AudioIMU: Enhancing Inertial Sensing-Based Activity Recognition with Acoustic Models

Dawei Liang
University of Texas at Austin
Austin, USA

Guihong Li
University of Texas at Austin
Austin, USA

Rebecca Adaimi
University of Texas at Austin
Austin, USA

Radu Marculescu
University of Texas at Austin
Austin, USA

Edison Thomaz
University of Texas at Austin
Austin, USA

ABSTRACT

Modern commercial wearable devices are widely equipped with inertial measurement units (IMU) and microphones. The motion and audio signals captured by these sensors can be used for recognizing a variety of user physical activities. Compared to motion data, audio data contains rich contextual information of human activities, but continuous audio sensing also poses extra data sampling burdens and privacy issues. Given such challenges, this paper studies a novel approach to augment IMU models for human activity recognition (HAR) with the superior acoustic knowledge of activities. Specifically, we propose a teacher-student framework to derive an IMU-based HAR model. Instead of training with motion data alone, an advanced audio-based teacher model is incorporated to guide the student HAR model. Once trained, the HAR model only takes as inputs motion data for inference. Based on a semi-controlled study with 15 participants, we show that an IMU model augmented with the proposed framework outperforms the original baseline model without augmentation (74.4% versus 70.0% accuracy) for recognizing 23 activities of daily living. We further discuss a few insights regarding the difference of model performance with and without our framework and possible trade-offs for actual deployment.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

KEYWORDS

Activity Recognition, Knowledge Distillation, Wearable Sensing

ACM Reference Format:
Dawei Liang, Guihong Li, Rebecca Adaimi, Radu Marculescu, and Edison Thomaz. 2022. AudioIMU: Enhancing Inertial Sensing-Based Activity Recognition with Acoustic Models. In *The 2022 International Symposium on Wearable Computers (ISWC '22), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3544794.3558471

ISWC '22, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9424-6/22/09…$15.00
https://doi.org/10.1145/3544794.3558471

1 INTRODUCTION

Automated human activity recognition (HAR) has many applications, such as health monitoring [9], context awareness [16], and assisted technologies [7]. Over the years, the use of inertial measurement units (IMU) has been mainstream for HAR studies [3, 17, 22, 33]. While IMUs are powerful for capturing human body movements, many human activities of daily living are associated with unique sound fingerprints that are hard to be sensed from motion data alone. As a result, thanks to the ubiquity of microphones in mobile devices, acoustic sensing has been increasingly explored in recent years in HAR applications [15, 16, 20, 21].

Compared to motion data, however, continuous audio sensing raises important practical challenges. For example, passive recording of sounds may capture audio and speech that might pose privacy concerns [12, 19]. Additionally, audio recording is often captured at a high sampling rate, which results in significant power consumption overhead; this is a major issue for wearable and edge devices with compact and small batteries.

To mitigate these concerns, we study a novel approach that makes use of the rich contextual knowledge expressed in human activity sounds while relying on inertial sensor data for modeling. The approach is based on a teacher-student framework for training an IMU-based HAR model with guidance from an acoustic model during the model development phase. At inference, the HAR model only takes motion data as input. Our experiments show that our framework offers superior performance over a model trained only with inertial data. The specific contributions of this work are:

- A method to augment HAR IMU models with a teacher-student framework; the models can benefit from acoustic teacher guidance without needing audio data for inference.
- An evaluation of the method based on 23 finely-grained daily activities captured from 15 participants with a commercial smartwatch. Our results show an improvement of HAR performance from 70.0% accuracy with inertial-based models to 74.4% accuracy with the proposed framework.
- Public access to our source code and study data to encourage validation and further development of our approach.

2 RELATED WORK

2.1 IMU- and Audio-Based Activity Recognition

HAR based on motion and audio data has been extensively studied. For example, researchers have explored the usage of motion signals captured by wrist-worn devices to recognize human activities

related to hand movements [17, 22, 30, 33]. Kwapisz et al. [14] used accelerometer data collected from a cell phone to recognize coarse body movements such as walking or biking. By using sounds collected from the body, Yatani et al. [36] proposed the Bodyscope system to recognize four throat activities with an accuracy of 71.5%. Thomaz et al. [31] detected eating based on sounds collected from the wrist with an accuracy over 80%. By using audio signals collected from smartphones [15, 20, 21] or smart speakers [1, 16, 34], it is also possible to infer various human activities and contexts. Recent studies have shown the benefits of combining IMU and audio inputs for HAR [4, 6, 26, 37]. The most recent work proposed by Bhattacharya et al. [6] fused acoustic signatures of human gestures with motion signals for hand activity recognition. Richoz et al. [25] also studied the fusion of motion, audio, and vision signals for transportation mode recognition.

