Landmark Detection Based on Human Activity Recognition for Automatic Floor Plan Construction

Zhao Huang[1], Stefan Poslad[1], Qingquan Li[2], Jianping Li[3], and Chi Chen[3](✉)

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK
{zhao.huang,stefan.poslad}@qmul.ac.uk
[2] School of Architecture and Urban Planning, Shenzhen University, Shenzhen 58000, China
liqq@szu.edu.cn
[3] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China
{lijianping,chichen}@whu.edu.cn

Abstract. Landmark detection technology has a wide range of applications in people's lives, including map correcting, localization and navigation, etc. Besides, landmarks are also utilized to label different areas for automatic floor plan construction. Currently, vision-based landmark detection methods have some limitations, such as light, camera shaking, and privacy-invasive. In addition, deep learning-based methods increase the time consumption of marking labels due to the huge requirement for data. Targeting the above challenges, our work first proposes a landmark detection approach based on Human Activity Recognition (HAR) for automatic floor plan construction, which introduces a self-attention model to recognize various landmarks by walker's daily activities due to their strong correlation. First, the accelerometer and gyroscope sensor data are extracted and eliminated by a Gaussian filter and are divided into the same length segments by slide window. Next, it is input into the self-attention network to train a human activity recognition model. Finally, the corresponding relationship between human activities and landmarks is created to detect landmarks through the trained HAR model. Empirical results on two publicly available USC-HAD and OPPORTUNITY datasets show our proposed approach can recognize landmarks effectively.

Keywords: Landmark detection · Automatic floor plan construction · Human activity recognition · Sensors · Self-attention

1 Introduction

Recently, localization is the technique to determine the position of an object or a person [1]. An indoor localization system is a system that attempts to find the accurate position of the object inside a building, mall, etc. The popularity of mobile computing [2–6] stimulates extensive research on the localization of persons or assets. In the present era

H. Gao et al. (Eds.): CollaborateCom 2022, LNICST 461, pp. 459–477, 2022.
https://doi.org/10.1007/978-3-031-24386-8_25

of mobile devices, location information is crucial in a wide range of applications such as manufacturing, healthcare, etc. In order to meet the user's needs, the location information of persons or assets is required which can be provided by the indoor localization system, which tries to identify the position of moving devices with the help of some fixed nodes and some mobile computing devices [7]. However, floor plan construction is an important stage to achieve accurate indoor localization.

Currently, there are numerous approaches to automatic floor plan construction that have been proposed [8–10], because floor plan construction is an important stage to achieve accurate indoor localization. Most maps lack labels for different spaces, they just construct the outline of buildings [11]. However, it is necessary to name some special places, such as where is a bedroom, where is a stair, where is the door of the kitchen, etc. Thus, labeling these spaces using landmarks becomes a good choice. Landmarks that are frequently used include some special structures of buildings or roads, some symbolic objects, etc. In an indoor environment, daily landmarks include doors, stairs, beds, chairs, etc., which often cause humans to perform some special actions, such as walking upstairs or downstairs, opening or closing a door, and sleeping, as shown in Fig. 1.

Fig. 1. Landmarks in our daily life

A great deal of landmark detection methods has been published [12–15] due to their importance in marking areas for automatic floor plan construction and assisting in the construction of sparse maps. These landmark detection approaches can be divided into two categories: signal-based [16–18] and vision-based [19–21]. Vision-based are using computer vision technology to detect special objects as landmarks, but the limitation of these approaches includes two aspects, on one hand, the estimation accuracy is significantly affected by light and camera shaking. Meanwhile, the vision-based methods have a common privacy-invasive issue that are people concerned about. On the other hand, the data-driven landmark detection method based on deep learning algorithms [22, 23]

is more popular due to its automatic feature extraction capabilities rather than mandate domain knowledge to craft shallow heuristic features, but deep learning methods require massive training data, which means people have to spend a lot of time labeling a large amount of data for training a reliable model, as a result, it dramatically increases the workload of manually signing labels. Compared to vision-based methods, signal-based such as radar, lidar, wifi, etc. have the multipath effect issue, and the performance changes with surroundings, which causes an unstable system. Meanwhile, deployment is more complex than mobile sensors. For example, in [16], scattering information from the polarimetric radar is used to detect point-shaped landmarks, which don't consider line-shaped landmarks, meanwhile, the scattering information is unstable and easily changes with surroundings. In [21], drogue's landmark detection is developed for autonomous aerial refueling of unmanned aerial vehicles, but the jitter of the camera or drogue in the air and the change of light will have a great impact to measure the position. But mobile and wearable sensors, particularly tiny sensors not only avoid being disturbed by the shape and surroundings of landmarks but also overcome the limitation of light and shaking of the camera.

Nevertheless, landmark detection approaches based on mobile or wearable sensors avoid the huge workload of labeling, because the label is just recorded when collecting data rather than marking objects from pictures one by one. Meanwhile, it also does not need to extract features from pictures to detect landmarks. Motivated by the above reasons, we introduce a unified and novel landmark detection method for automatic floor plan construction based on human activity recognition (HAR) using the self-attention network, where HAR is firstly adopted to detect the landmark in an indoor environment owing to the close relationship between human daily activities and landmarks. The reason why we choose a self-attention network is that it not only gets the global and local connection synchronously within one step but also will not be limited by the sequence length for the capture of long-term dependence like the RNN network. Meanwhile, the results of each step do not depend on the previous step and can be made into a parallel mode. Compared with CNN and RNN, the parameters are fewer and the complexity of the model is lower. Compared to the other landmark detection methods, our approaches not only avoid the negative impact of light and camera shaking but also solve the privacy-invasive problem. Meanwhile, the workload of labeling data decreases significantly for experiment operators due to the convenience of only recording the labels when collecting sensor data of different activities. The process of our approach is as follows, firstly, the timestep records, accelerometer, gyroscope, and IMU data are input into a sensor modality attention to calculate the attention score, and then, these attention scores are utilized to infer the relative weights of each time-step in the sequence and transform the presentation of each time-step. Following that, global temporal attention is used to rank the importance when predicting the categories of human behaviors. What's more, the human activity model based on self-attention is trained. Finally, the trained HAR model is utilized to recognize human daily activities, which means the landmarks also are detected according to the corresponding relationship between human activities and landmarks, so the name of labels in the floor plan can be obtained. The main novelties of this paper include 1) firstly proposing to build the corresponding relationship between waypoints and human daily activities for the automatic landmark detection task, which

is different from vision-based methods in this domain, this approach is a scheme with fewer privacy issues for detecting landmarks automatically through human behaviors rather than using images. 2) developing a novel landmark detection based on human activity recognition for automatic floor plan construction. Currently, the state-of-the-art vision-based landmark detection approaches have common issues: camera shaking and light changing. For signal-based methods, the surroundings changes will cause landmark estimation error, which is unstable. But our proposed method overcomes the limitation of vision-based and signal-based methods and enhanced the robustness. 3) automatically labeling areas, such as where is the bedroom, kitchen, etc.

