Fine-grained Activities Recognition with Coarse-grained Labeled Multi-modal Data

Zhizhang Hu
University of California
Merced
Merced, California
zhu42@ucmerced.edu

Tong Yu
Carnegie Mellon University
Moffett Field, California
tong.yu@sv.cmu.edu

Yue Zhang
University of California
Merced
Merced, California
yzhang58@ucmerced.edu

Shijia Pan
University of California
Merced
Merced, California
span24@ucmerced.edu

ABSTRACT

Fine-grained human activities recognition focuses on recognizing event- or action-level activities, which enables a new set of Internet-of-Things (IoT) applications such as behavior analysis. Prior work on fine-grained human activities recognition relies on supervised sensing, which makes the fine-grained labeling labor-intensive and difficult to scale up. On the other hand, it is much more practical to collect coarse-grained label at the level of activity of daily living (e.g., cooking, working), especially for real-world IoT systems. In this paper, we present a framework that learns fine-grained human activities recognition with coarse-grained labeled and a small amount of fine-grained labeled multi-modal data. Our system leverages the implicit physical knowledge on the hierarchy of the coarse- and fine-grained labels and conducts data-driven hierarchical learning that take into account the coarse-grained supervised prediction for fine-grained semi-supervised learning. We evaluated our framework and CFR-TSVM algorithm on the data gathered from real-world experiments. Results show that our CFR-TSVM achieved an 81% recognition accuracy over 10 fine-grained activities, which reduces the prediction error of the semi-supervised learning baseline TSVM by half.

1 INTRODUCTION

Fine-grained human activity recognition using Internet-of-Things (IoT) systems enables various smart home applications, such as elderly care, patient care, and occupant behavior profiling [9, 15, 20]. Previous studies mainly focus on recognizing and monitoring human activities at a coarse-grained level that describes the categories of activities of daily living, e.g., cooking, cleaning. This granularity level is insufficient to profile detailed human information, such as behavioral changes [15]. For instance, to analyze the complexity of the cooking activity, the coarse-grained 'cooking' activity recognition is not enough. Therefore, fine-grained action-level activities (e.g., cutting, washing) recognition is necessary.

Previous studies have attempted to leverage multiple sensing modalities to achieve fine-grained activities recognition, such as the combination of electrical load and passive RFID sensors [5],

as well as the combination of vibration sensors and electrical load sensors [15]. These prior attempts are mainly based on well-labeled data (e.g., [16]) to train machine learning models. The high-quality fine-grained labeled multi-modal data provides a solid foundation for high-performance models. However, such (labeled) data is labor-intensive to acquire and hence difficult to scale up. Prior work handling limited labeled data learning utilizes a semi-supervised learning technique, which combines a small amount of labeled data with a large amount of unlabeled data to train the model [7, 21–23]. However, these studies focus on traditional coarse-grained activities recognition only. We design our fine-grained activity recognition system using coarse-grained and a small number of fine-grained labels with two heuristics (physical knowledge).

- (1) Compared to fine-grained labels, coarse-grained labels are less labor-intensive to acquire.
- (2) Co-located sensing modalities capture complementary information to enhance fine-grained activity data representation resolution.

As a result, we target two challenges in this paper: 1) how to utilize coarse-grained label with a small amount of fine-grained label to achieve fine-grained activity recognition, 2) how to utilize multi-modal data to achieve an accurate fine-grained activity recognition. Our system follows a hierarchical structure, it conducts coarse-grained supervised learning and takes into account the supervised prediction as input for semi-supervised learning of fine-grained activities recognition. The contributions of this work are as follows:

- (1) We introduce a multi-modal coarse-to-fine grained activity recognition framework that learns with coarse-grained labels and a small number of fine-grained labels.
- (2) We present a semi-supervised coarse-to-fine relationship learning algorithm CFR-TSVM that leverages physical knowledge of coarse- and fine-grained label in a data-driven approach to achieve accurate activity recognition.
- (3) We conduct real-world experiments to evaluate the system and the algorithm with two sensing modalities – the structural-vibration and the wearable.

The rest of the paper is organized as follows. Section 2 demonstrates the system and algorithm design details. Next, Section 3 presents the evaluation results and analysis. Then Section 4 presents related work and discusses their limitations in our target scenarios. Finally, we conclude the paper in Section 5.