Despite the benefits, continuous capture of audio or multi-modal signals at inference time raises several practical concerns, including privacy issues and data sampling burdens. Researchers such as Liang et al. [19] and Iravantchi et al. [10] have explored methods to preserve audio privacy, but power consumption requirements for continuous audio recording remains a challenge.

2.2 Teacher-Student Knowledge Transfer

Transferring supplementary knowledge from a source domain to a target domain can be useful for the target task [29]. Canonically, knowledge transfer across features of different dimensions includes feature mapping to a common subspace [28, 35] or direct feature transform [18, 32]. In recent studies, researchers have shown that knowledge transfer across feature types may also be realized based on a teacher-student architecture [5, 11, 38]. The basic idea is to enforce a compact neural network (student) to mimic the outputs of a sophisticated network (teacher) so that the student can obtain high-level abstraction of features without an extensive model size [8]. Zhao et al. [38] first explored the usage of the architecture for a radio-based pose estimator with supervision from vision models. Bhalla et al. [5] applied the architecture to minimize the data annotation efforts of Doppler signals based on annotated IMU datasets. Similarly, Islam et al. [11] presented a method for breath detection with unlabeled audio by using labels generated by synchronized IMU inputs. To enhance IMU-based HAR, a possible direction is to transfer knowledge from synchronized vision signals [27]. Different from the above efforts, our study aims to enhance IMU-based activity recognition in a relatively low-cost manner by using the acoustic modality and only in the model development phase.

3 METHODOLOGY

3.1 Overview

The procedure of our study is composed of two phases: 1) Development of the HAR models; and 2) Inference (Figure 1). Phase 1 develops two neural network classifiers, i.e., an audio-based teacher model and the target motion-based HAR model. The teacher model is a neural network taking as inputs either pure audio or multi-modal (audio + motion) data. Once the teacher model is developed, it is then used as a knowledge extractor to guide the motion-based HAR model, i.e., the student. The student model takes as inputs only synchronized motion data of the corresponding activities and is trained to minimize a loss function incorporating both the outputs of the teacher model and the ground truth labels of the activities. This process enables the student model to gain the extra information obtained by the audio-based teacher model. During the inference phase, the teacher model is removed, and only the motion-based student model is used for inference.

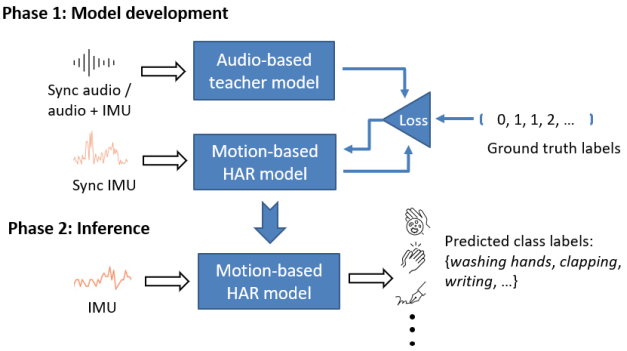

Figure 1: The overall procedure of our method. In phase 1, a motion-based HAR model is trained to minimize a loss function incorporating both the outputs of an audio-based teacher model and the ground truth. In phase 2, only the motion-based HAR model is used for inference.

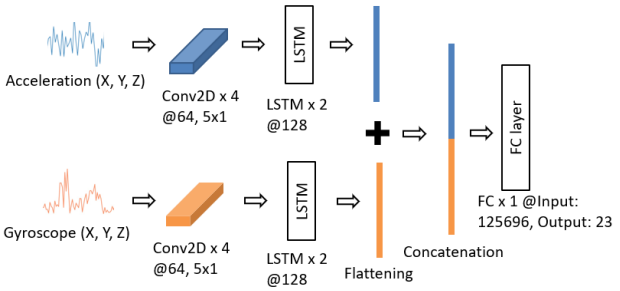

Figure 2: Our motion-based HAR model, inspired by [23]. It consists of two individual branches of DeepConvLSTM, taking raw acceleration and gyroscope inputs, respectively.