The rest of the paper is organized as follows. The related work is covered in Sect. 2. Section 3 will present the proposed methodology, including the framework, sensor modality attention, self-attention block, global temporal attention, and landmark detection. Finally, in Sect. 4, the performance of the proposed approach will be shown, and the following is the conclusion and future work.

2 Related Work

With the development of automatic floor plan construction technologies, more and more landmark detection methods are being used to assist in labeling different areas in a floor plan using these recognized landmarks. Typical landmark detection can be divided into three categories: image-based and signal-based. The image-based methods extract features from a set of pictures near the landmarks, while the signal-based is to receives the flection signals as the features to identify landmarks.

The majority of landmark detection is based on images, with machine learning or deep learning algorithms extracting features. For instance, Rous et al. [20] developed a natural landmark detection model based on a priori knowledge of the shape and functionality of searched structures, which combines region based as well as edge-based elements to detect indoor landmarks, especially for these have clear line structures and large homogenous color surfaces. But the model is unstable due to the lighting fluctuations. Zheng et al. [22] designed an efficient and robust landmark detection model using 3D deep learning in volumetric data, which greatly improves the detection speed and generalization capability, however, it is still more time-consuming than the current methods due to the double training process. Han et al. [24] proposed a multi-resolution regression-guided landmark detection frame to recover Haar-like appearance features from CT pictures and locate human organs, this framework overcame the problem of inaccurate matching due to the distant and changed corresponding structures. Unfortunately, the performance is poor for identifying larger organs. Likewise, A large-scale anatomical landmarks detection approach is presented by Zhang et al. [25], compared to other methods, this algorithm greatly reduces the requirement for training data, however, the fixed patch size used caused the high difference between various landmarks. Jheng et al. [26] demonstrated a convolutional neural network (CNN)-based algorithm (GUTAID) for landmark detection, which achieved and further characterize polyps for optical diagnosis. But the experimental images are not enough, and some images are not high-definition images. On the whole, the common limitations of these landmark detection approaches are the high requirement for light and heavy workload of data labeling.

Zhang, Z. et al. [27]. Designed an optimal facial landmark detection model based on Deep Convolutional Network (TCDCN), which successfully uses the back-propagation algorithm to enhance the generalization, unfortunately, the number of tasks is limited. Liu et al. [28] confuse millimeter-wave radar and camera to improve the performance of target recognition and the environmental awareness capabilities of the autonomous vehicle under severe weather conditions but the efficiency will face the challenge of one-way sensor failure. Wang et al. [29] introduce the fine-grained channel state information (CSI) from off-the-shelf WiFi to detect users' baggage, this scheme is low-cost and eligible for deployment, but the performance is easily disturbed by bag material. Beltrán et al. [30] first design an efficient LiDAR-based 3D object detection for driving environments by using a state-of-art CNN framework, which is suitable for on-board operation, however, the number of channels is limited, which causes the loss of some discriminative features and lower the performance of this method. Zhou et al. [11] introduce an Activity Landmarkbased Indoor Mapping system via Crowdsourcing (ALIMC), which constructs the landmark-based indoor map without any prior knowledge by connecting human activity patterns with the landmarks, but all these activities can not be recognized by a common model and the accuracies of the traditional detection algorithms are not better than deep learning methods. Zhou et al. [31] develop a fast, fine-grained, and low-cost floor plan construction system using sound signals suitable for heterogeneous microphones on commodity smartphones, which achieves good performance, unfortunately, the sound signals are easy to be distributed by noise from surroundings, which causes the bad robustness.

Many publications focus on human activity recognition [32, 33], including wearable sensor-based HAR and vision-based HAR. Varshney et al. [34] introduced the multiple CNN streams to recognize human activities from video by fusing spatial and temporal information, where the average and convolution fusion methods are discussed. Although this method outperforms other approaches, the model does not support multiple input modalities. Liu et al. [35] proposed a compound deep neural network including two sub-networks to generate optical flow images and extract the spatial-temporal information from RGB images respectively, and the spatial-temporal information is integrated to recognize human activities. Although this method achieves a good result, the complexity of network the remains to be improved. A human activity behavior based on stacked sparse autoencoder, and the history of binary motion image is shown by Gnouma et al. [36], which simplified the complexity of silhouette extraction, the limitation of this method is only some special behaviors can be identified. Snoun et al. [37] recognized activities using fine-tuning pre-trained CNNs via human skeletons from the frames, although it almost achieved good accuracy, the performance heavily relies on the result of pose estimation. Murad et al. [38] develop a HAR framework based on deep recurrent neural networks (DRNNs), which can process the variable-length sequences from the input layer, however, the data used is small scale and the generalization is poor. The embedding-based inception neural network and recurrent neural network landmark detection are developed by Xu et al. [39] to classify actions with multi-dimensional features, achieving high accuracy and good generalization, but the kernel size of the model remains optimized. A pattern-based HAR model is proposed by Zhang et al. [40], which established a correlation between signal variation from diffraction sensors

and human actions so as to match them. Similar to Zhang, Yan et al. [41] designed a WiAct to recognize activities through the correlations between body movement and the changes of Channel State Information (CSI) signals changes due to various human activities. However, the two approaches are disturbed heavily by noise and the signal is easy to be sheltered. Bashar et al. [42] produced a time-frequency-based human activities recognition model using activity-driven hand-crafted features, which achieve comparable accuracy and reduce the computation time, but the production of hand-crafted features is time-consuming.

