CMA: Cross-Modal Association Between Wearable and Structural Vibration Signal Segments for Indoor Occupant Sensing

YUE ZHANG, University of California Merced, USA

ZHIZHANG HU, University of California Merced, USA

URI BERGER, Yale University, Child Study Center, Anxiety and Mood Disorders Program, USA

SHIJIA PAN, University of California Merced, USA

Indoor occupant sensing enables many smart home applications, and various sensing systems have been explored. Based on their installation requirements, we consider two categories of sensors – on- and off-body – and we look into the combination of them for occupant sensing due to their spatial and temporal complementarity. We focus on an example modality pair of wearable IMU and structural vibration that demonstrate modality complementarity in prior work. However, current efforts are built upon the assumption that the knowledge of the signal segments from two modalities are known, which is challenged in a multiple occupants co-living scenario. Therefore, establishing accurate cross-modal signal segment associations is essential to ensure that a correct complementary relationship is learned.

We present *CMA*, a cross-modal signal segment association scheme between structural vibration and wearable sensors. We propose *AD-TCN*, a framework built upon a temporal convolutional network that calculates the amount of shared context between an structural vibration sensor and associated wearable sensor candidates from the parameters of the trained model. We evaluate *CMA* via a public multimodal dataset for systematic evaluation, and we collect a continuous uncontrolled dataset for robustness evaluation. *CMA* achieves up to 37% AUC value, 53% F1 score, and 43% accuracy improvement compared to baselines.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Cross-modal Association; Multimodal Sensing; Human Sensing

ACM Reference Format:
Yue Zhang, Zhizhang Hu, Uri Berger, and Shijia Pan. 2023. CMA: Cross-Modal Association Between Wearable and Structural Vibration Signal Segments for Indoor Occupant Sensing. In *The 22nd International Conference on Information Processing in Sensor Networks (IPSN '23), May 9–12, 2023, San Antonio, TX, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3583120.3586960

1 INTRODUCTION

Indoor occupant sensing enables many smart home applications, such as elderly care, building management, and personalized service. Various sensing modalities have been explored, and these systems fall into two categories based on whether it requires the occupant to carry extra devices: on-body and off-body sensing. Fusing on-body and off-body sensing is prevalent in indoor occupant sensing, given multimodal signals can provide complementary information for the same target, and therefore achieve robust information inference [7, 9, 12, 22, 36]. Among these combinations, wearable and structural vibration sensing have demonstrated efficient complementarity for various inference tasks [16, 22]. However, when the size of these IoT systems increases, they may sense multiple physical activities occurring

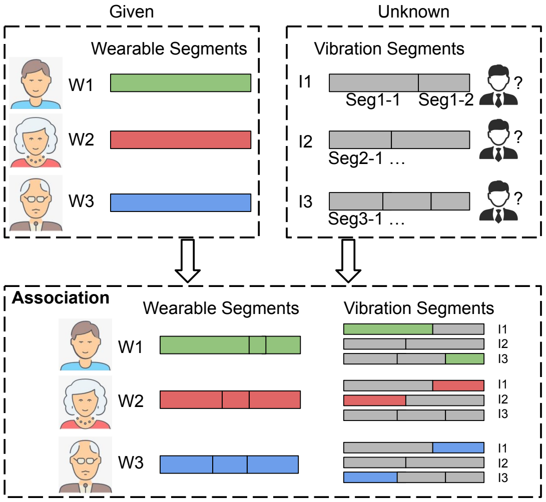


Fig. 1. Cross-modal signal segment association problem: given a set of segments from two sensing modalities collected during the same period, we aim to identify segment pairs that are associated. We refer to the segment pair that contains signals induced by the same physical activity as 'associated'.

at the same time. For example, for an IoT system deployed over different areas in a house, they may sense people doing different activities in different areas. It also means that for any pair of cross-modal sensors, the physical activity they are sensing may or may not be the same. If signal segments of two sensing modalities that capture different activities are used for inference, a spurious complementary relationship will be used. Therefore, it is of great importance to establish correct association relationships for signal segments from co-located sensors of different modalities.

This cross-modal association relationship is beneficial for multiple use cases: *1) User signal segment annotation.* When wearable and structural vibration sensors are used together, with this signal segment level association, the wearable sensors can be used as the identity annotation tool for the structural vibration sensors' signal segments, since the wearable is by nature associated with their user already. This could further advance the structural vibration sensing-based IoT system's usability and scalability as a zero-effort bootstrapping user annotation scheme. *2) Enhancing multimodal learning efficiency.* With a high-accuracy signal segment association, multimodal learning would be able to leverage this prior knowledge to achieve more accurate modeling, since falsely associated signal pairs may result in the spurious complementary relationship being modeled.

Therefore, we formulate this *cross-modal signal segment association* problem between wearable and structural vibration sensors [46–48] as illustrated in Figure 1. Given a set of segments from two co-located sensing modalities collected during the same period (e.g., Seg1-1,...Seg3-1), our goal is to learn a segment-level association cross modalities (e.g., Seg1-1: P1-I1, Seg2-1: P2-I1). However, this cross-modal signal segments association has the following **challenges**: *1) Indirect sensing leads to the lack of direct comparable information.* For indirect sensing systems of structural vibration and IMU, their raw measurements are often not directly interpretable, and therefore, can not be easily compared for shared context (signal examples in Figure 5 later). *2) Complementary leads to disassociation.* IoT systems that adopt multiple modalities often leverage their complementarity to achieve more efficient modeling. On the other hand, the more complementary the two modalities are, the less shared information they capture, and hence their signal segments are more difficult to be associated with. For example, prior work that conducts location association between the electric load sensor and microphone [16] requires longer measurements than that of the camera and IMU [37], because the latter leverages a clear shared context of acceleration. *3) Mobility variance leads to spatiotemporal variation.* For modalities with different levels of mobility, this association may vary over time. For example, occupants who carry an on-body sensor may move in the house and are captured by different off-body sensors. Therefore, this association relationship varies over time due to occupants' mobility. We form our **research question** as How do we learn the segmentation-level association relationship between wearable and structural vibration sensors with constrained shared context and without labeled data?

In this paper, we present *CMA*, a cross-modal signal segment association scheme between wearable and structural vibration sensors. To determine whether two signal segments from different modalities over the same period are associated, we calculate an *association probability* (AP). The intuitions to calculate this association probability are twofold: 1) as long as the sensors are capturing the same physical activity, there will be an implicit shared context between two signal segments, and 2) we assume that for the structural vibration signals that are segmented as one activity (e.g., five seconds), there will be only one wearable sensor associated to it. The temporal convolutional network (TCN) has shown efficient learning ability for the temporal representation features from time series signals [25]. We propose *AD-TCN*, a framework built upon TCN to calculate the amount of shared context between signal segments from different modalities. First, *AD-TCN* takes all candidate wearable segments and the vibration segment history values to predict the vibration segment's current time step value. Then we train the model and calculate the association probability between signal segments from two modalities based on the weights of the trained *AD-TCN*. The association probability reflects the contribution of one signal segment for predicting the other. If the contribution of a signal segment is higher than a threshold, we consider this wearable signal segment is associated with the vibration signal segment, i.e., they detect the same physical activity. In summary, the contributions of this work are as follows:

- We introduce *CMA*, a cross-modal sensing signals' segment-level association scheme for multimodal IoT systems.
- We present *AD-TCN* that learns the segment-level cross-model representation and uses the learned model parameters to calculate the amount of shared context between modalities.
- We evaluate *CMA* through both a public dataset and an uncontrolled real-world dataset for robustness analysis.

The rest of the paper is organized as follows. First, we investigate and compare *CMA* to related work in Section 2. Then, we illustrate the details of *CMA* in Section 3. Next, we demonstrate the experiments and analysis in Section 4 and Section 5. Finally, we discuss future directions in Section 6 and conclude the paper in Section 7.

2 RELATED WORK

We investigate prior research on device pairing/identification, and occupant identification, and discuss the research gap that we focused on in this paper.

