CCF Transactions on Pervasive Computing and Interaction
https://doi.org/10.1007/s42486-023-00132-x

REGULAR PAPER

Human activity recognition based on multi-modal fusion

Cheng Zhang[1] · Tianqi Zu[1] · Yibin Hou[1,2] · Jian He[1,2] · Shengqi Yang[1,2] · Ruihai Dong[3]

Received: 7 January 2023 / Accepted: 27 April 2023

Abstract
In recent years, human activity recognition (HAR) methods are developing rapidly. However, most existing methods base on single input data modality, and suffers from accuracy and robustness issues. In this paper, we present a novel multi-modal HAR architecture which fuses signals from both RGB visual data and Inertial Measurement Units (IMU) data. As for the RGB modality, the speed-weighted star RGB representation is proposed to aggregate the temporal information, and a convolutional network is employed to extract features; As for the IMU modality, Fast Fourier transform and multi-layer perceptron are employed to extract the dynamical features of IMU data. As for the feature fusion scheme, the global soft attention layer is designed to adjust the weights according to the concatenated features, and the L-softmax with soft voting is adopted to classify activities. The proposed method is evaluated on the UP-Fall dataset, the F1-scores are 0.92 and 1.00 for 11 classes classification task and fall/non-fall binary classification task respectively.

Keywords Human activity recognition · Multi-modal fusion · Fall detection · Convolutional network · Wearable device

1 Introduction

Human activity recognition (HAR) aims to detect and classify human activities, which is widely employed in intelligent security systems, health monitoring, virtual reality etc. Wan et al. (2020). According to the types of the data collection sensors, HAR methods can be categorized into 2 classes, which are ambient sensor-based methods and wearable sensor-based methods.

Ambient sensors are deployed in the surrounding environment of monitored subjects to record human activities. With the advancements of the machine learning technologies, ambient sensor-based methods achieve significant progress in the HAR fields. For instance, Luo et al. (2020) used 2-D LIDAR to perform multiple people activity recognition. They cluster signals into human and nonhuman classes and segmented the human trajectory by Kalman Filter, and then employ a long short-term memory (LSTM) network and a temporal convolutional network (TCN) to classify trajectory samples into 15 classes. Cippitelli et al. (2016) exploit skeleton data extracted by RGB-D sensors to compose feature vectors and then employ cluster algorithm and support vector machine (SVM) to classify human activities. Han and Bhanu (2005) investigate human repetitive activity properties with thermal infrared imagery to remove the affects of lighting conditions and colors of the human surfaces and backgrounds.

Although ambient sensors have many advantages, methods based on ambient sensors suffers from limitations and challenges. One of the major challenges is the occlusion problem. Usually, the installation positions of sensors such as RGB-D cameras and thermal cameras are fixed, and

✉ Jian He
jianhee@bjut.edu.cn

Cheng Zhang
zhangcheng402@emails.bjut.edu.cn

Tianqi Zu
tinccizu@gmail.com

Yibin Hou
ybhou@bjut.edu.cn

Shengqi Yang
shengqi.yang@gmail.com

Ruihai Dong
ruihai.dong@ucd.ie

[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

[2] Beijing Engineering Research Center for IoT Software and Systems, Beijing University of Technology, Beijing 100124, China

[3] Insight Centre for Data Analytics, University College Dublin, Dublin 4, Ireland

Published online: 22 May 2023

parts of the human body can be occluded according to the camera angles, human motions and body orientations etc. In such cases, it is difficult to accurately classify the actions. Another challenge is the complexity of the visual data, where different light conditions, background colors and moving background objects can affect the classification performance.