3 Methodology

This section describes the details of our landmark detection approach, the aim of this method is to train a landmark detection model based on human activity pattern recognition by introducing a self-attention mechanism without any recurrent architectures. The method includes four parts: sensor modality attention, self-attention block, global temporal attention, and landmark detection, the detailed specification is shown in the below subsequent sections.

3.1 The Framework of Our Proposed Method

The framework of the proposed approach is illustrated in Fig. 2, the accelerometer and gyroscope data are denoised through a Gaussian filter and then the sensor modality attention is utilized to calculate the weight of the two sensor's data according to their attention score respectively. Following this, the weighted sensor data are transferred into a fixed-size vector over single time steps through a 1-D convolution. Afterward, the values of the two math functions: sine and cosine, are added to the fixed vectors for encoding the position information of the samples in the sequence. Except for that, the feature presentation is scaled through the sqrt function and input into the self-attention block model.

In the self-attention block, the dot product is applied to obtain the new feature presentation of each time step and this new presentation is the input of the global temporal attention layer, and this layer generates the final feature presentation through learning parameters to set varying attention across the temporal dimension [43], Finally, the final feature presentation is utilized by the fully connected and soft-max layers.

3.2 Sensor Modality Attention

The aim of the sensor modality attention mechanism is to obtain the various contribution from the different modalities of sensors and then reduce the impact of lower contribution sensor data, taking sitting as an example, the data from these sensors placed on people's legs contain lower information, so the sensor modality attention mechanism is used to reduce the related weights. It means the sensor modality attention can capture the dependencies through learning such relationships using 2-D convolution across time-step and sensor values [43].

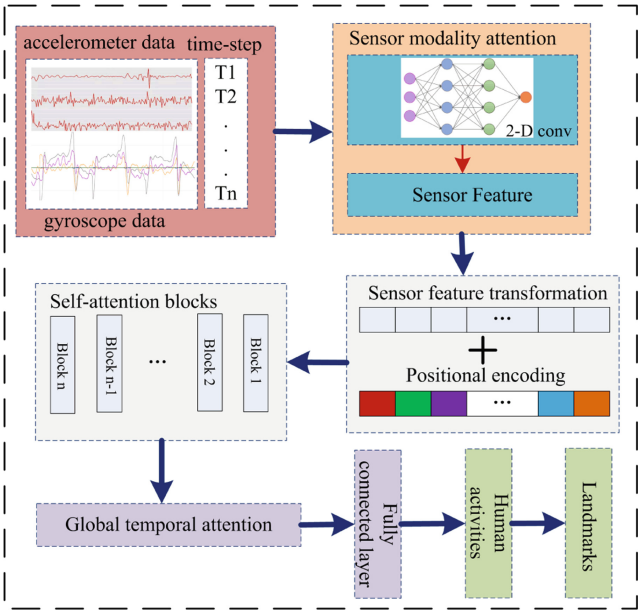


Fig. 2. The framework of our proposed method.

Firstly, the sensor data is converted to single-channel data, and then k convolutional filters are introduced to reshape the single-channel data into k channels data. Following this, the k channel data are converted back to one-channel data using a single convolution. To generate different attention scores for the various modality of sensors, the sensor-wise soft-max function is applied, the definition of this sensor-wise is as follows:

$$S_k^{(t_i)} = \frac{\exp(q_k^{t_i})}{\sum_k \exp(q_k^{t_i})} \tag{1}$$

In this equation, k is the label of sensors. Thus, the weight of the input can be calculated by this equation. Besides, the sensor modality attention also provides feature maps to show the interpretation of this model.

3.3 Self-attention Block

Self-attention layer is the core layer, which consists of self-attention blocks and each block also contains one multi-headed self-attention and one position-wise feed-forward layer [Saif Mahmud]. There are two main functions for this self-attention, one is to infer the relative weights of each time step in the sequence according to the similarity between this time step and all other time steps, and the other is that the feature presentation of each time step is transformed through building the relationship between the weights and the importance of information from other time-steps. Equation (2) show the process of

calculation [Attention Is All You Need].

$$A^{h_s}(Q, K, V) = soft\max(\frac{Q \cdot K^T}{\sqrt{d_k}})V \tag{2}$$

In Eq. (2), the (Q, K, V) indicates the learned linear transformation of input, Q, K, and V are key, query, and value respectively. In other words, the transformed vector of one specific time step can integrate the query, which is compared to the key vector at every other time step using a dot product. Following that, the dot product value is scaled by $\sqrt{d_k}$ and the soft-max function is used to normalize these values, the output of softmax is the attention score. Finally, the weighted representation of the value vectors for each of the time steps is inferred by the attention score.

To capture muti-aspects attention score, multi-headed self-attention is developed, and each head can parallel compute the attention value of the corresponding aspects. As shown in (2), the h_s presents the output of the head s, when all attention heads are concatenated to calculate the corresponding attention scores, these scores are converted back to the score dimension produced by the single attention head using the learned parameter. As shown in Eq. (3).

$$M = R \cdot concat(A^{h_1}, A^{h_2}, ...A^{h_{n-1}}, A^{h_n}) \tag{3}$$

In this equation, R is the learned parameter. In addition, every position in one block is corresponding to the position-wise feed-forward layer independently, which means the weights of the same block are consistent, but it is different across the blocks. Meanwhile, both sub-layers include one normalization layer and one residual connection.

3.4 Global Temporal Attention

TO rank the importance when predicting the categories of human behaviors, the attention score of each time-step output from the self-attention blocks and learned related parameters are input into the global temporal attention model. Firstly, the function C_s is built for capturing the temporal context when calculating the attention score, as shown in Eq. (4).

$$C^{(t_i)} = \tanh(R \cdot S^{(t_i)} + P) \tag{4}$$

In Eq. (4), R and P are the learned parameters generated from self-attention blocks.

$$\psi^{(t_i)} = \frac{\exp((C^{(t_i)})^T \cdot C_s)}{\sum_t \exp(C^{(t_i)} \cdot C_s)} \tag{5}$$

As for the ranking, which is calculated by Eq. (5) and utilized to produce a weighted average of the representations of all the time steps in an activity window [], meanwhile, it is input into the feed-forward layers as a feature vector to classify human behaviors.

$$N_i = \sum_t \psi^{(t_i)} S^{(t_i)} \tag{6}$$

Finally, the weights of all time steps are used to calculate the weighted summation by Eq. (6), which also forms a feature vector. To improve the efficiency of training, the dropout layer is added to the self-attention blocks and the fully connected layer, which is also used after positional encoding.