2.1 Cross-modal IoT Device Identification

Cross-modal IoT device pairing/identification is a relevant topic to cross-modal signal segment association. Prior work on cross-modal pairing relies on the shared context that can be sensed by both sensing modalities and compare the similarity of the acquired shared context to achieve the paring or identification. Ruiz et. al. leverages the shared 3D motion (spatial context) of human body parts captured by both camera and IMU sensor to achieve IoT device identification[32, 37]. Han et. al. utilize the shared context of activity start/end time (temporal context) to generate fingerprints for co-located device pairing [16]. However, these prior works do not directly apply to our target scenario

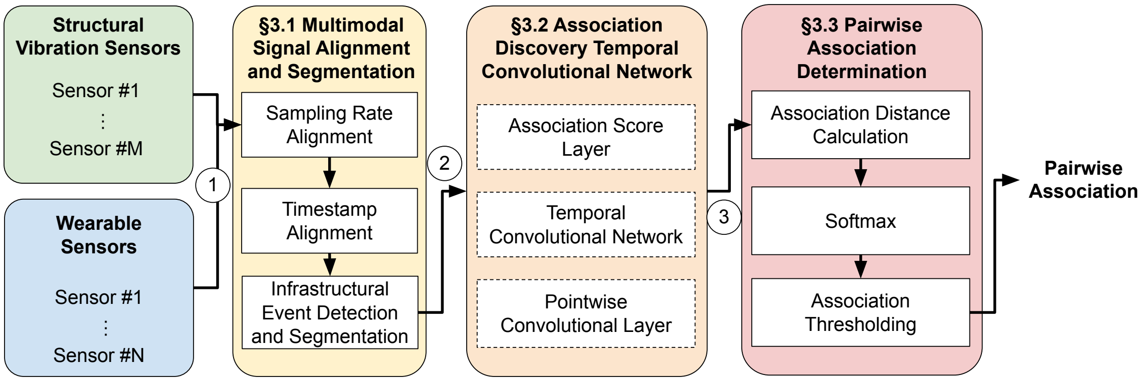

Structural Vibration Sensors
Sensor #1
⋮
Sensor #M
Wearable Sensors
Sensor #1
⋮
Sensor #N
1
§3.1 Multimodal Signal Alignment and Segmentation
Sampling Rate Alignment
Timestamp Alignment
Infrastructural Event Detection and Segmentation
2
§3.2 Association Discovery Temporal Convolutional Network
Association Score Layer
Temporal Convolutional Network
Pointwise Convolutional Layer
3
§3.3 Pairwise Association Determination
Association Distance Calculation
Softmax
Association Thresholding
Pairwise Association

Fig. 2. System overview. *CMA* consists of three modules to estimate the association relationship between the structural vibration and wearable sensors: 1) multimodal signal alignment, 2) Association Discovery Temporal Convolutional Network (*AD-TCN*), and 3) association probability estimation.

due to the challenges from *constrained shared context*. *CMA* solves these challenges by using the temporal convolutional network to efficiently discover the limited association information without an explicitly shared context.

2.2 IoT for Occupant Identification

The fundamental problem solved by this paper is to associate the infrastructure sensor signals with the person who induces it, which is also relevant to the sensor signal-based identification problem. Prior work on occupant identification has explored the possibility to identify the person based on how their behavior or interaction with the environment varies [17]. A more specific description of human behavior is the walking pattern or gait, which can be observed by a wide range of sensors [33, 42, 45]. Other biometrics are also explored to enable ubiquitous occupant identification in the smart home setting such as voice [27], human body's reflection, refraction, diffraction, and even absorption of radio signals [44]. However, all the identification systems require the occupant identity label to create the corresponding classifier model to achieve the identification. In our scenario, it is difficult and impractical to assume the availability of labeled data for each deployment. Instead, we leverage the wearable sensor and their nature association with individuals who wear them to 'label' the identity of the infrastructure sensing segment as a signal association problem.

3 *CMA* DESIGN

We present *CMA*, a cross-modal signal segment association scheme. Figure 2 describes *CMA*'s architecture, which consists of three modules. First, *CMA* aligns signals ① from all sensors by aligning their timestamp and sampling rate, so that these signals are comparable temporally (Section 3.1). Then a threshold-based event detection algorithm is applied to detect the valid events from the structural vibration data, and the timestamp of the structural vibration events are utilized to segment the wearable IMU data. The segmented multimodal events ② are then sent to the Association Discovery Temporal Convolutional Network (*AD-TCN*), where for each structural vibration sensor, a *AD-TCN* is trained and the weight values of the association score layer ③ are output (Section 3.2). Finally, *CMA* calculates the pairwise **association probability (AP)** between each structural vibration sensor and each wearable (Section 3.3). We consider the pair of the wearable and the structural vibration sensor with the association probability higher than a threshold is associated (i.e., they detect the same occupant).

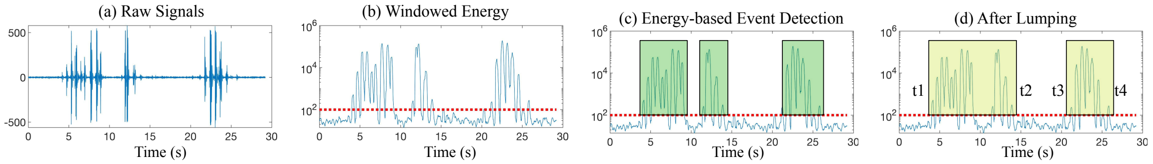


Fig. 3. Structural vibration event detection and activity segmentation signal examples. (a) depicts examples of raw signals of human-induced structural vibration. (b) shows the signal energy of the sliding window applied to the signal in (a). (c) we conduct an energy-based event detection on the windowed signal energy, where the detected events are marked by green boxes. (d) finally, events with intervals lower than a pre-selected threshold are lumped [16] as one activity, which is marked by the yellow box. For example, the segment from t1 to t2 contains signals of one activity.

3.1 Multimodal Signal Alignment and Segmentation

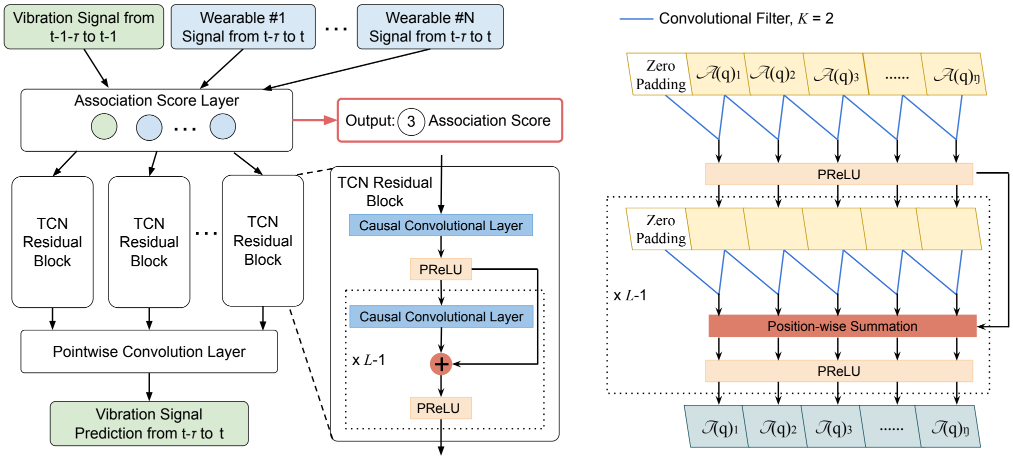
Due to the heterogeneity of the two sensing modalities, *CMA* first preprocesses the incoming signals by aligning and segmenting the signal of interest. Since different types of sensors are sampled at different rates, the number of samples in the same event duration may vary. Furthermore, since we utilize TCN architecture for association discovery (Section 3.2), the architecture takes the same length of time series data points as input and outputs. Therefore, it is important to ensure that all the sensor inputs have the same number of samples in each second, and samples over all the sensor inputs are temporally aligned (Section 3.1.1). In addition, since in our application scenarios, the wearable sensors are directly associated with the user identities and the structural vibration sensor signals need to be associated with the user identities, *CMA* only conducts association when there is vibration signal detected (Section 3.1.2).

3.1.1 Sampling Rate and Timestamp Alignment. To ensure accurate multimodal temporal information modeling, we first align the sampling rate over all the sensor inputs. We select the lowest sampling rate Q (reference) of all available sensors as the reference. Then we conduct resampling [39] on each of the other sensor inputs. Using a signal with an original sampling rate of P Hz as an example ($P \geq Q$, and P, $Q \in \mathbb{N}^{+}$). To resample the signal, first, the least common multiple (LCM) of P and Q is calculated. Then the linear interpolation is conducted to up-sampling the P Hz sampling rate data to LCM Hz. Next, a low-pass filter is applied to remove the higher frequency ($>P$) components in the up-sampling series. Finally, the up-sampling series is down-sampled to Q Hz [28].

Since the TCN leverages the temporal relationship between historical samples and current samples to establish models, it is important to have samples from all sensors time-aligned. Therefore, based on the periodically provided timestamp, *CMA* interpolates the timestamp for each sample for high-resolution alignment.

3.1.2 Structural Vibration Event Detection and Activity Segmentation. To detect the event of interest to conduct the temporal association on, we further conduct a threshold-based event detection algorithm on the vibration data. We first apply a sliding window on the time sequence data of the vibration sensor and calculate the energy of the windowed signal (Figure 3(b)). We characterize the ambient noise's windowed signal energy as Gaussian noise (μ_n, σ_n) [33]. Then, we select a lower bound θ_e as the energy threshold of the windowed signal. If the energy of this windowed signal is larger than ($\mu_n + \theta_e * \sigma_n$), we consider this window is an **event** (Figure 3(c)).

