2023 IEEE International Conference on Pervasive Computing and Communications (PerCom) | 978-1-6654-5378-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/PERCOM56429.2023.10099218

CHAR: Composite Head-body Activities Recognition with A Single Earable Device

Peizhao Zhu, Yongpan Zou, Wenyuan Li, Kaishun Wu
College of Computer Science and Software engineering, Shenzhen University
{zhupeizhao2017, liwenyuan2019}@email.szu.edu.cn, {yongpan, wu}@szu.edu.cn

***Abstract*—The increasing popularity of earable devices stimulates great academic interest to design novel head gesture-based interaction technologies. But existing works simply consider it as a singular activity recognition problem. This is not in line with practice since users may have different body movements such as walking and jogging along with head gestures. It is also beneficial to recognize body movements during human-device interaction since it provides useful context information. As a result, it is significant to recognize such composite activities in which actions of different body parts happen simultaneously. In this paper, we propose a system called CHAR to recognize composite head-body activities with a single IMU sensor. The key idea of our solution is to make use of the inter-correlation of different activities and design a multi-task learning network to extract shared and specific representations. We implement a real-time prototype and conduct extensive experiments to evaluate it. The results show that CHAR can recognize 60 kinds of composite activities (12 head gestures and 5 body movements) with high accuracies of 97.0% and 89.7% in user- dependent and independent cases, respectively.**

***Index Terms*—Composite activity recognition; Earable device; Multi-task learning**

I. Introduction

Human activity recognition (HAR) has become a hotspot in academia due to its great significance for various intelligent services in our daily life such as healthcare, personalized recommendation, and human-computer interaction (HCI). Researchers have proposed a variety of HAR methods with different sensing modalities including radio frequency (RF) signals [1]–[3], acoustic sensors [4]–[6], and inertial measurement unit (IMU) [7]–[10]. These works have considered a wide range of daily activities including walking [11], sleeping [12], eating [13], and *etc.*. In spite of good recognition performance, these works share a common shortcoming, that is, only considering singular activities and outputting a single activity type.

Singular activities usually involve a single body part such as a hand or head, or treat all the body parts as a whole. In singular activities recognition, researchers do not care about what actions are performed of different body parts. As a result, existing works usually assume that the whole body is static while a user is performing activities/gestures with her head or hand. But this is inconsistent with the characteristics of human activities, most of which consist of different actions of multiple body parts. For example, when people are walking, they may wave hands or nod heads to interact with others. In this process, walking and waving hands (or nodding heads) are independent activities of which both have meaningful and distinguishable semantic information. For another example, when people go upstairs, they may also pick up phone calls at the same time. Although such composite activities can be monitored with multiple sensors attached on different body parts as done in previous works [14]–[16], this is obviously not realistic and appealing in real-world scenarios as people may not be willing to wear multiple devices with them. Considering this, we pay attention to a more complex and practical HAR problem, that is, *how to recognize composite activities with a single wearable sensor*. Formally, we define composite activities as simultaneous actions of different body parts in this paper.

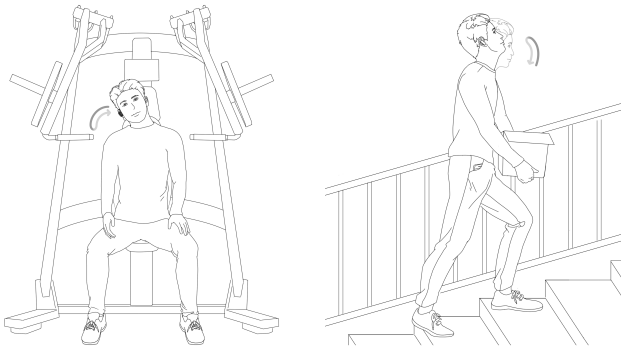


Fig. 1. The application scenarios of composite activities recognition with an earable device. Doing exercises while staying still with hands occupied (left) and going upstairs while carrying goods (right).

Composite activities have significant and practical value for human-computer interaction in real world. A typical example is shown in Fig. 1. When a user is doing exercises, she can utilize head gestures to interact with smart earphones such as turning up volume or picking up phone calls. At the same time, the smart earphones can recommend exercises-related musics or programs if doing exercises is recognized. Similarly, when a user goes upstairs with hands occupied, she can shake her head to turn up/down audio's volume of earphones, and turn on noise reduction mode. In a summary, the user performs composite activities in these examples with head gestures and body movements happening simultaneously. Consequently, we believe that it is meaningful to recognize both parts of activities (*i.e.*, head gestures and body movements) due to the following reasons. On one hand, we can see from the examples that both head gestures and body movements have

significant semantic information which can be used for human-device interaction. On the other hand, the commonalities and differences between different tasks are beneficial for boosting the recognition and generalization performance.

Considering the diversity of composite activities, we only focus on recognizing 60 kinds of head-body activities in this work including 12 head gestures and 5 body movements with a single earphone. There are two considerations. For one thing, a human being's head has a large degree of freedom and is very flexible to perform different gestures gently without causing any apparent discomfort. This makes head gestures appropriate for designing HCI interfaces. For another thing, compared with other body parts, head gestures happen together with body movements more naturally especially when a user is walking or jogging.

However, the solution to recognizing composite activities with a single sensor is not so straightforward due to two key obstacles. The first is the tight coupling of different tasks. Different actions will interfere with each other, which makes it difficult to recognize them. To be specific, sporadic head gestures produce notable measurements which are interference to body movements. The second obstacle is the indistinct measurements of body movements. Although an earable sensor is easy to capture head gestures, it is insensitive for body movements due to the stability of human's heads during activities. An intuitive idea is to treat each task independently and design separate learning networks to recognize different actions. But this approach ignores the correlation between head gestures and body movements which can be utilized for boosting performance. Inspired by multi-task learning, we have designed a composite activities recognition network for head gestures and body movements. Its key idea is to first extract general representation shared by coupled tasks with a shared bottom block, and then decouple them to further purify fine-grained task-specific features. As a result, it not only avoids over-fitting to a specific task, but also facilitates decision-making of each task.

Based on the above, we have built the first, as far as we know, composite head-body activities recognition system called CHAR with a single earable IMU sensor that recognizes 60 kinds of activities with high accuracies of 97.7% and 92.0% for head gestures and body movements, respectively.