Wearable sensors are another type of sensors widely employed in HAR methods (Ometov et al. 2021), which are able to collect human activity data continuously without the problem of occlusions and background noises. For example, Salehzadeh et al. (2020) used electroencephalogram (EEG) sensors to detect human activities, and developed a deep learning framework to classify human activities based on EEG artifacts (FCEA). He et al. (2019) designed an IMU-based wearable vest for fall detection, signals from the IMU sensors were converted into a 3-channel image feature. They proposed an FD-CNN network, which combines a convolutional neural network (CNN) with an LSTM network, to learn image features and classify human activities. Balli et al. (2019) integrated accelerometer, gyroscope, step counter, and heart rate sensors with the smartwatch to detect human movements. However, the data collected by wearable sensors are not as informative as that of the visual sensors. For example, falling using hands or knees are distinguishable by the visual data, but classifying them with IMUs are much harder, as these 2 actions produce similar inertial signals.

In order to increase the recognition accuracy of the HAR, some multi-modal approaches are proposed (Chen et al. 2017) to make different modalities complement each other. However, there are 2 challenges on these approaches: (1) The complexity of the data collected from ambient sensors, especially RGB cameras are high. Directly sending the data into neural networks can result in overfitting. Therefore, the handcrafted visual representation method needs to be designed to alleviate the overfit and presents useful information. (2) To make use of the features from different modalities, a fusion scheme needs to be applied. How to design the fusion scheme to reach higher performance on the classification still remains a challenge. In this paper, a multi-modal fused HAR method is proposed to take advantage of 2 types of sensors by combining RGB and IMU modalities, as shown in Fig. 1. The main contributions are as follows:

1) An speed-weighted star RGB algorithm is proposed to generate the trajectory map of human motion according to the RGB data, the high-velocity parts in the trajectory maps are emphasized.
2) The soft-attention ensemble is modified as the global soft-attention layer to fuse the features from different modalities.
3) The L-softmax algorithm with soft voting is employed to classify the activities.

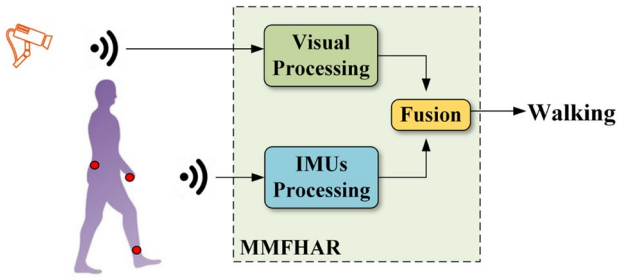

Fig. 1 The proposed method combines visual and IMU signals to recognize activities

2 Related works

Optical Flow (OF) and Motion History Image (MHI) (Bobick and Davis 2001) are widely used in vision-based human motion modelling. OF refers to the displacements of intensity patterns caused by relative motion between an observer and an object (Fortun et al. 2015). The typical Lucas-Kanade algorithm (Lucas et al. 1981) and Horn-Schunck algorithm (Horn and Schunck 1981) obtain the flow direction and density of each pixel at a specific moment by calculating the derivative of the time interval between two adjacent frames, but it is more time-consuming due to a large amount of computation in the derivation operation.

MHI stacks the differences and contour changing among a set of frames and outputs a trajectory (namely Motion Energy Image, MEI), which is widely studied in HAR tasks because of its time-efficiency and good performance (Ahad et al. 2012). For example, Barros et al. (2014) proposes a variant of MHI that stacks the trajectory into a gray-scale image. However, in this conversion process, some temporal sequential information is lost, and it leads to the misclassification of the activities, such as raising hands and putting down hands. To address this problem, dos Santos et al. (2020) proposes a method by amplifying the differences between frames and converting MHI to RGB images rather than gray-scale images so that the three channels can reflect the time series of motion.

In addition, pre-processing raw signals from sensors can help researchers extract apparent or interpretable features of human activities. For example, Mao et al. (2017) deploy IMU sensors on the shoulder, waist, and foot of the subjects, and convert the acceleration and angular velocity into Euler angle to predict whether a person has fallen according to a threshold of Euler angles.