Next, we conduct activity segmentation with an interval-based lumping method [16], where we segment the consecutive events that are less than the event interval threshold $\Delta\tau$ as one **activity segment (AS)** (Figure 3(d)). We segment the aligned IMU data consistently with the structural vibration sensor segments. The activity segment's start and end time is data-driven, therefore it does not have semantic meanings, and one segment maybe two people's


Vibration Signal from t-1-τ to t-1
Wearable #1 Signal from t-τ to t
...
Wearable #N Signal from t-τ to t
Association Score Layer
Output: 3 Association Score
TCN Residual Block
TCN Residual Block
TCN Residual Block
Pointwise Convolution Layer
Vibration Signal Prediction from t-τ to t
TCN Residual Block
Causal Convolutional Layer
PReLU
Causal Convolutional Layer
x L-1
PReLU
Convolutional Filter, K = 2
Zero Padding
𝒜(q)1
𝒜(q)2
𝒜(q)3
......
𝒜(q)η
PReLU
Zero Padding
x L-1
Position-wise Summation
PReLU
𝒯(q)1
𝒯(q)2
𝒯(q)3
......
𝒯(q)η

(a) Network Architecture

(b) Causal Convolution

Fig. 4. The architecture of the association discovery temporal convolutional network (*AD-TCN*). The network consists of the association score layer, the TCN residual block, and the pointwise convolution layer. The model is trained over multiple epochs, and the association score layers' node weights are the output of the *AD-TCN* model, marked as ③ in Figure 2.

activity occurring consecutively within $\Delta\tau$. To ensure efficient association, we further segment the activity segments into **association units (AU)** to unify the association signal with lower and upper bounds, τ_l and τ_u. We assume the association of the signals does not change within an AU. For a segmented AS, if the duration is shorter than τ_l, *CMA* discards it because there is not enough information to perform the association. On the other hand, if the duration is longer than τ_u, we will divide the AS into multiple AUs by the duration $\leq \tau_u$, and discards ones with a duration $< \tau_l$. The aligned AUs from two modalities are inputs for the next module.

3.2 Association Discovery Temporal Convolutional Network (*AD-TCN*)

In our cross-modal association problem, the wearable IMU measures the occupant's motion, which causes the structure to vibrate. Inspired by the prior work that utilizes the TCN architecture to infer Granger causality [30], we model the cross-modal signal association problem as a time series prediction problem and quantify the contribution of one segment (X) on the prediction of another segment (Y) as an indicator of such association relationship. In our model, for an AU of duration τ at time step t, we consider X is the raw signal of the wearable sensor between $t-\tau$ and t, and Y is the raw signal of the structural vibration sensor between $t-\tau$ and $t-1$. If X's past value at $t-\tau$ to t contributes to predict Y at t, then X and Y are associated with an association probability proportional to this contribution.

We present *AD-TCN*, an association discovery network built upon the TCN architecture to infer causal relationship [30] between pairs of multimodal sensing signals. Figure 4(a) shows the overview of the *AD-TCN*. The network has three parts, namely the association score layer, TCN residual block, and point-wise convolution layer. The network takes aligned AU of duration τ with $\eta = \tau \times Q$ samples from index η_0 as inputs. For wearable sensor signals, the input is signal indexing between η_0 and $\eta_0 + \eta$. For structural vibration sensor signals, the input ranges from $\eta_0 - 1$ to $\eta_0 - 1 + \eta$. The prediction output is the structural vibration sensor signal indexing from η_0 to $\eta_0 + \eta$. For each structural vibration sensor's AUs and n available wearable sensors' signal, an *AD-TCN* network is trained independently to estimate the association relationship.

3.2.1 Association Score Layer. We introduce a trainable association score layer to measure the weight put on each channel of sensor signals by the network. Figure 4(a) shows the architecture of the association score layer and its inputs and outputs. For a multimodal sensing system with M wearable sensors (the i^{th} sensor has C_i channels), the association score layer contains $h = 1 + \sum_{i=1}^{M} C_i$ nodes (shown as circles in Figure 4(a)), each contains a weight value. In the beginning, all nodes are initialized with the same weight value, i.e., each input equally contributes to structural vibration signal prediction. These weights are updated during model training by the gradient descent algorithm [35]. The association score is calculated from the weight via the softmax function as the layer's activation function. When the model training is finished, the final association score outputs to the Association Probability Estimation module (③ in Figure 2). A high association score indicates that this node's input has more contribution to predicting the structural vibration signal, and the input signal of this node is more likely associated with the structural vibration signal. On the other hand, during model training, the association scores are multiplied with their corresponding input signal as the output of the layer. For input multimodal signal segments with length $\eta = \tau \times Q$, the output of the association score layer is shown as follows:

$$\begin{aligned} \mathcal{A}(q) &= \alpha_q \cdot SE_q = \frac{\exp^{W_q}}{\sum_{j=1}^{h} \exp^{W_j}} \cdot SE_q \\ \mathcal{A}(q) &\in \mathbb{R}^{\eta \times 1}, q \in [1, h] \end{aligned} \tag{1}$$

Where $SE_q \in \mathbb{R}^{\eta \times 1}$ is the q th input, α_q and W_q are the association score and the weight of q th node, respectively.

3.2.2 Temporal Convolutional Network Residual Block. We adopt the temporal convolutional network (TCN) residual block [4] for its strong performance in time-series prediction. Transitional TCN is designed for univariate time-series prediction, i.e., predicting with one time-series data. However, *CMA* models the association problem as a time-series prediction problem with multiple time-series data inputs, i.e., multivariate time-series prediction. To adapt to the multivariate time-series prediction, we utilize a depthwise separable architecture to extend the univariate TCN architecture for multivariate prediction [8]. That is, for output from the association score layer's each node, they are separately sent to different TCN residual blocks, as shown in Figure 4(a). In total, there are h independent TCN residual blocks. Each block has the same architecture: L layers of 1-D causal convolutional network layers [41]. These layers have the same kernel size K. Figure 4(b) illustrates the mechanism of 1-D causal convolution in the TCN residual block. The 'causal' in this layer architecture name means that the prediction of time t data is generated only with data from time t and earlier. For instance, to predict $\mathcal{T}(q)_3$, only the data no late than $\mathcal{A}(q)_3$ is used. In this way, no future information is used in prediction i.e., no information leakage. Each causal convolutional layer has the same length (η) as the input time-series signal. Since only the history data can be used for prediction, in order to keep subsequent layers the same length as the first layer, a left zero-padding of size $K - 1$ is added. After each causal convectional layer, we adopt a Parametric Rectified Linear Unit (PReLU) [18] as the non-linear activation function, for its empirical strong performance on improving model fitting capability. A residual connection [19] is added before each PReLU activation result in the block, except the first one. The residual connection conducts a position-wise summation of the previous and current layers' results. This allows the block to learn modifications on the block input rather than the entire transformation, which has been shown to benefit scaling the network to very deep [4]. The set of calculations of the q th block can be described as follows:

$$
\begin{aligned}
\mathcal{T}_1(q) &= PReLU(G_q^1 * \mathcal{A}(q) + b_q^1)\\
\mathcal{T}_l(q) &= PReLU(G_q^l * \mathcal{T}_{l-1}(q) + b_q^l) + \mathcal{T}_{l-1}(q)\\
\mathcal{T}_1(q), \mathcal{T}_l(q) &\in \mathbb{R}^{\eta \times 1},\ l \in\ [2, L]
\end{aligned}
\tag{2}
$$

Where $\mathcal{T}_1(q)$ and $\mathcal{T}_l(q)$ are the output of the first layer and l th layer, $G_q^1, G_q^l \in \mathbb{R}^{K \text{x} 1}$ are weights of the convolution filters in the first layer and l th layer, and $b_q^1, b_q^l \in \mathbb{R}$ are bias terms of each layer. K is the kernel size of the convolution filter. $*$ denotes the convolution operator.

Receptive field is a term that describes how much history data is utilized in the prediction, and it has been proved the size of the receptive field has an impact on the prediction accuracy [41]. There are two hyper-parameters in the TCN residual block that jointly determine the receptive field size: L, number of causal convolutional layers; and K, kernel size of the 1-D convolution filter [4]. Additionally, we can achieve the same receptive field using a different composition of K and L, but the properties of the network may impact performance. For instance, a large L may make model training more difficult and cause overfitting [4]. The evaluation of receptive field size F and hyper-parameters setting (K and L) on the system performance is shown in Section 5.3.

3.2.3 Pointwise Convolution Layer. We apply a pointwise convolution layer to integrate the output of all h TCN residual blocks as the prediction of the structural vibration segment. The output of the pointwise convolution layer has the same length η of the input time-series signal segments. The calculation of the pointwise layer is as follows:

$$
\begin{aligned}
\hat{\mathcal{I}} &= \sum_{q=1}^{h} p_q \cdot \mathcal{T}_l(q)\\
\hat{\mathcal{I}} &\in \mathbb{R}^{\eta \times 1}
\end{aligned}
\tag{3}
$$

Where $p_q \in \mathbb{R}$ is the weight of the pointwise convolution filter for the qth TCN block output.