2 FRAMEWORK ARCHITECTURE

We present a multi-modal coarse-to-fine activity learning framework illustrated in Figure 1. The framework takes coarse-grained

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UbiComp/ISWC '20 Adjunct, September 12–16, 2020, Virtual Event, Mexico

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8076-8/20/09...\$15.00

<https://doi.org/10.1145/3410530.3414320>

Figure 1: Multi-modal coarse-to-fine activity recognition framework.

labeled data and a small amount of fine-grained labeled data to train the model, and test unlabeled data. The output of the system is the fine-grained activity prediction. The framework holds the following assumption:

- (1) There is a hierarchical relationship between the coarse- and fine-grained labels, i.e., a fine-grained label belongs to a coarse-grained label category.
- (2) There is sufficient coarse-grained labeled data for supervised learning and coarse-grained prediction.
- (3) The fine-grained activities are not overlapping in time, i.e., no time window has two activities occurring simultaneously.

In this section, we first introduce the details of the complementary information of multi-modal IoT sensing systems that enables accurate fine-grained activity information inference (Section 2.1). Then we explain our semi-supervised coarse-to-fine relationship learning that increases the learning accuracy utilizing the hierarchical relation between coarse- and fine-grained labels (Section 2.2).

2.1 Multi-modal IoT Sensing System

In this paper, we focus on the multi-modal combination of structural vibration sensing and wearable for the system design. The

Figure 2: System hardware. (left) the wearable sensing node with an IMU, and (right) the vibration sensing node with a geophone sensor.

vibration-based human sensing has been explored to enable long-term in-home monitoring for various smart home applications such as elderly care [4, 15]. On the other hand, more and more wearable devices, such as smartwatches or mobile phones, have entered people's lives for elder care and activity recognition [19]. Our system combines these two modalities by incorporating multiple structural vibration sensors that cover the designated activity area and an IMU sensor worn by the occupant, as shown in Figure 2.

To capture the spatial-temporal information from the multi-modal data, we place the vibration sensor on multiple surfaces such as table and floor. The vibration and IMU sensors generate a time sequence signal that reflects the investigated surface velocity and the body part acceleration, respectively. For each vibration sensor, there is one channel of signal, and for each IMU there are three. We apply a sliding window with overlapping on the in total N signal channels. For each window, we extract 36 key features from both the time and the frequency domains (details in Table 1). We refer to this 36-dimension feature vector as $\mathbf{x}_{channel}$. Then we concatenate $\mathbf{x}_{channel}$ from N channels to form the $36N$ -dimension feature vector $\mathbf{x}$ as the extracted feature vector of one timestamp.

2.2 Semi-Supervised Coarse-to-Fine Relationship Learning

To leverage the **implicit physical knowledge** of coarse- and fine-grained label relation, our system takes a **data-driven** approach by incorporating the coarse-grained prediction from supervised learning into the fine-grained semi-supervised learning. As shown in Figure 1, once our system obtains the coarse-grained label predictions, it combines this prediction and multi-modal features as a new feature vector. This new feature vector is then used as the input of semi-supervised learning, where the coarse-fine-grained relation between these labels is learned in a data-driven manner.

We select a commonly used supervised classifier Support Vector Machine (SVM) [2] as the algorithm example to explain the proposed framework. Limited fine-grained labeled data is insufficient to train a robust supervised learning model such as the SVM, thus we investigate SVM in the transductive learning settings to