In a summary, the contributions of this work can be summarized as follows:

- We consider a novel HAR problem in which composite head-body activities are recognized. Instead of decomposing a composite activity into individual tasks, we take full advantage of the correlations between different tasks and design a multi-task learning network to accomplish both goals simultaneously. Compared with traditional methods, our approach outperforms with higher accuracies.
- We design and implement an earphone-based real-time prototype system with low-cost hardware and a self-developed mobile application. Extensive experiments have shown that our system can recognize 60 composite activities with an high accuracy up to 89.7% even in user-independent case.

The remaining of this paper is organized as follows. Sec. II discusses the related work. Sec. III gives introduction to the system design. In Sec. IV and Sec. V, we mainly describe the details of system implementation, experimental settings, and performance evaluation. At last, Sec. VI concludes the paper.

II. Related Work

A. Human Activity Recognition

Vision systems [17], [18] use cameras for recognition, but they are considered invasive. Ambient sensors [2], [3], [19] require to be installed at fixed locations with extra expense. In addition to these, wearable sensors, which are usually built in mobile devices, have attracted the attention of researchers. Some researchers take advantage of microphone [5], [6], biological signal sensor [20], [21] for HAR. The works which utilize IMU to recognize sporadic [7]–[10] or periodic [11], [15], [22], [23] human acitivities are most related to ours. When it comes to sporadic activities, HulaMove [7] uses IMU of mobile phone for waist interaction. HeadGesture [8] proposes a hands-free input approach leveraging head gestures. GlassGesture [9] provides efficient gesture recognition and robust authentication with Google Glass. As for periodic movements, previous works [11], [24]–[26] use smartphone to recognize periodic movements and contribute their datasets. Other works propose some powerful networks such as DCNN [22], DeepSense [23], DeepConvLSTM [15] based on public datasets to recognize daily activities. Obviously, these works only concentrate on the recognition of singular activities. Nevertheless, human activity is usually a composite of many kinds of actions. As a contrast, our system makes use of multi-task learning to recognize sporadic head gestures and periodic body movements at the same time.

B. Multi-task Learning

Multi-task learning is a training paradigm which can deal with multiple tasks jointly. It has been successfully applied in many areas especially in natural language processing. However, multi-task learning has not been widely used in HAR. The works [27]–[29] make use of multi-task learning approach to realize gestures/activities recognition and user authentication jointly. CogAx [30] uses a contrastive and multi-task learning framework to correlate with underlying functional and cognitive health parameters of older adults. As the most relevant work, AROMA [14] defines complex activities as continuous and high-level semantic meaning activities which are combination of simple activities along temporal dimension. AROMA proposes a deep multi-task learning based method to facilitate complex activities recognition using features extracted from simple activities. In contrast, we concentrate on composite activities composed of simultaneous actions of different body parts. Consequently, we designed a multi-task recognition network which makes full use of shared information and further exploits task-specific representations of each task to facilitate decision-making.

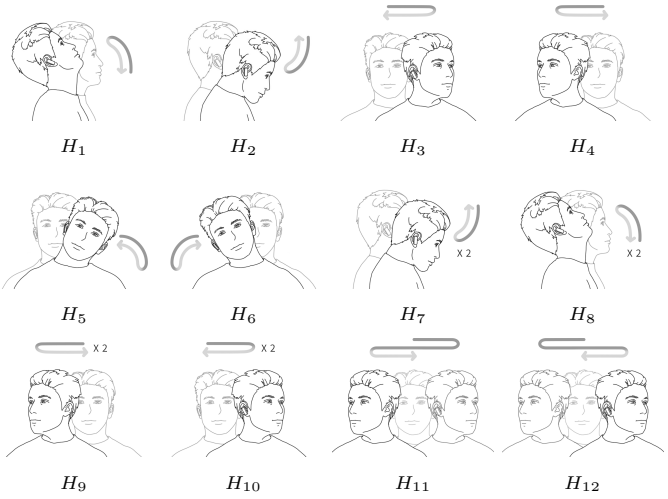


Fig. 2. The design of head gestures.

C. Earable Applications

Recently, researchers have shown great interest in developing earable applications. EarBuddy [31] is an on-face interactive system by recognizing sounds of on-face gestures. EarEcho [32] utilizes in-ear speaker and microphone to capture cannal echo for authentication. eBP [33] proposes a wearable system for blood pressure monitoring. While prior works have explored the use of microphone or physiological sensors in ear-worn devices, not many have explored the use of IMU on ear-worn devices for complex HAR. [34] investigates head motion tracking using the eSense [35] platform. MandiPass [36] leverages IMU to achieve continuous authentication function. Similarly, our system is built with low-cost and ubiquitous IMU which can be easily deployed in commercial earphones. In contrast, instead of concentrating on head or body activities, our earable device senses head-body composite activities simultaneously.

III. System Design

A. Composite Activities Taxonomy

In this paper, a composite head-body activity means the time aligned compound of a head gesture and a body movement. We refer to some HAR benchmark datasets of body movements recognition such as HHAR [24], UCI [25], MotionSense [26] and Shoaib [11], which have been commonly used in many works. Then we select 5 most common movements which are practical in various real-world scenarios and suitable for our single earable device interaction. To be specific, we consider going downstairs and upstairs, staying still, life-walking, jogging which are denoted by $B_1 \sim B_5$. As for the head part, we design a total number of 12 gestures as shown in Fig. 2 including raising head up and back for once and twice (H_1, H_8), nodding once and twice (H_2, H_7), shaking head to the left for once and twice (H_3, H_{10}), shaking head to the left first and then to the right (H_{11}), shaking head to the right for once and twice (H_4, H_9), shaking head to the right first and then to the left (H_{12}), leaning head to the left and right shoulders (H_5, H_6). The criteria of designing these head gestures is referred to previous works such as HeadGesture [8] and GlassGesture [9] since they are natural for users to perform. As a result, there are a total number of 60 composite head-body activities.

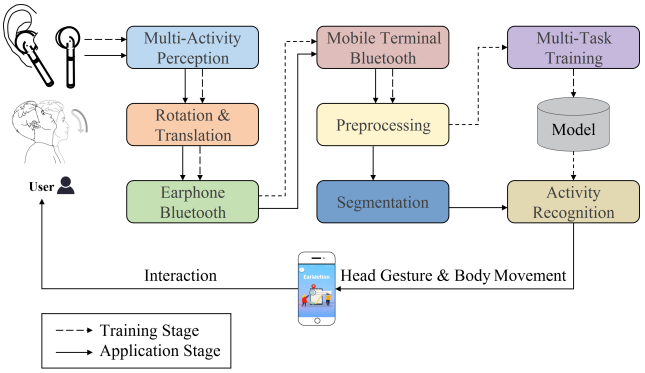


Fig. 3. The system overview of CHAR.