3.2.4 Loss Function. We use the mean square error (MSE) as the loss function to measure the difference between the raw vibration sequence (I) and the predicted sequence ($\hat{\mathcal{I}}$). The calculation of MSE is as follows:

$$
\mathcal{L} = \frac{\sum_{r=1}^{\eta}(I(r) - \hat{\mathcal{I}}(r))^2}{\eta}
\tag{4}
$$

Where η is the length of the AU. MSE reflects how similar the predicted sequence $\hat{\mathcal{I}}$ and the ground truth I. The optimization goal is to minimize $\mathcal{L}$ during the model training.

3.3 Pairwise Association Determination

To enable explainable association, *CMA* estimates an AP for each τ seconds multimodal data based on the *AD-TCN* output. The output – association score– is the attention value [3] of the neural network for each input, and cannot represent the association relationship directly. Furthermore, since *AD-TCN* is applied to each structural vibration sensor, the weight values of different *AD-TCN* are not comparable. Therefore, a common representation of the association relationship between the structural vibration and wearable sensors are needed. To do so, we first calculate a 'divergence' between the structural vibration sensor and all the available wearable sensors using the association score. Next, we apply a softmax function to convert the association divergence to the AP between the structural vibration sensors

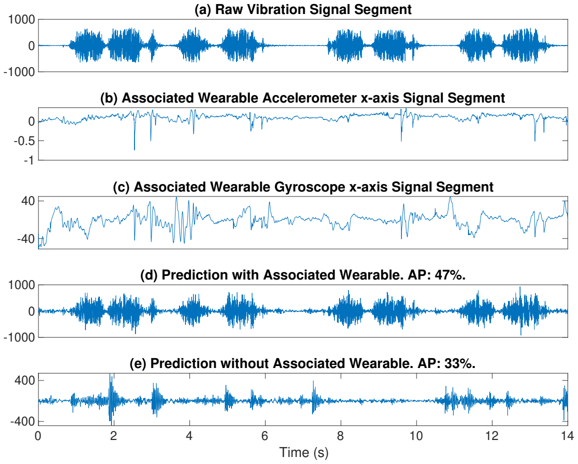


Fig. 5. One example of associated structural vibration (a) and wearable (b, c) signal segments, and the predicted structural vibration segment with (d) and without (e) associated wearable segment. We observe that the structural vibration segment predicted with the associated wearable's signal shows higher similarity to the raw structural vibration signal segments. *CMA* outputs AP of (d) and (e) as 47% and 33%, respectively. This AP difference indicates that the *AD-TCN* learns the implicit shared context between the structural vibration and wearable segments.

and wearable sensors. In this way, we find a common measurement of the association relationship between multiple structural vibration sensors and wearable sensors.

The association divergence measures the association relationship between the structural vibration sensor and wearable sensor. A low association divergence value means the IMU has less contribution on the prediction of the target vibration sensor, i.e., they have a lower probability to be associated. For the wearable sensor q with C channels, *CMA* outputs C values of association score, as a vector $\mathbf{W_q}$. *CMA* integrates the C channels of the association score into a divergence D_q as the square root of Euclidean norm [38] of the vector $\mathbf{W_q}$.

$$D_q = \sqrt{\sum_{i=1}^{C} W_q(i)^2} \tag{5}$$

Note that this D_q alone, or the vector $\mathbf{W_q}$ alone is not comparable to each other, because the association score for each structural vibration sensor are calculated individually by a neural network. Therefore, they cannot be directly compared to a global threshold. To allow explainable and comparable outputs, we further normalize this divergence by softmax [1], and output the *AP* as

$$AP = \frac{\exp^{(D_q)}}{\sum_{i=1}^{N} \exp^{(D_i)}} \tag{6}$$

CMA reports an association if the *AP* value is larger than a threshold θ_{AP}.

Figure 5 shows an example AU of duration $\tau = 14s$ with the structural vibration segment in (a) and wearable segments in (b,c). By directly comparing Figure 5(a) to (b,c), we do not observe a clear association between their waveforms. However, by using our *CMA* with this AU as inputs, the predicted structural vibration segment is shown in Figure 5 (d), which shows a high similarity to (a). On the other hand, if we replace the input of the wearable segment with a signal segment of the same dimension with value 0, i.e., a segment has no information, the predicted segment is as shown in Figure 5(e), which demonstrates a lower similarity to (a). The AP of the associated IMU sensor (#1 47%) is higher than that of the other two IMU sensors (#2 26% and #3 27%). The AP between all zeros sequence (33%) and unassociated IMU sensors (#2 35% and #3 32%) are similar to an even distribution (random guess 33%). Therefore, the association probability can reflect the association for cross-modal signals.

4 EXPERIMENT SETUP

We evaluate *CMA* from two aspects: 1) the association performance and system characterization on the public dataset and our collected uncontrolled dataset. 2) use case study for real application demonstration. We first conduct a set of controlled experiments for system characterization on the public dataset, including hyperparameter configuration, the impact of human activity category, and AP distribution. Then, we evaluate the performance of uncontrolled experiments for robustness verification. Finally, we implement two use cases on the public dataset to demonstrate how to adapt *CMA* in real applications, including occupant identification and multimodal human activity recognition.

In this section, we introduce the two datasets (one open-sourced and one real-world collected), ground truth, evaluation metrics, as well as the implementation of baselines, *CMA*, and two use cases. The experiments are conducted based on the guideline approved by the University Institutional Review Board (IRB) review.

4.1 Datasets Description

Public Dataset. The dataset [21] includes both structural vibration and wearable sensors – floor vibration sensors and on-wrist IMU (6-axis) sensors. The dataset is collected over two buildings with six human subjects with nine types of in-home activities of daily living. The nine types of in-home activities of daily living are keyboard typing, using mouse, handwriting, cutting food, stir-fry, wiping countertop, sweeping floor, vacuuming floor, open/close drawer. For each **scenario**, i.e., one building one human subject conducting nine types of activities, signals from four vibration sensors deployed in the house, and one IMU sensor deployed on the human subject's wrist are collected. Each human subject conducts the same set of activities in each **scenario** for 10 times and each time for approximately 15 seconds. The sampling rates of the vibration sensor and the IMU sensor are 6500 Hz and 235 Hz, respectively. The dataset also contains the ground truth of activity types, and start and end timestamps.

Continuous Uncontrolled Dataset. We adopt the same types of sensors, and sampling rate as the public dataset [21] and collect the continuous uncontrolled datasets over five houses. We recruit 11 human subjects in total, and maintain three subjects per house for the data collection. In each house, we deploy three vibration sensors on the surface of the furniture (desk, kitchen bar, etc.) to capture the subject induced vibration signals, including the kitchen area, living area, and dining area. Considering there are ~ 2.5 people per household on average in the United States in 2021 [6], we invite three participants to cohabit in each house, and each participant wears an IMU sensor on their wrist. We collect the six-axes IMU data (three-axes accelerometer and three-axes gyroscope) from three participants simultaneously. The duration of data collection in each house is around one hour. The participant conducts their daily activities in each area: cooking in the kitchen area, eating in the dining area, and watching TV or surfing on the Internet with a laptop in the living area. To reflect the diversity of participants' activities, the participant can do any activity in each area as natural as possible. For example, the subject can cook any food they like; some subjects cook potatoes, some cook sandwiches. In practice, the sampling rate of the vibration sensor and the IMU sensor are around 4000 Hz and 250 Hz, respectively. We also deploy a camera in each area to record which participant is active in this area.

4.2 Ground Truth of Pairwise Association and Dataset Preparation

The cross-modal association problem is described as determining if the signals from two sensing modalities for a given period are induced by the same physical event, which is the individual activity in our case. For an AU, the ground truth of the association between the vibration signal and the IMU signal is true if and only if the vibration signal is induced by the individual wearing the IMU.

Public Dataset. To utilize this dataset for evaluating *CMA* on the task of cross-modal association, we generate association ground truth based on the provided original activity ground truth. We first detect and segment each activity event based on the provided start and end timestamp of each activity event. For each activity segment with signals from four vibration sensors and one IMU sensor, we select the vibration sensor with the highest signal-to-noise ratio (SNR) as the signal associated with the corresponding IMU sensor. We go through the entire dataset and generate 1048 pairs of the cross-modal association data segments (each ~10s). For any two cross-modal segments $VibSig_i$ and $IMUSig_j$, the association labe is true if $i = j$, otherwire is false.

For each **trial**, we randomly select N of segment pairs from the candidate set (it can be the full set with 1048 pairs or a subset). We apply *CMA* on each $VibSig$ with all the $IMUSig_{1,...,N}$ and output N APs between the $VibSig$ and N $IMUSig$. To reflect the practical scenario of a home with parents and children, we set the default value for N as 3. For each **experiment**, we repeat this **trial** at least 100 times to reduce the random selection bias.