Table 1: Features of a zero mean windowed signal.¹

Feature ID	Feature	Feature Detail
1	Mean	$\mu = \frac{1}{K} \sum_{i=0}^K S_i$
2	STD	$\sigma = \sqrt{\frac{1}{K} \sum_{i=1}^K (S_i - \mu)^2}$
3	Min	Min(S).
4	Max	Max(S).
5	Mode	The most frequent value in S.
6	Value Range	Max(S) - Min(S).
7	Mean Crossing	Counts of signal cross the mean value.
8-17	Freq. Peak Values	Top 10 freq. spectrum peak values.
18-27	Top Frequencies	Freq. of top 10 freq. spectrum peaks.
28	Signal Energy	$\sum_{i=1}^K (S_i)^2$
29	Shape mean	$\mu_{shape} = \frac{1}{e} \sum_{i=1}^N i P_i$
30	Shape STD	$\sigma_{shape} = \sqrt{\frac{1}{e} \sum_{i=1}^N (i - \mu_{shape})^2 P_i}$
31	Shape Skewness	$\gamma_{shape} = \frac{1}{e} \sum_{i=1}^N \left(\frac{i - \mu_{shape}}{\sigma_{shape}} \right)^3 P_i$
32	Shape Kurtosis	$\zeta_{shape} = \frac{1}{e} \sum_{i=1}^N \left(\frac{i - \mu_{shape}}{\sigma_{shape}} \right)^4 P_i - 3$
33	Amplitude mean	$\mu_{amp} = \frac{1}{N} \sum_{i=1}^N P_i$
34	Amplitude STD	$\sigma_{amp} = \sqrt{\frac{1}{N} \sum_{i=1}^N (P_i - \mu_{amp})^2}$
35	Amplitude Skewness	$\gamma_{amp} = \frac{1}{N} \sum_{i=1}^N \left(\frac{P_i - \mu_{amp}}{\sigma_{amp}} \right)^3$
36	Amplitude Kurtosis	$\zeta_{amp} = \frac{1}{N} \sum_{i=1}^N \left(\frac{P_i - \mu_{amp}}{\sigma_{amp}} \right)^4 - 3$

train. Instead of finding a decision boundary with maximal margins over labeled data by SVM, transductive SVM (TSVM) aims to train a boundary with maximal margins over both labeled data and unlabeled data [11]

As a baseline, we introduce the TSVM classifier $g : \mathbb{R}^d \rightarrow \{+1, -1\}$ to predict the fine-grained label y , with the input feature vector $\mathbf{x}$, where d is the number of features. Given two-class training data $(y_1, \mathbf{x}_1), \dots, (y_l, \mathbf{x}_l)$ and unlabeled data $\mathbf{x}_{l+1}, \dots, \mathbf{x}_{l+u}$, TSVM aims to find the maximum-margin hyperplane w and bias term b by minimizing the loss function

$$\min_{w, b} \frac{\|w\|^2}{2} + C_1 \sum_{i=1}^l \max(1 - y_i(w^T \phi(\mathbf{x}_i) + b), 0) + C_2 \sum_{i=l+1}^l \max(1 - |w^T \phi(\mathbf{x}_i) + b|, 0), \quad (1)$$

where C_1 and C_2 are the penalty parameters balancing regularization term and training losses over labeled data and unlabeled data, respectively. Intuitively, TSVM tends to find boundaries in regions where there is less labeled and unlabeled data. Thus, it is a low-density separation method studied in semi-supervised learning problems [11]. Similar to [18], we extend the binary TSVM to handle a multi-class problem using a one-against-one strategy [8].

To efficient leverage (i) the relationship between fine-grained labels and coarse-grained labels and (ii) rich coarse-grained labels but limited fine-grained labels, we propose the CFR-TSVM approach. There are two steps to train the CFR-TSVM classifier. First, we train a SVM classifier f to predict the coarse-grained labels z with the input $\mathbf{x}$ in the supervised learning setting [2]. Second, we train

¹S: signal with length K ; S_i : the i th value of the signal; P : power spectral density of S , whose length is N ; P_i : the i th element of P ; $e = \sum_{i=1}^N P_i$.

Figure 3: Experimental deployments.**Table 2: Coarse- and fine-grained activity labels**

Coarse-grained Label	Fine-grained Label	Fine-grained Label ID
Working	Typing	1
	Hand Writing	2
	Reading Book	3
Cooking	Cutting	4
	Stir-frying	5
	Washing a Fruit	6
Cleaning	Using the vacuum	7
	Sweeping	8
	Wiping table	9
	None Specific	10

another TSVM classifier g' to predict the fine-grained labels y . Similar to the training of g , we train g' by minimizing the loss function defined in Equation 1. Instead of using $\mathbf{x}$ as the input feature vector, we use $[\tilde{f}(\mathbf{x}), \mathbf{x}]$, where $\tilde{f}(\mathbf{x})$ is the one-hot encoding of $f(\mathbf{x})$ and $[\tilde{f}(\mathbf{x}), \mathbf{x}]$ is the concatenation of $\tilde{f}(\mathbf{x})$ and $\mathbf{x}$.