B. Overview

Fig. 3 shows the overview of training and testing pipelines of our system. When a user performs head gestures accompany with body movements, the earable device such as a smart earphone with an IMU sensor senses the rotational and translational movements, and transmits the data to a mobile terminal through Bluetooth. After that, signal instances corresponding to each composite activity are extracted precisely with an adaptive segmentation method. In the training stage, the extracted signal segments are used for building a self-designed multi-task learning network which is trained on a remote server. In the application stage, the trained network is deployed on a mobile device such as a smartphone, takes in signals of a composite activity, and outputs labels of a head gesture and a body movement simultaneously. According to different head-body composite activities, the system will accomplish HCI by controlling multimedia applications such as picking up phone calls, adjusting volume, and *etc.*. In the following, we shall give details of each key part of our system design.

C. Activity Segmentation

Activity segmentation refers to detecting the starting point (SP) and ending point (EP) of an activity. In our head-body activity recognition task, since body movement is periodic and persistent, we only need to acquire head gesture segment, which contains the composite head-body activity. Other previous works set a fixed energy threshold for activity segmentation which can not apply to different practical scenarios. In our work, we propose an adaptive segmentation method which dynamically adjusts the threshold according to different scenarios. As a result, user's head gestures with different body movements will be accurately segmented. The activity segmentation algorithm is shown in Algorithm 1. The following section gives details of this method.

Algorithm 1: Activity segmentation algorithm

Input: 3-axis signal $Gyro = \{G_x, G_y, G_z\}$; Window length for EP and SP detection L_{end}, L_{start}; Framing stride S; Threshold for EP detection θ; Scaling coefficient for SP detection k

Output: SP; EP

W_{signal} = Frame($Gyro$, $(L_{end} + L_{start})$, S);
W_{end} = $W_{signal}(L_{start}\ :,\ :,\ :)$;
W_{cont} = $W_{signal}(:\ L_{start},\ :,\ :)$;
for *i = 1 to FrameNum(W_{end})* **do**
  for *j = 1 to ChannelNum(W_{end})* **do**
    $E_{ch}(i,\ j)$ = Energy($W_{end}(i,\ j,\ :)$)
  end
  $ch(i)$ = FindMaxIndex($E_{ch}(i,\ :)$);
  $E_{end}(i)$ = Energy($W_{end}(i,\ ch(i),\ :)$);
  E_d = Diff(E_{end});
  EP = FindFirstIndexLess($E_d(i)$, θ);
  if *EP is not None* **then**
    E_{cont} = Energy($W_{cont}(i,\ ch(i),\ :)$);
    W_{start} = Frame($W_{end}(i,\ ch(i),\ :)$, L_{start}, S);
    for *j = 1 to FrameNum(W_{start})* **do**
      $E_{start}(j)$ = Energy($W_{start}(j,\ :)$);
    end
    SP = FindFirstIndexGreater(E_{start}, $k \times E_{cont}$);
  end
end
return SP, EP

1) Signal Processing: In order to segment activity from data stream, we use 3 sliding windows W_{end}, W_{cont} and W_{start} to frame the signal with 1.0 seconds stride. Since W_{end} needs to cover entire activity for EP detection, the size of W_{end} is set to 4.0 seconds, which is slightly larger than the duration of all head gestures. W_{cont} captures the context information of the activity which follows W_{end} in time dimension and the size of W_{cont} is empirically set to 0.5 seconds. When the EP of an activity is detected, W_{start} further frames the current W_{end} with the same size of W_{cont}. After framing data stream, we calculate the average energy E by the formula as follows:

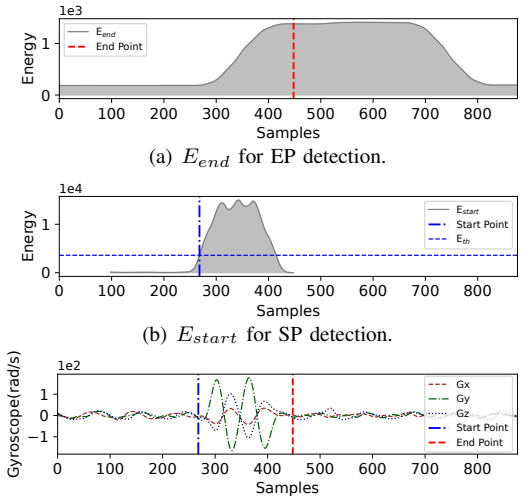
$$E = \frac{1}{T}\sum_{t}^{T}(x_t)^2 \tag{1}$$

where x is the signal and t is the timestamp of x.

2) Channel Selection: In our system, we get 6 channels signal of the IMU. Since gyroscope varies significantly and its noise is relatively low, we select the signal of it as the original signal for activity segmentation. Different head gestures have different variations in the 3-axis of gyroscope. We need to get the best segmentation channel in real time. Therefore, We select the axis with the largest average energy as the best segmentation channel ch as follows:

$$ch = \arg\max_i E_i \tag{2}$$

where i is channel index and E_i is the energy of each channel.



(a) E_{end} for EP detection.

(b) E_{start} for SP detection.

(c) Origin signal and segmentation result.

Fig. 4. Example of activity segmentation.

3) EP and SP Detection: Fig. 4 shows an example of activity segmentation. We utilize E_{end} for EP detection as shown in Fig. 4(a). When the user performs an head gesture, E_{end} will increase first and then remain stable. Because the context has the same background noise no matter what kind of scenario the head gesture is in. The EP of the activity is detected when the variation of E_{end} is less than θ, which is set to 4. We utilize W_{start} for SP detection as shown in Fig. 4(b). Since W_{end} contain the entire activity, the SP will be included in it. When $E_{start} > E_{th}$, the SP of the activity is detected. Context information $E_{th} = kE_{cont}$ is utilized as dynamic threshold, where k is a scaling coefficient, which is empirically set to 10. Different from previous works, E_{th} changes dynamically according to the background noise. When the user stays still and goes downstairs, the dynamic threshold becomes lower and higher, respectively. After EP and SP detection, we will get the segmentation result in origin signal, as shown in Fig. 4(c). Finally, we calculate the midpoint of SP and EP, and then segment 3.0 seconds length data in order to acquire fix length input data for the network.

