SoHAM: A Sound-Based Human Activity Monitoring Framework for Home Service Robots

Ha Manh Do, *Member, IEEE*, Karla Conn Welch, *Member, IEEE*, and Weihua Sheng, *Senior Member, IEEE*

***Abstract*— Monitoring daily activities is essential for home service robots to take care of the older adults who live alone in their homes. In this article, we proposed a sound-based human activity monitoring (SoHAM) framework by recognizing sound events in a home environment. First, the method of context-aware sound event recognition (CoSER) is developed, which uses contextual information to disambiguate sound events. The locational context of sound events is estimated by fusing the data from the distributed passive infrared (PIR) sensors deployed in the home. A two-level dynamic Bayesian network (DBN) is used to model the intratemporal and intertemporal constraints between the context and the sound events. Second, dynamic sliding time window-based human action recognition (DTW-HaR) is developed to estimate active sound event segments with their labels and durations, then infer actions and their durations. Finally, a conditional random field (CRF) model is proposed to predict human activities based on the recognized action, location, and time. We conducted experiments in our robot-integrated smart home (RiSH) testbed to evaluate the proposed framework. The obtained results show the effectiveness and accuracy of CoSER, action recognition, and human activity monitoring.**

***Note to Practitioners*—This article is motivated by the goal to develop companion robots that can assist older adults living alone. Among many capabilities, monitoring human daily activities is an essential one for such robots. Though computer vision or wearable sensors-based methods have been developed by other researchers, they are not practical due to the privacy concern and intrusiveness. Sound-based daily activity recognition can address these concerns and offer a viable solution. In this regard, our proposed method adopts microphones on the robot and a small set of motion sensors distributed in the home. The proposed theoretical framework was tested in a small-scale mock-up apartment with promising results. Before such companion robots can be deployed to real homes for elderly care, there is a need to improve the robustness of the algorithms. More thorough tests in various realistic home environments should be conducted to fully evaluate the performance of the robots. In addition, privacy concern related to audio capture should be further mitigated.**

***Index Terms*— Activity monitoring, context-awareness, elderly care, sound event.**

I. Introduction

THE elderly population is steadily rising around the globe. The population of 60-and-older people is projected to increase from 900 million in 2015 to over 2 billion in 2050 [1]. This trend leads to both economical and sociological challenges in elderly care [2]. On the other hand, many older adults prefer to stay in their homes rather than move to nursing homes, although their daily living activities may become more challenging [3]. In fact, more than a third of the older adults in the USA live alone in their homes [4], which poses serious risks to them in situations such as falling or medical emergencies. Therefore, assistive technologies, such as smart homes and home service robots, are currently being developed for elderly care.

As a critical part of assisted living, human activity monitoring has received great interest in recent years. Camera-based human activity monitoring has been developed for many applications such as surveillance and healthcare [5], [6]. Although the vision system on a robot provides abundant information, it is not always possible to observe the resident due to occlusion or poor lighting. In addition, the use of cameras raises significant privacy concerns. Recently, wearable sensor-based human activity monitoring has been studied, especially for health care, military, and security applications [7]–[9]. However, wearable sensors are intrusive and inconvenient if the users are required to wear them all the time. On the other hand, we know that most human daily activities generate sounds, such as eating, cooking, using the toilet, and having a shower. Therefore, it is highly desirable to equip home service robots with not only speech understanding but also sound understanding capabilities. Home sound understanding, which recognizes home sound events in the context of human daily activities, helps the robot not only monitor older adults' activities but also detect anomalies happening in the homes. Such a human-aware capability frees the robot to do its daily routine work while being able to care for the elderly more proactively and effectively.

Although sound event recognition has received much attention over the years, it is still a very challenging problem. The main reason is that event sounds are diverse, unstructured, and nonstationary. Understanding human activity using sound

Manuscript received March 11, 2021; accepted April 28, 2021. This article was recommended for publication by Associate Editor H. Wang and Editor B. Vogel-Heuser upon evaluation of the reviewers' comments. This work was supported in part by the National Science Foundation (NSF) under Grant CISE/IIS/1427345, Grant CISE/IIS/1910993, Grant DUE 1928711, Grant CISE/IIS 1838808, and Grant OIA 1849213 and in part by the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China, under Grant ICT2021B14. *(Corresponding author: Weihua Sheng.)*

Ha Manh Do is with Plume Design, Inc., Palo Alto, CA 94306 USA (e-mail: ha.do@okstate.edu).

Karla Conn Welch is with the Louisville Automation and Robotics Research Institute (LARRI), University of Louisville, Louisville, KY 40292 USA (e-mail: karla.welch@louisville.edu).

Weihua Sheng is with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078 USA (e-mail: weihua.sheng@okstate.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2021.3081406.

Digital Object Identifier 10.1109/TASE.2021.3081406

events is even more difficult because of the diversity of the sounds associated with the same events. For example, even the same event of a person falling on the floor can create different sounds, depending on where the fall occurs. Moreover, there are many different types of event sounds in home environments. Based on sound source features, they can be vocal or nonvocal. Based on acoustic features, they can be harmonic or nonharmonic. In addition, multiple sound events can occur at the same time. One advantage that allows humans to distinguish sound events is their knowledge of the context, which helps them form predictions and adapt their perception to the environment [10]. Context-aware sound event recognition (CoSER) allows a robot to associate contextual information with sound events, which enhances the performance of sound event recognition.

This article aims to develop a human activity monitoring framework for home service robots. This article has the following contributions:

1) A new framework for sound-based human activity monitoring (SoHAM) is proposed and developed which allows home service robots to better understand human daily activities.
2) A new method of CoSER is developed based on Dynamic Bayesian Networks (DBNs). This method improves recognition accuracy by considering contextual information estimated from multiple distributed sensors in a home environment.
3) A conditional random field (CRF)-based model is proposed to recognize human activities using the recognized action, location, and time. This method effectively overcomes the difficulties associated with the nondeterministic nature of complex daily activities.
4) We conducted experimental validation and evaluation of the proposed SoHAM framework in a smart home testbed using a custom-built home service robot.

The rest of this article is organized as follows. Section II reviews the related works in SoHAM and sound event recognition. Section III gives an overview of the human activity monitoring platform for home service robots. Section IV presents the method for CoSER using DBNs. Section V develops the algorithm of the dynamic sliding time window for human action recognition based on recognized sound events. Section VI describes CRFs-based human activity monitoring. Section VII gives the experimental results. Section VIII concludes this article and discusses the future work.

II. Related Works

A. Sound-Based Human Activity Monitoring

In recent years, research on SoHAM has received much attention. For example, an automated bathroom activity monitoring system based on acoustics was developed in [11]. In that project, an infrared door sensor was set up to detect the human entering or leaving the bathroom, and sound was recorded by a microphone. Six bathroom sound events were collected and classified by hidden Markov models (HMMs) with Mel-frequency Cepstral Coefficients (MFCCs) features. A support vector machine (SVM)-based system was introduced in [12] to detect and recognize human activities in meeting rooms using acoustic signals. In [13], the authors proposed SoundSense, a scalable framework for modeling and recognizing meaningful sound events that occur in users' everyday lives using mobile phones. SoundSense uses a combination of supervised and unsupervised learning techniques to classify both general sounds (e.g., music, voice) and discover novel sound events specific to each individual user. In [14], the authors proposed a novel recognition approach, non-Markovian ensemble voting (NEV), which was able to robustly recognize 22 different event sounds related to human activities in a bathroom and a kitchen. An acoustic-based activity recognition system inspired by the framework of three mental structures in cognitive psychology was proposed in [15], which consists of a sensory store, a working memory, and permanent memory modules. Sound features that include formant, intensity, pitch, and duration are extracted by the sensory store module and analyzed in the working memory module using the reasoning by similarity (RBS) and reasoning by elimination (RBE) strategies. The framework was tested on nine dining activities with an average accuracy of 83.2%. In [16], a framework for online activity recognition from event sounds and home sensors was proposed and evaluated on two existing smart home datasets using different probabilistic models including HMM, CRF, and Markov logic network (MLN). More recently, several studies introduced deep learning-based methods for activity recognition from sound events. The authors in [17] proposed a deep neural network (DNN)-based system for daily activity recognition using environmental sounds and body acceleration signals. A 5-layer DNN was trained by a dataset of ten activities and achieved a frame accuracy rate of 85.5% and a sample accuracy rate of 91.7%. Another convolution neural network (CNN)-based sound recognition model to detect occupant behavior and possible emergency events in single-person households was developed in [18]. This model successfully monitored 12 sequential events of acoustic sounds with an F1-score of 83.9%. However, most deep learning-based methods require the collection of a very large dataset of labeled sounds, which is time-consuming and costly.

It is clear that sound event recognition plays an important role in the above acoustic-based activity recognition systems. These systems directly inferred each human activity from a single sound event that is recognized. The sound itself may not be sufficient to infer human daily activities, unless the human context (location, time) is taken into consideration. For example, the same water-running sound may be generated by different activities, such as doing the morning routine in the bathroom or cooking in the kitchen.

B. Sound Event Recognition

Various approaches have been developed for sound event recognition (SER). Most stationary SER techniques are derived from the research on speech and music recognition using stationary features. Recent research on SER has explored nonstationary features of event sounds. More recently, deep

learning-based SER techniques have been receiving growing interest.

Inspired by the success of speech and music recognition, several parametric techniques using supervised learning have been adopted for sound event recognition, for example, HMMs with MFCCs features [19], Gaussian mixture models (GMM) with linear frequency Cepstral coefficient (LFCC) features [20], GMM with MFCC and other spectral features [21], [22], and HMM with MPEG7 features [23]. The speech recognition techniques work well, in practice, but the results on SER have not been satisfactory [24]. One reason is that the event sounds are less structured, have no subword dictionary in the same way as in speech and contain a wider range of characteristic and nonstationary effects.

Recently, research on SER has focused on exploring nonstationary features of event sounds to maximize information content related to signal's temporal and spectral characteristics [25]. In [26], the authors used the discrete chirplet transform (DChT) and the discrete curvelet transform (DCuT) along with several other common features such as MFCC and zero-crossing rate (ZCR). The matching pursuit (MP)-based features for SER was proposed in [27]. Several other nonstationary features have been proposed for SER, such as MP-Gabor features [28], image features of the subband power distribution [29], and stabilized auditory image (SAI) [30]. Nonparametric learning methods have also been developed, such as the technique based on the sparse coding of SAIs [31]. Recently, principal component analysis (PCA) and linear discriminate analysis (LDA) are applied to the scale-frequency map to generate the features for sound event classification based on the segment-level multiclass SVMs [28]. In [32], the approach interprets the sound event as a 2-D spectrogram image, with the two axes as the time and frequency dimensions, and adopts spectrogram image processing-based methods for sound event recognition. In [33], a method based on the multiview representation that combines auditory image-based visual features and cepstral features was proposed for sound event recognition using SVMs. This approach resulted in improved performance over other state-of-the-art traditional methods for Environmental Sound Classification - 50 (ESC-50), Detection and Classification of Acoustic Scenes and Events - 2016 (DCASE2016) Task 2, and DCASE2018 Task 2 datasets.

