LASO: Exploiting Locomotive and Acoustic Signatures over the Edge to Annotate IMU Data for Human Activity Recognition

Soumyajit Chatterjee
IIT Kharagpur, India
sjituit@gmail.com

Avijoy Chakma
UMBC, USA
achakma1@umbc.edu

Aryya Gangopadhyay
UMBC, USA
gangopad@umbc.edu

Nirmalya Roy
UMBC, USA
nroy@umbc.edu

Bivas Mitra
IIT Kharagpur, India
bivas@cse.iitkgp.ac.in

Sandip Chakraborty
IIT Kharagpur, India
sandipc@cse.iitkgp.ac.in

ABSTRACT

Annotated IMU sensor data from smart devices and wearables are essential for developing supervised models for fine-grained human activity recognition, albeit generating sufficient annotated data for diverse human activities under different environments is challenging. Existing approaches primarily use human-in-the-loop based techniques, including active learning; however, they are tedious, costly, and time-consuming. Leveraging the availability of acoustic data from embedded microphones over the data collection devices, in this paper, we propose *LASO*, a multimodal approach for automated data annotation from acoustic and locomotive information. *LASO* works over the edge device itself, ensuring that only the annotated IMU data is collected, discarding the acoustic data from the device itself, hence preserving the audio-privacy of the user. In the absence of any pre-existing labeling information, such an auto-annotation is challenging as the IMU data needs to be sessionized for different time-scaled activities in a completely unsupervised manner. We use a change-point detection technique while synchronizing the locomotive information from the IMU data with the acoustic data, and then use pre-trained audio-based activity recognition models for labeling the IMU data while handling the acoustic noises. *LASO* efficiently annotates IMU data, without any explicit human intervention, with a mean accuracy of 0.93 (±0.04) and 0.78 (±0.05) for two different real-life datasets from workshop and kitchen environments, respectively.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings**; • **Human-centered computing** → *Ubiquitous computing*.

KEYWORDS

human-in-the-loop; smart-environments; labeling human activity

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418826

ACM Reference Format:
Soumyajit Chatterjee, Avijoy Chakma, Aryya Gangopadhyay, Nirmalya Roy, Bivas Mitra, and Sandip Chakraborty. 2020. LASO: Exploiting Locomotive and Acoustic Signatures over the Edge to Annotate IMU Data for Human Activity Recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418826

1 INTRODUCTION

Over the last two decades, many research works and commercial applications have used data from inertial motion units (IMU) for human activity recognition [25, 30, 37, 38]. These approaches mostly use labeled data to train a model to capture the personalized traits of individuals during various activities. Although more elaborate models can be designed to capture more complex activities, the major bottleneck becomes the availability of labeled IMU data from individuals. The standard approach is to recruit human volunteers who independently collect and annotate IMU data for different activities; however, such a human-in-the-loop approach has limitations in terms of scalability [1], selection of proper annotators [29], time and cost of annotation [19] and noisy and conflicting labeling [43]. In recent times, such problems have been tackled to some extent by the use of *Active Learning* [11] that chooses the most uncertain instances from the entire data and queries the annotators only for those instances [19]. However, in this approach, (i) a seed set of labels is essential, (ii) the problem of human-in-loop persists [43].

The automatic IMU data annotation process mostly relies on an auxiliary data source, such as video, which needs to be uploaded on a cloud/server for processing [8, 12, 32]. This may significantly compromise with the information privacy along with the associated network usage cost. Annotating the IMU data at the edge device itself may substantially mitigate the issues mentioned above. In this paper, we develop a multimodal approach to annotate IMU data at the edge (devices near the data source) with the help of locomotive and acoustic information. Notably, data collection devices like smartphones and many smart wearables also embed a microphone that can capture the sounds from the environment. Acoustic being an extremely rich data-source has already been used widely for identifying fine-grained human activities [7, 22, 23, 44]. Exploiting this, we develop **an opportunistic approach** to label the IMU data locally at the edge with the help of acoustic information. As only the labeled IMU data gets generated from the edge-device while

the acoustic data are discarded after processing, this approach does not compromise the environment's audio-privacy.

Opportunities and challenges: Although various human activities generate distinct sounds [17, 23, 26], we still need a supervised model to identify the activity label from an input audio clip. Incidentally, a rich source of labeled audio data and corresponding pre-trained models are publicly available, such as *YouTube-8M* [2], *Urbansound* [33], *SoundNet* [6], etc. which can identify human activities at different environments with a high accuracy (e.g. accuracy of *SoundNet* is > 85%). However, the major challenge of labeling IMU data with the acoustic information is that a microphone captures sounds from different sources in the environment; therefore, the audio data not only captures the sounds generated by target activities but also contains other noises. Hence, existing audio-based activity recognition models [17, 22, 23, 26, 42] cannot be applied directly in our context. To alleviate this, we observe that human activities are time-sequenced, and the activity duration helps us to resolve the conflict. For example, if the acoustic information indicates three overlapping activities, say, *typing*, *operating microwave* and *pressing doorbell*, within the duration $[t_1, t_2]$, $[t_3, t_4]$ and $[t_5, t_6]$ respectively, and we infer that the target activity captured by the IMU data occurred within $[\sim t_3, \sim t_4]$, then with high confidence, we can label those IMU instances with the activity label *operating microwave*. As the unlabeled IMU data is the only source for inferring the activity duration, the challenge here is to design an unsupervised approach to extract the time instance when a subject changes the activity.

Our contributions: Owing to the above challenges, in this paper, we propose *LASO*, a multimodal approach that couples the locomotive information, obtained from the IMUs, with the acoustic information to extract the specific time segment when a target activity has occurred and use that segmenting information to map the acoustic label with the IMU label. *LASO* leverages the idea that the distribution of the accelerometer readings change significantly when the subject changes from one activity, say *handling utensils*, to another, say *operating blender* in a *kitchen* environment. Accordingly, we develop a fully unsupervised mechanism based on the computation of change-point scores from the accelerometer readings and using a k-means clustering mechanism to determine which change-point scores correspond to a change in the activities. We segment the IMU data based on the detected change-points, and the corresponding audio segments within the same time duration is used to map the activity labels to the IMU data instances.