3.5 Landmark Detection

Compared to image-based and appearance-based landmark detection technology, our approach estimates landmarks through human activity recognition based on a self-attention mechanism, the advantage of our method is that it is not affected by the light and appearance of objects. However, how constructing the relationship between human behaviors and landmarks is a key stage. As we know, the area of human activities and landmarks are the same place when people take activities, which indicates human actions and landmarks are interchangeable. Thus, the correlation between human behaviors and landmarks of our work is constructed and shown in Table 1. 12 kinds of human daily activities, containing: walking upstairs and downstairs, up and down elevators, sitting and sleeping, opening and closing doors, opening and closing fridges, and opening and closing dishwashers, are correlated with 6 common landmarks: stair, elevator, chair, bed (bedroom), door, fridge, and dishwasher (kitchen).

From Table 1, it is obvious that different human activities have corresponding landmarks. So, these landmarks are recognized easily when the related human behaviors are classified accurately through the self-attention mechanism.

Table 1. The correlation between human behaviors and landmarks.

Behavior	Landmark	Behavior	Landmark
Up/downstairs	Stair	Open/close a door	Door
Up/down elevators	Elevator	Open/close fridges	Fridge
Sitting	Chair	Open/close dishwashers	Dishwasher (kitchen)
Sleeping	Bed (bedroom)	\	\

To estimate the performance of different algorithms for landmark detection, this paper introduces the macro average F1-score (MAF1-score) as the metric, as shown in Eq. (7).

$$MAF1\_score = \frac{1}{|N|}\sum_{j=1}^{N}\frac{2 * P_j * \mathrm{Re}_j}{P_j + \mathrm{Re}_j} \tag{7}$$

In this equation, N is the class quantity of human activities, j is the label of each class, P_j and Re_j indicates the precision and recall of the j class.

4 Experiment Results

To verify the efficiency of this landmark detection algorithm based on human activity recognition, the two publicly available USC-HAD and OPPORTUNITY datasets are chosen to carry out experiments, because various daily activities data is included in both datasets, and the hardware of this experiment is a computer with an intel i7-9750H CPU, and the working frequency of this CPU is 2.6G Hz.

4.1 Dataset and Preprocessing

4.1.1 Dataset Description

1) USC-HAD Dataset: The USC human activity (USC-HAD) dataset [44] collected a triaxial accelerometer and gyroscope data using the MotionNode sensing platform, which contains 12 general human activities, and invites 14 participants to install the MotionNode sensors on their front right hip to collect data. Meanwhile, the sampling frequency is set to 100 Hz, and everyone repeats each activity five times, the parameters include: the activity's name, subject number, etc. are recorded by a nearby observer. The human daily activities include walking downstairs/upstairs, turning left/right going along with a circle, sitting, etc.
2) OPPORTUNITY Dataset: The OPPORTUNITY dataset [45] is a public dataset, which is published on the UCI Machine Learning repository and records both static/periodic and sporadic activities from wearable, objects, and ambient sensors in daily living. This dataset records 4 subjects performing 16 types of activities, and each person undertakes one ADL session and one drill session five times, and the drill run consists of 20 repetitions activities. The difference between the ADL session and the drill run session is that the ADL collects a series of human morning activities, which is continuous, and the drill run records some repetitive activities, including opening/closing a door, opening/closing a fridge, sleeping, etc. The frequency sampling of the Drill run is 32 Hz.

4.1.2 Data Preprocess

1) USC-HAD Dataset Preprocessing

   In this dataset, there are six types of human daily activities: sitting and sleeping, walking upstairs and downstairs, up-elevators, and down-elevators, which are chosen to perform our experiments. These activities are corresponding to four common landmarks in our daily life: chairs, beds, stairs, and elevators. To accurately estimate the four landmarks through the six actions, Firstly, the raw data of the six activities from the accelerometer and gyroscope are extracted and input into a gaussian filter for reducing the noise produced from collecting data. Then, these filtered data are divided into many segments by the fixed windows with 50% overlapping, the window size is 32. Finally, the Leave-One-User-Out (LOOCV) strategy is adopted to split the data of 14 users into 14 different groups respectively.

Figure 3 shows the signal changes of the accelerometer and gyroscope when users perform taking up-elevator and down-elevator, walking upstairs and downstairs, as can be seen from these pictures, the accelerometer and gyroscope signals have a strong periodicity totally. Compared to (a) and (b), the pictures from (c) and (d) show more regular changes. In Fig. 3 (a) and (b), the value of the z-axis decreases or increases obviously because taking the up-elevator and down-elevator only causes the acceleration changes in the vertical direction, but (c) and (d) are different due to producing acceleration changes in three directions, which indicates that up-elevator and down-elevator are harder to recognize than walking upstairs and downstairs.

2) OPPORTUNITY Dataset Preprocess

The six kinds of activities from the OPPORTUNITY dataset: opening and closing a door, opening and closing a fridge, and opening and closing a dishwasher are introduced to identify three types of landmarks. The reason why we choose the six activities is that the three corresponding landmarks with the six actions are an indispensable part of our life, which exist in our office, home, supermarket, etc. Firstly, an interpolation algorithm with the "movmedian" method is utilized to fill in the missing data due to the incomplete collected data from IMUs, where the window size is set to 10. Then, the Gaussian filter is adapted to relieve the negative impact of noise and the size of the window is designed to be 20. Followed by this, a fixed window with 50% overlapping divides these signals into the 32 points segments in a row. Finally, the accelerometer and gyroscope data of five Activities of Daily Living (ADLs) and the Drill from Subject 1, the Drill data and the data range from ADL1 to ADL3 of subjects 2 and 3, are extracted from the public dataset for training data, and the rest data from subject2 and subject3 are abstracted for testing the efficiency of the trained model [46].