Recent years have seen new methods proposed to tackle several challenges for sound event recognition: the adverse effects such as noise, distortion, and multiple sources as well as the poorly defined characteristics of acoustic events. Several works have recently applied DNNs for polyphonic sound event recognition such as multilabel DNNs [34], a novel spiking neural network system that combines a robust spike coding of local spectrogram features with an artificial neural network using a cost function [35], and a sound event classification framework that evaluates the DNNs with a different spectrogram image-based front features such as Google-style SAI features and time-frequency domain spectrogram image features (SIF) features [36]. In [37], the authors presented an approach to polyphonic sound event detection in real-life recordings based on bidirectional long short term memory (BLSTM) recurrent neural networks (RNNs). The authors in [38] combined existing pretrained CNN models in computer vision applications and SVM for domestic multichannel audio classification. Their method achieved an F1-score of around 89% on the dataset of 9 activities in the DCASE 2018 Task 5 challenge [39]. An end-to-end approach for environmental sound classification based on a 1-D CNN was proposed in [40]. The discussed deep learning approaches require sound datasets to be fully labeled, which incurs time-consuming annotation.

Although these DNNs have proved effective in several SER tasks, they are ineffective for monitoring human activities in home environments. There are two reasons: First, these DNNs were trained on public datasets of general environmental sounds, with very few home event sounds. Second, the sound itself is not sufficient to infer human daily activities, unless the human context is considered.

Context-awareness has been initially exploited in speech recognition. Different methods have been studied to include contextual information as prior knowledge to improve the recognition of phonemes, words, and sentences [41]–[43]. CoSER is still at its early stage compared with context-aware speech recognition. Niessen *et al.* [44] modeled the context in audio recognition by investigating the role of the dynamic network model to improve automatic audio identification and simultaneously reduce the search space of low-level audio features. The context-aware level describes more general information about an audio device such as location [45], time [46], weather [47], and even user-dependent states like emotion [48]. Heittola *et al.* [49] proposed a context-dependent sound event detection system. The context information is recognized from the audio stream by applying GMMs. However, HMM-based event detection models the contexts by a 3-state left-to-right HMM. The recognition still faces difficulties because of the great number of possible event combinations and the transitions among them. Lu *et al.* [50] proposed a context-based environmental audio event recognition framework that applies a two-level HMM for the acoustic scene recognition. Their work is the latest publication of context-aware sound event recognition. In the experiment part, we conducted a comparison of our proposed method and their method.

The above works have mainly targeted the environmental sound events in general and have not taken into account the correlation between the sound events associated with the human's daily activities and the contextual information, such as the human's location in indoor environments. In this article, we propose a novel context-based method for sound event recognition using a DBN that can model intratemporal and intertemporal constraints among the context and sound events. Then human activity monitoring is realized based on recognized sound events.

III. System Overview

This section gives an overview of human activity monitoring for home service robots. Our goal is to monitor human activities over time in a home environment using the audio data captured by the auditory system on the robot and the human location data estimated by the home sensor network.

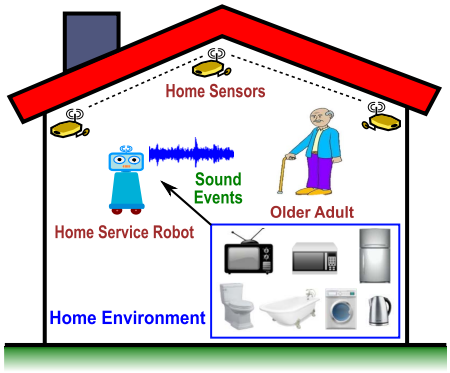

Fig. 1. Home service robot monitors daily activities of the older adult through sound events and locational context.

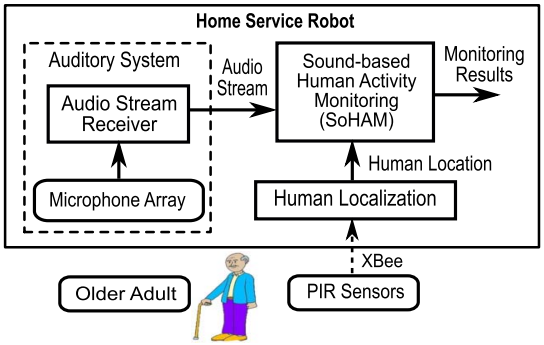

Fig. 2. Overview of the human activity monitoring system on the home service robot.

As illustrated in Fig. 1, a home service robot is integrated into a smart home equipped with distributed sensors that can provide information regarding human locations. This robot can capture event sounds in the home environment through its microphone array. The event sounds are recognized through a local classifier on the robot. Based on the recognized sound event and context information estimated from the home sensor network, the robot can accurately recognize human activities. The overview of the human activity monitoring system is shown in Fig. 2. The home service robot, the human localization module, and the modeling of human activity monitoring are presented in Sections III-A–III-C as follows.

A. Home Service Robot

As shown in Fig. 3, the home service robot that was developed in our Laboratory for Advanced Sensing, Computation and Control (ASCC) was built on a Pioneer P3-DX base with an approximately 1.5 m-long aluminum frame holding up a touch screen monitor which is used for video communication and graphic user interface [51]. The robot is equipped with various sensors and devices. The auditory system is built by extending the built-in microphone array of a PS3eye camera [52]. It features four microphones and employs technologies for echo cancellation, background noise suppression, and multidirectional sound source tracking. This allows the auditory system to be used for speech recognition, sound localization, and sound separation in noisy environments. The microphone array operates with each channel processing 16-bit samples at a sampling rate of up to 48 kHz per channel and a large dynamic range of signal-to-noise ratio up to 90 dB.

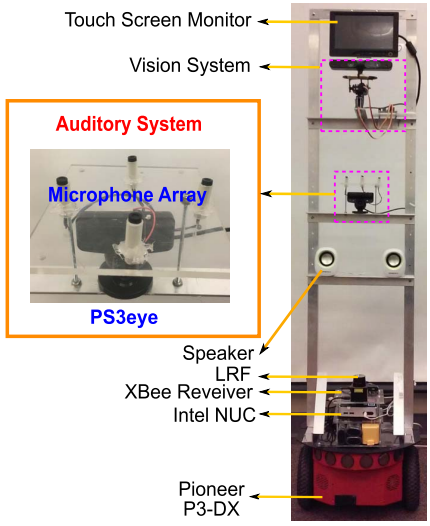

Fig. 3. Home service robot platform.

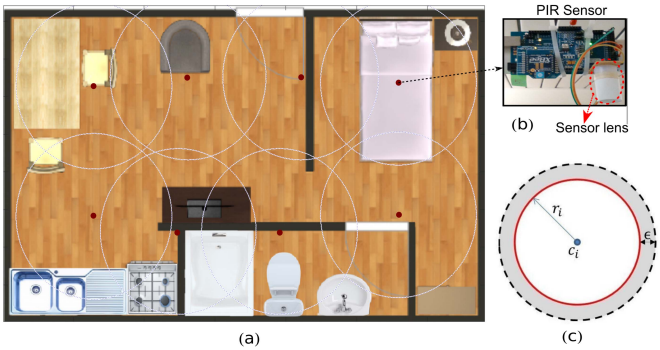

Fig. 4. (a) Configuration of the PIR network in the bestbed. (b) PIR sensor node. (c) Sensing region of a PIR node.

The software for the robot was developed on Robot Operating System (ROS) [53] which was installed in the Ubuntu operating system (OS) on the Intel Next Unit of Computing (NUC) minicomputer. We utilized exiting packages from ROS repositories to develop the device drivers that interface with the robot base, the laser rangefinder (LRF), and the RGB-D camera. Based on the drivers, we also implemented several robot services including Simultaneous Localization and Mapping (SLAM) and navigation.

B. Human Localization

The human localization module estimates the rough human location by using the passive infrared (PIR) sensor network deployed in the home environment. As shown in Fig. 4(a), the PIR sensor network consists of eight sensor nodes that are placed on the ceiling at a height of 8 feet and the coverage of each PIR sensor node is set to be a circle with a radius of 3.6 feet using a cylindrical lens cover. Data from these nodes are transmitted through the XBee protocol to the robot. Each PIR node detects the human motion inside its sensing region. Therefore, the human location is approximately estimated to be within the sensing region once the sensor gives a high output. To achieve that, a new PIR sensor observation model is developed based on the existing model in [54]. Our

new PIR sensor model is expressed as follows:

$$P\left(z_k^{\mathrm{PIR},i}|s_k\right) = \begin{cases} p^{z_k^{\mathrm{PIR},i}}(1-p)^{1-z_k^{\mathrm{PIR},i}}, & \text{if } |s_k - C_i| \le r_i \\ q^{z_k^{\mathrm{PIR},i}}(1-q)^{1-z_k^{\mathrm{PIR},i}}, & \text{if } r_i \le |s_k - C_i| \le r_i + \epsilon \\ 1 - z_k^{\mathrm{PIR},i}, & \text{if } |s_k - C_i| \ge r_i + \epsilon \end{cases} \quad (1)$$

where p is the probability of detection; q is the probability of false alarm; $z_k^{\mathrm{PIR},i}$ is the binary output {0,1} from PIR sensor i at time k; s_k is the human state which is the 2-D location; C_i and r_i are the center and the radius of the sensing region of PIR sensor i, respectively. We discovered that false alarms may occur when the human is not in the sensing range, but not too far away from the sensor, which is depicted by the gray area ϵ as shown in Fig. 4(c). Inside the gray area, those probabilities estimated from our measurements are $p = 0.9$ and $q = 0.05$. If the human is out of the dashed circle, the false alarm rate q becomes 0. However, in order to simplify the human localization task, the human's location L_H can be estimated from PIR sensors using the naive Bayes classifier

$$L_H = \arg\max_{L=L_1,\ldots,L_k}\{P(L|\mathrm{IR})\} \quad (2)$$

where $P(L|\mathrm{IR})$ is the probability of the semantic area L where the human is inside given the PIR data vector IR that is created by the outputs of all PIR sensors.

C. Modeling Human Activity Monitoring

This section presents definitions, the modeling of the human activity monitoring task and an overview of the SoHAM framework for home service robots.

