, Glen Nowak, and Elizabeth Ferrell. 2000. Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy & Marketing* 19, 1 (2000), 27–41.

[75] Benjamin Planche, Ziyan Wu, Kai Ma, Shanhui Sun, Stefan Kluckner, Oliver Lehmann, Terrence Chen, Andreas Hutter, Sergey Zakharov, Harald Kosch, et al. 2017. Depthsynth: Real-time realistic synthetic data generation from cad models for 2.5 d recognition. In *2017 International Conference on 3D Vision (3DV)*. IEEE, 1–10.

[76] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel. 2013. Whole-home gesture recognition using wireless signals. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. 27–38.

[77] Bhiksha Raj, Kaustubh Kalgaonkar, Chris Harrison, and Paul Dietz. 2012. Ultrasonic doppler sensing in hci. *IEEE Pervasive Computing* 11, 2 (2012), 24–29.

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

[79] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1010–1019.

[80] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. 2015. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119* (2015).

[81] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* 28 (2015), 802–810.

[82] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[83] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. 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*. 51–56.

[84] Google ATAP Soli. 2020. Google ATAP Soli. Retrieved 2020 from https://atap.google.com/soli/

[85] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2016. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *arXiv preprint arXiv:1611.06067* (2016).

[86] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[87] Bo Tan, Karl Woodbridge, and Kevin Chetty. 2016. Awireless passive radar system for real-time through-wall movement detection. *IEEE Trans. Aerospace Electron. Systems* 52, 5 (2016), 2596–2603.

[88] Emmanuel Munguia Tapia, Stephen S Intille, and Kent Larson. 2004. Activity recognition in the home using simple and ubiquitous sensors. In *International conference on pervasive computing*. Springer, 158–175.

[89] Emmanuel Munguia Tapia, Stephen S Intille, and Kent Larson. 2007. Portable wireless sensors for object usage sensing in the home: Challenges and practicalities. In *European Conference on Ambient Intelligence*. Springer, 19–37.

[90] Gül Varol, Ivan Laptev, and Cordelia Schmid. 2017. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1510–1517.

[91] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 109–117.

[92] Chunyu Wang, Yizhou Wang, and Alan L Yuille. 2013. An approach to pose-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 915–922.

[93] Limin Wang, Yu Qiao, and Xiaoou Tang. 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4305–4314.

[94] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[95] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. 2016. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 851–860.

[96] Xiaolong Wang and Abhinav Gupta. 2018. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*. 399–417.

[97] Jason Wu, Chris Harrison, Jeffrey P Bigham, and Gierad Laput. 2020. Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[98] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. 2015. Joint action recognition and pose estimation from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1293–1301.

[99] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. 2020. EventCap: Monocular 3D Capture of High-Speed Human Motions using an Event Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4968–4978.

[100] Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. 2018. STAIR Actions: A Video Dataset of Everyday Home Actions. *arXiv preprint arXiv:1804.04326* (2018). http://arxiv.org/abs/1804.04326

[101] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4694–4702.

[102] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).

[103] Chenyang Zhang and Yingli Tian. 2012. RGB-D camera-based daily living activity recognition. *Journal of computer vision and image processing* 2, 4 (2012), 12.

[104] Hansong Zhang and Kenneth E Hoff III. 1997. Fast backface culling using normal masks. In *Proceedings of the 1997 symposium on Interactive 3D graphics*. 103–ff.

[105] Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*. 8778–8788.

[106] Yang Zhao, Zicheng Liu, Lu Yang, and Hong Cheng. 2012. Combing rgb and depth map features for human activity recognition. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1–4.

[107] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3550–3558.

[108] Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao. 2016. A key volume mining deep framework for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1991–1999.

[109] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. 2017. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2904–2913.