Conventional machine learning methods in the HAR field has the advantage of working with small dataset, but the performances are limited. With the rapid advancement of deep learning techniques in recent years, there

has been an increased interest in combining conventional data representation techniques with deep neural networks in HAR tasks (Demrozi et al. 2020). For example, Ravi et al. (2005) used FFT to extract the frequent features of IMU data and employed base-level and meta-level machine learning methods to classify human activities. Steven Eyobu and Han (2018) apply the short-time Fourier method to analyze IMU data to obtain the spectral features, and adopt an LSTM network to capture sequential patterns, then feed the outputs into an MLP layer to recognize human activities. Rivera et al. (2017) proposed an RNN-based network to fuse the data from multiple IMU sensors. Zimmermann et al. (2018) further utilize the benefits of both CNN and RNN networks to extract features of the data from multiple IMU sensors deployed in lower limbs, and also adopted an MLP layer to classify human activities. Similarly, Stoeve et al. (2021) embedded IMU sensors on shoes and employ CNN and LSTM networks in their solution to classify the football shooting and passing activities.

Studies show that directly applying machine learning solutions on handcrafted features for multi-modal fusion on HAR (Zhu et al. 2019) suffers from kinematic model errors and noises (Li and Wu 2015). In recent years, researchers have explored building end-to-end models, which utilize the advantages of deep learning technologies to extract features from raw data automatically, and integrate with classification tasks seamlessly. For example, Jian et al. introduced a convolutional pose machine and LSTM combined with two kinds of handcrafted features (namely relative bone length and angle with gravity) to recognize traffic police gestures (He et al. 2020). Lu and Velipasalar (2019) use a feature-level (Brena et al. 2020) concatenation to fuse the visual and the dynamic features. Feichtenhofer et al. (2016) introduced the two-stream algorithm to fuse the appearance and motion features of human activities simultaneously. dos Santos et al. (2020) used two CNN networks to extract two kinds of gesture features, and introduced a local soft-attention mechanism to fuse those features so as to classify human activities. In general, convolutional fusion is suitable for the modalities with the same semantic, while local soft-attention is suitable for modalities only considering a subset of features to solve the narrow-sight problem (Luong et al. 2015).

Researchers also study multi-modal fusion to address the flaws within single modality methods. Hwang et al. (2017) combined a fixed camera with a wrist-mounted inertial sensor to collect image and dynamic data of human activities. CNN and RNN were introduced to extract features and classify human activities. The experimental result showed that it solved the problems of spatial constraints, occlusions in images. Mallat et al. (2018) employed a device integrated with a camera, an IMU sensor, and a Wii Balance board to capture human kinematics data of human activities. Meanwhile, extended Kalman filter was introduced to assess the kinetics motion. Lu and Velipasalar (2019) combined an egocentric camera with an IMU sensor to collect the data of fine-grained activities (such as eating, drinking, etc.). Meanwhile, they developed a framework that consists of a capsule network and an LSTM network to classify the fine-grained activities. Abebe and Cavallaro (2017) develop a framework integrating CNN with LSTM and transfer learning to extract features of data both from IMU sensor and camera to recognize egocentric prospective activities.

In conclusion, single-modal approaches in either IMU modality or visual modality face many problems such as resource constrain, occlusion, fine-grained classification and constrained working range, etc. Previous researches have apparently shown that fusing both modalities could offset the aforementioned problems and preserve the merits of each modality. But detailed problems such as MHI could not reflect speed, effective IMU data extraction, automatic modality fusion method are urged to be solved.

3 Methodology

This paper adopts the star RGB algorithm which achieves high performance in gesture recognition into the field of activity recognition, and optimize it by adding a weight term that represents the motion speed to further emphasize the expressive power of the human motion trajectory maps. Meanwhile, an FFT followed by MLP is employed to extract IMU features collected with wearable sensors. Then, a global soft attention mechanism is introduced to fuse the visual and the dynamic features both from the camera and IMU sensors. Last, the large margin Softmax (L-Softmax) combined with a soft voting algorithm is adopted to robustly classify human activities. The overall process is shown in Fig. 2, which contains four parts:

1) Visual feature extraction. The visual features of human activities are extracted from the video captured by an RGB camera, an RGB image of human motion trajectory map is generated from each video fragment, which distinguishes the motion speed and gives prominence to faster movements. Then, the trajectory map is input into a CNN to obtain the visual feature of human activities.
2) IMU feature extraction. The acceleration and angular velocity data from the wearable IMU sensors placed on different parts of the human body are collected, the frequent features of the human activities from IMUs in each period are generated by FFT, and are processed by the MLP to generate lower dimensional dynamic features of human activities.