1) Definitions: Actions: An action is the operation a subject does with or without an object. The set of actions in human monitoring is denoted as $S_a = \{a_1, a_2, \ldots, a_m\}$ where m is the total number of actions. The *criticalness of action* a, $Cr(a) \in (0, 1)$, reflects how important immediate attention is while an action is detected. For example, an action of "*falling on the floor*" requires immediate attention.

Activities: A human activity is usually composed of a sequence of actions with temporal constraints [55]. The set of activities in human monitoring is denoted as $S_A = \{A_1, A_2, \ldots, A_M\}$ where M is the total number of activities. The activities can be daily activities (*eating, cooking, using the toilet, sleeping, having a shower, watching television (TV)*, etc.) and abnormal activities (*coughing, crying for help, falling on the floor*). The *criticalness of activity* A, $Cr(A) \in (0, 1)$, reflects how important immediate attention is while an activity is detected.

Monitoring Task: The monitoring task is to find out an estimate of the most likely action a or activity A based on observed data $D_{1:n} = [d_1, d_2, \ldots, d_n]$ collected by the microphone sensors and the distributed PIR sensors. The estimate is typically a posterior probability distribution $p(a|D_{1:n})$ or $p(A|D_{1:n})$, from which a decision can be made.

Quality of Monitoring (QoM): The quality of monitoring is a measure of the confidence of decision regarding the current human action or activity. In this work, we mainly evaluate the quality of action monitoring QoM(a). This measure can be defined as a function of the entropy of predicted actions $a = (a_1, a_2, \ldots, a_m)$ as follows:

$$\begin{aligned} \mathrm{QoM}(a) &= 1 - H(a|D_{1:n})/\log_2 m \\ &= 1 + \left[\sum_{i=1}^{m} p(a_i|D_{1:n}) \log_2(p(a_i|D_{1:n}))\right] / \log_2 m \end{aligned} \quad (3)$$

where m is the number of actions defined previously. This function maps the QoM(a) to a value in $[0, 1]$. When each action is predicted with the equal confidence $P(a_i|D_{1:n}) = 1/m$, which means the robot has no knowledge regarding the actions a, $H(a|D_{1:n}) = H_{\max} = \log_2 m$ [56] and QoM(a) = 0. When any action a_i is recognized at the confidence $P(a_i|D_{1:n})$ of 1.0, $H(a|D_{1:n}) = 0$ and QoM(a) = 1. Maintaining a certain level of quality of monitoring QoM(a) requires that the entropy of predicted actions a be less than a certain threshold Hth, or in other words, the robot is confident about the predicted actions or activities.

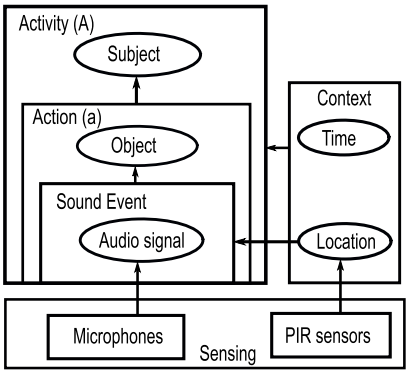

Fig. 5. Generic model of sound-based activity monitoring.

2) Generic Model of Sound-Based Activity Monitoring: Inspired by the work [57], we develop a generic model of the SoHAM as shown in Fig. 5. This model shows the relationship between the components of the SoHAM. Sound events are recognized by fusing audio signal and contextual information estimated from the human location. Action is determined by a sound event and an object. However, some actions have no object involved, such as *falls, cough, heavy-breathing*. To simplify the action recognition task, in this work, we label each sound event by the name of corresponding action. Activities can be recognized from sequences of actions that occur in a certain location and at a certain time. Some activities can be recognized by only one action, such as *watching TV, falls, and drinking*. Therefore, the activity monitoring task is to recognize both actions and activities in each location in a home environment over the time.

3) Overview of the SoHAM Framework: We propose a SoHAM framework as shown in Fig. 6. The framework consists of three modules: CoSER, dynamic sliding time window-based human action recognition (DTW-HaR), and CRF-based human activity monitoring (CRF-HAM). The CoSER utilizes human location data as the context information to improve the performance of sound event recognition. The DTW-HaR post-processes the sequence of recognized sound event labels from the CoSER to filter out false negative and

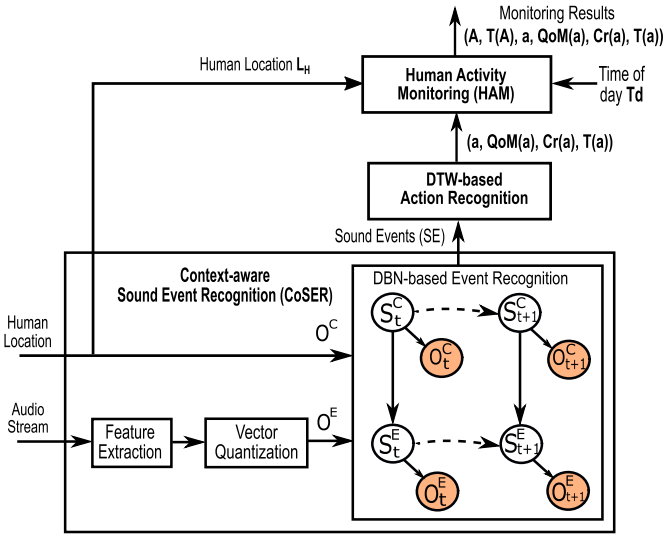

Fig. 6. SoHAM framework.

false positive, detect acoustic event segments with their labels, durations, and confidences, then recognizes actions based on such acoustic event segments. The CRF-HAM uses location, time, and actions to recognize human activities. The monitoring results are the tuple $(A, T(A), a, \text{QoM}(a), Cr(a), T(a))$, where A is the current activity of the human; $T(A)$ is the duration of the activity; a is the current action; $\text{QoM}(a)$ is the quality of monitoring with respect to action a; $Cr(a)$ is the criticalness of action a; $T(a)$ is the duration of the action. Such monitoring results enable the robot to make proper decisions in taking care of the elderly living alone in their homes.

IV. CoSER Using DBNs

In this work, the context of sound events defined as the semantic area O_C where the action occurs, which is the human location L_H that is estimated through the PIR sensors as discussed in Section III-B. We propose the CoSER framework as shown in Fig. 6. The framework features four main modules: 1) feature extraction, 2) vector quantization (VQ), and 3) DBN-based event recognition. The audio stream captured by the auditory system on the robot is decomposed into frames. These frames are extracted into feature vectors which are then assigned the labels of codewords by the VQ. The sound event observations O^E are fused with their contextual information O^C using the DBN framework for final labels at the frame level.

A. Feature Extraction

Statistical features are calculated for each audio frame. The length of the frame ranges from 5 to 150 ms and the overlap between two adjacent frames is set to half of the frame size. We use the 31-dimension feature vectors that consist of 12 MFCCs [58], 12 delta-MFCCs, log energy, zero cross rate, entropy, centroid, spread, skewness, and kurtosis [59].

B. Vector Quantization

VQ is to compress a dataset into a small set of representatives, which reduces the space to store data but still maintains sufficient information. Given a k-dimension feature vector $x \in R^k$ in a vector space S, after VQ, x is classified to a vector subspace S_j

$$q(x) = S_j \tag{4}$$

where $q(.)$ is the quantization operator. The whole vector space is $S = S_1 \cup S_2 \cup \cdots \cup S_M$. Each partition S_j forms a nonoverlapping region and is characterized by its centroid vector z_j. Set $Z = \{z_1, z_2, \ldots, z_M\}$ is called a codebook and z_j is the jth codeword that is labeled as j. M is the size of the codebook. The error between a vector and a codeword $d(x, z)$ is called the distortion error. A vector is always assigned to the region with the smallest distortion. Therefore, the observed feature vector can be assigned to the label of a codeword as follows:

$$O^E = \arg\min_{j=1,2,\ldots,M}\{d(x, z_j)\}. \tag{5}$$

Linde–Buzo–Gray (LBG) algorithm [60] is applied to design the codebook in VQ. The LBG algorithm is a hierarchical clustering algorithm, which first starts with a 1-level codebook, then uses a splitting method to obtain a 2-level codebook, and continues until an M-level codebook is acquired.

C. Hierarchical Context and Sound Event Model

To represent the sound event-human location correlation, the given map of a home environment is segmented into different areas with corresponding probabilities of sound events. In the time domain, the transition of human locations follows certain patterns. For example, the human always walks from one area to another adjacent area and there is a probability distribution according to the floor plan and personal preference. We assume the transition of locations is a discrete and first-order Markov process. In this article, the locational contexts that are to be recognized are the rooms in a house, including the living room, dining room, kitchen, bathroom, and bedroom. The locations of the human given by the PIR-based human localization module are mapped into N_L semantic areas.

In an indoor environment, human locations and sound events have both intratemporal causal relationships and intertemporal constraints, which are modeled by the two-level DBN model shown in Fig. 6. The individual nodes in this graphical model represent hidden states and the shaded nodes represent observations. The solid arcs correspond to causal dependencies between the nodes in the same time slice, while the dashed arcs correspond to the temporal dependencies between two slices at time t and $t + 1$.

The higher level of the model represents the hidden states of sound context S^C. The lower level represents the hidden states of sound events S^E. As discussed above, the human locations estimated from the PIR network are clustered into the context observation O^C. The audio stream captured by the robot is decomposed into frames which are extracted into feature vectors and then assigned to the labels of codewords by the LBG-based VQ. These labels are used as the observation O^E of event sounds. In our model, the dependencies between the nodes in the DBN include both spatial and temporal

components. The observation O_t^C and O_t^E are dependent on the corresponding intratemporal hidden state S_t^C and S_t^E. The context S_{t+1}^C at time $t+1$ depends on the previous context at time t. The sound event S_{t+1}^E at time $t+1$ depends on the previous sound event at time t, the context at time $t+1$.

D. Implementation of the DBN

1) Mathematic Representations: In the two-level DBN model, the superscript of states and observations represents the levels including the locational context (top) and the sound event (bottom), while the subscript represents the time index. Each level has three basic elements as follows:

a) The state transition probability distribution: The state transition probability distribution in each level reflects the intratemporal dependence in Fig. 6.