Continuous Uncontrolled Dataset. For the continuous uncontrolled dataset, we first apply the event detection and activity segmentation (introduced in Section 3.1.2) on each vibration sensor. The vibration segment and other segmented IMU segments combine a AU. We determine the association ground truth of this AU by watching the recorded video in the vibration sensor deployed area, and we consider the human subject who appears in this area during this event period as the inducer of this event. For each experiment, we use all detected AUs in one house to evaluate the performance of *CMA* in real-world experiments and evaluate the robustness of *CMA* by comparing the performance variation in different houses.

4.3 Evaluation Metric

We consider two metrics in the evaluation: 1) the ROC curve and its AUC value to evaluate the performance in all thresholds, 2) F1 score and accuracy to evaluate the performance in a selected threshold. In this work, we usually use the former metric to evaluate *CMA* and the baseline methods, and use the latter metric to provide an intuitive evaluation of the overall performance in the public dataset and continuous uncontrolled dataset.

4.3.1 ROC Curve and AUC value. In our sensor signal association problem, both the true positive (i.e., the structural vibration sensor's signal is associated to the wearable sensor that causes vibration) and false positive (i.e., the structural vibration sensor's signal is not associated to the non-causal wearable sensor) are important performance indicators. Therefore, we adopt ROC (Receiver Operating Characteristic) curve and AUC (Area under the ROC Curve) [23] to evaluate each **experiment**. ROC curve is a probability curve that systematically depicts the performance (true and false positive rates) change across the entire range of thresholds [13]. To generate the ROC curve, we apply different AP thresholds θ_{AP} and calculate the true positive and false positive rate. AUC measures the quality of the association irrespective of threshold values [20]. The higher AUC value indicates a better performance.

Table 1. Signal similarity metrics for signals X, Y of length l.

Metric	Equation
MCC	MCC(X, Y) = $\max_{k=0,\ldots,l} \frac{\sum_{i=1}^{l} x_i \cdot y_{i+k}}{\sqrt{\sum_{i=1}^{l} x_i^2} \cdot \sqrt{\sum_{i=1}^{l} y_i^2}}$
Cosine Similarity	CS(X, Y) = $\frac{\sum_{i=1}^{l} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{l} x_i^2} \cdot \sqrt{\sum_{i=1}^{l} y_i^2}}$
Surface Similarity	SS(X, Y) = $\frac{\sqrt{\sum_{i=1}^{l} (x_i - y_i)^2}}{\sqrt{\sum_{i=1}^{l} x_i^2} + \sqrt{\sum_{i=1}^{l} y_i^2}}$

4.3.2 F1 score and Accuracy. Since the final output of *CMA* is a pairwise association between two modalities, we further threshold the AP and calculate the F1 score [40] and accuracy. For each AU, if the IMU segment association matches with the ground truth, we consider it as a true positive (TP). If the associated IMU ID does not match with the association ground truth, we consider it is a false positive (FP); and vice versa, for a false negative (FN). The precision and recall are calculated as $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$. The F1 score is a function of precision and recall, $F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$. The accuracy is the percentage of correctly determined association cases and unassociation cases over all cases.

4.4 Baseline Methods

We consider measuring the shared context or similarity between cross-modal signals as baselines, so we evaluate *CMA* against three commonly used signal similarity metrics [37]. For vibration data segments $VibSig_i$ and IMU data segments $IMUSig_j$, we calculate 1) Cosine similarity (CS), 2) max cross-correlation (MCC), 3) Surface similarity (SS) between them as shown in Table 1. For IMU signals with six axes, we calculate the signal similarity between each axis' and the vibration signal and report the highest similarity over all six axes. For all the baseline methods, the higher value between $VibSig_i$ and $IMUSig_j$ means that the vibration segment i is more likely to be associated with IMU segment j.

4.5 *CMA* Implementation

Multimodal Signal Alignment. Since the sampling rate for the vibration sensor and the IMU sensor are different in the two datasets, we resample the vibration sensing data from 6500 Hz to 235 Hz for the public dataset and resample the vibration sensing data from 4000 Hz to 250 Hz for the continuous uncontrolled dataset to align the multimodal signal inputs. We utilize the resample function [28] in Matlab to re-sample the data. We use the recorded timestamp to align the vibration sensing data with the IMU sensing data for the uncontrolled dataset. We empirically set the energy threshold θ_e as eight, and the threshold of event interval $\Delta\tau$ as four seconds. We set the upper bound of activity segments τ_u as 20 seconds and the lower bound of activity segments τ_l as eight seconds.

Association Discovery. Then for the *AD-TCN* model training, we use the Stochastic Gradient Descent algorithm, and ADAM [24] as optimizer. We set the maximum training epochs as 6000. To avoid the impact of over-fitting or under-fitting of *AD-TCN*, we apply the early stopping method to automatically stop the training based on the loss decrease [43]. We use *ReduceLROnPlateau* function [11] which is integrated into PyTorch [34] to implement early stopping and set the factor and patience parameter as 0.5 and 4, respectively. We terminate the training when the learning rate drops to less than 0.001 (initially 0.01). Parameters of dilation and stride in *Conv1d* [10] are both set as 1.

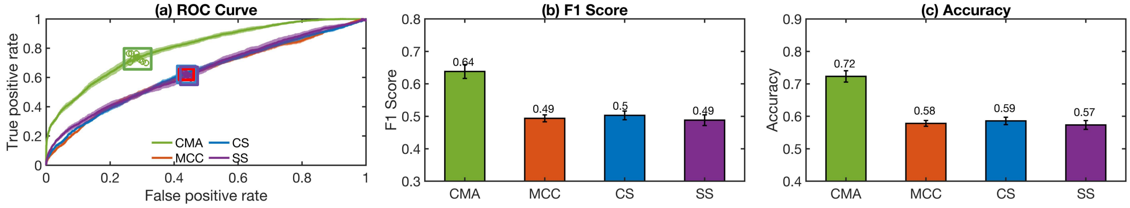


Fig. 6. Association performance with the public dataset. (a) shows the average ROC curve and the standard deviation (width of the curve) of false positive rate and true positive rate in 10 experiments. (b) and (c) shows the F1 score and accuracy calculated from the circled data points in (a), respectively.

Association Threshold. We consider the output of the softmax function (section 3.3) as the estimated AP between N IMU segments. If all IMU segments are not associated with the vibration segment, the ideal distribution of AP should be a uniform distribution. So we select $1/N$ as the association threshold for *CMA*. For the baseline methods, we select the mean value over all detected events in each experiment set (100 trails in the public dataset) as the threshold to determine the association. Once the baseline values (CS, MCC, SS) between the vibration segment and the IMU segment is larger than this threshold, we report they are associated.

4.6 Use Case Study

We implement the two aforementioned uses cases on the public dataset [22] due to the availability of the identity and activity labels. We consider the use case scenario of three participants co-habit in a house. We investigate three association conditions: 1) Ideal association (ground truth). The pair of IMU and vibration data are of their true associations. 2) *CMA* association. The pair of IMU and vibration data are based on the *CMA*'s output. 3) Random association (baseline). The pair of IMU and vibration data are randomly assigned. For learning models, we randomly select 80% data for training, and the rest for testing.

Occupant Identification. In scenarios of vibration-based in-home elderly or patient monitoring [17, 33], it is challenging to acquire the identity labels of each occupant's vibration signals to bootstrap the learning model in the real-world deployment. We envision a temporary setup with the IMU sensor could be used with our *CMA* association scheme to provide initial identity labels for the learning model for a household of three people. We run *CMA* to acquire the identity label of the structural vibration signal segments, and then we train an SVM model [17] on these segments with pseudo label from the association. We report the identification accuracy values over the three association scenarios.

Multimodal Human Activity Recognition (HAR). In this use case, we conduct the multimodal human activity recognition (HAR) [22] to depict the cross-modal association's importance. Instead of directly fuse two types of sensor data with random association, or provide manual label of this association (ideal), we leverage *CMA* to provide this information. Then this association will determine the input IMU-vibration signal pair to the multimodal learning training and testing for activity recognition. We use a same fully connected neural network as classifier to recognize the occupant activity [22]. The model is trained with a cross entropy loss and the Adam optimizer. We report the accuracy of nine activities recognition over the three association scenarios.

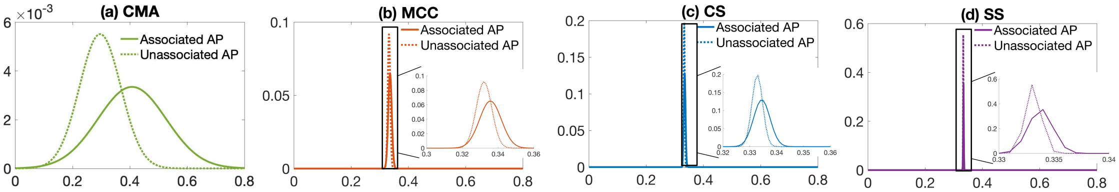


Fig. 7. The distribution of associated and unassociated AP of *CMA* and baselines.