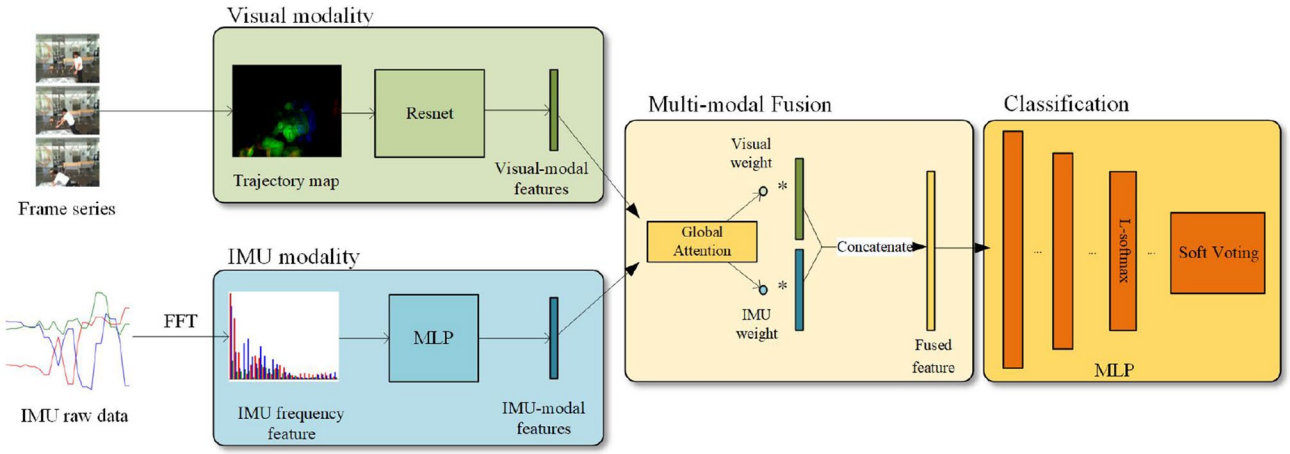

Fig. 2 The overall process of the proposed method

3) Multi-modal fusion. Global soft attention mechanism is employed to fuse the visual and dynamic features of human activities.
4) Activity classification. The L-Softmax (Liu et al. 2016) which can actively expand the distance among features of training samples to identify more implicit differences between different samples is introduced to classify human activities according to the fused multi-modal features. Meanwhile, a soft voting algorithm is adopted to improve the robustness of L-Softmax.

3.1 Visual feature extraction

The conventional MHI-based method assigned a fixed motion strength to each foreground point, and then updated it with a small constant for the background point. It causes body parts with different movement speeds to have similar intensity, and may generate indistinguishable MHI patterns (Tsai et al. 2015). To alleviate this problem, this paper proposes a speed-weighted mask that gives more weight to fast-moving objects based on the star RGB MHI to better represent the motions.

The star RGB representation is proposed by dos Santos et al. (2020), which calculate a human motion trajectory map from a sequence of continuous frames according to Eq. 1.

$$M_k(i,j) = \left(1 - \frac{\lambda}{2}\right) \cdot |(\|I_k(i,j)\|_2 - \|I_{k+1}(i,j)\|_2)| \tag{1}$$

Where $M_k(i,j)$ is the trajectory map generated by k and $k+1$ frames at pixel position (i,j), as shown in Fig. 3. $I_k(i,j)$ is the k frame's RGB vector at position (i,j), $k = 1, 2, 3, \cdots, N-1$. N is the number of frames. λ is calculated according to Eq. 2

$$\lambda = 1 - \cos\theta = 1 - \frac{I_k(i,j)^T I_{k+1}(i,j)}{\|I_k(i,j)\|_2 \cdot \|I_{k+1}(i,j)\|_2} \tag{2}$$