D. Network Design

1) Multi-task Learning: To deal with multi-task problems, the most intuitive idea is to split the problem into several independent single-task subproblems that are solved separately and then recombined. However, this method has deficiency for two reasons. For one thing, there is shared information between head gestures and body movements. Making good use of shared information is beneficial to extract features for each task. For another thing, head gestures and body movements recognition are not completely separate of each other. Taking full advantage of the correlations between them by joint training facilitates the decision-making of each task. As a result, we turn our attention to multi-task learning.

Multi-task learning is a training paradigm where learning model is trained with data from multiple tasks simultaneously.

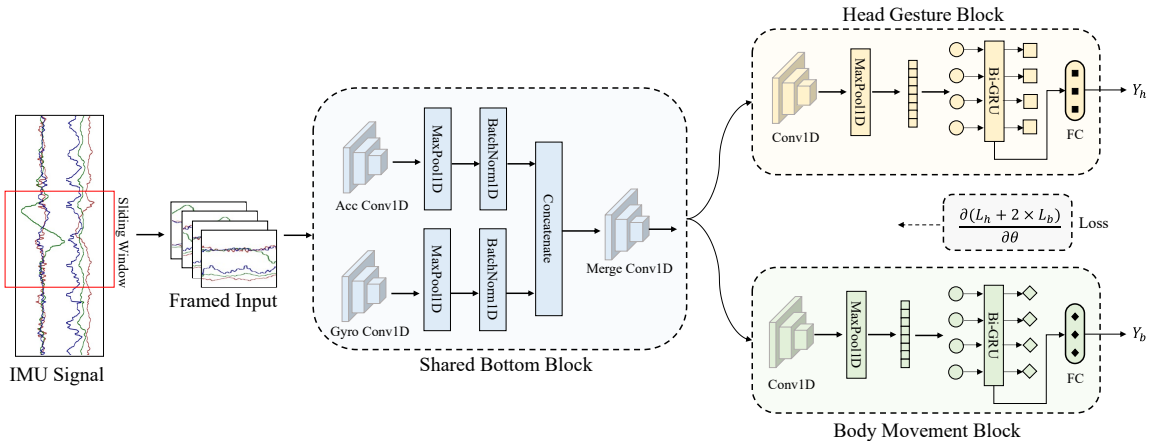


Fig. 5. The architecture of our composite-activity recognition network.

Inspired by the insight of multi-task learning, we design a composite activities recognition network (CARN) for sporadic head gestures and periodic body movements recognition. In evaluation section, we compare the performance of some classical single-task networks for HAR and our multi-task network. The experimental results show the superiority of our network, which indicates that multi-task learning plays an important role in composite activities recognition. Fig. 5 shows the architecture of CARN, which consists of shared bottom block and specific top blocks.

2) Design of Shared Bottom Block: The bottom layers are shared across tasks to extract task-shared representation which facilitates both head gestures and body movements recognition. This kind of shared structure substantially reduces the risk of overfitting. For example, in backpropagation, the contribution of head gesture task to the aggregated gradient can be regarded as noise for body movement task. In other words, one task can be used as a noise source to avoid overfitting of another. Shared bottom block improves generalization performance by learning head gesture and body movement recognition tasks in parallel while using shared representation.

Inspired by LRCN [37], we treat the 3.0 seconds length segment as frame-by-frame data. To be specific, we use a 0.5 seconds length sliding window to frame the segments with 0.1 seconds stride and then feed all frames to the network. To begin with, the signals of each sensors have different signal-noise patterns and representation capability. Therefore, we need to extract intra-modality representation across time and channels of each sensor separately. Meanwhile, we pay attention to the correlation between different modalities. Thus, we further extract cross-modality representation across different sensors. These two operations avoid mutual interference between the two modalities. For implementation details, we utilize two different 1D-CNN to extract the spatial features of accelerometer and gyroscope seperately. Max Pooling and BatchNorm are applied on the each results of 1D-CNN. After that, we concatenate the intra-modality features along channel dimension and feed them to a 1D-CNN for multi-modality fusion and cross-modality features extraction. In this way, we extract intra-modality and cross-modality spatial features as shared representation for specific tops.

3) Design of Specific Top Blocks: The specific top layers are explicitly utilized to excavate specific and fine-grained representation of each task. Since specific tasks aim to recognize diverse kind of activities, we separately adopt two different feature extractors to acquire task-specific representations. The reason is that different tasks favor different spatial and temporal information. Specifically, head gestures and body movements recognition tasks prefer to rotation and translation spatial information separately. Besides, it is obvious that head gestures and body movements are sporadic and periodic signals, respectively. Consequently, these two recognition tasks also require different temporal information.

The specific tops consist of head gesture block and body movement block. To extract specific and deeper representations, we employ two separate 1D-CNN to extract the spatial features of rotation and translation for different tasks. Max Pooling is applied to reduce the dimensionality along time. Afterwards, we flatten the outputs and empirically adopt a 256-unit 1-layer Bi-GRU and a 128-unit 1-layer Bi-GRU to profile the temporal relationships for head gestures and body movements, respectively. It is worth mentioning that the time steps of Bi-GRU corresponds to the input frames. Finally, we feed the last time step of Bi-GRU into the fully-connected layers of specific task blocks to get Y_h and Y_b. We separately define cross-entropy loss L_h and L_b for head gesture and body movement task. At multi-task learning, we use $L = L_h + \lambda L_b$ for joint training, where λ is empirically set to 2.

IV. Implementation and Experiments

In this section, we describe the implementation of CHAR and give details about how data are collected for evaluation.

A. Implementation

As commercial earphones have no access to sensor's data, we build a prototype with low-cost hardware including a

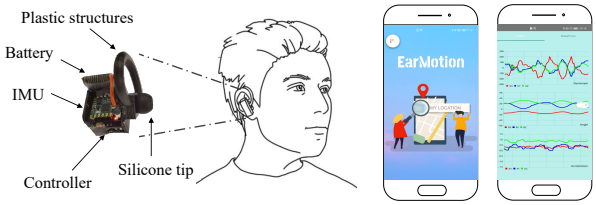
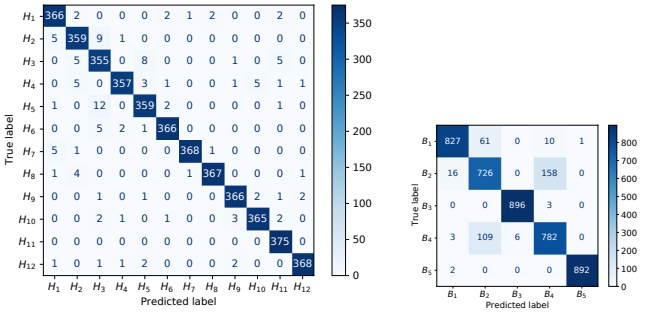


Fig. 6. The hardware and mobile application of CHAR.