However, for a few activities having similar locomotive information, the above approach fails to work, and we get confounded activities where *LASO* returns multiple activity labels (say, *opening drawer* and *picking up a fork*) for the same segment of the IMU data. In order to address this issue, *LASO* implements a feedback mechanism where an explicit unsupervised segmentation over the audio data is used to identify & resolve confounded activities to the single correct activity. Finally, we evaluate (Section 5) *LASO* on two real-life datasets over *workshop* and *kitchen* environments. Experimental results show that *LASO* can efficiently label a good volume of data with an appreciable accuracy for both the datasets. Further, we develop a proof of concept (PoC) implementation using Raspberry Pi 3 (Model B) to show that *LASO* works perfectly over an edge setup. In summary, the major contributions in this paper are as follows. a) We develop a fully unsupervised mechanism based on the locomotive information to segment the IMU data for determining a change in the activity labels, which is used to map the label from the audio data. b) To overcome the problem when more than one consecutive activities exhibit similar locomotive information, we develop a feedback mechanism based on the unsupervised segmentation of the audio. c) Resource consumption of *LASO* has been analyzed thoroughly over an edge testbed.

2 RELATED WORK

In the context of detecting complex activities using IMU sensors, the problem of unlabeled data has been a well-studied topic where the majority of the works use human annotators to label the data. However, various recent researches have pointed out the problems with human annotations, such as – (a) time and cost of annotation [29] and (b) noisy labeling [43, 45]. One of the most accepted solutions to tackle all these problems has been to choose annotators based on interactions between the users and the annotators [29]. However, one of the major bottlenecks in this scheme is its dependence on the social relationship with the annotators. Besides, other techniques like *Experience Sampling* [13, 24], have also tried to mitigate these concerns by allowing the system to probe the subject for providing labels on the fly. However, in this case, a critical requirement is that the subjects involved in the process must be capable enough to provide such inputs, which can be a concern when we deal with smart-environments for the elderly assistance [16].

In the past, several works have tried to solve this in different ways. Out of these, the most successful ones are implemented using *Active Learning* [11, 18, 19], which helps reduce the volume of labeling tasks by choosing the most informative samples for annotation. Although this reduces the volume by a significant amount, it lacks in two aspects – (a) it usually demands partially labeled set and (b) human intervention is still required for annotating the chosen samples. In many cases, obtaining partially labeled data is also very challenging because of constraints like privacy concerns [16, 19]. Besides this, a few other works have also looked into approaches that may allow automatic annotations of sensor data. One of the earliest approaches in this domain is through the application of *Abductive* Reasoning [4] for annotating medical monitoring sensors using public repositories of knowledge. Although this approach might not be handy for any general-purpose sensors, it was one of the very initial ideas that tried to automatically label sensor data.

Subsequent works, like [8], developed techniques to extract human key points from [9] and use deep neural networks to annotate physical activities like walking, standing, etc. using videos as an auxiliary information source. However, one of the main limitations of this work is that it is restricted to simpler activities. Similar works like [32] have developed schemes using regression models that can be trained using the video data from YouTube and subsequently use monocular RGB videos to obtain the activity label for more straightforward fitness activities. We pose a brief comparison of all these works with our proposed framework, *LASO*, in the Table 1.

3 PROBLEM STATEMENT AND DATASETS

Let T be the entire duration of the data collection session of a subject. Let $\mathcal{I}_t$ be an unlabeled IMU data collected from device $\mathcal{D}$ at time instance $t \in [0, T]$, and $\mathcal{X}_t$ be the respective audio data

Table 1: Summary of Related Work on Automatic Annotation of Sensor Modalities

Paper	Primary Modalities	Auxiliary Modalities	External Labeling Source	Brief Methodology	Label General-Purpose Activities?
Alirezaie et. al. [4]	Medical Monitoring Sensors	NA	Open-Linked Knowledge Sources	1. Use *Abductive* reasoning.	No
Benndorf et. al. [8]	IMU Sensors	Video	OpenPose [9]	1. Key-points extraction. 2. Annotate sensors from videos.	Limited
Rey et. al. [32]	IMU Sensors	Monocular RGB videos	YouTube Videos	1. Extracting 2D poses from Video Frames 2. Based on a regression model	Limited
LASO	IMU Sensors	Audio Information	YouTube-8M [2]	1. Change-point detection 2. Unsupervised audio segmentation 3. Audio-based activity recognition.	**Yes**

Table 2: Activity Labels used in *LASO*

Activity Labels	Context
Using Drill	Workshop
Chopping	Kitchen
Door-In-Use	General
Water-Running	Kitchen/Bathroom
Knocking	General
Operating Microwave Oven	Kitchen
Using Shaver	Bathroom
Using Toothbrush	Bathroom
Operating Blenders	Kitchen
Pressing Doorbell	Entrance
Flushing Toilet	Bathroom
Using Hair-Dryer	Bathroom
Typing	Office
Hammering	Workshop
Using Saw	Workshop
Cooking or Handling Utensils	Kitchen

obtained from the microphone within time duration $[t, t+1]$ over the device $\mathcal{D}'$. $\mathcal{D}$ and $\mathcal{D}'$ are either the same device or paired through Bluetooth or WiFi, while synchronizing the time using NTP or RTC [16, 35]. We consider a label space $\mathcal{L} = \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_m\}$ where m is the total number of available unique labels. The objective of this paper is to develop a framework *LASO* to annotate the collected IMU data stream $\mathcal{I}_t$ with the activity label $\mathcal{L}_i \in \mathcal{L}$. Our implementation of *LASO* uses the label space $\mathcal{L}$ given in Table 2, which we fix from the audio-based activity detection framework adopted in [23]. We consider the following two scenarios. (a) The IMU and the microphone data are collected from the same device, say a smartphone; in this case, *LASO* may run on the smartphone to process and label $\mathcal{I}_t$. (b) The subject performs an activity, say *cooking*, in a smart home, where the IMU data is collected from a wrist-worn smartwatch, and the microphone data is recorded over a smartphone in her pocket. In such a scenario, typically, the smartwatch is paired with the smartphone, and the IMU data from the smartwatch is collected through the smartphone. Here, *LASO* can either run on the smartphone or an edge device within the smart home, like the data gateway, to label $\mathcal{I}_t$.

3.1 Key Idea and Challenges

The broad idea behind the design of *LASO* is as follows. Considering highly-accurate pre-trained models available publicly for recognizing activities from acoustic signatures, like *YouTube-8M* [2] and *Soundnet* [6], we obtain the label of the audio data segment $\mathcal{X}_t$, say $\mathcal{L}_i \in \mathcal{L}$, from such a model $\mathcal{M}$ and assign $\mathcal{L}_i$ to $\mathcal{I}_t$. However, a microphone not only captures the sound generated from the target activity but also captures other environmental sounds. For example, when the subject is *operating a microwave*, another person might *knock the kitchen door*; in this case, the microphone captures both the sounds. Therefore, model $\mathcal{M}$ may return multiple labels for an input $\mathcal{X}_t$, and hence we fail to assign any of those labels reliably and uniquely to $\mathcal{I}_t$.