However, the generated motion trajectory map only represents the temporal change in pixel level, The moving speed of the pixels are not represented. This paper further considers the part moving velocities at the regional level, and uses a weight term to emphasize the speed information. For a

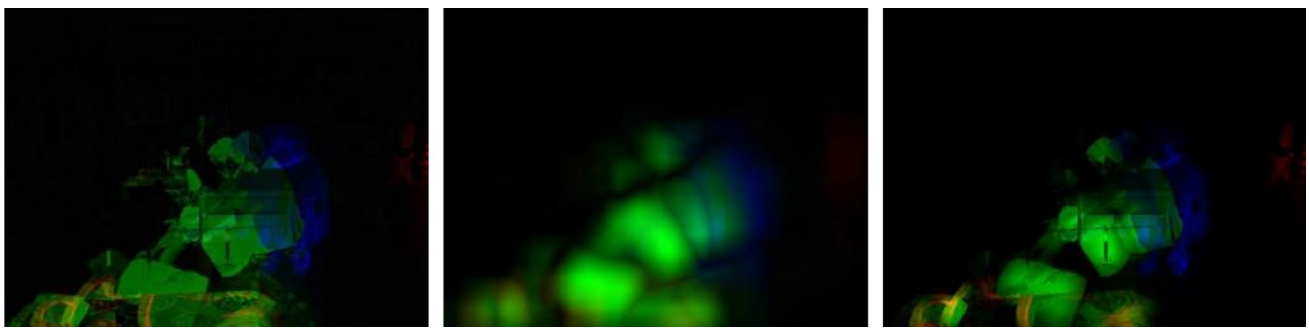

Fig. 3 The heatmaps of star RGB trajectory, speed weight and the weighted trajectory respectively

pixel position (i, j), a region H is defined as the surrounding area of the pixel. In this region, if the total number of the pixels changes dramatically between frames, then this pixel is deemed to have a large motion speed. For such pixels, larger weights are given. Therefore, the weight for a pixel is decided by the pixel differences on the motion image of its surrounding area. The above method is implemented by convolution operation, shown in Eq. 3

$$W_k(i,j) = \left| \sum_H I_{k+1}(i,j) - \sum_H I_k(i,j) \right| \tag{3}$$

Where $W_k(i,j)$ is the speed weight of the pixel at position (i, j) generated according to k and $k+1$ frame. H is the region around pixel (i, j). Enumerating (i, j) within the image by $\sum_H I_k(i,j)$ outputs a regional speed information map at k frame, in which each pixel represents the corresponding surrounding region information. The speed representation, namely speed-weighted mask, is obtained by subtracting the regional speed information map between k and $k+1$ frame. The subtraction operation computes the regional difference between two continuous frames, which reflects the motion speed. Fig. 3 shows the appearance of the speed-weighted mask.

In order to preserve the pixel information from the star RGB trajectory map, the weight term should be normalized into range [0,1]. Let $\min(\cdot)$ and $\max(\cdot)$ be the minimum and maximum matrix, which have the same size as the speed-weighted mask, and they capture elements' minimum and maximum value, respectively. The normalized D_k could be calculated according to Eq. 4.

$$D_k = M_k \cdot \left(\frac{W_k - \min(W_k)}{\max(W_k) - \min(W_k)} \right) \tag{4}$$

Where $\cdot$ is element-wise matrix multiplication.

Then, the temporal information is aggregated. Computing each pair of frames within N frames produces $N-1$ trajectory maps, The star RGB image uses RGB channels to represent the sequential information of these trajectory maps. Let $D = \{D_1, D_2, ..., D_{N-1}\}$ be the trajectory set, a 3-channel RGB image T is generated by aggregating D.

Last, the 3-channel weighted trajectory map T is sent into the Resnet101 network (He et al. 2016) to extract visual features as one of the inputs in the subsequent fusion stage.