(a) Head gestures. (b) Body movements.

Fig. 7. Confusion matrices of head gestures and body movements recognition in user-independent cases.

JY901 IMU, an ESP32 as the microcontroller, a 3.7 V lithium battery as the power supply, and a 3D printed plastic enclosure shaped like an earphone as shown in Fig. 6. The JY901 module consists of an accelerometer, a gyroscope, and a magnetometer with sampling rate set to be 100 Hz during experiments. The ESP32 controls IMU data collection and transmission to a mobile device (*e.g.*, a smartphone) via Bluetooth.

We have also developed an Android application on the mobile device to receive and process data, and run activities recognition model, following the pipeline as introduced Sec. III. Specifically, after training the network on a server with an Intel(R) Xeon(R) Platinum 8260 CPU and NVIDIA GeForce RTX 2080Ti GPU, we deploy it on a HUAWEI Mate40 Pro with a HI-Silicon Kirin 9000 CPU, 8GB RAM, and 256GB ROM. Based on the output of the recognition model, developers can design different HCI applications.

B. Experiments

To evaluate CHAR, we recruit 15 participants denoted by $P_1 \sim P_{15}$ aged from 18 to 25 years old. Before experiments, we instruct them to use our system and tell them necessary details. We then request each of them to perform every composite head-body activity for 5 times, and finally get a total number of 4500 (*i.e.*, $15 \times 12 \times 5 \times 5$) raw instances, since there are 12 head gestures and 5 body movements as aforementioned. After data collection, we annotate the SP and EP of each raw instance as ground truth in evaluation. As the duration between SPs and EPs are different among instances, we calculate the midpoint of SP and EP, and use a fixed length window of 3.0 second to segment the raw instances.

In the following evaluation, we consider user- dependent and independent cases. In the former cases, we train and test the system with 15-fold cross-validation method, while in the latter ones, we train and test it with 'leave-one-user-out' strategy. In both kinds of cases, the results are averaged over different testing rounds. Note that segments produced by raw instances do not overlap. Thus there is no information leakage between training and testing datasets.

V. Evaluation

A. Overall Performance

1) Activity Segmentation: We manually segment signal sequences by marking SPs and EPs of activities as ground truths, and compare them with the results obtained by our proposed segmentation method (*i.e.*, Algorithm 1). We utilize missed detection rate (MDR) and false detection rate (FDR) as the metrics for evaluation. Formally, MDR is defined as the percentage of missing activities that are not detected by the algorithm. FDR is defined as the percentage of mistakenly detected activities outside the ground truth. Experiments show that the MDR and FDR with Algorithm 1 are 1.8% and 1.2%, respectively. This means that more than 97.0% of head gestures can be detected correctly. In addition, the differences between segmentation results and the ground truth will also affect the final recognition performance of CHAR. We will evaluate this point in Sec. V-A3.

2) Activity Recognition: In the user- dependent and independent cases, CHAR recognizes composite activities with high accuracies of 97.0% and 89.7%, respectively. Since users prefer to use a system immediately without retraining, we set the user-independent condition as our baseline case in the following evaluation. Fig. 7 shows the confusion matrices that CHAR recognizes head gestures and body movements in the baseline case. As we can see, the recognition accuracies of head gestures and body movements reach 97.7% and 92.0%, respectively. This indicates that CHAR can recognize composite head-body activities with high accuracy. This is mainly due to our designed network which decouples their relationship. In addition, we can also see that walking (B_1) and going upstairs (B_2) are easier to be confused, which is mainly originated from user diversity. There is user diversity of some activities in HAR, which leads to poor performance in user-independent recognition. We carefully check these data of walking and going upstairs. It was found that these two movements are relatively similar across different users. Firstly, walking and going upstairs produce periodic signals with similar periods, which are longer than going downstairs and jogging. Then, walking on rough ground tends to produce z-axis acceleration signal analogous to that of going upstairs, which leads to confusion between these two movements. Nevertheless, accuracies of CHAR on body movements recognition is still up to 98.0% and 92.0% in user- dependent and independent situations, which is sufficient for practical needs.

TABLE I
ACCURACY USING DIFFERENT KINDS OF DATA SEGMENT.

Training / Testing	Head gesture	Body movement
D_h / D_h	97.7%	92.0%
D_h / D_a	93.7%	91.9%
$\boldsymbol{D_a}$ / $\boldsymbol{D_a}$	**94.9%**	**89.8%**

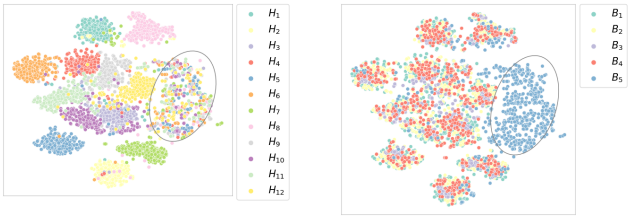
3) Cascade Performance: According to Fig. 3, segmentation is followed by the recognition network. Hence, its result has impact on the network's classification performance. To evaluate this, we consider two segmentation methods, namely, by Algorithm 1 and by human (*i.e.*, ground truth). Correspondingly, we denote data obtained by them as D_a and D_h, respectively. Then we train and test the network with different combinations of D_a and D_h as shown in Table I and get corresponding results. We can see that using D_a as the testing set causes accuracy drops of 4.0% and 0.1% for head gestures and body movements recognition, respectively. This indicates that our segmentation algorithm has some impact on head gestures, but has little on body movements. This is because some instances of the combined head gestures (such as H_7, H_9, H_{11}) are disjointed, so that the algorithm segments part of these gestures. Partial segmentation leads to recognition error because head gesture is sporadic. Differently, body movement is periodic, and slight segmentation differences have little affect on recognition.