5 RESULTS AND ANALYSIS

In this Section, we first introduce the overall performance (Section 5.1) of *CMA*, the impact from data (Section 5.2), and *CMA* configuration (Section 5.3) over the public dataset. Then we further shows the performance in the continuous uncontrolled dataset (Section 5.4). Finally, we demonstrate the performance of *CMA* in two user cases (Section 5.5).

5.1 Overall Performance

In the overall performance experiment, we randomly select three pairs of segments out of the full set (1048 pairs) to conduct the overall performance evaluation with the experiment procedure introduced in Section 4.2. Figure 6(a) shows the ROC curve of *CMA* and baseline methods. The solid line presents the average value of the ROC curve, and the area around the line presents the standard deviation of the 10-repetition experiments. We observe that the ROC curve of *CMA* is always above those of the baseline methods, which indicates a better association accuracy. If we consider a tolerable false positive rate of 0.2, the average true positive rate for *CMA* can achieve 0.63, which is up to 1.5× (50% improvement) of the baselines (MCC 0.39, CS 0.40, SS 0.41). The average AUC value of *CMA* achieves 0.80, which is up to 30% improvement compared to the baselines (MCC 0.63, CS 0.62, and SS 0.64). Figure 6(b) and (c) show the F1 score and accuracy calculated from the circled data points in Figure 6(a). The average F1 score of *CMA* and baselines achieve 0.64, 0.49 (MCC), 0.50 (CS), and 0.49 (SS), respectively. *CMA* achieves 1.3× F1 score value of the baseline methods (up to 31% improvement). The average accuracy of *CMA* and baselines achieve 0.72, 0.58 (MCC), 0.59 (CS), and 0.57 (SS), respectively. The accuracy of *CMA* achieves up to 26% improvement than the baseline methods.

AP Distribution. We also demonstrate the distribution of associated and unassociated AP to further analyze the performance of *CMA* and baselines. For the baselines, we adopt the softmax function to converter the metric values between two cross-modal segments to association probability (Equation 6). Figure 7 shows the AP distribution of *CMA* and baselines. We can observe that the distribution of associated and unassociated AP of *CMA* has less overlapping than the baselines, which indicates the estimated AP value of *CMA* is more separable.

5.2 Impacts of Data: Activity Category and Association Levels

One potential factor that may impact the association performance is the type of activities. Because the association level varies for different activities. For some activities, the motion measured by the wearable also directly induces structural vibration. For example, when people cut food, their wrist motion (measured by the IMU) **directly** causes the knife to impact the cutting board (measured by vibration sensors). On the other hand, for some activities, the motion measured by the wearable does not directly associate with the structural vibration. For example, vacuuming the floor causes the floor to vibration due to motor vibration, which does not directly **indirectly** cause structural vibration via wrist

Table 2. Types of activities and cross-modal association levels.

Assoc. Levels	Activities
direct	cutting food, stir-fry, open/close drawer
indirect	keyboard typing, handwriting, vacuuming
semi-direct	using mouse, wiping countertop, sweeping

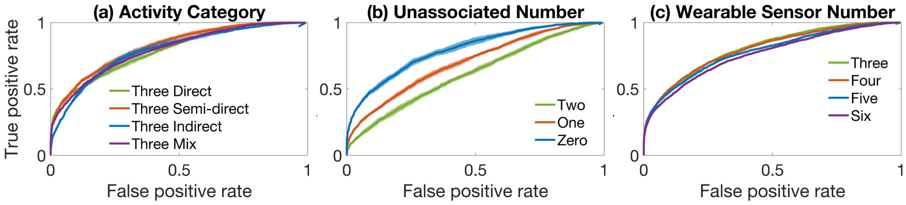


Fig. 8. Roc curve of *CMA* in the public dataset analysis. (a), (b), and (c) show the performance of *CMA* for impact of activity category, unassociated number, and wearable sensor number, respectively.

motions. Therefore, we categorize the nine types of activities into three levels of association – direct, indirect, and semi-direct – in Table 2.

5.2.1 Activity Category Combinations. To demonstrate *CMA*'s robustness over types of activities with different association levels, we randomly select four pairs of segments out of subsets of pairs with different types of activities – direct associated activities, indirect associated activities, semi-direct associated activities, and mixed activities. Then we follow the same experiment procedure in Section 4.2.

Figure 8(a) depicts the average ROC curve on 10 repetition experiments. The average AUC value of *CMA* is 0.79 (direct), 0.82 (semi-direct), 0.79 (indirect), 0.80 (mix). Three baselines depict overall lower than 0.7 AUC values. *CMA* achieves the best performance in all activity category combinations. Furthermore, *CMA* demonstrates robustness over different activity categories, while the baselines have inconsistent performance with the AUC value varying between 0.6 and 0.7.

5.2.2 Unassociated Combinations. To better understand how *CMA* performs in the real scenario, we further evaluate when some of the vibration signals are generated by occupants without an IMU sensor. We randomly select three pairs of signals ($VibSig_i$ and $IMUSig_j$) from the full set of pairs (1048) and investigate the scenario where for 0/1/2 of them $i \neq j$ and the rest $i = j$. Then we follow the same experiment procedure in Section 4.2 and compare the AUC values when there are different numbers of unassociated pairs among the three.

Figure 8(b) shows the average ROC curve of *CMA* on 10 repetition experiments. Overall, when the number of unassociated pairs increases, the AUC value decreases. This could be because the prediction of the unassociated infrastructural signal is done with multiple IMU signals equally not associated, which results in similar APs that is not efficient for distinguishing the association relationship. When there is one unassociated signal pair, *CMA* achieves an AUC value over 0.7, while the baselines only achieve 0.57, 0.56, and 0.58, respectively (random selection's AUC value is 0.5). *CMA* also achieves the best performance than baselines.

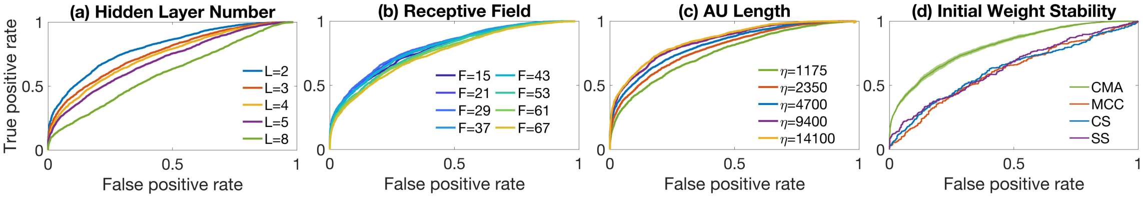


Fig. 9. Impact of *CMA* configures and repeatability of *CMA*. (a), (b), and (c) show the performance of *CMA* under the different configurations of hidden layer number, receptive field, and AU length, respectively. (d) shows the ROC curve of *CMA* and baseline methods when we repeat *CMA* 10 times on the same experiment set.

5.2.3 Wearable Sensor Number. To better understand the scalability of *CMA*, we further evaluate *CMA* when the number of wearable devices N is larger than 3. In this experiment, we first randomly select three pairs of signal segments from the full set of pairs. Then we further randomly select extra numbers of $IMUSig$ and apply *CMA* to associate $M = 3$ number of $VibSig$ and N number of $IMUSig$, where $N = 3, 4, 5, 6$. Then we follow the same experiment procedure in Section 4.2.

Figure 8(c) shows the average ROC curve of *CMA*. When the number of wearable sensor N increase from 3 (the same as the number of vibration sensors M) to 6, the average AUC values decreases slightly (≤ 0.05) with the number of wearable increase. This is because the difficult of find the associated IMU segment increases when the number of IMU segments N increase. In summary, *CMA* also works for the scenario that is more than three people.

5.3 Impacts of *CMA* Configuration

We further explore the impact of the hyper-parameter configuration of *CMA* on the performance. As introduced in Section 3.2, *CMA* contains three hyper-parameters: 1) hidden layer number L, 2) receptive field F (adjusted by kernel size K), and 3) input AU length η. The default values for these hyper-parameters are shown in Table 3. We randomly select three pairs of segments from the full set (1048 pairs) and conduct experiments with the procedure introduced in Section 4.2 with varying *AD-TCN* hyper-parameters.