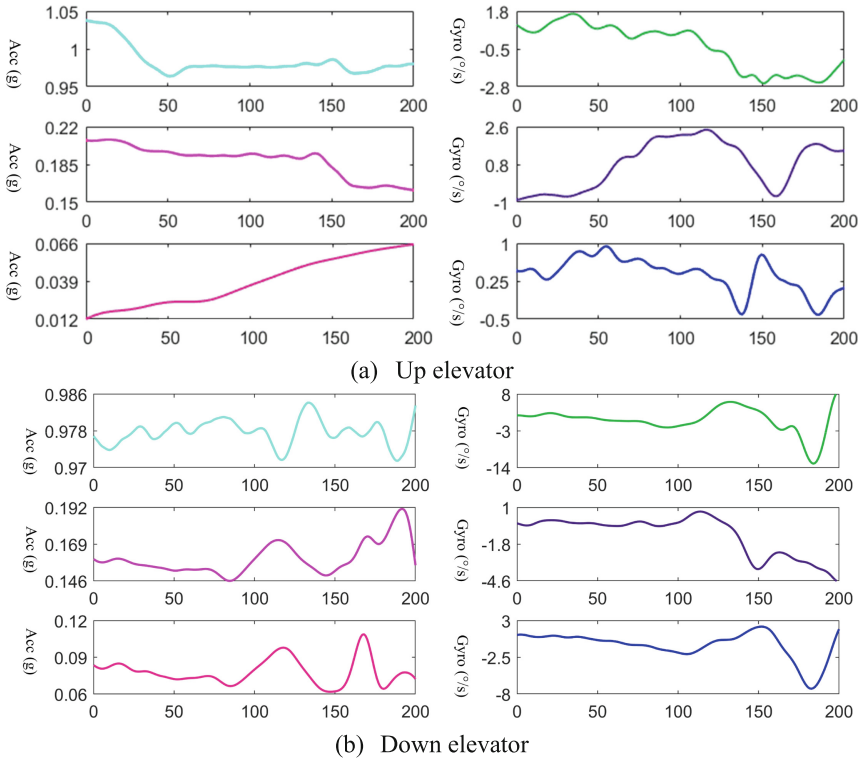
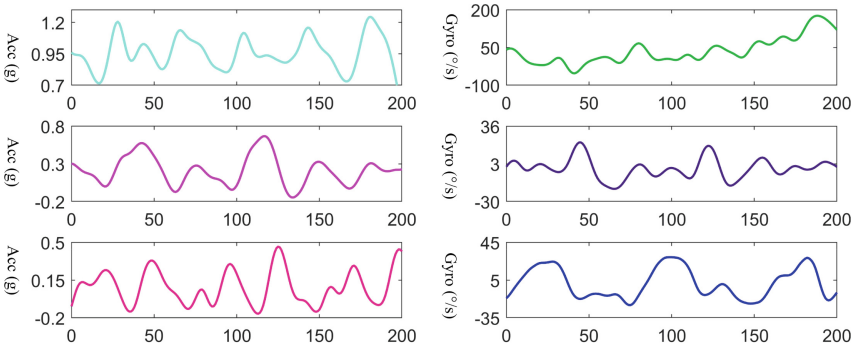
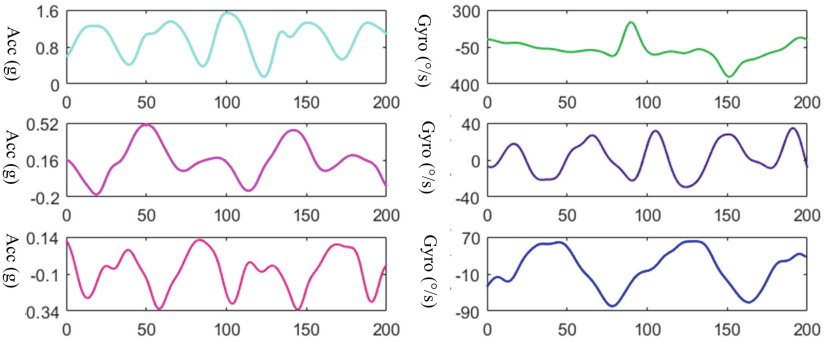


Fig. 3. Accelerometer and Gyroscope signal changes of different activity patterns, (a)–(d) present different activities using different color curves, X-axis presents the number of sampling points, Y-axis presents sensor value. (a) up elevator. (b) down elevator. (c) walk upstairs. (d) walk downstairs.



(c) Walk upstairs



(d) Walk downstairs

Fig. 3. (*continued*)

The signal changes of the accelerometer and gyroscope to different patterns from the sensors on the right wrist are shown in Fig. 4. As can be seen in this picture, the four types of daily activities are not sensitive to the accelerometer and gyroscope, and the value of three axes change less. Compared to walking upstairs or downstairs, the signal changes are more irregular than the signals from other activities due to the complexity of hand gestures when opening or closing a door, fridge, etc.

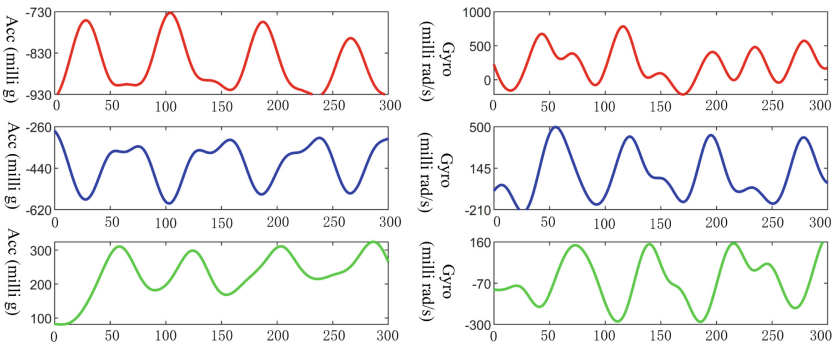
4.2 Experiment Result Analysis

We perform extensive experiments based on the two benchmark datasets discussed in the last section with Convolutional Autoencoder Long Short-Term Memory (ConvAE-LSTM) [46], and Deep Convolutional LSTM (DeepConvLSTM) [47]. The experiment results are listed in Table 2.