Importantly, activity duration plays a key role here. If we can somehow figure out that the *target activity* has been performed in the interval $[\sim\tau, \sim\tau + \Delta t]$, then among the various overlapping activities detected by the microphone data $\mathcal{X}_t$, we can uniquely identify the activity whose time duration is $[\sim\tau, \sim\tau + \Delta t]$ and assign that label to $\mathcal{I}_t$ with high confidence. However, as IMU is the primary data source that has to be labeled, such *target activity* interval needs to be extracted from $\mathcal{I}_t$ itself. Hence the challenge is, **without having any explicit labeling information associated with $\mathcal{I}_t$, how do we determine that $\{\mathcal{I}_t | t \in [\tau, \tau + \Delta t]\}$ indicates a *single activity instance* performed by the subject?** We develop, explain, and evaluate our methodology based on the observations from two different datasets from two different environments, as follows.

3.2 Dataset Details

Following is the summary of the dataset.

3.2.1 In-House Dataset: Workshop Environment. We collected an in-house dataset for a workshop environment with two significant activities – (a) *using a saw* and (b) *using a hammer*. We involved 5 participants and asked them to perform activities like (i) cutting a wood brick or an aluminum pipe with a saw and (ii) hammer nails or metal sheets over a wooden plank. No explicit constraints were posed on the participants except that they have to repeat each activity at least 2 times, and the overall time spent performing all the activities should be at least 6 minutes.

For collecting the data, we used Moto 360 smartwatches, worn by the participants on the wrist (either left or right depending on the handedness of the participants) to capture the IMU data (sampled at 50Hz) and used an OnePlus 3 smartphone for capturing the audio data (sampled at 44.1kHz). These two devices were paired with each other over Bluetooth to ensure time synchronization between the two modalities. For collecting the ground-truth label, we captured videos (with a frame rate of 30 fps) with a watermarked timestamp (in order of milliseconds). Subsequently, we engaged an annotator, who independently labeled those videos from the aforementioned two activities, as the performed activities were straightforward.

3.2.2 CMU-MMAC Dataset: Kitchen Environment. In addition to the in-house dataset, we further evaluate the performance of the framework on real-life smart-home datasets from the CMU-MMAC kitchen dataset [1]. In this setup, each subject wears 5 wired and wireless IMU devices in different parts of the body (elbow, knees, and other anchor points). We extract a subset of the entire dataset (only *Brownie recipe*) with 10 subjects for which multiple IMU sensor data and the microphone data along with the ground-truth activity annotations are available. Notably, the data also contains

[1] http://kitchen.cs.cmu.edu/ Last Accessed: September 6, 2020

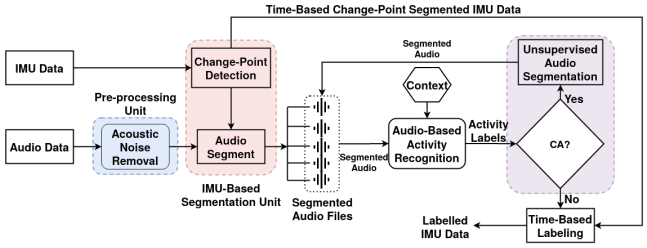


Figure 1: *LASO* Framework; CA ~ Confounded Activities

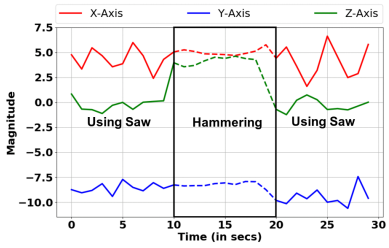


Figure 2: Locomotive Signature Indicating Activity Changes

videos from 3 high resolution, 2 low resolution, and 1 wearable cameras for ground truth annotation. We choose the IMU data from the 5 IMU sensors (sampled at 125Hz), and for the audio data (sampled at 44.1kHz), we consider the data captured by 5 balanced microphones placed in the environment. All the modalities in the CMU-MMAC kitchen setup are time-synchronized using NTP; although, the polling of different IMU sensors is asynchronous.

4 METHODOLOGY

Figure 1 shows the overall framework of *LASO*. After preprocessing the collected data, we use a two-step approach to find out the time duration $[\tau, \tau + \Delta t]$ for segmenting the IMU data, such that $\{\mathcal{I}_t | t \in [\tau, \tau + \Delta t]\}$ indicates a *single activity instance* (based on the label space $\mathcal{L}$). First, we compute *change-point scores* from $\{\mathcal{I}_t | t \in [0, T]\}$ over consecutive time windows of length ω, which indicates probable changes in the user activities. In the second step, we use an unsupervised clustering-based approach to obtain the change-point scores which indicate a change in the activity label; thus, we obtain the time duration $[\tau, \tau + \Delta t]$ for IMU segmentation. We then segment the acoustic data based on $[\tau, \tau + \Delta t]$ and determine the label from the input audio clip $\{\mathcal{X}_t | t \in [\tau, \tau + \Delta t]\}$; this label is tagged with the corresponding IMU segment. At this stage, we observe some *confounded activities* (activities having similar locomotive signatures) which we resolve based on a feedback mechanism. The details follow.

4.1 Data Preprocessing

The first task is to preprocess the acoustic and the IMU data for noise removal. The audio signal contains various types of noise generated by non-human sources such as air conditioning system or computers which can collude the acoustic signature. As shown in [23], audio signals in between 50Hz to 16kHz typically contains the acoustic signatures generated from human activities; therefore, we apply a Butterworth bandpass filter (order = 5) to extract the signals in between above frequencies. In addition, we also perform a noise-profiling using Audacity audio processing toolbox [27] to remove complex background noises from the audio data.

A noise filtering over the IMU data may result in a loss of other useful information [3]; therefore, we process the raw data directly. However, many complex activity detection techniques rely on more than one IMUs attached with different anchors (like elbows, knees, etc.) [21, 35]. Even though these IMUs are usually time-synchronized, they might not get polled at the same time instance, which is essential for detecting an activity change. In *LASO*, we resolve this issue by first obtaining the earliest time, say t_0 at which one of the IMUs is polled and then subsequently use t_0 to create a fixed-duration window δ over the entire data, collected from all the IMU sensors. If a sensor is polled within $[t_0, t_0 + \delta]$, we map that data point with the time instance t_0 and construct a combined IMU data $\{\mathcal{I}_{t_0}^1, \mathcal{I}_{t_0}^2, ..., \mathcal{I}_{t_0}^N\}$ where N is the number of such sensor instances.