The top level location transition probability represents the topology of the layout of the home environment

$$a_{i,j}^C = P\big(S_{t+1}^C = j | S_t^C = i\big). \tag{6}$$

The bottom level sound event transition probability depends on the context

$$a_{i,j,p}^E = P\big(S_{t+1}^E = j | S_t^E = i, S_{t+1}^C = p\big). \tag{7}$$

b) The observation symbol probability distribution: Since the observed variables only depend on the corresponding states in the same level, the observation symbol probability distribution can be expressed as follows:

$$b_{i,j}^C = P\big(O_t^C = j | S_t^C = i\big) \tag{8}$$

$$b_{i,j}^E = P\big(O_t^E = j | S_t^E = i\big). \tag{9}$$

c) The initial state distribution: Since the intratemporal dependence exists from the beginning of the sequence, the initial state distribution also follows the relationship of the links between levels in Fig. 6

$$\pi_i^C = P\big(S_1^C = i\big) \tag{10}$$

$$\pi_{j,i}^E = P\big(S_1^E = j | S_1^C = i\big). \tag{11}$$

The above probability distributions are trained by using Expectation Maximization (EM) as proposed in [61].

The Viterbi algorithm [62] is applied to estimate the probability of state sequence given the observation sequence recursively.

2) Short-Time Viterbi Algorithm for Online Smoothing: The standard Viterbi algorithm retrieves the state sequence, which maximizes the belief value [62]. The retrieved state sequence has the maximum likelihood given the observation sequence from time 1 to T. The computational complexity of the standard Viterbi algorithm makes it not suitable for real-time implementation. The short-time Viterbi algorithm can solve this problem and enhance the efficiency [63], which has three steps: initialization, recursion, and path smoothing.

Initialization

$$\delta_1(i,j) = P\big(S_1^C = i\big) P\big(O_1^C | S_1^C = i\big) P\big(S_1^E = j | S_1^C = i\big) P\big(O_1^E | S_1^E = j\big) \tag{12}$$

$$\psi_1(i,j) = [0,0]. \tag{13}$$

Algorithm 1 Short-Time Viterbi for Smoothing in DBN

Initial Viterbi sequence length T, δ_1, and ψ_1 using (12) and (13)
for each new observation $O_t = [O_t^C, O_t^E]$ **do**
- Obtain $\delta_t(i,j)$ and $\psi_t(i,j)$ using (14) and (15)
- Obtain current state estimate q_t^* from $\delta_t(i,j)$ using (16)
- Backward one step and calculate the path (previous state estimate) using (17)
- Correct previous state output if q_{t-1}^* changes
- Save current $\delta_t(i,j)$ for next loop.

end for

Recursion

$$\begin{aligned}\delta_t(i,j) &= \max_{p,q}\big[\delta_{t-1}(p,q) a_{pi}^C b_{pi}^C a_{qij}^E b_{qj}^E\big] \\ &= \max_{p,q}\big[\delta_{t-1}(p,q) P\big(S_t^C = i | S_{t-1}^C = p\big) b_{pi}^C \\ &\qquad P\big(S_t^E = j | S_{t-1}^E = q, S_t^C = i\big) b_{qj}^E\big]\end{aligned} \tag{14}$$

$$\psi_t(i,j) = \arg\max_{p,q} \delta_t(i,j) \tag{15}$$

$$q_t^* = \arg\max_{i,j} \delta_t(i,j). \tag{16}$$

Path Smoothing

$$q_{t-1}^* = \psi_t\big(q_t^*\big). \tag{17}$$

The short-time Viterbi algorithm is shown in Algorithm 1.

V. Dynamic Sliding Time Window (DTW)-Human Action Recognition

As discussed above, the CoSER receives the audio stream of sound events and the locational context information to recognize the event labels. The recognition results are obtained for every sequence of T consecutive frames based on the maximum likelihood given the sequence of audio frames. In real-time monitoring, in order to reduce the computational complexity of state search in the DBN inference, the sequence size T is set to be fixed and small. Due to the diversity of audio signals, environmental noise, and reverberation, the sound event recognition results have false positive and false negative labels. Therefore, the robot has to improve the confidence of its decisions. To this end, we implemented the DTW-HaR. The DTW-HaR divides the sequence of sound event labels into acoustic event segments based on the assumption that the time window of an acoustic event is greater than a minimal time window $T_{\min}^a$ and the gap between active acoustic events is greater than a minimal quiet time window $T_{\min}^g$. In other words, every recognized acoustic event that has a duration less than $T_{\min}^a$ is considered as false positive and every recognized gap with a length less than $T_{\min}^g$ is considered as false negative. Based on the dynamic sliding time window approach, the DTW-HaR detects the duration of the current action. Then the probabilities of sound event labels in this duration are estimated and the activity label is assigned to the sound event that has the maximum probability. Finally, the criticalness of action and the quality of monitoring are

Algorithm 2 Dynamic Time Sliding Window for Human Action Recognition

+ Input: Sequence of sound event SE
+ Output: $(a, QoM(a), Cr(a), T(a))$
Initialize a, $T(a)$, T^a_{min}, T^g_{min}, T_{step}
Initialize the sliding window W_a and the observed sequence of sound event labels SE
for each new sequence SE of T_{step} labels from the CoSER **do**
  Update SE and W_a:
  $SE = [SE \;\; S_E]$
  $W_a = [W_a \;\; W_E]$ where W_E is the time window of SE
  {***Remove false positives***}
  Detect sound event segments $\{D_a\}$ in the sliding window W_a
  for each segment d_i in $\{D_a\}$ **do**
    if length$(d_i) < T^a_{min}$ **then**
      $W_a(d_i) = 0$
    end if
  end for
  {***Remove false negatives***}
  Detect background segments $\{D_g\}$ in the sliding window W_a
  for each segment d_i in $\{D_g\}$ **do**
    if length$(d_i) < T^g_{min}$ **then**
      $W_a(d_i) = 1$
    end if
  end for
  $a = \arg\max_{a=a_1,\ldots,a_m}\{P(a|SE(W_a))\}$
  Compute $Cr(a) = CrMatrix(a, Ts)$
  Compute $QoM(a)$ using 18
  if *full event detected* **then**
    Update $T(a) = length(W_a)$
    Store $(a, QoM(a), Cr(a), T(a))$
    Reset a, W_a, $T(a)$, SE
  end if
end for

computed. The quality of monitoring is computed as follows:

$$\begin{aligned}\text{QoM}(a) &= 1 - H(a|D_{1:n})/\log_2 m \\ &= 1 + \left[\sum_{i=1}^{m} P\left(a_i|O^C_{1:T_a}, O^E_{1:T_a}\right)\right. \\ &\quad \left.\log_2\left[\left(P\left(a_i|O^C_{1:T_a}, O^E_{1:T_a}\right)\right]\right/\log_2 m \qquad (18)\end{aligned}$$

where m is the number of activities of interest defined previously; $T(a)$ is the duration of action.

The criticalness of activity is computed based on the criticalness matrix of activities over the time $\text{CrMatrix}(a, t)$, which is predefined. For example, $Cr(\text{fall}) = 1$, $Cr(\text{having breakfast in the morning}) = 0.5$, and $Cr(\text{having shower in the evening}) = 0$. The online monitoring results are updated over the time window step T_{step} based on Algorithm 2.

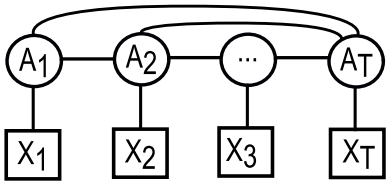

Fig. 7. Conditional random filed-based activity monitoring; A_i and X_i are the activity hidden state and the observation at the time step i, respectively.

VI. CRF-Based Activity Monitoring (CRF-HAM)

In this section, we develop a method that fuses location, time, sequences of actions to recognize human activities.

Recognizing human activities is challenging because human activities are complex, diverse, interleaved, and ambiguous [64]. The order of actions to perform an activity is different for different people. In practice, it is common that a person stops an activity to do another one then resumes the previous one again. Similar actions can be interpreted as different activities. Recognizing activities using sound events is even harder. The same sound event may be generated by different activities. For example, a *water-running* sound event can belong to several activities, such as *washing-hands*, *washing-dishes*, or *filling-water*. Thus, the models for sound-based activity recognition must address these challenges. In practice, many activities may have nondeterministic natures [57], where some actions of an activity may be performed in any order and not every action generates sound. In this work, we design a CRF model [65] to implement human activity monitoring.

CRF [65] has recently gained popularity in activity recognition [16], [66]. CRF is a discriminative model that uses an undirected graphical model to represent conditional probability $P(Y|X)$ of a sequence of labels Y (hidden states) given a sequence of observations X. CRF can model any dependencies between different labels and flexibly capture arbitrary dependences amongst observation variables. This model can capture long-range dependences, therefore, has a potential to recognize complex activities. Additionally, CRF does not consider the order constraint of the hidden states, therefore, can be used for human activity recognition, where many activities may have nondeterministic features, i.e., some actions can be conducted in any order with ambiguous and interleaved activities. We implement a skip-chain CRF model for activity recognition based on sound events, which is shown in Fig. 7. This model represents the conditional likelihood $P(A|X)$ of the sequence of activity labels A, $A = [A_1, A_2, \ldots, A_T]$, given the sequence of observations X, $X = [X_1, X_2, \ldots, X_T]$, where X_i includes the human location L_i, the time of the day Td_i, the recognized action a_i, and its duration da_i, thus, $X_i = [x_1, x_2, ..., x_k] = [L_i, Td_i, a_i, da_i]$. The activity labels are predicted as follows:

$$A^* = \arg\max_A\{P(A|X)\}. \qquad (19)$$

The probability $P(A|X)$ is computed as follows:

$$P(A|X) = \frac{1}{Z_X}\prod_{t=1}^{T}\Psi_t(A_t, A_{t-1}, X)\prod_{t=1}^{T-1}\Psi_{tT}(A_t, A_T, X) \qquad (20)$$

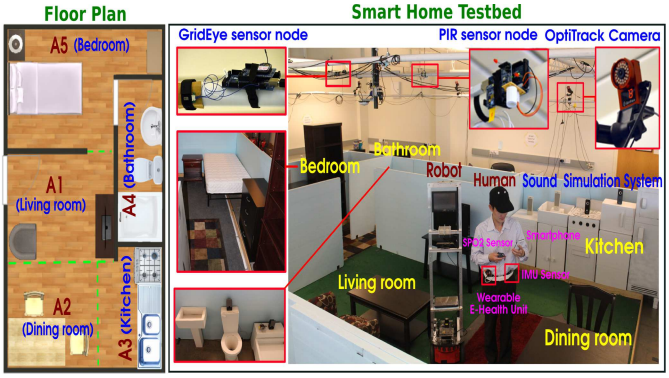

Fig. 8. Floor plan (left) and the smart home testbed (right).