Besides, we explore the possibility of directly annotating instances using Algorithm 1 without human effort, by training and testing the model with D_a. It can be seen that CHAR can recognize head gestures and body movements with accuracies of 94.9% and 89.8%, respectively. Compared with the D_h/D_h case, the accuracies decrease by 2.8% and 2.2%, respectively. It validates that with the proposed segmentation method, CHAR can still reliably detect the SP and EP of an activity instance, and achieve very close performance to that of ground-truth segmentation. Hence, our proposed system can perform well in practical application scenarios and reduce the burden of annotating data.

B. Embedding Visualization

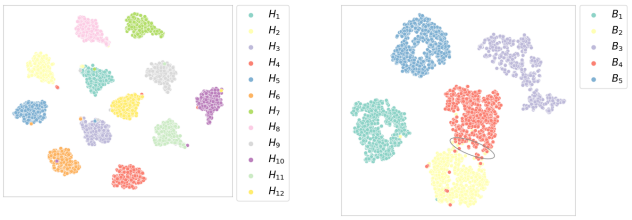
To better dissect the internal mechanism of multi-task learning, we show the embeddings of CARN in two-dimension space with t-SNE algorithm.

1) Shared Bottom Embedding: Fig. 8 shows the shared feature representations extracted by the shared bottom block. Note that feature representations shown in Fig. 8(a) and Fig. 8(b) are totally the same but are colored in terms of head gestures and body movements, respectively. Fig. 8(a) shows that head gestures are grouped at a coarse-grained level except for the circled data points. This means that the shared bottom block can extract distinct features for most instances. Samples mixed together in the circled zone represent jogging movements as shown in Fig. 8(b). This is because gentle body movements



(a) Colored by head gestures. (b) Colored by body movements.

Fig. 8. Embedding visualization with t-SNE algorithm for features extracted by Shared Bottom Block of CARN. The figure corresponds to the same features and has been colored in terms of different types of activities.



(a) Head gesture embedding. (b) Body movement embedding.

Fig. 9. Embedding visualization with t-SNE algorithm for features further extracted by Head Gesture Block (a) and Body Movement Block (b) of CARN, respectively.

such as being still, walking, going upstairs and downstairs have relatively slight effect on the measurements of head gestures. But when users are jogging, the IMU measurements of head gestures tend to be overwhelmed by it.

Accordingly, from Fig. 8(b), we can find that only the jogging instances are well grouped and other body movements are easily mixed together after shared bottom block. The reason is that head gestures seriously deform periodic signals generated by body movements. What is more, periodic signals generated by body movements require a model to extract temporal features, while the shared bottom block only uses CNNs to extract shared spatial features. In order to better group body movements, we use GRUs to extract specific time-domain information following the shared body bottom.

2) Specific Task Embedding: Fig. 9 shows the feature representations of head gestures and body movements extracted by corresponding blocks. We can observe from Fig. 9(a) that all head gestures (including in jogging movement) are well grouped. Moreover, compared to the results shown in Fig. 8(a), the distances between different categories is obviously larger. It means that the head gesture block extracts unique features for better recognition on the basis of shared feature block. Similarly, it can be seen from Fig. 9(b) that, except for a few data points in the circled zone, the body movements can be well grouped. It indicates that even though the shared bottom block cannot directly group some of the body movements, the relevant part of the network has the ability to extract unique periodic features corresponding to body movements.

TABLE II
ACCURACY OF BENCHMARKS AND OUR NETWORK.

Network	Head gesture	Body movement	Composite activity
DCNN	91.2%	81.0%	73.4%
DeepSense	92.4%	76.5%	69.8%
DeepConvLSTM	88.1%	84.9%	74.4%
CARN	**97.7%**	**92.0%**	**89.7%**

C. Comparison with Benchmarks

In this section, we compare our proposed network CARN with several classical networks for HAR as benchmarks including DCNN [22], DeepSense [23], and DeepConvLSTM [15]. DCNN [22] designs a convolutional neural network to learn features from raw IMU input signals which outperforms traditional machine learning methods such as support vector machine and deep belief network. DeepSense [23] integrates CNNs and RNNs to combine different sensing modalities of mobile sensors and extract temporal features for HAR. DeepConvLSTM [15] proposes a deep activity recognition framework composed of convolutional and LSTM recurrent layers, which is suitable for the recognition of static/periodic activities and sporadic activities. As these networks deal with single-task HAR, we train them for head gestures and body movements recognition tasks separately. In contrast, CARN is trained for both tasks simultaneously. Also, in order to make our dataset suitable for those networks, we slightly modify their input and output layers.

Table II gives comparative results of CARN and other benchmarks in both recognition tasks. From our previous analysis, it can be known that sporadic head gestures can be well recognized by only using CNNs to extract spatial features. Since head gestures are accompanied by periodic body movements, using RNNs to extract temporal features will interfere with head gestures recognition instead. Therefore, the head gestures recognition accuracy of DCNN reach 91.2% with CNNs only. On the contrary, DeepConvLSTM using RNNs has the worst performance on the head gestures recognition task, but its accuracy of body movements recognition reaches 84.9%. In addition, it can be seen that CARN outperforms benchmarks in both head gestures and body movements recognition tasks. Accuracies of head gestures and body movements recognition tasks are 5.3% higher than DeepSense and 7.1% higher than DeepConvLSTM, respectively. This is because irrelevant task will interfere the target task, which results in poor recognition performance. However, our designed multi-task learning framework fully learns the commonalities of two tasks and then decouples them to learn the differences. It reduces the risk of overfitting and improves generalization ability of recognition model.

D. Impact of Composite Activity Types

Without specification, we train the network with all types of head-body composite activities (*i.e.*, 60 types of composite activities). A natural question is whether the network can be trained with a part of composite activities, but can still recognize the whole set accurately. The intuitive rationale is that unseen composite activities are also composed of the same basic head gestures and body movements which appear in the training activities. To quantify the impact of the number of composite activities, we randomly select a certain number of composite activities types in the training set, and test CHAR's performance on the whole set. We vary the number of composite activities types in the training set from 12 to 60, and obtain the results as shown in Fig. 10. It can be seen that as the number of composite activities increases, the recognition accuracy first increases and then remains stable. The recognition performance of the system tends to be stable when more than 36 composite activities are used for model training. Experiments demonstrate that the network can be trained using data from some composite activities and applied to all, thereby reducing the burden of data collection.