4.2 Segmenting IMU Data

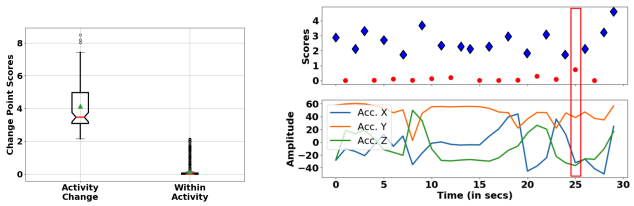
Our objective is to divide the IMU data $\{\mathcal{I}_t | t \in [0, T]\}$ into multiple segments $\left\{\{\mathcal{I}_t | t \in [\tau, \tau + \Delta t]\} \mid \sum_{\forall \tau} [\tau, \tau + \Delta t] = [0, T]\right\}$, such that $\{\mathcal{I}_t | t \in [\tau, \tau + \Delta t]\}$ contains the IMU instances corresponding to a single activity instance from the label space $\mathcal{L}$. To perform this segmentation in a complete unsupervised way, we observe that the locomotive information (obtained from the accelerometer) changes as the subject migrates from one activity to another (Figure 2). Accordingly, we use the concept of *change-point scores* to extract the change-points where the previous activity ends and a new activity (or no activity) commences.

4.2.1 Computing Change-point Scores. Formally, a change-point is defined as the instance where a stochastic process or time-series changes its probability distribution [3]. In order to compute the change-point scores, we consider the 3-axis accelerometer data $s(t)$ at time instance t obtained from the IMU sensor stream I_t. For a window of size ω, the matrix $S(t) = [s(t), s(t+1), \ldots, s(t+\omega-1)]$ forms the corresponding *Hankel* matrix [20]. Here ω is one of the most important hyper-parameters in the design of *LASO*; whose impact is analyzed in Section 5. The change point score $c(t)$ is computed by measuring the dissimilarity between two consecutive windows of $\mathcal{I}_t$. For measuring this dissimilarity, we use *Pearson Divergence Estimation* (PE) [31] represented as follows.

$$D(P_t | P_{t+\omega}) = \frac{1}{2} \int p'(S) \left(\frac{p(S)}{p'(S)} - 1 \right) \times dS \tag{1}$$

Here P_t and $P_{t+\omega}$ are the probability distributions corresponding to the samples in $S(t)$ and $S(t+\omega)$ respectively. In this context, $p(S)$ and $p'(S)$ form the probability densities. The values of $p(S)$ and $p'(S)$ are unknown in real-time; however, computing the *relative density ratio* is considered to be simpler than estimating the true densities [36]. For estimating the relative density ratio, we use the existing *relative density-ratio estimator* (RuLSIF) algorithm [41], and thus the α-*relative* PE-Divergence for the distributions is represented as,

$$D_\alpha(P_t | P_{t+\omega}) = \frac{1}{2} \int p'_\alpha(S) \left(\frac{p(S)}{p'(S)} - 1 \right) \times dS, \alpha \in (0, 1) \tag{2}$$



(a) (b)

Figure 3: (a) Clustering of Change Points, (b) Impact of Noise on Clustering of Change-Points. Blue Diamonds – Activity Changes and Red Dots – within an Activity

Finally, the change-point score $c(t)$[2] is obtained using the following equation,

$$c(t) = D_\alpha(P_t|P_{t+\omega}) + D_\alpha(P_{t+\omega}|P_t) \quad (3)$$

The change-point scores indicate the relative change in the data points within consecutive windows in the time-series distribution of $\mathcal{I}_t$. However, the value of the accelerometer changes even within a single activity (Figure 2), so we typically get a non-zero change-point scores between each consecutive windows. Therefore, the important question here is – **Which values of the change-point scores actually indicate a change between two activities?**

To find out the actual change-points from the change-point scores, existing approaches like [3] primarily used a supervised model by recruiting volunteers who perform the activities from the label space $\mathcal{L}$. This ground-truth data is used to train the supervised model to find out the change-point scores corresponding to the actual activity changes. Consequently, the methodology becomes applicable only for a pre-defined $\mathcal{L}$. Further, as *LASO* does not want to involve any human annotators, we cannot use such supervised approach to determine the change-point scores corresponding to the activity changes.

4.2.2 Detection of Activity Change-points. Although we cannot use a supervised approach as adopted in the existing studies, we observe that *LASO* does not require the change-point scores for detecting a change between each individual activities. Rather, it needs a threshold in the change-point scores, which can separate the change-points from *"within an activity"* to *"activity changes"*. To find out this threshold dynamically in an automated way, we first cluster the change-point scores using k-means clustering with $k = 2$. Once the clustering is done, we analyze the statistical significance of these clusters by performing *Welch's t-Test* [39] with a confidence interval of 95%. Subsequently, if the clusters are found to be statistically significant (when p-value < 0.05), we compute the means of the individual clusters for comparison. As change-point scores are meant to measure the changes in the time-series, we assume that the higher change-point scores correspond to the actual changes in the activity; hence, we mark the cluster with higher mean as the cluster with the change-point scores specifying an actual change in the activity pattern.

A typical analysis of the clusters for a randomly chosen user from CMU-MMAC dataset is shown in Figure 3a. From the figure, we can see that the medians of both the clusters are appreciably separated, thus proving the clustering to be significant in separating the change-point scores. We can also observe that the means of both the clusters coincide with their respective medians signifying the symmetrical distribution of the clusters. We use these means as thresholds in the change-point score to segment $\mathcal{I}_t$.