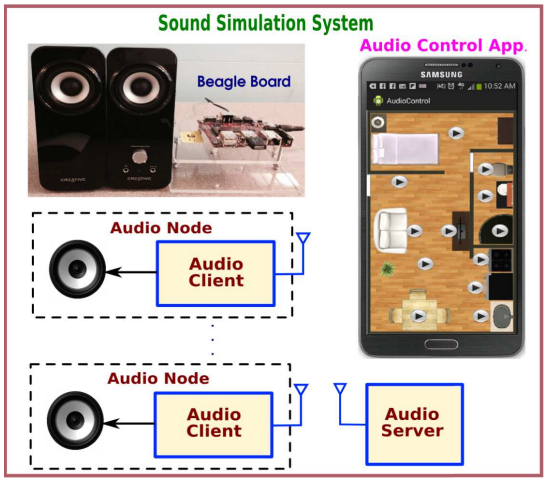

Fig. 9. Sound simulation system.

where Z_X is the normalization constant computed by aggregating the numerator as shown in 20; Ψ_t is the potential function that considers the relationship between observations over a period of time and activities; and Ψ_{tT} is the transition feature function that considers the relationship between past activities and future activities; These functions are computed as follows:

$$Z_X = \sum_{A} \prod_{t=1}^{T} \Psi_t(A_t, A_{t-1}, X) \prod_{t=1}^{T-1} \Psi_{tT}(A_t, A_T, X) \tag{21}$$

$$\Psi_t(A_t, A_{t-1}, X) = \exp\left\{ \sum_{k}^{K_t} \lambda_k f_k(A_t, A_{t-1}, X, t) \right\} \tag{22}$$

$$\Psi_{tT}(A_t, A_T, X) = \exp\left\{ \sum_{k}^{K_f} \mu_k g_k(A_t, A_T, X, t, T) \right\} \tag{23}$$

where $\theta_t = \{\lambda_k\}_{k=1}^{K_t}$ and $\theta_f = \{\mu_k\}_{k=1}^{K_f}$ are the parameters of the proposed skip-chain CRF; f_k and g_k are the feature function and transition function for each state pair and each observation-state pair, which are defined as follows:

$$f_k(A_t, A_{t-1}, X, t) = \delta(A_t, \tilde{A}_t)\delta(A_{t-1}, \tilde{A}_{t-1})q_k(X_t) \tag{24}$$

$$g_k(A_t, A_T, X_t, t, T) = \delta(A_t, \tilde{A}_t)\delta(A_T, \tilde{A}_T)q'_k(X_t, X_T) \tag{25}$$

where $\delta(y, i)$ is the indicator function and takes the value 1 if $y = i$ and 0 otherwise; $\tilde{A}_t$ is the single output configuration value of A_t; $q_k(X_t)$ is the feature representation function of X_t; $q'_k(X_t, X_T)$ is the feature representation function that combine the features of X_t and X_T. The parameters of the model are trained based on maximum likelihood technique [67].

VII. Experiments and Results

We conducted physical experiments to test and evaluate the proposed framework. A smart home testbed was set up in our lab which has an area of 16 ft × 22 ft as shown in Fig. 8. It simulates a small apartment, which includes a living room, a bedroom, a kitchen, a dining room, and a bathroom. A sound simulation system was developed to simulate the multiple sound events like those heard in a typical house [68]. The sound simulation system includes multiple audio nodes, an audio server, and an audio control application as shown in Fig. 9. The audio nodes were developed using Beagleboard minicomputers and speakers. The sound events in the bathroom, kitchen, dining room, living room, and bedroom were recorded in real environments and collected from the Google AudioSet [69] and freesound.org. These sound samples form the ASCC sound library (ASCCsoundLib). Currently, it has 43 sound events and each has 3 or 4 clips, including 6 dining room sounds, 5 living room sounds, 5 bathroom sounds, 16 kitchen sounds, 2 bedroom sounds, and 12 other sounds. Sound events and their labels are as follows:

Dining Room: *2:preparing-dining-table, 3:using-fork-spoon-plate, 4:eating (cereal, pizza, snack), 5:pouring-water-in-glass, 6:drinking-water, 7:clean-dining-table.*

Living Room: *5:pouring-water-in-glass, 6:drinking-water, 8:opening-door, 9:closing-door, 10:TV.*

Bathroom: *11:brushing-teeth, 12:having-shower, 13:washing-hands, 14:flushing-toilet, 15: filling-sink.*

Kitchen: *16:using-faucet, 17-using-blender, 18-sifting-flour, 19-chopping, 20-boiling (water, cooking), 21-frying-pan, 22:teapot-whistle, 23:mixing-sauce, 24:controlling-microwave, 25:microwave-running, 26:closing-microwave, 27:opening-microwave, 28:making-coffee, 29:cooker-hood, 30:dish-washer-machine, 31:washing-dishes.*

Bedroom: *32:heavy-breathing-sleeping, 33:snoring.*

Other Sounds: *1: background-sound, 34-laughing, 35:eructation, 36:heavy-breathing, 37:cough, 38:yawning, 39:footsteps, 40:speaking, 41:falling-sounds, 42:dropping, 43:vacuum, 44:other-sounds.*

The audio control program on a smartphone can trigger the playback of a sequence of sound events associated with human activities or multiple simultaneous sound events using different audio nodes placed at different locations. This approach allows the robot to collect sound data with strong labels quicker than manual annotation. We conducted experiments in our smart home testbed to evaluate the CoSER, DTW-HaR, and CRF-HAM as well as show the capabilities of the home service robot in monitoring the elderly in the smart home environment. We moved the robot around the home testbed to collect audio data at different locations with different setups. The robot was stationary when it recorded sound events. The experimental audio data were collected with 5 graduate student subjects doing various activities in the testbed multiple times.

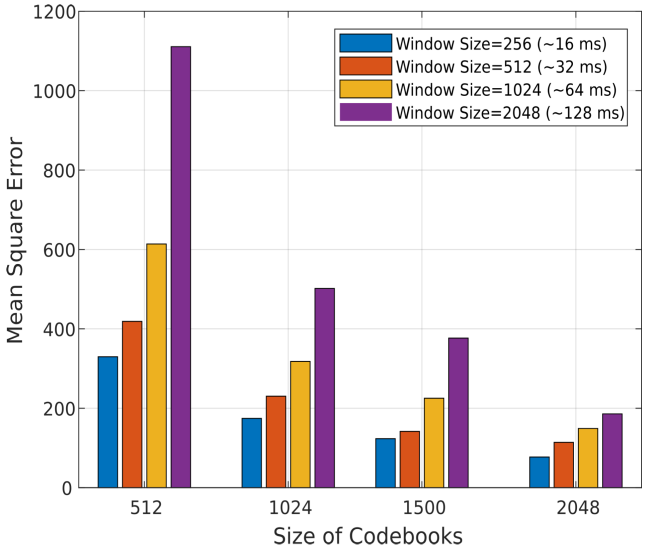

Fig. 10. Mean square distortion error of the codebooks.

The audio control program on the smartphone triggered the playback of some event sounds which cannot be collected in the home environment, such as flushing the toilet. Actions happened from about 0.5 s (e.g., falling, closing-microwave, etc.) to around 30 s (e.g., preparing-dining-table, and having a shower). These audio data were recorded by the robot at a sampling rate of 16 kHz. The total recording time is approximately 200 min.

A. Context-Aware Sound Event Recognition

The above experimental audio data were used to train the codebooks with different codebook sizes and window sizes. The sliding steps were set to half of the window sizes. As shown in Fig. 10, the mean square distortion errors of codebooks decrease almost linearly with the numbers of codewords in the codebooks but increase with the window sizes. However, using a large number of codewords and the small number of window size incurs more computation. Those parameters also affect the accuracy of sound event recognition. Therefore, they were also considered in the experiments to evaluate the CoSER.

We also used the above recorded audio files and context observations provided by the PIR network to train and evaluate the DBN using threefold cross-validation. The sound event data were partitioned into three equal-sized subsets. All frames in the same event have the same label. Sequentially one subset was tested using the DBN trained on the other 2 subsets. The cross-validation accuracy is the average of the rate of audio frames that are correctly classified in 3 testing subsets. To evaluate the performance of the proposed sound event recognition model, evaluation measures are computed from the confusion matrices. These measures include the true positive TP, true negative TN, false positive FP, false negative FN, recall R, precision Pr, $F1$ score F_1, accuracy acc, and overall accuracy ACC—the average accuracy of the total number $N_$tests of tested instances for every classes.

In order to evaluate the context-aware SER (CoSER), context-independent SER was implemented by using the

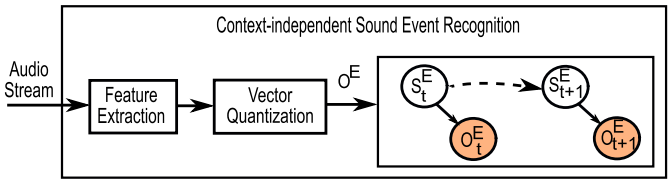

Fig. 11. Context-independent sound event recognition.

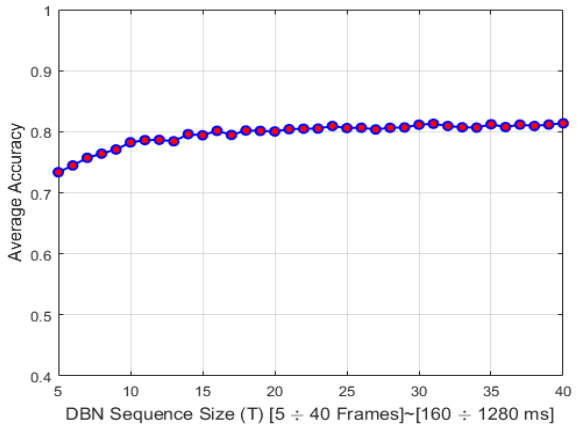

Fig. 12. Average accuracy of the CoSER with respect to the sequence sizes.

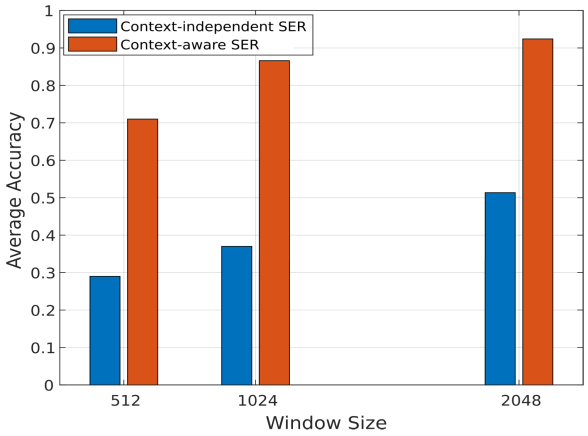

Fig. 13. Average accuracy of CoSER with respect to the window sizes.

one-level DBN as shown in Fig. 11. In this model, O_t^E is dependent on the corresponding intratemporal hidden state S_t^E. The sound event state S_{t+1}^E at time $t+1$ depends on the previous sound event state at time t. The testing data were used to test both SERs with the same parameters of the VQ and DBNs. In order to reduce the effect of the codebook's distortion error on the SER performance, the codebook size was set to 2048. Therefore, the numbers of states S^E, O^E, S^C, and O^C are 44, 2048, 5, and 5, respectively.