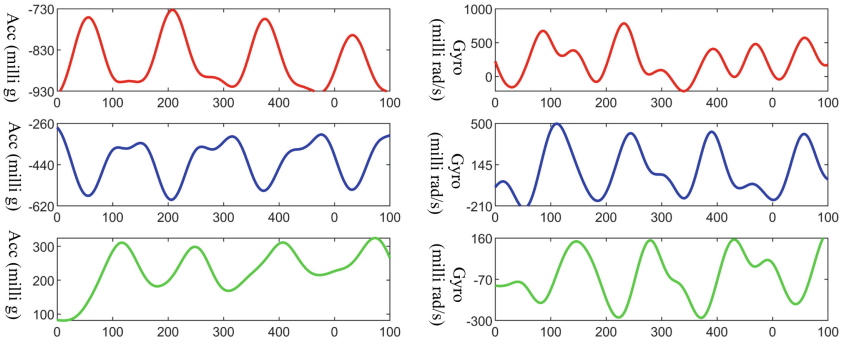
As can be seen from this table, the landmark detection approach proposed in this paper outperforms DeepConvLSTM and ConvAE. Compared to the two methods, the Macro F1-score of our method improves from 0.69 (DeepConvLSTM) and 0.76 (ConvAE) to 0.84 respectively for the USC-HAD dataset and increases by 0.08 (DeepConvLSTM) and 0.05 (ConvAE-LSTM) for OPPORTUNITY dataset.

Figure 5 presents the classification details of different human activities from the USC-HAD dataset when the self-attention model is used. As shown in the figure, the four landmarks can be recognized totally through the corresponding human activities. However, sleeping and upstairs have the highest f1-score (more than 0.94), while downstairs has the lowest F1-score (0.61) with the greatest misclassification (0.26). To measure the performance of this landmark detection algorithm, the average F1-score of the corresponding activities from the same landmark is applied to present the f1-score of this landmark. Thus, based on the correlation between human behaviors and landmarks (Table 1), we can conclude that the bed (bedroom) has the highest recognition efficiency, while the elevator and stairs have lower F1 scores, which means the bedroom is easy to label automatically when constructing the floor plan.

The classification of human daily activities from the OPPORTUNITY dataset is shown in Fig. 6. It can be seen from these pictures, Opening the door1 and the fridge outperform other activities, and the F1-score is up to 0.985 and 0.9 respectively, but closing the dishwasher has the highest misclassification (up to 0.4). The average F1 scores of doors1, doors2, fridge, and dishwasher are 0.91, 0.66, 0.79, and 0.69, separately.

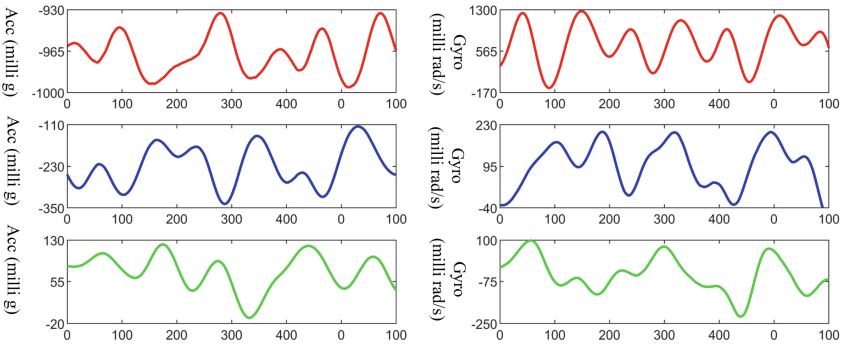


(a) Open a fridge

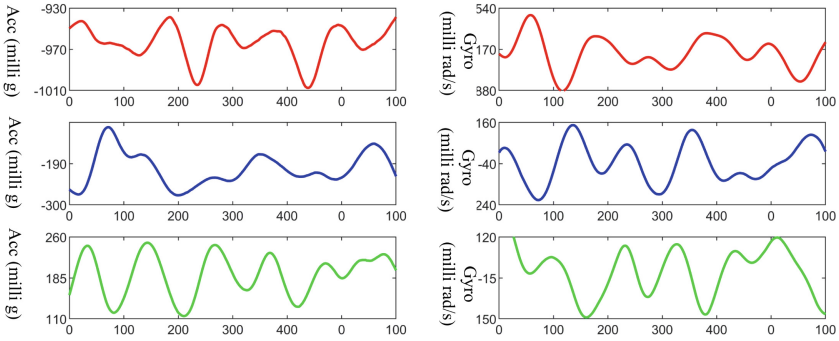


(b) Close a fridge

Fig. 4. Accelerometer and Gyroscope signal changes of different activity patterns, (a)–(d) present different activities using various color curves, X-axis presents the number of sampling points, Y-axis presents sensor value. (a) open a fridge. (b) close a fridge. (c) open a door. (d) close a door.



(c) Open a door



(d) Close a door

Fig. 4. (*continued*)

Table 2. Macro F1-score for landmark detection

Dataset	Proposed method	DeepConvLSTM	ConvAE-LSTM
USC-HAD	0.84	0.69	0.76
OPPORTUNITY	0.76	0.68	0.71

Thus, door1 and the fridge are easier to identify than others, in contrast, door2 is difficult to be categorized. In total, all of these landmark classification results are considerable and beneficial to constructing floor plans.

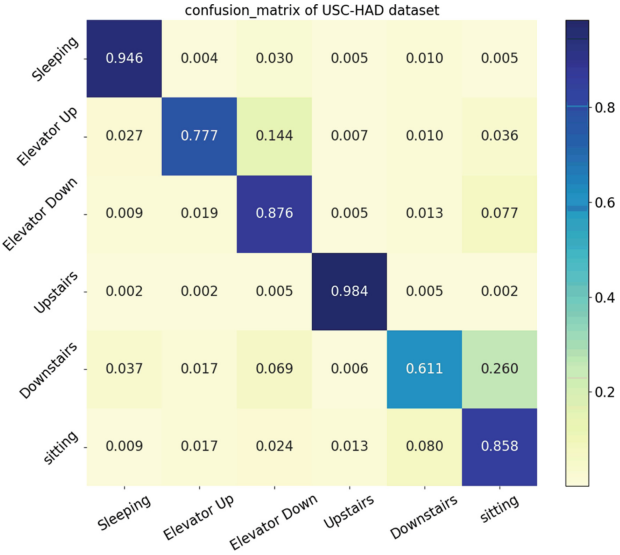


Fig. 5. The confusion matrix of the USC-HAD dataset.