[2]only absolute values are considered for the D_α

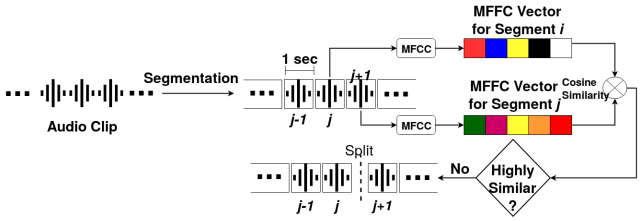


Figure 4: Unsupervised Audio Segmentation

4.3 Audio-based Activity Recognition

Once these change-points are clustered into separate clusters marking *"within an activity"* and *"activity changes"*, we check the corresponding time windows $[\tau, \tau + \Delta t]$ to segment the audio data $\mathcal{X}_t$ [15]. For smart-environments, like the one described in [35], where there is more than one source (also defined as *tracks*) of audio information placed in different strategic positions, there is a need to identify which audio track captures the activity information, corresponding to an IMU window, in the best possible way. We determine this by segmenting all the tracks at the same instance and then subsequently checking the *Power Spectral Density* (PSD) [3] for each of the audio segments. Finally, we choose the audio segment with the highest PSD and mark it as the audio source for that particular IMU window. Once the audio data $\mathcal{X}_t$ is segmented as $\{\mathcal{X}_t | t \in [\tau, \tau + \Delta t]\}$, *LASO* borrows a concept from [22, 23]to use *pre-trained* models from a large corpus of labeled sound files provided by sources like Youtube-8M [2] and SoundNet [6] to find out the label of $\{\mathcal{X}_t | t \in [\tau, \tau + \Delta t]\}$ eliminating the need of any external human annotators. For this purpose, we use the approach discussed in [23] which, in an ideal scenario, outputs either one of the labels from the label space $\mathcal{L}$ or, if the audio clip does not contain any relevant activity information or has no audio information, then this module returns a label '*Invalid*' to mark such instances. However, for some cases, it returns more than one label for an input audio segment, where we use a feedback mechanism to resolve such multi-label instances, as discussed next.

4.4 The Feedback Mechanism

As the IMU data is segmented based on a complete unsupervised technique, there is a high probability that there can be multiple activities that stay confounded within the same segment because of the noise in the data [5]. For example, an analysis of a user from CMU-MMAC dataset, as shown in Figure 3b [4], reveals the noise in the IMU data coupled with the unsupervised clustering can lead to marking an *"activity change"* as *"within an activity"* when the locomotive signatures of two consecutive activities are almost similar (for example, *opening a drawer* followed by *picking up a fork*).

[3]https://www.cygres.com/OcnPageE/Glosry/SpecE.html Last Accessed: September 6, 2020

[4]Calculated using the logs from eWatch [34] attached to the active hand

This can lead to confounding of activities where a IMU segment $\{\mathcal{X}_t | t \in [\tau, \tau + \Delta t]\}$ actually contains two different activities from $\mathcal{L}$. This problem gets further escalated for IMU data obtained from multiple units of IMU sensors; because, in that case, we need an additional level of windowing to synchronize the data.

To resolve the above concern, we introduce a feedback mechanism where we revisit those instances only where more than one labels are obtained from the audio segment. Borrowing the idea from [14, 40], we first segment the entire audio clip $\{\mathcal{X}_t | t \in [\tau, \tau + \Delta t]\}$ into multiple 1 sec segments and compute the *Mel Frequency Cepstral Coefficients* (MFCC) vectors for each of these segments. We then compute the cosine similarity of the MFCC vectors (with the number of coefficients = 20) obtained from two consecutive segments $\{\mathcal{X}_t | t \in [\tau + j - 1, \tau + j]\}$ and $\{\mathcal{X}_t | t \in [\tau + j, \tau + j + 1]\}$. If the similarity score is high ($\geq \cos 15^\circ$), we conclude that the subject continues to perform the same activity as there is no significant change in the audio components. If not, we split the audio at that time instance $\tau + j$ and create two new segments (see Figure 4) $\{\mathcal{X}_t | t \in [\tau, \tau + j]\}$ and $\{\mathcal{X}_t | t \in [\tau + j, \tau + \Delta t]\}$. The labels of these two audio segments are assigned as the label of the corresponding IMU segments $\{\mathcal{I}_t | t \in [\tau, \tau + j]\}$ and $\{\mathcal{I}_t | t \in [\tau + j, \tau + \Delta t]\}$.

Through this scheme, we eliminate most of the multi-label instances, albeit a few activities with very similar audio signatures may remain. In this context, if any intermediate segment is silent, then also the algorithm may find it different from its previous segment (which may contain some valid audio). In such cases, we ignore the intermediate silent audio segments as neither the audio nor the corresponding IMU segment points to a relevant change.

5 EVALUATION

As *LASO* sometime returns multiple labels while encountering confounded activities, we compute the accuracy of *LASO* over the two datasets (details in Section 3) in terms of *Dice Similarity Coefficient score* ($\mathcal{D}$) which measures the similarity between two sets as follows. Let A and B be the sets of activities in the ground-truth and the ones detected by *LASO*, respectively. Then $\mathcal{D} = \frac{2 \times |A \cap B|}{|A| + |B|}$. We compute the average Dice-Coefficient across all the instances.

5.1 Preparation of the Ground-Truth

For the in-house dataset, we directly obtain the ground-truth annotations from the video data. On the other hand, although the CMU-MMAC dataset provides activity labels annotated by human annotators, we cannot use those as the ground truths because of the following reason. The activity label space in CMU-MMAC is unrestricted; therefore, the annotators have used very granular labels where multiple similar labels may indicate a single activity. For example, both the labels *'take-fork'* and *'take-scissors'* have been used in CMU-MMAC, but even the IMU data may not be able to detect activities at this granularity; so we use a single activity label, for example, *'handling utensils'* replacing both the above labels. Therefore, we obtain the ground-truth, by re-annotating as follows.

We start with the 45 unique labels provided by the CMU-MMAC dataset annotators across the 10 subjects. Corresponding to each of these 45 unique labels, we look into the individual video instances (available with the dataset, see Subsection 3.2) of different activities. As the videos are multi-angled, to reduce the cognitive load of the annotators, we clip these videos and create a set of 27 unique animated images based on the primary view-port for individual activities. We pose the annotation task as an online survey where we ask the annotators to annotate these images with the activity labels from Table 2 with the context *'Kitchen'* and *'General'* with an additional label *'None-of-the-Above'*. From the survey, we have received the annotations from 15 independent annotators. We first start by assessing the quality of the annotations by evaluating *Cohen's* κ statistics [28]. We observe a fair inter-annotator agreement with an average of κ-statistics ≈ 0.40. This is further comforting for us to know that after taking the majority agreement, we find only 15% of the activities mapped to *'None-of-the-Above'*. Once these annotations are obtained, we prepare the ground-truth for all the 10 subjects, taking the majority consensus for each activity annotation. We also filter out the activities marked as *'none'* in the CMU-MMAC annotations and consider only the activities of interest.