We also evaluated the SERs on different DBN parameters which include the sliding step and Viterbi sequence length. First, the window size was set to 512 and the sequence length was changed from 5 to 50. The average accuracy of the CoSER with respect to the sequence sizes is shown in Fig. 12. The accuracy increases slightly as the sequence length increases. The longer sequence is, the more confidence the CoSER has in long-time sound events such as cough sounds and bushing-teeth sounds, but less confidence in short-time sound events such as falling sounds. In order to test the SERs in

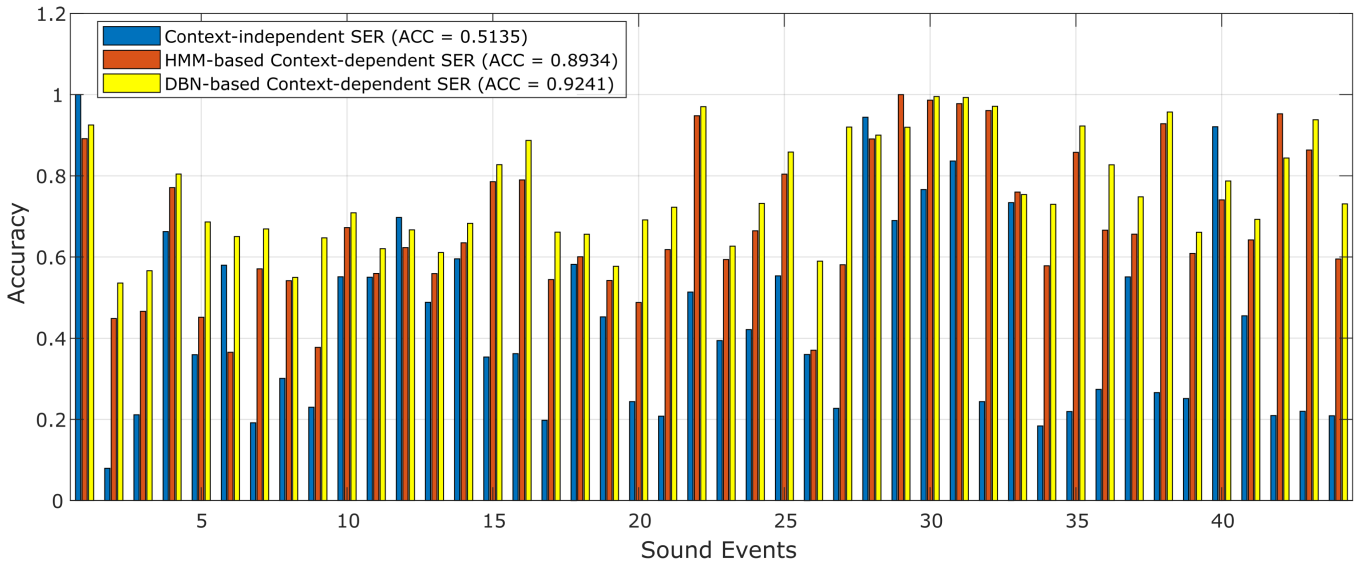

Fig. 14. Recognition accuracy of sound events at the frame level of the context-independent SER, HMM-based CoSER, and DBN-based CoSER with the window size of 1024 and the codebook size of 2048.

real-time, the Viterbi sequence length is set to 10, which corresponds to 160, 320, or 640 ms, given the window step is 256, 512, or 1024, respectively. The overlap between the two adjacent sequences is 1. The average accuracy of both SERs at the frame level is shown in Fig. 13. The context-aware SER produces much better results than the context-independent SER. Its average accuracy is about 70% with a window size of 512 and achieves more than 90% with a window size of 2048. Although larger window sizes have less time-domain resolution, they give better frequency resolution, which is more efficient in classifying sound events that have a wide range of frequencies. The recognition accuracy of sound events at the frame level with a window size of 1024 is shown in Fig. 14. While the context-independent SER has poor performance, the CoSER produces the results at accuracy rates of more than 80% for over half of the sound events. Only five sound events have an accuracy rate of between 50% and 50%. Besides, the evaluation result of each class is shown in Table I.

In addition, the recognition accuracy of DBN-based CoSER was compared with that of the HMM-based CoSER that was proposed in [50]. The HMMs were implemented and trained for each context by threefold cross-validation as the DBN model. As shown in Fig. 14, the DBN-based method has a better performance in more than two thirds of sound events and has a higher average accuracy rate than the HMM-based method.

B. DTW-Based Human Action Recognition (DTW-HaR)

The DTW-HaR was evaluated by the above testing data with a window size of 1024, a cookbook size of 2048, and a Viterbi sequence length of 10. We tested the DTW-HaR with different values of $T^a_{\min}, T^g_{\min}, T_{\text{step}}$. We found the best values via grid-search on $T^a_{\min}, T^g_{\min} = 5, 6, 7, 8, 9, 10, 11, 12$ and $T_{\text{step}} = 10, 12, 14, 16, 18, 20, 22, 24$. Various settings are tried and the best one is $(T^a_{\min}, T^g_{\min}, T_{\text{step}}) = (9, 9, 18)$. The average accuracy of action recognition reaches approximately 95.62% at the frame level. Moreover, the robot can update the

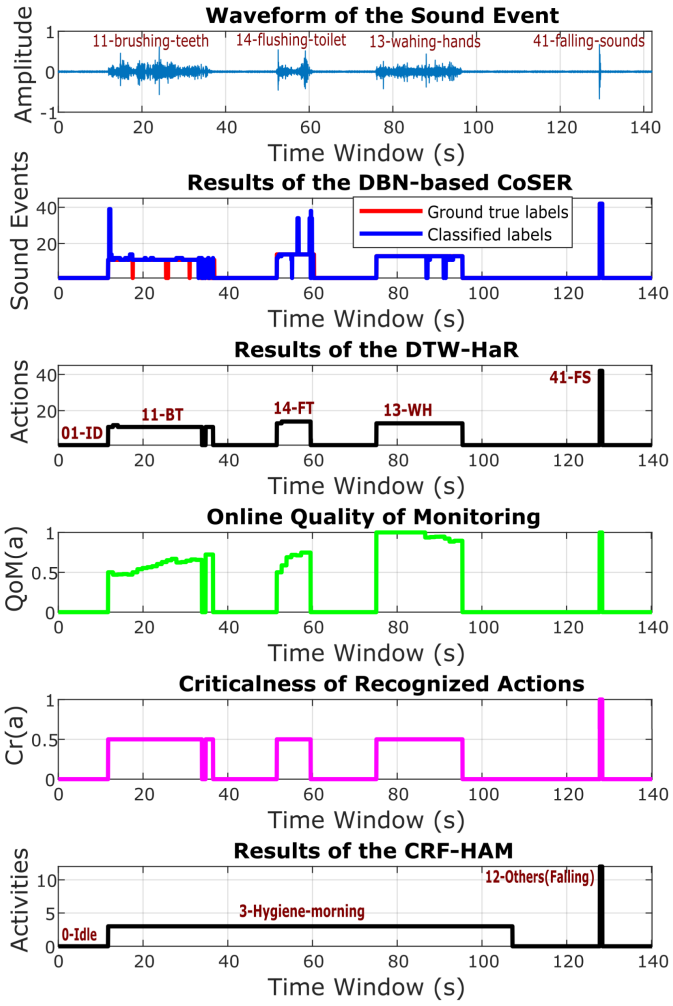

Fig. 15. Monitoring results of the SoHAM.

QoM(a) and $Cr(a)$ in realtime after every 18 frames, which corresponds to 576 ms. As shown in Fig. 15, the DTW-HaR produces smoother sound event segments. Each segment is

TABLE I
EVALUATION MEASURES OF COSER PERFORMANCE (TOTAL NUMBER OF TESTED AUDIO FRAMES: 378507)

Sound Event/Action	Action Index	TP	FN	FP	R	Pr	F1	acc
01-background-sound/01-Idle	01-ID	190225	8007	7395	0.960	0.963	0.961	0.925
02-preparing-dining-table	02-PDT	403	238	111	0.629	0.784	0.698	0.536
03-using-fork-spoon-plate	03-UFSP	1138	287	585	0.799	0.660	0.723	0.566
04-eating (cereal, pizza, snack)	04-ES	13121	786	2403	0.943	0.845	0.892	0.804
05-pouring-water-in-glass	05-PWG	2691	645	586	0.807	0.821	0.814	0.686
06-drinking-water	06-DW	898	168	315	0.842	0.740	0.788	0.650
07-clean-dining-table	07-CDT	669	254	77	0.725	0.897	0.802	0.669
08-opening-door	08-OD	182	123	26	0.597	0.875	0.710	0.550
09-closing-door	09-CD	143	68	10	0.678	0.935	0.786	0.647
10-television	10-TV	1924	528	263	0.785	0.880	0.829	0.709
11-brushing-teeth	11-BT	3929	1177	1227	0.769	0.762	0.766	0.620
12-having-shower	12-HS	5779	1054	1835	0.846	0.759	0.800	0.667
13-washing-hands	13-WH	2472	869	705	0.740	0.778	0.759	0.611
14-flushing-toilet	14-FT	4705	969	1218	0.829	0.794	0.811	0.683
15-filling-sink	15-FS	2787	344	237	0.890	0.922	0.906	0.827
16-using-faucet	16-UF	3498	127	318	0.965	0.917	0.940	0.887
17-using-blender	17-UB	1619	566	264	0.741	0.860	0.796	0.661
18-sifting-flour	18-SF	3747	206	1759	0.948	0.681	0.792	0.656
19-chopping	19-CH	1662	375	843	0.816	0.663	0.732	0.577
20-boiling (water, cooking)	20-BL	743	222	110	0.770	0.871	0.817	0.691
21-frying-pan	21-FP	2082	588	212	0.780	0.908	0.839	0.722
22-teapot-whistle	22-BL	3580	100	10	0.973	0.997	0.985	0.970
23-mixing-sauce	23-MS	1450	570	294	0.718	0.831	0.770	0.627
24-controlling-microwave	24-CTM	393	38	106	0.912	0.788	0.845	0.732
25-microwave-running	25-MR	2185	246	114	0.899	0.950	0.924	0.859
26-closing-microwave	26-CM	138	43	53	0.762	0.723	0.742	0.590
27-opening-microwave	27-OM	230	20	0	0.920	1.000	0.958	0.920
28-making-coffee	28-MC	8620	410	546	0.955	0.940	0.947	0.900
29-cooker-hood	29-CH	4000	350	0	0.920	1.000	0.958	0.920
30-dish-washer-machine	30-DWM	3930	10	9	0.997	0.998	0.998	0.995
31-washing-dishes	31-WD	4780	10	25	0.998	0.995	0.996	0.993
32-heavy-breathing-sleeping	32-HBS	1845	20	35	0.989	0.981	0.985	0.971
33-snoring	33-SN	477	72	84	0.869	0.850	0.859	0.754
34-laughing	34-LA	7332	2003	715	0.785	0.911	0.844	0.730
35-eructation	35-EW	11400	950	6	0.923	0.999	0.960	0.923
36-heavy-breathing	36-HB	3431	454	263	0.883	0.929	0.905	0.827
37-cough	37-CO	6790	1560	728	0.813	0.903	0.856	0.748
38-yawning	38-YA	6100	50	224	0.992	0.965	0.978	0.957
39-footsteps	39-FS	4525	1105	1218	0.804	0.788	0.796	0.661
40-speaking	40-SP	10876	859	2086	0.927	0.839	0.881	0.787
41-falling-sounds	41-FS	1433	189	447	0.883	0.762	0.818	0.693
42-dropping	42-DR	800	125	23	0.865	0.972	0.915	0.844
43-vacuum	43-VC	15085	460	537	0.970	0.966	0.968	0.938
44-other-sounds	44-OS	5958	1487	710	0.800	0.894	0.844	0.731
Overall accuracy ACC: 0.9241								