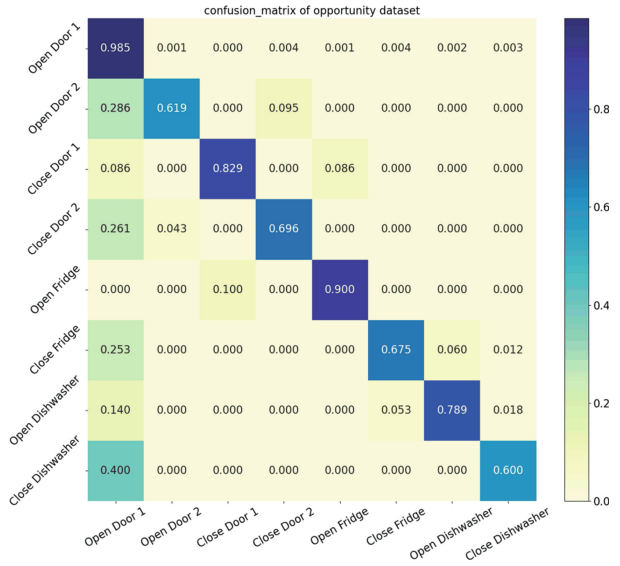


Fig. 6. The confusion matrix of the OPPORTUNITY dataset.

5 Conclusion and Future Work

Landmark detection technology is growing in popularity due to its widespread use in localization and mapping research, and image-based landmark detection methods have

achieved cutting-edge performance. The limitation of image-based approaches, however, is that detection accuracy is severely hampered by light. Meanwhile, it necessitates a massive amount of training data, which significantly increases the workload of labeling labels. Therefore, an algorithm capable of balancing the trade-off between performance and labeling workload is required. In this paper, we first propose a novel landmark detection method for floor plan construction based on human daily activity recognition using the self-attention mechanism. Unlike vision-based methods, our approach estimates landmarks around people based on a strong correlation between landmarks and human activities that extends beyond the limits of light. The proposed landmark detection methods are evaluated on two public HAR benchmark datasets: USC-HAD and OPPORTUNITY datasets, achieving considerable performance. On the whole, the landmark detection algorithm not only estimates landmarks accurately in low-light environments but also significantly reduces the workload of data labeling, which is advantageous for landmark detection with strict latency constraints, therefore, these landmarks can be used to label some places in automatic floor plan construction.

Although our proposed landmark detection algorithm overcomes the constraints of light conditions and the camera shaking problem, several issues remain to be considered in the future. Some human activities that are performed, may have landmarks that can be ad hoc. For example, to improve the estimation efficiency, not all landmarks need to be detected, which is a challenge to decide what types of landmarks should be detected. What's more, the physical design of the same type of landmark can vary. For example, the doors of many public places, such as transport hubs, school libraries, supermarkets, etc., may be push or pull, have different types of handles, or be automatic, so estimating these different types of landmarks, e.g., doors through human activities become more challenging in real life.

References

1. Yanying, G., Lo, A., Niemegeers, I.: A survey of indoor positioning systems for wireless personal networks. IEEE Commun. Surv. Tutorials **11**, 13–32 (2009)
2. Forman, G.H., Zahorjan, J.: The challenges of mobile computing. Computer **27,** 38–47 (1994)
3. Barry, B., et al.: Educating for mobile computing: addressing the new challenges. In: Proceedings of the Final Reports on Innovation and Technology in Computer Science Education 2012 Working Groups Haifa, Israel: ACM, pp. 51–63 (2012)
4. Kakousis, K., Paspallis, N., Papadopoulos, G.A.: A survey of software adaptation in mobile and ubiquitous computing. Enterp. Inf. Syst. **4**, 355–389 (2010)
5. Ladd, D., Alan, D., Avimanyu, S., et al.: Trends in mobile computing with in the is discipline: a ten-year retrospective. Commun. Assoc. Inf. Syst. **27**, 285–316 (2010)
6. Gay, G.: Context-aware mobile computing: affordances of space, social awareness, and social influence. Synthesis Lectures on Human-Centered Informatics. Morgan and Claypool Publishers, San Rafael. vol. 2, pp. 1–62 (2009)
7. Sana.: A survey of indoor localization techniques. IOSR J. Electr. Electron. Eng. (IOSR-JEEE). **6**, 69–76 (2013)
8. Alzantot, M.: Youssef, M.: Crowdinside: automatic construction of indoor floorplans. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, New York, United States, pp. 99–108 (2012)

9. X. Zhang, Y. Jin, et al. CIMLoc: A crowdsourcing indoor digital map construction system for localization. In 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Singapore, pp. 1–6, IEEE (2014)
10. Elhamshary, M., Alzantot, M., Youssef, M.: JustWalk: a crowdsourcing approach for the automatic construction of indoor floorplans. IEEE Trans. Mob. Comput. **18**(10), 2358–2371 (2018)
11. Zhou, B., Li, Q., Mao, Q., Tu, W., et al.: ALIMC: activity landmark-based indoor mapping via crowdsourcing. IEEE Trans. Intell. Transp. Syst. **16**(5), 2774–2785 (2015)
12. Amarasinghe, D., Mann, G.K., Gosine, R.G.: Landmark detection and localization for mobile robot applications: a multisensor approach. Robotica **28**(5), 663–673 (2010)
13. Alansary, A., Oktay, O., et al.: Evaluating reinforcement learning agents for anatomical landmark detection. Med. Image Anal. **53**, 156–164 (2019)
14. Nilwong, S., Hossain, D., et al.: Deep learning-based landmark detection for mobile robot outdoor localization. Machines **7**(2), 25 (2019)
15. Wang, Z., Vandersteen, C., Raffaelli, C., Guevara, N., Patou, F., Delingette, H.: One-shot learning for landmarks detection. In: Engelhardt, S., et al. (eds.) Deep Generative Models, and Data Augmentation, Labelling, and Imperfections. Lecture Notes in Computer Science, vol. 13003, pp. 163–172. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88210-5_15
16. Weishaupt, F., Will, P.S., et al.: Robust point-shaped landmark detection using polarimetric radar. In: 2021 IEEE Intelligent Vehicles Symposium (IV), pp. 859–865, IEEE (2021)
17. Narayana, K., Goulette, F., Steux, B.: Planar landmark detection using a specific arrangement of LIDAR scanners. In: IEEE/ION Position, Location and Navigation Symposium, pp. 1057–1069, IEEE, May 2010
18. Ravankar, A., Hoshino, Y., Kobayashi, Y.: Robust landmark detection in vineyards using laser range sensor. In: The Proceedings of JSME annual Conference on Robotics and Mechatronics (Robomec), pp. 1A1-E03 (2019)
19. Sun, S., Yin, Y., et al.: D. Robust landmark detection and position measurement based on monocular vision for autonomous aerial refueling of UAVs. IEEE Trans. Cybern. **49**(12), 4167–4179 (2018)
20. Rous, M., Lupschen, H., et al.: Vision-based indoor scene analysis for natural landmark detection. In: Proceedings of the 2005 IEEE International conference on Robotics and Automation, Barcelona, Spain, pp. 4642–4647 (2005)
21. Sun, S., Yin, Y., Wang, X., Xu, D.: Robust landmark detection and position measurement based on monocular vision for autonomous aerial refueling of UAVs. IEEE Trans. Cybern. **49**(12), 4167–4179 (2018)
22. Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D.: 3D deep learning for efficient and robust landmark detection in volumetric data. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9349, pp. 565–572. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_69
23. Schwendicke, F., et al.: Deep learning for cephalometric landmark detection: systematic review and meta-analysis. Clin. Oral Invest. **25**(7), 4299–4309 (2021). https://doi.org/10.1007/s00784-021-03990-w
24. Han, D., Gao, Y., Wu, G., Yap, P.-T., Shen, D.: Robust anatomical landmark detection for MR brain image registration. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. Lecture Notes in Computer Science, vol. 8673, pp. 186–193. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_24