5.3.1 Hidden Layer Number L. Hidden layer number directly impacts the complexity of the neural network. Therefore we investigate how the model acts at different levels of complexity for the cross-modal time series prediction. We increase L from 2 to 8, and demonstrate the average ROC curve of *CMA* in Figure 9(a). The average AUC value of each configuration are 0.81, 0.76, 0.74, 0.71, 0.61, respectively. We observe that *CMA* achieves the highest AUC value when the L is set to 2. This result indicates that a shallow architecture is more suit for the cross-modal association task. It could be because the association discovery task is fundamentally a binary classification task, and the model can be presented with a simple network architecture sufficiently. A large L value may cause the network to overfit [4]. When

Table 3. *CMA* hyper-parameters

Parameters	Default	Controlled Experiment Values
L	2	2,3,4,5,8
F	29	15, 21, 29, 37, 43, 53, 61, 67
η	2350 to 3055	1175, 2350, 4700, 9400, 14100

the overfit occurs, the network cannot generalize to test data, hence is not able to make accurate prediction [15]. Under this circumstance, the calculated association score is not reliable for the association discovery.

5.3.2 *Receptive Field F.*

The receptive field F is determined by both the hidden layer number L and the causal convolutional layer's kernel size K as $F = (K - 1) \cdot L + 1$ [31]. It describes how 'far' the model can 'see' to predict the current samples [26]. For example, Figure 4(b) shows an example of a causal convolutional layer with a kernel size $K = 2$. If layer number $L = 2$, then receptive field $F = (2 - 1) \cdot 2 + 1 = 3$.

Figure 9(b) shows the average Roc curve in different receptive field configurations. When F increases, the average AUC value first increases then decreases (0.79, 0.80, 0.81, 0.80, 0.79, 0.78, 0.76, and 0.75 for F from 15 to 67). *CMA* demonstrates a stable performance and achieves the highest average AUC value when F is 29. One explanation for *CMA* achieves the highest AUC with $F = 29$ is that the time duration for 29 samples is approximately 0.1 second, which is approximately the duration for an arm motion to cause an impulsive vibration signal. Therefore, this amount of 'history' data is most helpful for the prediction of current sample value.

5.3.3 *Input AU Length η.*

The input AU length η determines how much data is available to calculate AP and determine the association relationship. Intuitively, the longer the observation data is, the more accurate the time-series prediction model is, and hence the network parameter that describes the association relationship is more accurate.

Figure 9(c) shows the average ROC curve of *CMA* when the input AU length η varies from 1175 ($\tau = 5$ seconds) to 14100 ($\tau = 60$ seconds). With the increase of η, the performance of all evaluated methods increases. We select η taking into account the trade-off between the prediction accuracy and the data practicality. Since our assumption is that the signal association within η is invariant, it means the higher the η, the more unlikely the assumption holds. For the public dataset, we consider the default value of η is 2350 ($\tau = 10$) because the duration of activity from the public dataset is in the range of 10 to 15 seconds.

5.3.4 *AD-TCN Initial Weight Stability.*

The initial weight assignment can directly impact the neural network model and it's performance [14]. Therefore, we also investigate the repeatability of *AD-TCN* with different random initial weights. We randomly select three pairs of segments out of the full set, and conduct the *AD-TCN* training with different initial randomization 10 times. We repeat this random selection 110 times to avoid sampling bias.

Figure 9(d) show the average ROC curves of *CMA* and baselines when we train *AD-TCN* on the same dataset 10 times with different random initial weights. The green line shows the average false positive rate and true positive rate, and the green area around the green line shows the standard deviation of 10 times of weight initialization. *CMA* demonstrates a stable performance when the weights of the neural network module is initialized differently.

5.4 Robustness in Uncontrolled Deployment

Figure 10 (a) shows the average ROC curve and the standard deviation of false positive rate and true positive rate of *CMA* and baselines in five houses dataset. We can observe the performance of *CMA* is better than the baselines and the false positive rate and true positive rate is more stable. The average AUC value of *CMA*, and baselines are 0.85, 0.64 (MCC), 0.56 (CS), and 0.64 (SS), respectively. The AUC value of *CMA* achieves 0.85, which is up to 37% improvement compared to the baselines.

The circle marks in Figure 10(a) indicate the false positive rate and true positive rate under the selected association threshold (introduced in Section 4.5). Figure 10(b) and (c)demonstrates the F1 score and accuracy, respectively. The average F1 score of *CMA* and baselines achieve 0.69, 0.45 (MCC), 0.51 (CS), and 0.51 (SS), respectively. The F1 score of

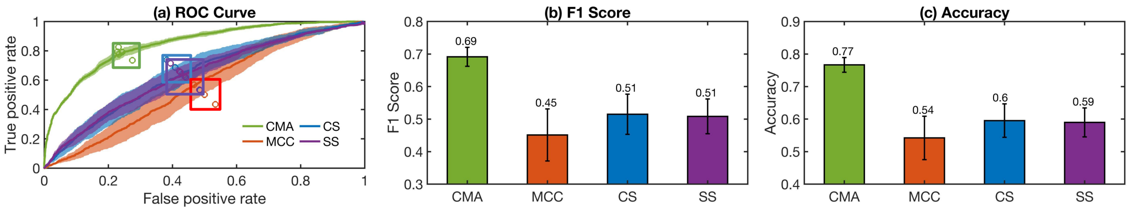


Fig. 10. Overall performance with the uncontrolled dataset. (a) shows the average ROC curve and the standard deviation (width of the curve) of false positive rate and true positive rate in different houses dataset. The circle on the curve indicates the false positive rate and true positive rate when *CMA* operates with the selected threshold. (b) and (c) show the F1 score and accuracy under the selected association threshold, respectively.

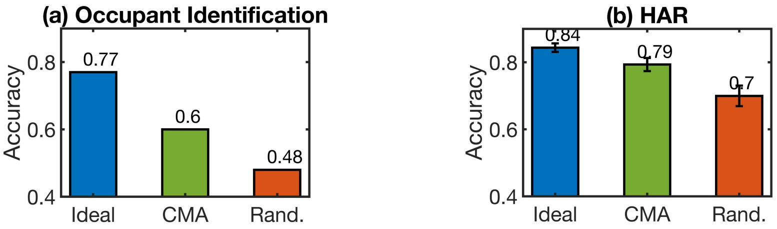


Fig. 11. The performance of two use cases with cross-modal association information provided by *CMA*.

CMA achieves 0.69, which is up to 53% improvement compared to the baselines. The average accuracy of *CMA* and baselines achieve 0.77, 0.54 (MCC), 0.60 (CS), and 0.59 (SS), respectively. The accuracy of *CMA* achieves 0.77, which is up to 43% improvement compared to the baselines.

We also observe that compared with the performance in the public dataset, the performance of *CMA* in the uncontrolled dataset is 0.05% better (AUC value 0.8 vs. 0.85, F1 score 0.64 vs. 0.69, accuracy 0.72 vs. 0.77). This might be because in the uncontrolled dataset, the three human subjects are more likely to conduct different types of activity at the same time than in the public dataset. Finding the association relationship from the same type of activity is more difficult since the IMU segments of the same type of activity are more similar to each other.

5.5 Use Case Performance

Figure 11 shows the accuracy of *CMA* compared with baselines for two use cases. The blue, green and red bars represent of ideal association (ground truth), *CMA* association, and random association (baseline), respectively. We observe that with the association provided by *CMA*, both use cases demonstrate an improvement in accuracy compared to the baseline. For occupant identification, the system achieves a 12% accuracy increase with the pseudo label provided by *CMA* without any manual label. For HAR, *CMA* achieves approximate 10% accuracy improvement compared to without the association information, and it is only 5% lower than the accuracy with ideal association. Such improvement is promising, considering that it is made with leveraging the pervasive wearable IMU data, and without requirements of any label data.

6 DISCUSSION

Temporal Overlapping and Activity Segmentation. In this work, we focus on the cross-modal segment-level association problem with the assumption of no temporal signal overlapping of multiple sources at one structural vibration sensor. If one structural vibration sensor captures overlapped signals from multiple activities, the implicit shared context can be learned for association purposes will be more constrained than what has been investigated in this work and therefore more challenging. In the future, we plan to explore either leveraging hierarchical temporal information over different time resolutions, or combining frequency domain analysis to tackle the signal temporal overlapping challenge.

Activity segmentation is another important aspect of indoor occupant sensing. In this work, we adopted the lumping algorithm [16]. Our uncontrolled experiment result inherits the segmentation error from the lumping algorithm. In the future, we will explore incorporating other activity segmentation schemes. Furthermore, we will explore jointly conducting the separation and segmentation with *CMA* to further improve the robustness.

Association-Aware Multimodal Learning. With the segment-level association learned for each segment, we can further use this learned information to enhance the existing multimodal learning. For example, the association can be used as a dynamic sensor selection criteria to allow the inference models to adapt to input channels, as well as a regularization to reduce the chance of learning a spurious relationship between input channels and data labels. For graph neural network-based models, this association may be used as the prior knowledge to establish the graph, ensuring a more efficient and robust inference [29].