3.2 Motion-Based HAR Model

The activity recognition model we used was inspired by the DeepConvLSTM architecture, as it demonstrates promising performance in activity recognition with inertial data [23]. In our study, both acceleration and gyroscope inputs were included, so we deployed separate branches of DeepConvLSTM for each input type. The outputs of the branches were then concatenated for model prediction. Figure 2 shows the detailed design. The model fits with raw acceleration and gyroscope data based on 10s sliding windows with 50% overlap. The window size is chosen empirically for better synchronization between our motion and acoustic model inputs. As shown in Figure 2, each input branch consists of four 2D convolutional layers and a long short-term memory (LSTM) layer. The stride of the convolutional layers is 1, with no padding added to the outputs.

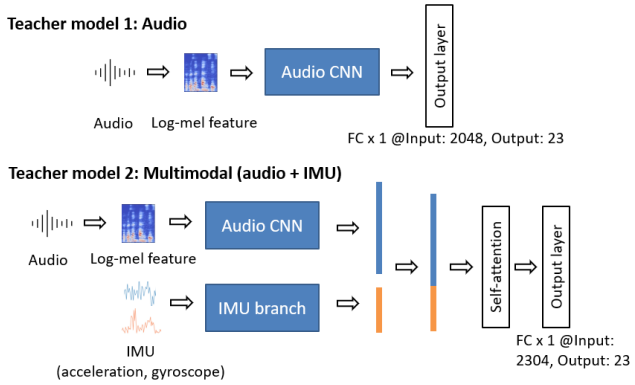

Figure 3: The teacher models. We tested two teacher models in our study, one with audio inputs only and the other with multi-modal (audio + IMU) inputs.

Each convolutional layer also comes with batch normalization. The output of the branches is flattened and concatenated before passing to a final fully-connected (FC) layer.

3.3 Audio-Based Teacher Model

As previously stated, the motivation of introducing a teacher model is to provide additional information gained from audio that the original HAR model is not able to learn by itself based on the IMU data. Once a teacher model is developed, it is then fixed as a knowledge extractor for the HAR model. In our experiments, we explored two teacher architectures - a neural network taking pure audio inputs and the other taking synchronized multi-modal (audio + motion) inputs (Figure 3). We adopted a VGG-like convolutional neural network (CNN) from prior work [13] to encode the audio inputs, which we refer to as *Audio CNN* in the paper. Similarly, we segmented the audio into 10s clips with 50% overlap. We then extracted the log-mel spectrogram features from the audio segments using a sliding window size of 1024, a hop size of 320, and 64 bins. Details of Audio CNN are as follows:

Input → Conv[64] → Conv[128] → Conv[256] → Conv[512] → Conv[1024] → Conv[2048] → FC[2048] → Average pooling

where ConvX[K] denotes a convolutional block of two 2D convolutional layers, each with K channels, and intermediary average pooling layers at the end of the block. The kernel size of each convolutional layer is (3×3). FC[K] denotes a fully-connected layer of size K activated by the ReLU activation [2]. The output of the last convolutional layers is not flattened, but globally pooled after the FC layer, resulting in 1D outputs of shape 2048 from Audio CNN.

For the teacher model 2, we directly adopted the design of our motion-based HAR model to encode the motion inputs, which is referred to as the IMU branch of the teacher model. To augment the performance, we applied the same extra temporal attention modules following the LSTM outputs of the acceleration and the gyroscope branches. The outputs of Audio CNN and the IMU branch are then concatenated and passed to a self-attention layer to learn the relative importance of the modality outputs.