3 EVALUATION

Our proposed system is evaluated with real-world experiment deployments and multi-modal sensing data.

3.1 Data Collection Setup

We conducted a real-world data collection to evaluate the system. Figure 2 demonstrates the hardware settings of the system. A wearable sensing device that consists of an Arduino Feather M0 board and an IMU ADXL337 module is placed on the wrist of the user's dominant hand. Structural vibration sensing devices that consist of an Arduino zero board with an LMV385 OpAmp and a geophone SM-24 sensor are placed on the floor and the table/counter top.

As shown in Figure 3, we deployed one vibration sensor on the floor (floor sensor) and one vibration sensor on the table (table sensor). The sampling rate of the vibration sensor is 6500 Hz. The participant also wears an IMU model on the wrist of their dominant hand. The sampling rate of the IMU model is 800 Hz. We record

Proposed System, Accuracy: 80.83%

Output Class \ Target Class	Typing	Writing	Reading	Cutting	Stir-frying	Washing Fruit	Using Vacuum	Sweeping	Wiping the table	None
Typing	80.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.0%
Writing	16	1	0	0	0	0	0	0	0	3
Reading	5.0%	85.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.3%
Cutting	1	17	0	0	0	0	0	0	0	2
Stir-frying	10.0%	10.0%	65.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.0%	6.7%
Washing Fruit	2	2	13	0	0	0	0	0	1	4
Using Vacuum	0	0	0	95.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Sweeping	0	0	0	19	0	0	0	0	0	0
Wiping the table	0	0	0	0	75.0%	0.0%	0.0%	0.0%	0.0%	0.0%
None	0	0	0	0	15	65.0%	0.0%	0.0%	0.0%	0.0%
Typing	5.0%	0.0%	0.0%	0.0%	2	13	100.0%	5.0%	0.0%	1.7%
Writing	1	0	0	0	0	0	20	1	0	1
Reading	0	0	0	0	0	0	0	90.0%	0.0%	3.3%
Cutting	0	0	0	0	0	0	0	18	0	2
Stir-frying	0	0	0	0	0	0	0	80.0%	1.7%	0
Washing Fruit	0	0	0	0	0	0	0	16	1	0
Using Vacuum	0	0	0	0	0	0	0	15.0%	78.3%	0
Sweeping	0	0	0	0	0	0	0	7	0	0
Wiping the table	0	0	0	0	0	0	0	0	0	0
None	0	0	0	0	0	0	0	0	0	0

(a) Semi-supervised coarse-to-fine relationship learning: 120 fine-grained label, 1612 coarse-grained label, 240 unlabeled.

Baseline, Accuracy: 65.42%

Output Class \ Target Class	Typing	Writing	Reading	Cutting	Stir-frying	Washing Fruit	Using Vacuum	Sweeping	Wiping the table	None
Typing	70.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.7%
Writing	14	0	0	0	0	0	0	0	0	1
Reading	10.0%	70.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.7%
Cutting	2	14	0	0	0	0	0	0	0	4
Stir-frying	0	0	0	0	0	0	0	0	0	0
Washing Fruit	0	0	0	0	0	0	0	0	0	0
Using Vacuum	0	0	0	95.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Sweeping	0	0	0	19	0	0	0	0	0	0
Wiping the table	0	0	0	0	75.0%	0.0%	0.0%	0.0%	0.0%	0.0%
None	0	0	0	0	15	20.0%	100.0%	25.0%	0.0%	8.3%
Typing	5.0%	0.0%	0.0%	0.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Writing	1	0	0	0	1	4	20	5	0	5
Reading	0	0	0	0	0	0	0	70.0%	0.0%	0.0%
Cutting	0	0	0	0	0	0	0	14	0	0
Stir-frying	0	0	0	0	0	0	0	0	80.0%	0.0%
Washing Fruit	0	0	0	0	0	0	0	0	16	0
Using Vacuum	0	0	0	0	0	0	0	0	0	0
Sweeping	0	0	0	0	0	0	0	0	0	0
Wiping the table	15.0%	30.0%	10.0%	0.0%	0.0%	0.0%	0.0%	0.0%	15.0%	13.3%
None	3	6	2	0	0	0	0	0	3	8
Typing	0	0	0	0	0	0	0	0	0	0
Writing	0	0	0	0	0	0	0	0	0	0
Reading	0	0	0	0	0	0	0	0	0	0
Cutting	0	0	0	0	0	0	0	0	0	0
Stir-frying	0	0	0	0	0	0	0	0	0	0
Washing Fruit	0	0	0	0	0	0	0	0	0	0
Using Vacuum	0	0	0	0	0	0	0	0	0	0
Sweeping	0	0	0	0	0	0	0	0	0	0
Wiping the table	0	0	0	0	0	0	0	0	0	0
None	0	0	0	0	0	0	0	0	0	0