Modality Generalizability. In this paper, we evaluate *CMA* with the combination of structural vibration sensing and wearable on-wrist IMU sensing. *CMA* is designed for general time series sensing modalities, and in the future, we plan to explore more modalities (e.g., acoustic, event camera, electricity load, physiological sensors) combination to further understand its limitation and generalizability. For the high-dimension sensing data, we can build an encoder to convert the high-dimension data to one-dimension sequences, such as data2vec [2]. On the other hand, association learning is more challenging for modalities with a latent and longer dependency. For example, when the occupant turns on the heater, the indoor temperature becomes warmer, and the occupant's heart rate will slowly go higher [5]. In this case, the association between the electricity load sensor and physiological sensors (heart rate monitor) data is latent and potentially requires a new framework for association learning.

Computational Requirements of CMA. In our experiment, the time consumption of *CMA* for one AU is around 10 seconds in an Apple MacBook Pro 2022 using CPU only. In this work, we focus on providing a data-driven method to discover the association relationship between two modalities without the requirements of label data. However, the time consumption can be decreased by optimizing multiple factors, such as the code implementation framework, and adopting parallel computing. The current computation is on the server side, and in the future, we would also consider offloading the computation to the nearby devices with an event-driven design on the embedded platform side.

7 CONCLUSION

We present *CMA*, a cross-modal signal segment association scheme between wearable and structural vibration sensors. We introduce *AD-TCN*, a TCN-based framework, to calculate the amount of shared context between signal segments from two modalities. After training the network, we calculate the association probability based on the weights of the trained *AD-TCN*, and determine the pairwise segment association. We evaluate *CMA* via a public multimodal dataset for systematic evaluation, and we collect a continuous uncontrolled dataset for robustness evaluation. *CMA* achieves up to 37% AUC value, 53% F1 score, and 43% accuracy improvement compared to baselines.

ACKNOWLEDGEMENT

We sincerely thank the anonymous shepherd and reviewers for their constructive suggestions. This research was supported by an American Psychological Foundation (APF) 2021 Drs. Rosalee G. and Raymond A. Weiss Research and Program Innovation Grant, a 2022 Seed Fund Award from CITRIS and the Banatao Institute at the University of California.

REFERENCES

[1] 2019. Softmax function. https://deepai.org/machine-learning-glossary-and-terms/softmax-layer

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*. PMLR, 1298–1312.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[5] Stewart S Bruce-Low, David Cotterrell, and Gareth E Jones. 2006. Heart rate variability during high ambient heat exposure. *Aviation, space, and environmental medicine* 77, 9 (2006), 915–920.

[6] Published by Statista Research Department and Feb 22. 2022. Average size of households in the U.S. 2021. https://www.statista.com/statistics/183648/average-size-of-households-in-the-us/

[7] Maojian Chen, Ying Li, Xiong Luo, Weiping Wang, Long Wang, and Wenbing Zhao. 2018. A novel human activity recognition scheme for smart health using multilayer extreme learning machine. *IEEE Internet of Things Journal* 6, 2 (2018), 1410–1418.

[8] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.

[9] Seungeun Chung, Jiyoun Lim, Kyoung Ju Noh, Gague Kim, and Hyuntae Jeong. 2019. Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning. *Sensors* 19, 7 (2019), 1716.

[10] PyTorch Contributors. [n.d.]. CONV1D. https://pytorch.org/docs/stable/generated/torch.nn.Conv1d.html

[11] PyTorch Contributors. [n.d.]. Reducelronplateau. https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html

[12] Lang Deng, Jianfei Yang, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2022. GaitFi: Robust Device-Free Human Identification via WiFi and Vision Multimodal Learning. *IEEE Internet of Things Journal* (2022).

[13] Peter A Flach. 2016. ROC analysis. In *Encyclopedia of machine learning and data mining*. Springer, 1–8.

[14] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[16] Jun Han, Albert Jin Chung, Manal Kumar Sinha, Madhumitha Harishankar, Shijia Pan, Hae Young Noh, Pei Zhang, and Patrick Tague. 2018. Do you feel what I hear? Enabling autonomous IoT device pairing using different sensor types. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 836–852.

[17] Jun Han, Shijia Pan, Manal Kumar Sinha, Hae Young Noh, Pei Zhang, and Patrick Tague. 2017. Sensetribute: smart home occupant identification via fusion across on-object sensing devices. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*. 1–10.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Zhe Hui Hoo, Jane Candlish, and Dawn Teare. 2017. What is an ROC curve? , 357–359 pages.

[21] Zhizhang Hu, Yue Zhang, and Shijia Pan. 2022. *Multimodal Fine-grained Human Activity Dataset*. https://doi.org/10.5281/zenodo.6519052

[22] Zhizhang Hu, Yue Zhang, Tong Yu, and Shijia Pan. 2022. VMA: Domain Variance-and Modality-Aware Model Transfer for Fine-Grained Occupant Activity Recognition. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 259–270.

[23] Jin Huang and Charles X Ling. 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering* 17, 3 (2005), 299–310.

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

[25] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 156–165.

[26] Yang Lin, Irena Koprinska, and Mashud Rana. 2021. Temporal convolutional attention neural networks for time series forecasting. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[27] Chris Xiaoxuan Lu, Hongkai Wen, Sen Wang, Andrew Markham, and Niki Trigoni. 2017. SCAN: learning speaker identity from noisy sensor data. In *2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 67–78.

[28] Matlab. [n.d.]. resample uniform or nonuniform data to new fixed rate - matlab. https://www.mathworks.com/help/signal/ref/resample.html

[29] Shenghuan Miao, Ling Chen, Rong Hu, and Yingsong Luo. 2022. Towards a Dynamic Inter-Sensor Correlations Learning Framework for Multi-Sensor-Based Wearable Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–25.

[30] Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction* 1, 1 (2019), 312–340.

[31] Studio Otwarte. 2022. Temporal convolutional networks and forecasting. https://unit8.com/resources/temporal-convolutional-networks-and-forecasting/

[32] Shijia Pan, Carlos Ruiz, Jun Han, Adeola Bannis, Patrick Tague, Hae Young Noh, and Pei Zhang. 2018. Universense: Iot device pairing through heterogeneous sensing signals. In *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*. 55–60.

[33] Shijia Pan, Tong Yu, Mostafa Mirshekari, Jonathon Fagert, Amelie Bonde, Ole J Mengshoel, Hae Young Noh, and Pei Zhang. 2017. Footprintid: Indoor pedestrian identification through ambient structural vibration sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–31.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[35] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016).

[36] Carlos Ruiz, Shijia Pan, Adeola Bannis, Ming-Po Chang, Hae Young Noh, and Pei Zhang. 2020. IDIoT: Towards ubiquitous identification of iot devices through visual and inertial orientation matching during human activity. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 40–52.

[37] Carlos Ruiz, Shijia Pan, Adeola Bannis, Xinlei Chen, Carlee Joe-Wong, Hae Young Noh, and Pei Zhang. 2018. Idrone: Robust drone identification through motion actuation feedback. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–22.

[38] Fred Szabo. 2015. *The linear algebra survival guide: illustrated with Mathematica*. Academic Press.

[39] Northwestern University. [n.d.]. http://www.ece.northwestern.edu/local-apps/matlabhelp/toolbox/signal/resample.html

[40] Vincent Van Asch. 2013. Macro-and micro-averaged evaluation measures. *Belgium: CLiPS* 49 (2013).

[41] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. *SSW* 125 (2016), 2.

[42] Jin Wang, Mary She, Saeid Nahavandi, and Abbas Kouzani. 2010. A review of vision-based gait recognition methods for human identification. In *2010 international conference on digital image computing: techniques and applications*. IEEE, 320–327.

[43] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation* 26, 2 (2007), 289–315.

[44] Jie Yin, Son N Tran, and Qing Zhang. 2018. Human identification via unsupervised feature learning from UWB radar data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 322–334.

[45] Jin Zhang, Bo Wei, Fuxiang Wu, Limeng Dong, Wen Hu, Salil S Kanhere, Chengwen Luo, Shui Yu, and Jun Cheng. 2020. Gate-ID: WiFi-based human identification irrespective of walking directions in smart home. *IEEE Internet of Things Journal* 8, 9 (2020), 7610–7624.

[46] Yue Zhang, Zhizhang Hu, Susu Xu, and Shijia Pan. 2021. AutoQual: task-oriented structural vibration sensing quality assessment leveraging co-located mobile sensing context. *CCF Transactions on Pervasive Computing and Interaction* 3, 4 (2021), 378–396.

[47] Yue Zhang, Shijia Pan, Jonathon Fagert, Mostafa Mirshekari, Hae Young Noh, Pei Zhang, and Lin Zhang. 2018. Occupant activity level estimation using floor vibration. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 1355–1363.

[48] Yue Zhang, Lin Zhang, Hae Young Noh, Pei Zhang, and Shijia Pan. 2019. A signal quality assessment metrics for vibration-based human sensing data acquisition. In *Proceedings of the 2nd Workshop on Data Acquisition To Analysis*. 29–33.