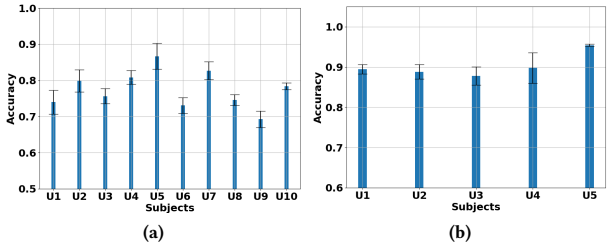


Figure 5: Accuracy: (a) CMU-MMAC and (b) In-house Dataset

5.2 Labeling Performance

From the results shown in Figures 5a and 5b, we see that for most of the users, *LASO* performs with an average Dice-Coefficient of ~ 0.78 (± 0.05) for the CMU-MMAC dataset and ~ 0.93 (± 0.04) for the in-house dataset. We get a little low accuracy over the CMU-MMAC dataset compared to the in-house dataset, as CMU-MMAC uses 6 different labels in contrast to the 3 different labels used in the in-house dataset. We also investigate the drop in the accuracy for some subjects and find that in a few cases, the audio-based activity recognition model gets confused between the labels where the audio frequencies of the generated sound are similar. For example, cutting a metal pipe rapidly using a saw generates audio signals similar to those generated by a drill (the activity *'using drill'* is also there in the label space, see Table 2); therefore an error gets introduced when the subject cuts the pipe rapidly.

5.3 Eliminating Multi-label Instances

Next, we observe the volume of multi-labeled instances as a result of confounded activities after the first pass (before feedback) and the second pass (after feedback) of *LASO*. Figure 6a shows the percentage of multi-labeled instances after the first pass and the second pass over the CMU-MMAC dataset. Similarly, Figure 6b plots the same after the first pass over the in-house dataset; notably, almost all the multi-label instances have been resolved after the second pass. From both the figures, we observe that the proposed feedback mechanism over *LASO* can eliminate the majority of the multi-label instances by resolving the confounding of activities.

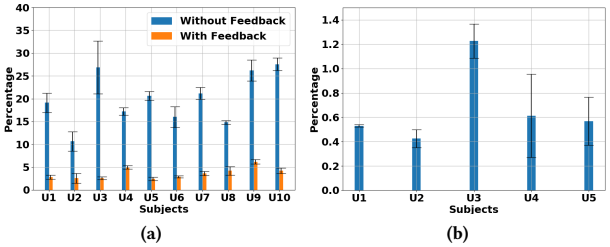


Figure 6: Percentage of Multi-labeled Instances in (a) CMU-MMAC Dataset, (b) In-house Dataset

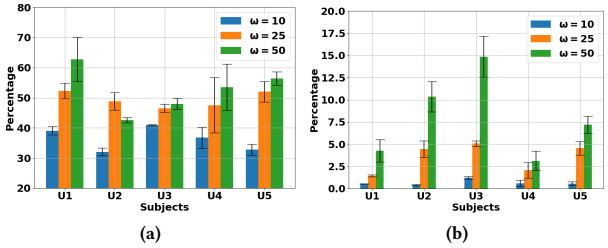


Figure 7: Impact of ω over In-house Dataset: (a) Data Volume Labeled and (b) Percentage of Confounded Instances

Notably, these results have been generated using $\omega = 10$; we next analyze the impact of ω over *LASO* design.

5.4 Impact of ω: Annotated Data Volume vs Amount of Multi-labeled Instances

ω defines the window size based on which we compute the change-point scores (Subsection 4.2). In Figure 7a, we show the amount of labeled IMU instances (in terms of percentage of the total available data) with different ω values for the 5 users over the in-house dataset. Figure 7b plots the percentage of multi-labeled instances over the in-house dataset after the first pass of *LASO*. We observe that the volume of labeled data increases with the increase in the ω value, whereas the amount of multi-label instances due to the confounded activities also increases. Although we have not shown the result for the CMU-MMAC dataset explicitly due to space limitation, we observe similar trends in that dataset as well. A large value of ω may cause a miss in the actual activity change-points, resulting in confounded activities and multi-labeled instances. Notably, an increasing number of multi-labeled instances after the first pass of *LASO* increases the feedback processing overhead (Subsection 4.4) over the edge-device. However, a small value of ω may produce frequent activity change-points. As the IMU and thereafter the audio segmentation are done at the detected change-points, the generated audio segments are likely to be too small to detect any meaningful activity label, thus resulting in an increased number of unlabeled IMU instances. Besides this, the audio signal's background noise also affects, which is likely to be more prominent over a small segment of the audio clip. For example, many times, the audio-based model returns an activity like *'person talking'*; consequently, we have to ignore such segments for IMU labeling. Therefore, we observe that a large window size is more suitable; although, it escalates multi-label instances. Thus, a suitable choice of ω depends on the processing capability of the edge-device.

Interestingly, we observe that even though for most of the users, *LASO* can label 55–60% of the data when the ω value is large; however, an appreciable portion is still left unlabeled. Investigating this, we find that at times, the human annotators labeled the data in a stretch, causing a good number of irrelevant activities to creep in within the valid labels. For example, when the subject is *hammering a wooden plank with nails*, for which he often *picks the nails*. As these intermediate activities (like *picking a nail*) is less than a few seconds, the annotators usually skip these and include them in the correct ground truth time with appropriate labels like *'using a hammer'*. We

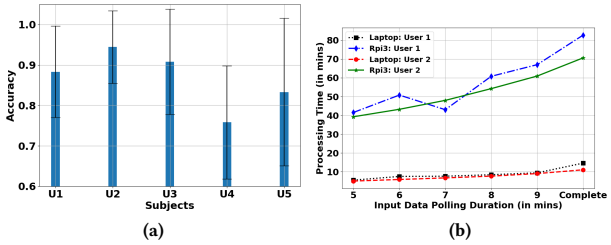


Figure 8: (a) Classification Accuracy on the Labeled Data and (b) Time Required for Labeling at the Edge

observe the same behavior for the CMU-MMAC dataset, where the annotators annotate the entire activities of *'putting the pan in the baking oven'* and *'switching on the oven'* as *'put-baking_pan-into-oven'* (we discuss this in Section 6). However, most of these small intermediate activities contain no relevant acoustic signatures, thus providing no valid labels. Therefore, a significant portion of the unlabeled data that *LASO* skips in labeling are actually due to such noises in the dataset itself. Thus, we observe that *LASO* firmly ensures that the labels are provided only for those instances where the framework is confident that the target activities have occurred with appreciable acoustic signatures.