(b) Baseline TSVM: 120 fine-grained label and 240 unlabeled.

Figure 4: Fine-grained activity recognition confusion matrix for our method and the semi-supervised learning baseline.

the vibration data and the IMU data at the same time, and the IMU data and the vibration data both contain the timestamp data. We achieve microsecond-level data synchronization between the vibration sensor and the IMU sensor via Network Time Protocol (NTP)[13]. The participant repeats each fine-grained activity last at least 15 seconds and repeated 15 times. Between two fine-grained activities, there are 5 seconds with no specific activities.

3.2 Coarse- and Fine-grained Labeling

In this section, we introduce how the coarse- and fine-grained label is assigned to the investigating sliding window signals. Table 2 shows the target nine common in-home fine-grained activities as our recognition interests. These nine fine-grained activities fall into three coarse-grained activities of daily living, which are often used for the elderly's ability of independent living evaluation. However, for the signal segments labeled in a coarse-grained manner but are not associated with any fine-grained activity, we consider them the tenth label of 'None Specific.'

The camera is used to collect the ground truth. We use the frame-level timestamp start and stop of an activity to label the data. Given the signal between the time window $[t_{start}, t_{end}]$, we label a coarse-grained activity with the ground-truth from the camera. Within the window $[t_{start}, t_{end}]$, we further label N fine-grained activities. Specifically, in this window, we identify N non-overlapping sub-windows $[t_{start}^i, t_{end}^i]$ with fine-grained activities, where $1 \leq i \leq N$. The signal within each $[t_{start}^i, t_{end}^i]$ is labeled with a fine-grained activity.

3.3 Evaluation Experiments

For the aforementioned dataset, after applying sliding windows, we have 1972 windows. We did the cross validation, where in each iteration we sample 240 target unlabeled windows, We consider the rest of the 1612 window with coarse-grained labels and among

them 120 are with fine-grained labels. Our system utilizes a support vector machine (SVM) to conduct the coarse-grained classification for the unlabeled target data [2]. Since we have multiple classes, we adopt the one-against-one strategy to conduct multi-class classification[3].

To understand the impact of the multi-modal design as well as the efficiency of the semi-supervised coarse-to-fine relationship learning, we consider the following baselines.

- (1) Baseline 1: only use semi-supervised learning TSVM trained with 120 fine-grained labeled windows and tested on the 240 unlabeled windows.
- (2) Baseline 2: use supervised learning SVM trained with 1732 fine-grained labeled windows and tested on the 240 unlabeled windows.

Note that Baseline 2 would be significantly (15x) more labor-intensive to acquire the fine-grained labels compared to our methods.

To compare our system's performance to the baseline fairly, we compute the classification accuracy for each fine-grained class for assessment. For evaluating the system performance, we calculate the average classification accuracy. Both performance metrics are calculated at window-wise level.

3.4 System Performance Analysis

We focus on two aspects of evaluation 1) the efficiency of our coarse-to-fine relationship learning algorithm, and 2) the efficiency of the multi-modal IoT data combination.

3.4.1 Semi-Supervised Coarse-to-fine Relationship Learning. Figure 4 illustrates the fine-grained activity recognition confusion matrix of our proposed system and the baseline. The averaged prediction accuracy and macro F-1 score for baseline is 67.08% and 67.45%. With the proposed framework, the performance can be boosted to an 81.67% averaged accuracy and 83.51% macro F-1 score, which is an approximately 1.8x error reduction.