3.4 Loss Function

Inspired by [8], the design of our loss function aims to incorporate both the teacher outputs and the ground truth activity labels. In our paper, we denote the original output of the audio-based teacher model, the output of the motion-based HAR model, and the ground truth labels as y_t, y_s, and y, respectively. The model outputs are first converted into a smoothed form:

$$q_s = ln(\frac{exp(y_s/T)}{\sum_j exp(y_{s_j}/T)}) \qquad q_t = \frac{exp(y_t/T)}{\sum_j exp(y_{t_j}/T)} \tag{1}$$

where j is the activity class index, and temperature T controls the smoothness of the outputs. Then, the loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{CE}(y_s, y) + \alpha * T^2 \mathcal{L}_{KL}(q_s, q_t) \tag{2}$$

where $\mathcal{L}_{CE}$ and $\mathcal{L}_{KL}$ denote the cross-entropy loss and the KL-divergence loss, respectively; α controls the effects of the teacher outputs on the student model. In our experiments, α ranges from 0.1 to 0.9, with a step of 0.1; T ranges from 1 to 8, with a step of 1.

3.5 Training Setup

The HAR models were trained with an initial learning rate of 0.001, decaying every 10 epochs by a factor of 0.9. We used a batch size of 256 split on two GPUs, except for evaluating the teacher model 2 where the batch size was halved due to system instability. We used the Adam optimizer with betas (0.9, 0.999) and an epsilon of 10^{-8}. The maximum learning epoch was 100, and we applied early stopping if no improvement of accuracy was observed for 20 consecutive epochs on the evaluation set. The teacher models were trained following a similar strategy, but the learning rate was fixed at 10^{-4}. Besides, we applied an early stopping of 10 epochs instead of 20. All models were developed in PyTorch [24]. Our source code and study data is publicly accessible[1].

4 DATA COLLECTION

The data for our study was collected via an IRB-approved *semi-naturalistic* user study with 15 participants performing a set of daily activities in their own homes. A custom Android application running on a Fossil Gen 4 smartwatch was designed to collect accelerometer, gyroscope, and acoustic data synchronously and store it locally on the watch. Inertial data was sampled at 50Hz, while acoustic data was sampled at 22.05KHz.

The participants' age ranged from 23 to 64. The 23 study activities included *writing, drawing, cutting paper, typing on keyboard, typing on phone, browsing on phone, clapping, shuffling cards, scratching, wiping table, brushing hair, washing hands, drinking, eating snacks, brushing teeth, chopping, grating, frying, sweeping, vacuuming, washing dishes, filling water, using microwave.* Two sessions of data collection were conducted for each of the 23 activities, and each activity was performed for a minimum of 30 seconds.

5 RESULTS AND DISCUSSIONS

We used leave-one-participant-out (LOPO) cross-validation to evaluate the performance of the HAR models. Following this approach, 14 of the 15 participants were used for model training, and the remaining one was used to derive checkpoint models. For each of

[1]https://github.com/Human-Signals-Lab/AudioIMU

Table 1: Leave-one-participant-out evaluation performance for 15 participants based on different settings.

Setting	Accuracy	F1 score
Baseline HAR model	70.0%	67.9%
HAR (guided by teacher 1)	**74.4%**	**72.4%**
HAR (guided by teacher 2)	**74.0%**	**71.6%**

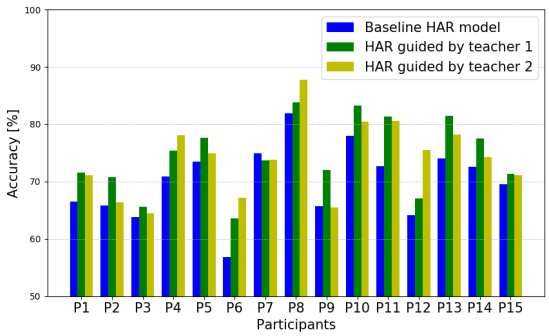

Figure 4: Participant-wise results based on different settings. The range of accuracy is shown from 50% to 100% to better visualize the performance gain.

the participants, we first obtained the two teacher models independently. Then, we derived and reported the best checkpoint HAR models per participant based on evaluation accuracy. We examined three training settings: 1) baseline HAR models trained on the IMU data only; 2) HAR models trained on the IMU data and guided by outputs of the teacher model 1; and 3) HAR models trained on the IMU data and guided by outputs of the teacher model 2. For 1), the HAR models were trained with standard cross-entropy loss. For 2) and 3), we tested all sets of α and T for model development and reported the best performing checkpoint models per participant.