25. Zhang, J., Liu, M., Shen, D.: Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Trans. Image Process. **26**(10), 4753–4764 (2017)
26. Jheng, Y.-C., et al.: A novel machine learning-based algorithm to identify and classify lesions and anatomical landmarks in colonoscopy images. Surg. Endosc. **36**(1), 640–650 (2021). https://doi.org/10.1007/s00464-021-08331-2
27. Zhang, Z., Luo, P., et al.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision, Part II, Zurich, Switzerland, pp. 94–108, 6–12 Sep 2014
28. Liu, Z., et al.: Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions. IEEE Trans. Intell. Transp. Syst. **23**(7) 6640–653 (2021)
29. Wang, C., Liu, J., Chen, Y., et al.: Towards in-baggage suspicious object detection using commodity wifi. In: 2018 IEEE Conference on Communications and Network Security (CNS), pp. 1–9. IEEE, May 2018
30. Beltrán, J., Guindel, C., Moreno, F.M., et al.: BirdNet: a 3D object detection framework from lidar information. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3517–3523. IEEE, November 2018
31. Zhou, B., Elbadry, M., Gao, R., Ye, F.: Towards scalable indoor map construction and refinement using acoustics on smartphones. IEEE Trans. Mob. Comput. **19**(1), 217–230 (2019)
32. Dubois, A., François, C.: Human activities recognition with RGB-Depth camera using HMM. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, Osaka, Japan, 3–7 Jul 2013
33. Wang, K., He, J., Zhang, L.: Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors. IEEE Sens. J. **19**(17), 7598–7604 (2019)
34. Varshney, N., Bakariya, B.: Deep convolutional neural model for human activities recognition in a sequence of video by combining multiple CNN streams. Multimedia Tools Appl. **81**, 1–13 (2021). https://doi.org/10.1007/s11042-021-11220-4
35. Liu, Z., Han, Y., Chen, Z., Fang, Y., Qian, H., Zhou, J.: Human activities recognition from videos based on compound deep neural network. In: Liu, Qi., Liu, X., Shen, T., Qiu, X. (eds.) The 10th International Conference on Computer Engineering and Networks. Advances in Intelligent Systems and Computing, vol. 1274, pp. 314–326. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-8462-6_37
36. Gnouma, M., Ladjailia, A., Ejbali, R., Zaied, M.: Stacked sparse autoencoder and history of binary motion image for human activity recognition. Multimedia Tools Appl. **78**(2), 2157–2179 (2018). https://doi.org/10.1007/s11042-018-6273-1
37. Snoun, A., Jlidi, N., Bouchrika, T., Jemai, O., Zaied, M.: Towards a deep human activity recognition approach based on video to image transformation with skeleton data. Multimedia Tools Appl. **80**(19), 29675–29698 (2021). https://doi.org/10.1007/s11042-021-11188-1
38. Murad, A., Pyun, J.Y.: Deep recurrent neural networks for human activity recognition. Sensors **17**(11), 2556 (2017)
39. Xu, C., et al.: InnoHAR: a deep neural network for complex human activity recognition. IEEE Access **7**, 9893–9902 (2019)
40. Zhang, F., et al.: Towards a diffraction-based sensing approach on human activity recognition. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **3**(1), 1–25 (2019)
41. Yan, H., et al.: WiAct: a passive WiFi-based human activity recognition system. IEEE Sens. J. **20**(1), 296–305 (2019)
42. Bashar, S.K., Abdullah, A.F., Ki, H.C.: Smartphone based human activity recognition with feature selection and dense neural network. In: 42nd Annual International Conference of the

IEEE Engineering in Medicine and Biology Society (EMBC), Montreal, Canada, pp. 20–24 (2020)
43. Mahmud, S., Tonmoy, M.: et al.: Human activity recognition from wearable sensor data using self-attention. arXiv preprint arXiv:2003.09018. (2020)
44. Zhang, M., Sawchuk, A.A.: USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, USA, pp. 1036–1043 (2012)
45. Roggen. D., Calatroni, A., et al:. Collecting complex activity datasets in highly rich networked sensor environments. In: 2010 Seventh International Conference on Networked Sensing Systems (INSS), Kassel, Germany, pp. 233–240, IEEE (2010)
46. Thakur, D., Biswas, S., Ho., et al.: ConvAE-LSTM: convolutional Autoencoder Long Short-Term Memory Network for Smartphone-Based Human Activity Recognition. IEEE Access **10**, 4137–4156 (2022)
47. Lim, X.Y., Gan, K.B., et al.: Deep ConvLSTM network with dataset resampling for upper body activity recognition using minimal number of IMU sensors. Appl. Sci. **11**(8), 3543 (2021)