labeled with the action index which corresponds to the dominant sound event in that segment. All action indexes are shown in the second column of Table I. The DTW-HaR can produce the sequences of actions with their duration, quality of monitoring QoM(a), and criticalness of action $Cr(a)$ as shown in Fig. 15.

C. CRF-Based Activity Monitoring

This section evaluates the performance of the CRF-HAM using the above sound dataset. We simulated the activities in the simulated smart home testbed, which are composed of sequences of actions that generate sound events. The activities are 12 popular daily activities: *0-Idle, 1-Leaving-home, 2-Entering-home, 3-Hygiene-morning (personal hygiene in the morning), 4-Preparing-breakfast, 5-Having-breakfast, 06-Preparing-meal, 7-Having-lunch, 8-Having-dinner, 9-Watching-TV, 10-Cleaning, 11-Hygiene-evening (personal hygiene in the evening), 12-Others*, where *0-Idle* is no action

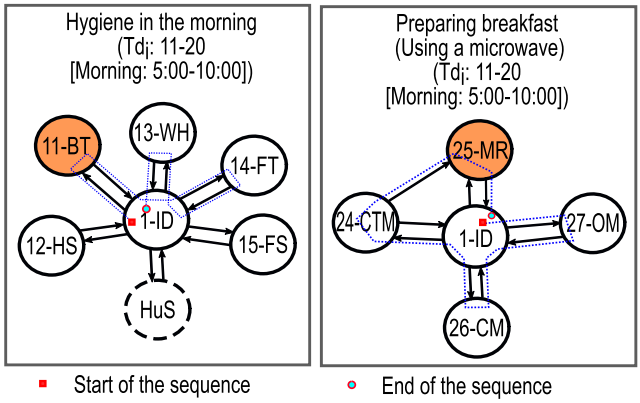

Fig. 16. Examples of semantic activity graphs.

and *12-Others* is the activity that consists of single action only. We defined each of the 10 other activities by using a semantic activity graph that represents the likely transitions

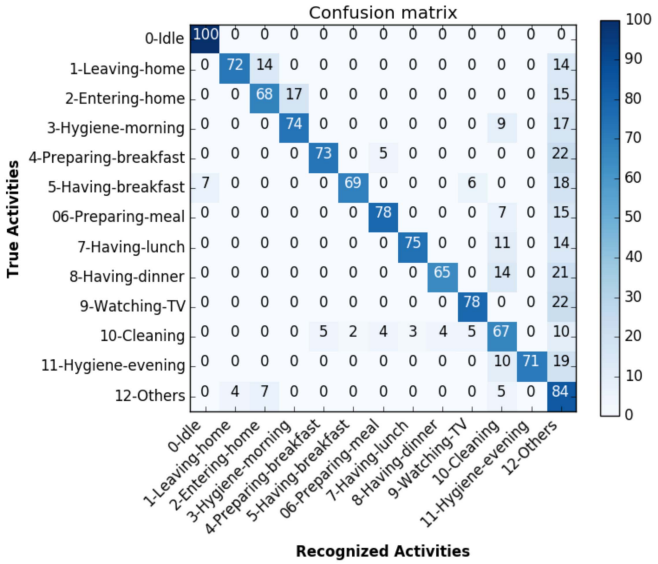

Fig. 17. Confusion matrix of the activity recognition (overall accuracy ACC = 0.75).

between actions. For example, the semantic activity graphs of *3-Hygiene-morning* and *4-Preparing-breakfast (Using a microwave)* is are shown in Fig. 16, where a shaded node represents a core action that is essential for identifying an activity; a dash node represents an unexpected action; and the other nodes represent those actions that are not essential. An example of the unexpected action in the activity of hygiene in the morning is whatever generates a human sound event (HuS) such as *cough, footsteps, speaking, breathing, snoring, or laughing*. We randomly generated 100 sample sequences for each activity based on its graph. The time of the day Td_i is divided into the discretized values of every 30 min. The value are binned into the following ranges: *0-10* (Night: 0:00-5:00), *11-20* (Morning: 5:00-10:00), *21-26* (Midday: 10:00-13:00), *27-34* (Afternoon: 13:00-17:00), *35-44* (Evening: 17:00-22:00), *45-48* (Night: 22:00-24:00). The CRF model was trained and evaluated by these datasets using threefold cross-validation. The confusion matrix of activity recognition is shown in Fig. 17. The true positives of activity recognition are around 70%. However, most of the false negatives are in *12-Others*. In such cases, the recognized actions are mainly considered as the activities.

VIII. Conclusion and Future Work

In this research, we proposed and developed a framework of SoHAM for home service robots. The framework consists of the CoSER module, the dynamic time window-based human action recognition (DTW-HaR) module, and the CRF-based activity monitoring module. In the CoSER, the locational context of sound events associated with human daily activities is recognized based on the PIR network. The audio stream of sound events in the home environment is captured by the robot and extracted into feature vectors. Based on the context and sound event observations, the robot can recognize the sound events in real-time through a two-level DBN model using the short-time Viterbi algorithm. We tested and evaluated the framework with different parameters of the VQ and the DBN. We also proposed an algorithm called DTW-HaR to observe the sequence of sound event labels from the CoSER to estimate the current action, the duration of activity, the quality of monitoring, and the criticalness of action. A CRF model was implemented to recognize human activities based on the sequences of recognized actions. Experimental evaluation verified that our proposed framework can improve sound-based activity monitoring significantly.

In the future, we will address some limitations in the current work. First, we will keep collecting new sound data and build a larger home sound dataset with more data variations, including sound events with very short durations. This dataset will be made available to the research community. Second, machine learning methods for identifying the criticalness of each activity will be investigated, which can reduce human involvement to the minimum. Third, the proposed CRF-HAM will be enhanced to handle multiple unexpected transitions. Fourth, the proposed SoHAM framework can also be applied to distributed microphones in a smart home setting. We will compare the performance of these two microphone setups. Fifth, we will investigate how to leverage robot mobility to improve activity recognition accuracy. Finally, we will develop applications that can deliver elderly care services in response to recognized human activities.

References

[1] WHO. *World Health Organization: 10 Facts on Ageing and the Life Course*. Accessed: May 1, 2021. [Online]. Available: http://www.who.int/features/factfiles/ageing/en/

[2] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 81–94, Mar. 2016.

[3] H. M. Do, W. Sheng, E. E. Harrington, and A. J. Bishop, "Clinical screening interview using a social robot for geriatric care," *IEEE Trans. Autom. Sci. Eng.*, early access, Jun. 12, 2020, doi: 10.1109/TASE.2020.2999203.

[4] J. Vespa, J. M. Lewis, and R. M. Kreider, "America's families and living arrangements: 2012 population characteristics," *Current Population Rep.*, vol. 20, pp. 1–34, Aug. 2013.

[5] M. Hu, Y. Wang, Z. Zhang, D. Zhang, and J. J. Little, "Incremental learning for video-based gait recognition with LBP flow," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 77–89, Feb. 2013.

[6] M.-J. Tsai *et al.*, "Context-aware activity prediction using human behavior pattern in real smart home environments," in *Proc. IEEE Int. Conf. Automat. Sci. Eng. (CASE)*, Aug. 2016, pp. 168–173.

[7] C. Zhu, W. Sheng, and M. Liu, "Wearable sensor-based behavioral anomaly detection in smart assisted living systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 4, pp. 1225–1234, Oct. 2015.

[8] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE Sensors J.*, vol. 15, no. 3, pp. 1321–1330, Mar. 2015.

[9] H. M. Do, M. Pham, W. Sheng, D. Yang, and M. Liu, "RiSH: A robot-integrated smart home for elderly care," *Robot. Auto. Syst.*, vol. 101, pp. 74–92, Mar. 2018.

[10] M. Bar, "The proactive brain: Using analogies and associations to generate predictions," *Trends Cognit. Sci.*, vol. 11, no. 7, pp. 280–289, Jul. 2007.

[11] J. Chen, A. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on sound," in *Proc. Int. Conf. Pervasive Comput.* Berlin, Germany: Springer, May 2005, pp. 65–76.

[12] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognit. Lett.*, vol. 30, no. 14, pp. 1281–1288, Oct. 2009.

[13] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "SoundSense: Scalable sound sensing for people-centric applications on mobile phones," in *Proc. 7th Int. Conf. Mobile Syst., Appl., Services (Mobisys)*, Jun. 2009, pp. 165–178.

[14] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-Markovian ensemble voting," in *Proc. 21st IEEE Int. Symp. Robot Hum. Interact. Commun. (IEEE RO-MAN)*, Sep. 2012, pp. 509–514.

[15] J. M. Sim, Y. Lee, and O. Kwon, "Acoustic sensor based recognition of human activity in everyday life for smart home services," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 9, Sep. 2015, Art. no. 679123.