E. Impact of Data Length

As mentioned in Sec. III-C, after detecting the EP and SP of an activity, we extract a fixed-length data segment and feed it into the network. The data length is a critical parameter for recognizing activities accurately. Intuitively, it should be large enough to cover an activity as much as possible. To determine its proper value, we first obtain a statistical histogram of head gestures' durations as shown in Fig. 11. It can be seen that more than 90% of the head gestures last less than 3.0 seconds, which indicates a proper range of data length. Although a longer segment contains more information of activities, it also brings about heavier computational overhead and more energy consumption. To achieve good trade-off, we test CHAR's performance with different data lengths from 1.0 second to 4.0 seconds. As shown in Fig. 12, the recognition accuracies rise with the data length. But the improvement is very limited after the data length exceeds 3.0 seconds. In addition, only about 50% of the head gestures duration are less than 2.0 seconds from Fig. 11. However, setting the data length to 2.0 seconds can also achieve 96.1% and 88.7% recognition accuracies for head gestures and body movements, respectively. As a result, for mobile devices with limited resources, we suggest appropriately reducing the data length to ease computing overhead and energy consumption.

F. Impact of Data Size

We also evaluate the impact of training dataset size by randomly sampling a whole dataset at different percentages. Since the evaluation is conducted with 'leave-one-user-out' strategy, a whole dataset contains 4200 samples in total. Fig. 13 shows the results. With the percentage increasing, the accuracies of recognizing head gestures, body movements, and composite activities initially increase fast and then remain relatively stable, after the percentage exceeds 0.8. It means that with 80% of the dataset, CHAR's performance is close to the optimal. Moreover, compared with body movements, head gestures can be recognized more accurately with less

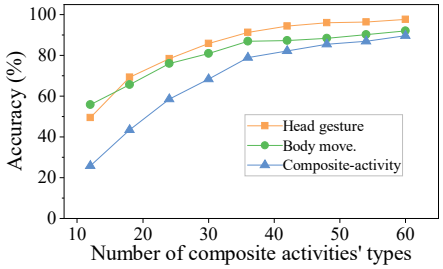


Fig. 10. Recognition accuracies with different number of training activity types.

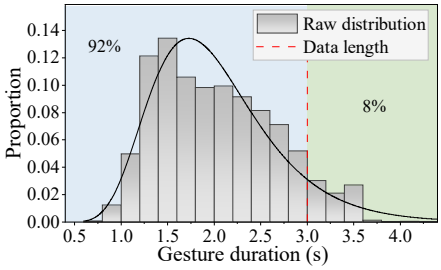


Fig. 11. Statistical histogram of head gestures performing time.

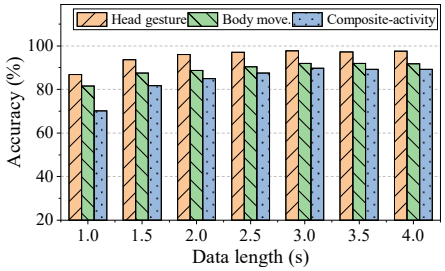


Fig. 12. Recognition accuracies with different data lengths.

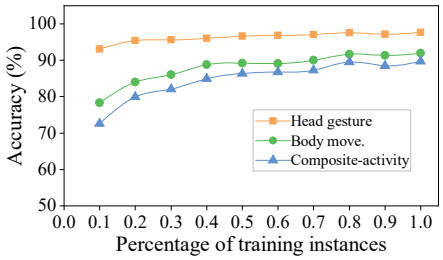


Fig. 13. Recognition accuracies with different training dataset sizes.

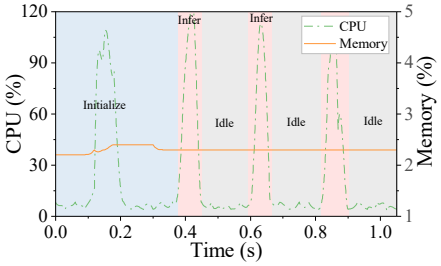


Fig. 14. CPU and memory occupation.

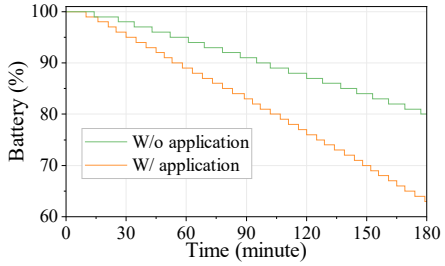


Fig. 15. Energy consumption.

training data. This is because IMU measurements caused by head gestures are more obvious than those body movements. As a result, even a small amount of data enables the model to learn feature patterns of head gestures. In contrast, body movements cause less notable sensor measurements which are easier to be interfered. Hence, it requires more training data to recognize body movements (and thus composite activities) with high accuracy.

G. System-running Performance

In this section, we evaluate CHAR's real-time performance including CPU and memory occupation, response time, and energy consumption. Specifically, we utilize Android Debug Brige to acquire CPU and memory occupation when CHAR runs at different stages including initialization, inference, and being idle. We can observe from Fig. 14 that the memory occupation is only about 2.4% in average and slightly increases during initialization stage where the recognition model is loaded. The CPU occupation is at a low level of about 10% when CHAR is idle, but varies rapidly during the inference stage. In addition, we insert a piece of codes in the Android application to measure the response time of CHAR which is the duration of outputting the result after performing an activity. The experimental results show that CHAR can recognize gestures within 52 ms in average on a mobile device.

We also measure the energy consumption through Android BatteryManager by continuously running the testing pipeline with data fed into the application for three hours. During this process, we turn off all the other applications and set the screen brightness to the medium level. Fig. 15 shows how the battery level varies with running time when CHAR is turned on (yellow line) and off (green line), respectively. As we can see, when CHAR is running, the average energy consumption per minute can be estimated by $\frac{4400\times(100\%-63\%)}{180}$ which equals 9.07 mAh. Note that this energy consumption includes the part of lighting the screen which can be calculated by $\frac{4400\times(100\%-80\%)}{180}$ equalling 4.89 mAh. Consequently, our system consumes 4.18 mAh per minute which is relatively low especially considering the much lower interference frequency in real-world usage cases.

VI. Conclusion

Composite activities are rather common and significant for human beings, but has not been carefully investigated yet. In this paper, we put forward a composite activity recognition system called CHAR that can recognize a variety of head-body activities based on a single IMU measurement output by an earphone device. The high-level idea of our solution is to take full advantage of the inter correlation between head gestures and body movements, and design a multi-task learning network to extract shared and task-specific feature representations. We have implemented a real-time prototype and conduct extensive experiments to evaluate its performance. The results show that CHAR can recognize 60 head-body composite activities with a high accuracy even in the user-independent case. We envision that our approach can be also utilized for other composite activities recognition.