5.5 Classification Quality Analysis

As the labeled data generated by *LASO* is likely to be used subsequently for supervised learning, we check the classification accuracy using the *LASO*-generated labeled IMU data. As there is a certain amount of imbalance in the duration of each activity being performed by the users, we first balance the dataset by *oversampling* it using *Synthetic Minority Oversampling Technique (SMOTE)* [10]. We then perform a 5-fold cross-validation with stratified random sampling on this balanced data for evaluating the labeling accuracy using a Random Forest-based model (number of estimators = 10). From Figure 8a, we observe that the labeling accuracy is quite high (for in-house dataset), albeit with a high variance because of the restricted volume of the overall dataset.

5.6 Resource Consumption over the Edge

In *LASO*, we chose audio over other modalities like the video to ensure that the processing overhead is less enough for executing the framework over an edge-device. Here, we benchmark the resource consumption behavior of *LASO* over a Raspberry Pi 3 Model B

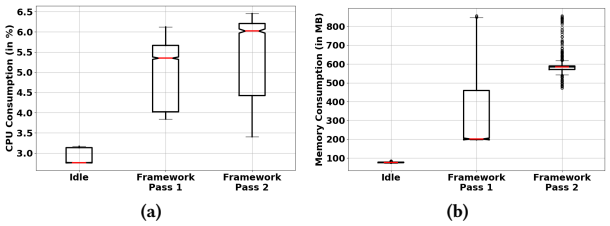


Figure 9: Resource Consumption over the Edge – (a) CPU Consumption and (b) Memory Consumption

development board (running Raspbian OS with Linux kernel version 4.19.75-v7+). The device runs on an ARMv7 processor with primary memory of 1GB. Here we show the results for the in-house dataset only due to the space limitation; we observe a similar resource consumption behavior for the CMU-MMAC dataset.

We start by analyzing the total execution time of *LASO*. For this, we record the execution time for two subjects with the maximum volume of IMU data. Furthermore, to get a clear idea, we compare this time with the total execution time of *LASO* on a standard Dell Inspiron laptop with Intel Core i7 processor and 16GB primary memory (running Ubuntu 18.04 with Linux kernel version 4.15.0−99−generic). From Figure 8b, we see that *LASO* takes higher execution time on Raspberry Pi in comparison to the laptop. However, in both cases, we observe that the execution time increases with an increase in the data volume. Therefore, we can understand that there is a trade-off between the execution time, periodicity of labeling activity, and the specification of the edge device, which can be exploited judiciously depending on available infrastructure.

To delve deeper, we also observe the CPU consumption (from `/proc/stat`) and primary memory consumption (using `free`) by *LASO*. From Figures 9a and 9b, we see that *LASO* consumes more resources both in terms of computing and memory while executing the feedback mechanism. This is mostly because the feedback mechanism involves more computation, including complex signal processing tasks to obtain MFCC vectors corresponding to the audio segments. However, in none of the cases, *LASO* chokes the resources on the edge-device and can successfully label the data.

6 DISCUSSION

Here we summarize some interesting observations as experienced during the development of *LASO*.

6.1 Granularity of Labeling

One of the primary advantages of human-in-the-loop based annotation is that it allows fine-grained activity labels to be tagged to the sensor data. However, as *LASO* depends on the acoustic context, it is restricted to the granularity of labeling. Nevertheless, as *LASO* leverages the open pre-trained models which tend to get richer with the increasing availability of new acoustic signatures, *LASO* is expected to improve over time in terms of the granularity of the activities. Albeit these labels might not be as fine-grained as annotated by a human annotator; however, at times, it may put forward more accurate labels as well. For example, during experimentation on the CMU-MMAC dataset, we observe that for the activities *"putting the baking pan into the oven"* and subsequently *"switching it on"*, the corresponding human-annotated label (provided by CMU-MMAC annotators) is *'put-baking-pan-into-oven'*. In contrast, the framework detected two activities *'Handling Utensils'* and *'Microwave-in-Use'*. This is because the IMU change-points recorded the changes in activities, and subsequently, (a) the noise of the baking pan and (b) the beeping sounds caused by pressing the oven-switches further confirmed two separate labels.

6.2 Labeling Silent Activities

As *LASO* depends on the availability of audio signatures, one scenario that the framework does not explicitly tackle is regarding the labeling of 'silent activities' that do not generate significant audio. For example, we observe in the CMU-MMAC kitchen dataset that there are instances where the subject is *'walking-to-the-fridge'*, and the corresponding audio segment does not capture anything relevant to infer this activity. Despite this limitation, this framework can efficiently detect change-points in the IMU data stream through the variations in the change-point scores. This opportunity can be utilized to obtain the activity labels only for these instances by applying schemes similar to [13], where we can directly probe the user after a specific change in activity is observed, although no significant audio signature is recorded. Moreover, we also find that the number of such instances is usually not that huge in volume; thus, involving some "human-in-the-loop" might not be too expensive.

7 CONCLUSION

Data annotation for human activity recognition is a significant challenge, where myriads of sensors continuously generate data, albeit a majority of the data remains unlabeled. Existing frameworks like active learning try to reduce human interventions by selectively probing the human annotators for labeling the ambiguous data points; however, such frameworks involve tedious and time-consuming "human-in-the-loop" approaches. In contrast to this, in this paper, we have developed an edge-based framework to fully automate the data annotation process with the help of locomotive and audio data, both of which are widely available in various smart-home environments. Being edge-based, the framework allows the entire process to be non-intrusive in terms of user's audio privacy. Rigorous experimentation of *LASO* over two real-life datasets shows that *LASO* can label significant portions of the IMU data without involving any human annotator while maintaining appreciable accuracy. Also, a PoC implementation over a resource-constrained edge-device shows that the framework is capable of successfully performing these annotations over the edge.

ACKNOWLEDGMENT

This work has been partially supported by the MHRD funded SPARC collaborative project 'SFE_SKI-1220' along with the partial support from SERB Early Career Research Award ECR/2017/000121 (18/07/2017), funded by Department of Science and Technology, Government of India. N. Roy, A. Chakma, and A. Gangopadhyay acknowledge NSF CAREER Award # 1750936, ONR under grant N00014-18-1-2462, and Alzheimer's Association, Grant/Award # AARG-17-533039.

REFERENCES

[1] Alaa E Abdel Hakim and Wael Deabes. 2019. Can People Really Do Nothing? Handling Annotation Gaps in ADL Sensor Data. *Algorithms* 12 (2019), 217.