Figure 5: Fine-grained activity recognition performance, where the x-axis is the sensing modality combination and the y-axis is the learning accuracy. The gray bars represent the accuracy using TSVM only with 6% data (120 windows) as fine-grained labeled training data. The green bars represent the accuracy using our CFR-TSVM with 6% data (120 windows) as fine-grained labeled data and 82% data (1612 windows) as coarse-grained labeled data. The orange bars represent the accuracy using supervised SVM with 88% fine-grained labeled data.

The improvement of performance mainly comes from the accurate coarse-grained prediction. For example, the using vacuum was often confused for the washing fruit when the baseline semi-supervised learning is used alone. This is because the amount of fine-grained labeled data is insufficient to train a robust classifier to make a delicate classification. With accurate coarse-grained prediction, the semi-supervised coarse-to-fine learning captures this hierarchical relationship and hence increased the fine-grained prediction accuracy.

3.4.2 Multi-modal Learning for Fine-grained Activity Recognition.

We further investigate the effectiveness of multiple sensing modalities. Figure 5 depicts the system’s activity recognition accuracy with each type of individual sensor and the multi-modal combination. When only use the vibration based method, the baseline 1 TSVM only approach’s accuracy is 8%. The baseline 2 supervised learning SVM with significantly more fine-grained labeled data achieved 72% accuracy. Our CFR-TSVM achieved 48% accuracy, which is a 6x improvement compared to baseline 1 and only yields 20% accuracy with 15x less fine-grained labels. Similarly, when only use the IMU data, the baseline 1, CFR-TSVM, and the baseline 2 achieved 52%, 71%, and 85% accuracy. When both modalities are used, the baseline 1, CFR-TSVM, and the baseline 2 achieved 67%, 81%, and 94% accuracy. Our multi-modal coarse-to-fine relationship learning approach achieved comparable accuracy to a single-modal fully supervised approach with only 1/15 fine-grained labeled data.

The multi-modal achieved the highest accuracy due to the complementary information in their data stream. For instance, from

the body motion perspective, wiping the table and reading both consist of a similar arm limb’s back and forth motion, of which the nuance difference is challenging to be distinguished by IMU sensors. On the other hand, from the structural vibration perspective, the excitation induced different types of waves – flipping pages induce impulsive signals that are dominated by Rayleigh-lamb waves while wiping induce slip-pulse signal dominated by body waves [17] – which can be distinguished. As a result, the complementary information provided by multiple sensing modalities can effectively benefit fine-grained activities recognition. It is also suggested that the proposed system can effectually leverage the captured spatial-temporal relationship between multiple sensing modalities to facilitate fine-grained activities recognition.

4 RELATED WORK

Learning models on various sensor data, such as vibration, electrical load, accelerometer, and RFID, enables a large number of context-aware applications such as daily life logging, activities of daily living (ADL) monitoring, and human activity recognition [5, 15, 16]. In this work, we focus on fine-grained activity recognition by utilizing rich coarse-grained labels and a limited number of fine-grained labels. We review the related works in these areas, as follows.

4.1 Coarse-to-Fine Classification

For coarse-medium-fine feature extraction and fusion, a coarse-fine convolutional deep-learning strategy is proposed for human activity recognition [1]. The ideas of learning and predicting coarse and fine classes have been explored in computer vision. To obtain a more accurate feature representation, Lin *et al.* [12] introduces a coarse-to-fine network which extracts shared deep features at different action class granularity. Huo *et al.* [10] proposed a Multi-linear Pooling with Hierarchy model for the combinations of fine- and coarse-grained tasks in image classification and retrieval.

Different from these coarse to fine classification tasks, we focus on *developing robust classifiers with very limited IoT sensing data*. Instead of developing deep learning models on rich training data, we develop models that generalize well with limited training data. To achieve high accuracy in the coarse-to-fine activity recognition task, we propose semi-supervised coarse-to-fine relationship learning models that leverage data from multiple modalities and labels of multiple levels granularity for accurate learning.