5.1 Overall Results

We computed performance results using both accuracy and F1 score. For the teacher models, we obtained an average of 81.0% / 78.9% recognition accuracy / F1 for the teacher model 1, and 83.7% / 82.2% accuracy / F1 for the teacher model 2. Table 1 shows the averaged LOPO results for the HAR models, with and without teacher guidance. Overall, HAR models guided by either type of teacher output outperformed the baseline HAR models without guidance. The guiding performance of the teacher model 2 was marginally worse than that of the teacher model 1, probably because of the reduction of the batch size during student training. For a more thorough validation, we further repeated the baseline LOPO test 10 times. The average mean absolute deviation of accuracy / F1 per participant was 0.88% / 1.01%, and the best LOPO accuracy / F1 was 70.7% / 68.2%. The consistency of the results better demonstrates the difference of performance with and without the teacher guidance.

5.2 Discussions

5.2.1 Further analysis of the student performance. Figure 4 visualizes the activity recognition performance for individual participants, with and without teacher guidance. For 13 out of 15 participants, the activity recognition performance is improved with both types of teacher design. To better understand how the teacher models improve training, we analyzed the class-wise performance for sample participants that consistently benefited from the teachers (e.g., P4, 5 and 11) and obtained a few insights. First, rather than a uniform improvement of all classes, the teacher enhancement was mostly addressed on specific classes. For example, we observed an improvement of at least 20% class accuracy for the top three classes of the student models that benefited most from the acoustic teachers, but the improvement of the overall accuracy for those subjects was only around 6%. In other words, the teacher boost on most of the remaining classes was mild. Besides, the student HAR models could perform even better than both the baseline motion models and the teacher models for some classes. For example, the student / teacher / baseline class accuracy of *drawing* for P4 guided by both types of teachers was 70% / 20% / 60% and 90% / 40% / 60%, respectively. This indicates that the teacher models may enable a better generalization of the IMU models even if the class patterns are not well reflected in the teacher representations.

5.2.2 Trade-offs of introducing teacher models: Although we experimentally showed that the performance of motion-based HAR can be improved by incorporating audio-based teacher guidance, there are also extra burdens for deployment. First of all, introduction of the acoustic teacher models can bring extra computational burdens. For example, the teacher models 1 and 2 have 79,720,919 and 96,404,312 trainable parameters respectively, whereas the baseline HAR model only has 3,610,007 trainable parameters. Besides, the introduction of our framework brings extra time costs for model training. Specifically, the average duration required to train an epoch is 6.18s for our baseline HAR models (batch size 64, two NVIDIA Titan Xp GPUs), but it becomes 13.08s with the teacher model 1 and 15.98s with the teacher model 2, more than twice as much as originally needed for the accuracy gain from 70% to 74%. Most importantly, our experiments show that the optimal hyper-parameters of the loss function are highly sensitive to the model architecture, participants, and the learning parameters (e.g., learning rate). An inappropriate selection of the hyper-parameters can even degrade the student performance. In our test, the optimal α value mostly remains between 0.7 to 0.9, while the optimal temperature tends to be 2, 3 and 5. Searching for these parameters is critical yet extremely time-consuming. Hence, despite the benefits for HAR performance with the proposed framework, such trade-offs should be carefully considered and handled in practical cases.

6 CONCLUSION

In this paper, we present a framework to augment IMU-based HAR by introducing acoustic knowledge to the model. Based on a teacher-student framework, the model benefits from acoustic information during training while relying on only inertial sensor data for inference. This approach is compelling because it mitigates the privacy risks and high data sampling burdens of continuous audio recording. Based on a study with 15 participants in which they performed 23 activities in their homes, we show the extent to which our framework improved performance and discuss the practical considerations for deploying our framework.

REFERENCES

[1] Rebecca Adaimi, Howard Yong, and Edison Thomaz. 2021. Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–24.

[2] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

[3] Ahmad Akl, Chen Feng, and Shahrokh Valaee. 2011. A novel accelerometer-based gesture recognition system. *IEEE Transactions on Signal Processing* 59, 12 (2011), 6197–6205.

[4] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: combining audio and motion sensing for gesture recognition on smartwatches. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 10–19.