[2] 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).

[3] Mohammad Arif Ul Alam, Nirmalya Roy, Aryya Gangopadhyay, and Elizabeth Galik. 2017. A smart segmentation technique towards improved infrequent non-speech gestural activity recognition model. *Pervasive and Mobile Computing* 34 (2017), 25–45.

[4] Marjan Alirezaie and Amy Loutfi. 2013. Automatic Annotation of Sensor Data Streams using Abductive Reasoning.. In *KEOD*. 345–354.

[5] Samaneh Aminikhanghahi, Tinghui Wang, and Diane J Cook. 2018. Real-time change point detection with application to smart home time series data. *IEEE Transactions on Knowledge and Data Engineering* (2018), 1010–1023.

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*. 892–900.

[7] 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*. ACM, 10–19.

[8] Maik Benndorf, Frederic Ringsleben, Thomas Haenselmann, and Bharat Yadav. 2017. Automated Annotation of Sensor data for Activity Recognition using Deep Learning. *INFORMATIK 2017* (2017).

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Int. Res.* (2002), 321–357.

[11] David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning* 15 (1994), 201–221.

[12] Federico Cruciani, Ian Cleland, Chris Nugent, Paul McCullagh, Kåre Synnes, and Josef Hallberg. 2018. Automatic annotation for human activity recognition in free living using a smartphone. *Sensors* 18 (2018).

[13] Mihaly Csikszentmihalyi and Reed Larson. 2014. Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology*. Springer, 35–54.

[14] Alain De Cheveigné and Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111 (2002), 1917–1930.

[15] Alexander Diete, Timo Sztyler, and Heiner Stuckenschmidt. 2018. Exploring semi-supervised methods for labeling support in multimodal datasets. *Sensors* 18 (2018), 2639.

[16] Anthony Fleury, Michel Vacher, and Norbert Noury. 2010. SVM-based Multimodal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms, and First Experimental Results. *Trans. Info. Tech. Biomed.* 14, 2 (2010), 274–283.

[17] P. Haubrick and J. Ye. 2019. Robust Audio Sensing with Multi-Sound Classification. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*. 1–7.

[18] H. M. Sajjad Hossain, M. D. Abdullah Al Haiz Khan, and Nirmalya Roy. 2018. DeActive: Scaling Activity Recognition with Active Deep Learning. *IMWUT* 2 (2018), 66:1–66:23.

[19] HM Sajjad Hossain, Md Abdullah Al Hafiz Khan, and Nirmalya Roy. 2017. Active learning enabled activity recognition. *Pervasive and Mobile Computing* 38 (2017), 312–330.

[20] Yoshinobu Kawahara, Takehisa Yairi, and Kazuo Machida. 2007. Change-point detection in time-series data based on subspace identification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, 559–564.

[21] Matthew Keally, Gang Zhou, Guoliang Xing, Jianxin Wu, and Andrew Pyles. 2011. PBN: Towards Practical Activity Recognition Using Smartphone-Based Body Sensor Networks. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 246–259.

[22] 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*. ACM, 283–294.

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

[24] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 338:1–338:13.

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

[26] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3 (2019), 17:1–17:18.

[27] Dominic Mazzoni. 1999–2019. Audacity software is copyright 1999-2019 Audacity Team. The name Audacity is a registered trademark of Dominic Mazzoni. https://www.audacityteam.org/.

[28] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22 (2012), 276–282.

[29] Aditi Muralidharan, Zoltan Gyongyi, and Ed Chi. 2012. Social annotations in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1085–1094.

[30] Viswam Nathan, Sudip Paul, Temiloluwa Prioleau, Li Niu, Bobak J Mortazavi, Stephen A Cambone, Ashok Veeraraghavan, Ashutosh Sabharwal, and Roozbeh Jafari. 2018. A survey on smart homes for aging in place: Toward solutions to the specific needs of the elderly. *IEEE Signal Processing Magazine* 35 (2018), 111–119.

[31] Karl Pearson. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (1900), 157–175.

[32] Vitor Fortes Rey, Peter Hevesi, Onorina Kovalenko, and Paul Lukowicz. 2019. Let There Be IMU Data: Generating Training Data for Wearable, Motion Sensor Based Activity Recognition from Monocular RGB Videos. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 699–708.

[33] J. Salamon, C. Jacoby, and J. P. Bello. 2014. A Dataset and Taxonomy for Urban Sound Research. In *ACM MM*. Orlando, FL, USA, 1041–1044.

[34] Asim Smailagic, Daniel P. Siewiorek, Uwe Maurer, Anthony Rowe, and Karen P. Tang. 2005. eWatch: Context Sensitive System Design Case Study. In *Proceedings of the IEEE Computer Society Annual Symposium on VLSI: New Frontiers in VLSI Design*. IEEE Computer Society, 98–103.

[35] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. 2009. Temporal segmentation and activity classification from first-person sensing. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 17–24.

[36] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. 2012. *Density ratio estimation in machine learning*. Cambridge University Press.

[37] Catherine Tong, Shyam A Tailor, and Nicholas D Lane. 2020. Are Accelerometers for Activity Recognition a Dead-end?. In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*. 39–44.

[38] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119 (2019), 3–11.

[39] Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika* (1947), 28–35.

[40] Chenren Xu, Sugang Li, Gang Liu, Yanyong Zhang, Emiliano Miluzzo, Yih-Farn Chen, Jun Li, and Bernhard Firner. 2013. Crowd++: Unsupervised Speaker Count with Smartphones. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 43–52.

[41] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. 2013. Relative density-ratio estimation for robust distribution comparison. *Neural computation* (2013), 1324–1370.

[42] 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*. ACM, 341–350.

[43] Mattia Zeni, Wanyi Zhang, Enrico Bignotti, Andrea Passerini, and Fausto Giunchiglia. 2019. Fixing Mislabeling by Human Annotators Leveraging Conflict Resolution and Prior Knowledge. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2019), 32:1–32:23.

[44] Cheng Zhang, AbdelKareem Bedri, Gabriel Reyes, Bailey Bercik, Omer T. Inan, Thad E. Starner, and Gregory D. Abowd. 2016. TapSkin: Recognizing On-Skin Input for Smartwatches. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*. ACM, 13–22.

[45] Liyue Zhao, Gita Sukthankar, and Rahul Sukthankar. 2011. Incremental relabeling for active learning with noisy crowdsourced annotations. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 728–733.