4.2 Semi-Supervised Activity Recognition

Our task is closely related to the area where we learn models to predict user activities with limited training data. Unlike images or videos that human beings can easily label, strictly labeling long sequences of sensor data with different granularity of activity needs much more human labors [7, 21–23]. To extract information from weakly labeled sensor data by adaptively selecting a sequence of locations, a recurrent attention learning is proposed in [7]. Stikic *et al.* [21] proposes a novel annotation strategies that substantially reduce the required amount of annotation, and explore two learning schemes to effectively leverage such sparsely labeled data together with more easily available unlabeled data. To learn accurate representations with limited labeled data and rich unlabeled data, the

convolutional ladder network is proposed for human activity recognition in [23]. To effectively resist outliers in noisy sensor data and improve recognition performance by discerning underlying commonalities from activities, [22] proposes $l_{2,1}$ minimization on both loss function and regularization.

Traditionally, the machine learning models are usually trained on a small amount of activity data with labels and a large amount of activity data without labels. Instead, in our IoT sensing application setting, the coarse-grained labels are less labor intensive to acquire, but with insufficient information resolution required by target applications, e.g., behavior profiling. Therefore, our paper focuses on a novel setting, where we train the models on activity data with rich coarse-grained labels but only with a small amount of fine-grained labels.

4.3 Multi-modal Activity Recognition

Multi-modal data is widely available in various ubiquitous computing systems and IoT applications. Data from different modalities encode diverse information from different sources. Various approaches have been proposed to fuse knowledge between different data modalities and understand user activities. Pan et al. present fine-grained indoor activities of daily living inference through structural vibration sensing and electrical load sensing, which captures two important aspects of the indoor activities [15]. To capture local dependency over time and spatial dependency over sensors, a multi-modal CNN using 2D kernels in convolutional and pooling layers is proposed in [6]. Münzner et al. [14] explored the possibilities to use a CNN-based multi-modal sensor with specially designed techniques to tackle HAR domain-specific challenges (e.g. handling of multi-modal sensor data, lack of large labeled datasets).

5 CONCLUSION

In this paper, we present a multi-modal coarse-to-fine grained human activity recognition framework. The intuitions are twofold 1) multi-modal sensing data capture complementary information to enable finer-grained activities recognition, and 2) coarse-grained labels are easier to acquire and scale-up compared to fine-grained labels, and its hierarchical relation to fine-grained labels can enhance the fine-grained prediction accuracy. We present a coarse-to-fine grained activity recognition learning framework and a CFR-TSVM algorithm as an example of the framework. Then we conducted real-world experiments to evaluate our framework design and achieved 81% fine-grained activity recognition accuracy with only 6% fine-grained labeled data, which reduces the prediction error by half compared to the semi-supervised learning baseline.