[5] Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2021. IMU2Doppler: Cross-Modal Domain Adaptation for Doppler-based Activity Recognition Using IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–20.

[6] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.

[7] Kadian Davis, Evans Owusu, Vahid Bastani, Lucio Marcenaro, Jun Hu, Carlo Regazzoni, and Loe Feijs. 2016. Activity recognition based on inertial sensors for ambient assisted living. In *2016 19th international conference on information fusion (fusion)*. Ieee, 371–378.

[8] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[9] Yu-Jin Hong, Ig-Jae Kim, Sang Chul Ahn, and Hyoung-Gon Kim. 2010. Mobile health monitoring system based on activity recognition using accelerometer. *Simulation Modelling Practice and Theory* 18, 4 (2010), 446–455.

[10] Yasha Iravantchi, Karan Ahuja, Mayank Goel, Chris Harrison, and Alanson Sample. 2021. PrivacyMic: Utilizing Inaudible Frequencies for Privacy Preserving Daily Activity Recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[11] Bashima Islam, Md Mahbubur Rahman, Tousif Ahmed, Mohsin Yusuf Ahmed, Md Mehedi Hasan, Viswam Nathan, Korosh Vatanparvar, Ebrahim Nemati, Jilong Kuang, and Jun Alex Gao. 2021. BreathTrack: Detecting Regular Breathing Phases from Unannotated Acoustic Data Captured by a Smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–22.

[12] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring privacy concerns about personal sensing. In *International Conference on Pervasive Computing*. Springer, 176–183.

[13] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.

[14] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.

[15] Nicholas D Lane, Petko Georgiev, and Lorena Qendro. 2015. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 283–294.

[16] 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.

[17] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 321–333.

[18] Dawei Liang, Yangyang Shi, Yun Wang, Nayan Singhal, Alex Xiao, Jonathan Shaw, Edison Thomaz, Ozlem Kalinli, and Mike Seltzer. 2021. Transferring Voice Knowledge for Acoustic Event Detection: An Empirical Study. *arXiv preprint arXiv:2110.03174* (2021).

[19] Dawei Liang, Wenting Song, and Edison Thomaz. 2020. Characterizing the Effect of Audio Degradation on Privacy Perception And Inference Performance in Audio-Based Human Activity Recognition. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–10.

[20] 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.

[21] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2009. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. 165–178.

[22] Alessandra Moschetti, Laura Fiorini, Dario Esposito, Paolo Dario, and Filippo Cavallo. 2016. Recognition of daily gestures with wearable inertial rings and bracelets. *Sensors* 16, 8 (2016), 1341.

[23] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[25] Sebastien Richoz, Lin Wang, Philip Birch, and Daniel Roggen. 2020. Transportation mode recognition fusing wearable motion, sound, and vision sensors. *IEEE Sensors Journal* 20, 16 (2020), 9314–9328.

[26] Nabeel Siddiqui and Rosa HM Chan. 2020. Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist. *Plos one* 15, 1 (2020), e0227039.

[27] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence* (2022).

[28] Ben Tan, Yu Zhang, Sinno Pan, and Qiang Yang. 2017. Distant domain transfer learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[29] Chuanqi Tan et al. 2018. A survey on deep transfer learning. In *ICANN*.

[30] Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 1029–1040.

[31] Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D Abowd. 2015. Inferring meal eating activities in real world settings from ambient sounds: A feasibility study. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. 427–431.

[32] Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. 2016. Learning cross-domain landmarks for heterogeneous domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5081–5090.

[33] Hongyi Wen, Julian Ramos Rojas, and Anind K Dey. 2016. Serendipity: Finger gesture recognition using an off-the-shelf smartwatch. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3847–3851.

[34] 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.

[35] Tianwei Xing, Sandeep Singh Sandha, Bharathan Balaji, Supriyo Chakraborty, and Mani Srivastava. 2018. 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*. 37–42.

[36] Koji Yatani and Khai N Truong. 2012. Bodyscope: a wearable acoustic sensor for activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 341–350.

[37] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott Gilliland, Thomas Ploetz, Thad E Starner, Omer T Inan, and Gregory D Abowd. 2017. Fingersound: Recognizing unistroke thumb gestures using a ring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–19.

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