[16] P. Chahuara, A. Fleury, F. Portet, and M. Vacher, "On-line human activity recognition from audio and home automation sensors: Comparison of sequential and non-sequential models in realistic smart Homes1," *J. Ambient Intell. Smart Environ.*, vol. 8, no. 4, pp. 399–422, Jul. 2016.

[17] T. Hayashi, M. Nishida, N. Kitaoka, and K. Takeda, "Daily activity recognition based on DNN using environmental sound and acceleration signals," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 2306–2310.

[18] J. Kim, K. Min, M. Jung, and S. Chi, "Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition," *Building Environ.*, vol. 181, Aug. 2020, Art. no. 107092.

[19] V. Ramasubramanian, R. Karthik, S. Thiyagarajan, and S. Cherla, "Continuous audio analytics by HMM and viterbi decoding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 2396–2399.

[20] A. Fleury, N. Noury, M. Vacher, H. Glasson, and J.-F. Seri, "Sound and speech detection and classification in a health smart home," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2008, pp. 4644–4647.

[21] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Sep. 2007, pp. 21–26.

[22] Y. Sasaki, N. Hatao, K. Yoshii, and S. Kagami, "Nested iGMM recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 3930–3936.

[23] J.-F. Wang, J.-C. Wang, T.-H. Huang, and C.-S. Hsu, "Home environmental sound recognition based on MPEG-7 features," in *Proc. 46th Midwest Symp. Circuits Syst.*, vol. 2, Dec. 2003, pp. 682–685.

[24] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 1–5.

[25] S. Chachada and C.-C.-J. Kuo, "Environmental sound recognition: A survey," *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. 14, p. e14, Dec. 2014.

[26] B.-J. Han and E. Hwang, "Environmental sound classification based on feature collaboration," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2009, pp. 542–545.

[27] S. Chu, S. Narayanan, and C.-C.-J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Trans. Audio, Speech, Language Process*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[28] J.-C. Wang, C.-H. Lin, B.-W. Chen, and M.-K. Tsai, "Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 2, pp. 607–613, Apr. 2014.

[29] J. Dennis, H. D. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Trans. Audio, Speech, Language Process*, vol. 21, no. 2, pp. 367–377, Feb. 2013.

[30] T. C. Walters, "Auditory-based processing of communication sounds," Ph.D. dissertation, Dept. Physiol., Develop. Neurosci., Univ. Cambridge, Cambridge, U.K., 2011.

[31] R. F. Lyon, M. Rehn, S. Bengio, T. C. Walters, and G. Chechik, "Sound retrieval and ranking using sparse auditory representations," *Neural Comput.*, vol. 22, no. 9, pp. 2390–2416, Sep. 2010.

[32] J. W. Dennis, "Sound event recognition in unstructured environments using spectrogram image processing," Ph.D. dissertation, School Comput. Eng., Nanyang Technol. Univ., Singapore, 2014.

[33] S. Chandrakala, M. Venkatraman, N. Shreyas, and S. L. Jayalakshmi, "Multi-view representation for sound event recognition," *Signal, Image Video Process.*, vol. 139, pp. 1–9, Jan. 2021.

[34] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–7.

[35] J. Dennis, T. H. Dat, and H. Li, "Combining robust spike coding with spiking neural networks for sound event classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 176–180.

[36] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 3, pp. 540–552, Mar. 2015.

[37] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6440–6444.

[38] A. Copiaco, C. Ritz, S. Fasciani, and N. Abdulaziz, "Scalogram neural network activations with machine learning for domestic multi-channel audio classification," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2019, pp. 1–6.

[39] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 challenge–task 5: Monitoring of domestic activities based on multi-channel acoustics," 2018, *arXiv:1807.11246*. [Online]. Available: https://arxiv.org/abs/1807.11246

[40] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, Dec. 2019.

[41] K.-F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 38, no. 4, pp. 599–609, Apr. 1990.

[42] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[43] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, vol. 999. Cambridge, MA, USA: MIT Press, 1999.

[44] M. E. Niessen, L. Van Maanen, and T. C. Andringa, "Disambiguating sound through context," *Int. J. Semantic Comput.*, vol. 2, no. 3, pp. 327–341, Sep. 2008.

[45] L. Ma, D. Smith, and B. Milner, "Environmental noise classification for context-aware applications," in *Proc. Int. Conf. Database Expert Syst. Appl.*, Sep. 2003, pp. 360–370.

[46] J.-H. Su, H.-H. Yeh, P. S. Yu, and V. S. Tseng, "Music recommendation using content and context information mining," *IEEE Intell. Syst.*, vol. 25, no. 1, pp. 16–26, Jan. 2010.

[47] H.-S. Park, J.-O. Yoo, and S.-B. Cho, "A context-aware music recommendation system using fuzzy Bayesian networks with utility theory," in *Fuzzy Syst. Knowl. Discovery*. Berlin, Germany: Springer, 2006, pp. 970–979.

[48] S. Rho, B.-J. Han, and E. Hwang, "SVR-based music mood classification and context-based music recommendation," in *Proc. 17th ACM Int. Conf. Multimedia (MM)*, Oct. 2009, pp. 713–716.

[49] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 1, pp. 1–13, Dec. 2013.

[50] T. Lu, G. Wang, and F. Su, "Context-based environmental audio event recognition for scene understanding," *Multimedia Syst.*, vol. 21, no. 5, pp. 507–524, Oct. 2015.

[51] H. M. Do, W. Sheng, and M. Liu, "An open platform of auditory perception for home service robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 6161–6166.

[52] *PS3 Eye Camera*. Accessed: May 1, 2021. [Online]. Available: https://www.sony.co.in/product/

[53] M. Quigley *et al.*, "ROS: An open-source robot operating system," in *Proc. ICRA Workshop Open Source Softw.*, vol. 3, no. 3.2, May 2009, p. 5.

[54] S. Oh, L. Schenato, P. Chen, and S. Sastry, "Tracking and coordination of multiple agents using sensor networks: System design, algorithms and experiments," *Proc. IEEE*, vol. 95, no. 1, pp. 234–254, Jan. 2007.

[55] K. Kuutti, "Activity theory as a potential framework for human-computer interaction research," in *Context and Consciousness: Activity Theory and Human-Computer Interaction*, vol. 17. Cambridge, MA, USA: MIT Press, 1996.

[56] T. M. Cover and J. A. Thomas, *Grey Information: Theory and Practical Applications*. Hoboken, NJ, USA: Wiley, 2006.

[57] E. Kim and S. Helal, "Modeling human activity semantics for improved recognition performance," in *Proc. Int. Conf. Ubiquitous Intell. Comput.* Berlin, Germany: Springer, 2011, pp. 514–528.

[58] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Commun.*, vol. 54, no. 4, pp. 543–565, May 2012.

[59] K.-M. Kim, S.-Y. Kim, J.-K. Jeon, and K.-S. Park, "Quick audio retrieval using multiple feature vectors," *IEEE Trans. Consum. Electron.*, vol. 52, no. 1, pp. 200–205, Feb. 2006.

[60] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.
[61] K. P. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," M.S. thesis, Univ. California, Berkeley, CA, USA, 2002.
[62] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.
[63] J. Bloit and X. Rodet, "Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Mar. 2008, pp. 2121–2124.
[64] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern discovery," *IEEE Pervas. Comput.*, vol. 9, no. 1, pp. 48–53, Jan./Mar. 2010.
[65] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, 2001, pp. 282–289.
[66] M. Agarwal and P. Flach, "Activity recognition using conditional random field," in *Proc. 2nd Int. Workshop Sensor-Based Activity Recognit. Interact.*, Jun. 2015, pp. 1–8.
[67] H. Wallach, "Efficient training of conditional random fields," Ph.D. dissertation, School Cogn. Sci., Univ. Edinburgh, Edinburgh, U.K., 2002.
[68] H. M. Do, W. Sheng, and M. Liu, "Human-assisted sound event recognition for home service robots," *Robot. Biomimetics*, vol. 3, no. 1, pp. 1–12, Dec. 2016.
[69] Google. *A Large-Scale Dataset of Manually Annotated Audio Events*. Accessed: May 1, 2021. [Online]. Available: https://research.google.com/audioset/index.html

Ha Manh Do (Member, IEEE) received the B.Sc. degree in electronics and telecommunications from the Hanoi University of Science and Technology, Hanoi, Vietnam, in May 1999, and the M.S. and Ph.D. degrees in electrical engineering from Oklahoma State University (OSU), Stillwater, OK, USA, in May 2015 and December 2018, respectively.

He worked as a Postdoctoral Researcher with OSU in Spring 2019, at Colorado State University-Pueblo, Pueblo, CO, USA, from June 2019 to July 2020, and at the Louisville Automation and Robotics Research Institute (LARRI), University of Louisville, Louisville, KY, USA, from August 2020 to April 2021. He is currently a Senior Machine Learning Engineer with Plume Design, Inc., Palo Alto, CA, USA. His primary research interests include smart homes, home service robots, auditory perception, computer vision, spoken language understanding, human–robot interaction, applied artificial intelligence, and machine learning.

Karla Conn Welch (Member, IEEE) received the B.S. degree in electrical and computer engineering from the University of Kentucky, Lexington, KY, USA, in 2003, and the Ph.D. degree in electrical engineering and computer science from Vanderbilt University, Nashville, TN, USA, in 2009.

She is currently an Associate Professor of electrical and computer engineering with the Louisville Automation and Robotics Research Institute (LARRI), University of Louisville, Louisville, KY, USA, and the Director of the Machine Learning and Interaction Laboratory. Her research interests include human–machine interaction, affective computing, adaptive-response systems, and robotics.

Weihua Sheng (Senior Member, IEEE) received B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 1994 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from Michigan State University, East Lansing, MI, USA, in May 2002.

He is currently an Associate Professor with the School of Electrical and Computer Engineering, Oklahoma State University (OSU), Stillwater, OK, USA. He is the Director of the Laboratory for Advanced Sensing, Computation and Control (ASCC Lab, http://ascc.okstate.edu) at OSU. He is the author of more than 200 peer-reviewed articles in major journals and international conferences. Eight of them have won Best Paper or Best Student Paper Awards in major international conferences. His current research interests include social robotics, wearable computing, human robot interaction, and intelligent transportation systems. His research has been supported by US National Science Foundation (NSF), Department of Defense (DoD), Oklahoma Transportation Center (OTC)/Department of Transportation (DoT), etc.

Dr. Sheng served as an Associate Editor for IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING from 2013 to 2019. He is currently an Associate Editor for *IEEE Robotics and Automation Magazine*.