REFERENCES

- [1] Carlos Avilés-Cruz, Andrés Ferreyra-Ramírez, Arturo Zúñiga-López, and Juan Villegas-Cortéz. 2019. Coarse-fine convolutional deep-learning strategy for human activity recognition. *Sensors* 19, 7 (2019), 1556.
- [2] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 1–27.
- [3] Kai-Bo Duan and S Sathiya Keerthi. 2005. Which is the best multiclass SVM method? An empirical study. In *International workshop on multiple classifier systems*. Springer, 278–285.
- [4] Jonathon Fagert, Mostafa Mirshekari, Shijia Pan, Pei Zhang, and Hae Young Noh. 2020. Structural Property Guided Gait Parameter Estimation Using Footstep-Induced Floor Vibrations. In *Dynamics of Civil Structures, Volume 2*. Springer, 191–194.
- [5] Dany Fortin-Simard, Jean-Sébastien Bilodeau, Sébastien Gaboury, Bruno Bouchard, and Abdenour Bouzouane. 2014. Human activity recognition in smart homes: Combining passive rfid and load signatures of electrical devices. In *2014 IEEE symposium on intelligent agents (IA)*. IEEE, 22–29.
- [6] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi. 2015. Multi-modal convolutional neural networks for activity recognition. In *2015 IEEE International conference on systems, man, and cybernetics*. IEEE, 3017–3022.
- [7] Jun He, Qian Zhang, Liqun Wang, and Ling Pei. 2018. Weakly supervised human activity recognition from wearable sensors by recurrent attention learning. *IEEE Sensors Journal* 19, 6 (2018), 2287–2297.
- [8] Chih-Wei Hsu and Chih-Jen Lin. 2002. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks* 13, 2 (2002), 415–425.
- [9] Zhizhang Hu, Emre Sezgin, Simon Lin, Pei Zhang, Hae Young Noh, and Shijia Pan. 2019. Device-free Sleep Stage Recognition through Bed Frame Vibration Sensing. In *Proceedings of the 1st ACM International Workshop on Device-Free Human Sensing*. 39–43.
- [10] Yuqi Huo, Yao Lu, Yulei Niu, Zhiwu Lu, and Ji-Rong Wen. 2019. Coarse-to-Fine Grained Classification. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1033–1036.
- [11] Thorsten Joachims. 2003. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 290–297.
- [12] Weiyao Lin, Chongyang Zhang, Ke Lu, Bin Sheng, Jianxin Wu, Bingbing Ni, Xin Liu, and Hongkai Xiong. 2018. Action recognition with coarse-to-fine deep feature integration and asynchronous fusion. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [13] David L Mills. 1991. Internet time synchronization: the network time protocol. *IEEE Transactions on communications* 39, 10 (1991), 1482–1493.
- [14] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelwagen, and Robert Dürichen. 2017. CNN-Based Sensor Fusion Techniques for Multimodal Human Activity Recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers (Maui, Hawaii) (ISWC '17)*. Association for Computing Machinery, New York, NY, USA, 158–165. <https://doi.org/10.1145/3123021.3123046>
- [15] Shijia Pan, Mario Berges, Juleen Rodakowski, Pei Zhang, and Hae Young Noh. 2019. Fine-grained recognition of activities of daily living through structural vibration and electrical sensing. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 149–158.
- [16] Shijia Pan, Mario Berges, Juleen Rodakowski, Pei Zhang, and Hae Young Noh. 2020. Fine-Grained Activities of Daily Living Data with Structural Vibration and Electrical Load Sensing. <https://doi.org/10.5281/zenodo.3745900> This research was supported in part by Highmark, the National Science Foundation (under grants CMMI-1653550), Intel and Google. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of CMU, UCM, Stanford, UPitts, NSF, or the U.S. Government or any of its agencies.
- [17] Shijia Pan, Mostafa Mirshekari, Jonathon Fagert, Ceferino Gabriel Ramirez, Albert Jin Chung, Chih Chi Hu, John Paul Shen, Pei Zhang, and Hae Young Noh. 2018. Characterizing human activity induced impulse and slip-pulse excitations through structural vibration. *Journal of Sound and Vibration* 414 (2018), 61–80.
- [18] Shijia Pan, Tong Yu, Mostafa Mirshekari, Jonathon Fagert, Amelie Bonde, Ole J Mengshoel, Hae Young Noh, and Pei Zhang. 2017. Footprint: Indoor pedestrian identification through ambient structural vibration sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–31.
- [19] Paola Pierleoni, Alberto Belli, Lorenzo Maurizi, Lorenzo Palma, Luca Pernini, Michele Paniccia, and Simone Valenti. 2016. A wearable fall detector for elderly people based on ahrs and barometric sensor. *IEEE sensors journal* 16, 17 (2016), 6733–6744.
- [20] Parisa Rashidi and Diane J Cook. 2009. Keeping the resident in the loop: Adapting the smart home to the user. *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans* 39, 5 (2009), 949–959.
- [21] Maja Stikic, Diane Larlus, Sandra Ebert, and Bernt Schiele. 2011. Weakly supervised recognition of daily life activities with wearable sensors. *IEEE transactions on pattern analysis and machine intelligence* 33, 12 (2011), 2521–2537.
- [22] Lina Yao, Feiping Nie, Quan Z Sheng, Tao Gu, Xue Li, and Sen Wang. 2016. Learning from less for better: semi-supervised activity recognition via shared structure discovery. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 13–24.
- [23] Ming Zeng, Tong Yu, Xiao Wang, Le T Nguyen, Ole J Mengshoel, and Ian Lane. 2017. Semi-supervised convolutional neural networks for human activity recognition. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 522–529.