3.2 IMU feature extraction

FFT and MLP are adopted to extract IMU dynamic features. FFT converts IMU's temporal domain feature into frequency domain feature, and decomposes the signal into amplitudes in the corresponding frequency. Let A_k and G_k be the signals produced by the k-th accelerometer and gyroscope placed on human body, the signal values on each time points are denoted with

$$A_k = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_t | t \in T\} \tag{5}$$

$$G_k = \{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_t | t \in T\} \tag{6}$$

Where t is the t-th moment in the period T. $\mathbf{a}_t$ and $\mathbf{g}_t$ denote $(a_x^t, a_y^t, a_z^t)^T$ and $(g_x^t, g_y^t, g_z^t)^T$ respectively. Let M_k be the k-th IMU vector, and $M_k = (A_k, G_k)^T$. Then the IMU data matrix is obtained by concatenating K-th IMU data, where $U = (M_1, M_2, \cdots, M_k)^T$.

Let F be the feature matrix generated by decomposing the data matrix U by FFT as $F = \text{FFT}(U)$. Dimensional feature vector F_{flat} is then obtained by flattening the feature matrix F. The F_{flat} is sent into MLP as $F' = \text{MLP}(F_{\text{flat}})$ Where F' is the final IMU modal feature as one of the inputs in the subsequent fusion stage.

3.3 Multi-modal fusion

With the above processing steps, visual features and IMU features are extracted, the spatial and temporal dimensions of the data are aggregated, only the channel dimension remains. To classify the activities, the probability of these features belonging to each class needs to be computed, therefore the features from different modalities must be fused. The conventional fusion schemes are concatenation and averaging. As for the concatenation, the features are concatenated before sending into the MLP and the softmax layers. As for the averaging, features from different modalities are sent into the MLP and the softmax layers to produce the class probability distributions, then the distributions are averaged as the final result.

In this paper, a concatenation based multi-modal fusion scheme is proposed, where a global soft attention mechanism is employed during the fusion process to weight the features from different modalities. The soft attention mechanism is proposed by dos Santos et al. (2020) to fuse and weight the features from two networks, as shown in Eq. 7, where F denotes the features and F_{fuse} denotes the attention weighted features. The input is a single visual modality.

$$F_{\text{fuse}} = \sum_k^K F_k \cdot \text{softmax}(\text{MLP}(F_k)) \tag{7}$$

The original soft attention method combines features generated from the same modality. However, for the multi-modal activity recognition task, The inputs are collected from different modalities, which requires global attention mechanism. Attention mechanism are classified as local attention mechanism and global attention mechanism (Luong et al. 2015). The global attention mechanism generates weights

according to features from all modalities, the local attention mechanism generates weights according to single modality.

To solve the above problem, this paper provides a global attention based soft attention fusion scheme, the improved architecture is shown in Fig. 2. First, the concatenated features are employed to generate weight mask for features in each modality, as shown in Eq. 8. Then, features in each modality are weighted separately. Last, the weighted features are concatenated, as shown in Eq. 9.

$$(W_v, W_i) = \text{softmax}(\text{MLP}(F_v \oplus F_i)) \tag{8}$$

$$F_{\text{fuse}} = W_v \cdot F_v \oplus W_i \cdot F_i \tag{9}$$

Where F_{fuse} is the final feature vector of the two modalities, $\oplus$ is vector concatenation, W_v and W_i are the scalar weight of visual modality and IMU modality. F_v and F_i are the feature of visual modality and IMU modality.

3.4 Classification

The conventional architecture for classification an MLP followed by Softmax mapping layer in the last. However, this approach is not enough to encourage discriminative learning of features due to high similarities between some activity samples. Following the work of Liu et al. (2016), the cross-entropy loss, the Softmax function, and the fully connected layer are employed to enhance the intra-class compactness and the inter-class separability between learned features.