Acknowledgment

This research was supported in part by China NSFC Grant (62172286, 61872247, U2001207), Guangdong NSF Grant (2022A1515011509, 2017A030312008), and the Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603. Yongpan Zou is the corresponding author.

REFERENCES

[1] Y. Yang, J. Cao, and Y. Wang, "Robust rfid-based respiration monitoring in dynamic environments," *IEEE Transactions on Mobile Computing*, 2021.

[2] X. Li, Y. Zhang, I. Marsic, A. Sarcevic, and R. S. Burd, "Deep learning for rfid-based activity recognition," in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, 2016, pp. 164–175.

[3] Y. Wang and Y. Zheng, "Modeling rfid signal reflection for contact-free activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–22, 2018.

[4] Y. Jin, Y. Gao, Y. Zhu, W. Wang, J. Li, S. Choi, Z. Li, J. Chauhan, A. K. Dey, and Z. Jin, "Sonicasl: An acoustic-based sign language gesture recognizer using earphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–30, 2021.

[5] K. Ling, H. Dai, Y. Liu, A. X. Liu, W. Wang, and Q. Gu, "Ultragesture: Fine-grained gesture sensing and recognition," *IEEE Transactions on Mobile Computing*, 2020.

[6] Y. Ren, C. Wang, J. Yang, and Y. Chen, "Fine-grained sleep monitoring: Hearing your breathing with smartphones," in *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 1194–1202.

[7] X. Xu, J. Li, T. Yuan, L. He, X. Liu, Y. Yan, Y. Wang, Y. Shi, J. Mankoff, and A. K. Dey, "Hulamove: Using commodity imu for waist interaction," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.

[8] Y. Yan, C. Yu, X. Yi, and Y. Shi, "Headgesture: hands-free input approach leveraging head movements for hmd devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–23, 2018.

[9] S. Yi, Z. Qin, E. Novak, Y. Yin, and Q. Li, "Glassgesture: Exploring head gesture interface of smart glasses," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 2016, pp. 1–9.

[10] T. Hachaj and M. Piekarczyk, "Evaluation of pattern recognition methods for head gesture-based interface of a virtual reality helmet equipped with a single imu sensor," *Sensors*, vol. 19, no. 24, p. 5408, 2019.

[11] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors*, vol. 14, no. 6, pp. 10 146–10 176, 2014.

[12] C. Liu, J. Xiong, L. Cai, L. Feng, X. Chen, and D. Fang, "Beyond respiration: Contactless sleep sound-activity recognition using rf signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–22, 2019.

[13] S. Zhang, Y. Zhao, D. T. Nguyen, R. Xu, S. Sen, J. Hester, and N. Alshurafa, "Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 2, pp. 1–26, 2020.

[14] L. Peng, L. Chen, Z. Ye, and Y. Zhang, "Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, pp. 1–16, 2018.

[15] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[16] C. Wang, Y. Gao, A. Mathur, A. C. De C. Williams, N. D. Lane, and N. Bianchi-Berthouze, "Leveraging activity recognition to enable protective behavior detection in continuous data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–27, 2021.

[17] P. Voigt, M. Budde, E. Pescara, M. Fujimoto, K. Yasumoto, and M. Beigl, "Feasibility of human activity recognition using wearable depth cameras," in *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, 2018, pp. 92–95.

[18] R. Khurana, K. Ahuja, Z. Yu, J. Mankoff, C. Harrison, and M. Goel, "Gymcam: Detecting, recognizing and tracking simultaneous exercises in unconstrained scenes," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–17, 2018.

[19] Y. Wang, K. Wu, and L. M. Ni, "Wifall: Device-free fall detection by wireless networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 581–594, 2016.

[20] T. Zhao, J. Liu, Y. Wang, H. Liu, and Y. Chen, "Ppg-based finger-level gesture recognition leveraging wearables," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1457–1465.

[21] Q. Zhang, J. Jing, D. Wang, and R. Zhao, "Wearsign: Pushing the limit of sign language translation using inertial and emg wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–27, 2022.

[22] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[23] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 351–360.

[24] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen, "Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition," in *Proceedings of the 13th ACM conference on embedded networked sensor systems*, 2015, pp. 127–140.

[25] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, 2016.

[26] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *Proceedings of the international conference on internet of things design and implementation*, 2019, pp. 49–58.

[27] C. Li, M. Liu, and Z. Cao, "Wihf: Enable user identified gesture recognition with wifi," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 586–595.

[28] H. Kong, L. Lu, J. Yu, Y. Chen, X. Xu, F. Tang, and Y.-C. Chen, "Multiauth: Enable multi-user authentication with single commodity wifi device," in *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2021, pp. 31–40.

[29] L. Chen, Y. Zhang, and L. Peng, "Metier: A deep multi-task learning based activity and user recognition model using wearable sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–18, 2020.

[30] S. R. Ramamurthy, S. Chatterjee, E. Galik, A. Gangopadhyay, N. Roy, B. Mitra, and S. Chakraborty, "Cogax: Early assessment of cognitive and functional impairment from accelerometry," in *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2022, pp. 66–76.

[31] X. Xu, H. Shi, X. Yi, W. Liu, Y. Yan, Y. Shi, A. Mariakakis, J. Mankoff, and A. K. Dey, "Earbuddy: Enabling on-face interaction via wireless earbuds," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[32] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "Earecho: Using ear canal echo for wearable authentication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–24, 2019.

[33] N. Bui, N. Pham, J. J. Barnitz, Z. Zou, P. Nguyen, H. Truong, T. Kim, N. Farrow, A. Nguyen, J. Xiao *et al.*, "ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear," in *The 25th annual international conference on mobile computing and networking*, 2019, pp. 1–17.

[34] A. Ferlini, A. Montanari, C. Mascolo, and R. Harle, "Head motion tracking through in-ear wearables," in *Proceedings of the 1st International Workshop on Earable Computing*, 2019, pp. 8–13.

[35] F. Kawsar, C. Min, A. Mathur, and A. Montanari, "Earables for personal-scale behavior analytics," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 83–89, 2018.

[36] J. Liu, W. Song, L. Shen, J. Han, and K. Ren, "Secure user verification and continuous authentication via earphone imu," *IEEE Transactions on Mobile Computing*, 2022.

[37] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.