The distance between sample and parameter are decomposed into amplitude ones and angular ones with cosine similarity. Let W_n be the n-th row in the last fully connected layer's parameter matrix, x be the learned sample feature from the previous layers, θ_n is the angle between W_n and x, the cross-entropy loss is denoted as

$$L = -\log\left(\frac{e^{\|W_i\|\|x\|\phi(\theta_i)}}{e^{\|W_i\|\|x\|\phi(\theta_i)} + \sum_{j\neq i} e^{\|W_j\|\|x\|\cos\theta_j}}\right) \tag{10}$$

Where $\phi(\theta_i)$ is denoted as

$$\phi(\theta_i) = (-1)^k \cos m\theta - 2k \tag{11}$$

Where $\theta \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$, $k \in [0, m-1]$, and k is an integer, m is the separability margin. With larger m, the separability margin is larger, but the learning process is more difficult.

Last, the soft voting is adopted to improve the robustness of the architecture during the test stage, which considers the diversity within a video clip. It averages the diversity of a video clip, and offsets the influence of abnormal samples.

Let V be a video clip sample, take equal amounts frames M times with random interval on a video clip, this randomness only acting on visual modality, where M different trajectory maps are generated, namely v_m. As shown in Eq. 12,

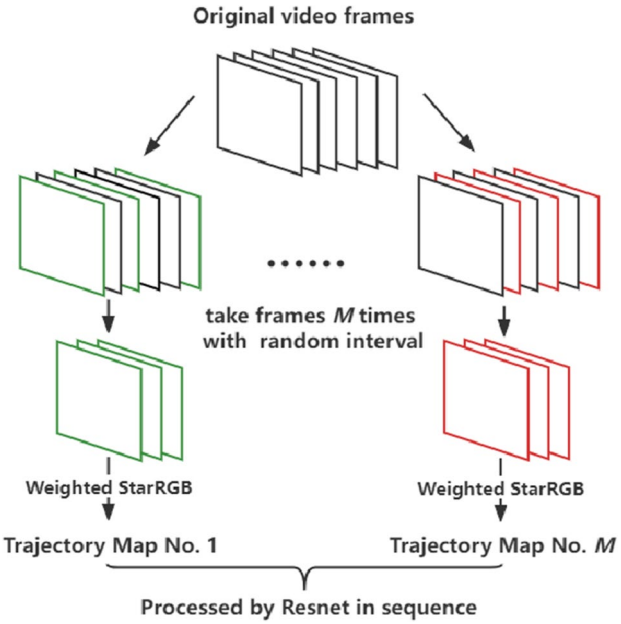

Fig. 4 The demonstration of frames selection in voting stage

a video clip is represented by a series of trajectory maps in the visual modality. The demonstration are shown in Fig. 4.

$$V = \{v_1, v_2, \cdots, v_m | m \in M\} \tag{12}$$

On the other hand, IMU modality data in this period is not changed. Let I be the IMU modality data in the corresponding period and S_k be the multi-modal sub-sample at the k-th frame-taking step, where $S_k = (v_k, I)$. After combining the two modalities, the set of M raw multi-modal sub-samples is used to represent this multi-modal human activity sample in the corresponding period, as Eq. 13.

$$S = \{S_1, S_2, \cdots, S_m | m \in M\} \tag{13}$$

Then, M sub-samples are classified by the trained model in turn, which generates M n-dimensional similarities prediction vectors, where n is the number of classes. M predicted similarities with corresponding regression lines of each class are summed up, and operated by Softmax to obtain an n-dimensional probability prediction vector P, as Eq. 14.

$$P = \frac{\sum_{m=1}^{M} \text{softmax}(\text{model}(S_m))}{M} \tag{14}$$

Finally, the class with the highest predicted probability is selected as the output, denoted as

$$c = \text{argmax}(P) \tag{15}$$

4 Experiments

4.1 Dataset

The UP-Fall dataset (Martínez-Villaseñor et al. 2019) presented for the Challenge UP Competition in 2019 (Ponce and Martínez-Villaseñor 2020) is selected for the experiment. The dataset consists of the data of human falls and daily activities. The sensors used in the UP-Fall dataset included RGB camera, IMUs, illumination sensor, infrared sensor, and EEG sensor. 17 persons (age 18–24 years old, height 1.57−1.75 ms, and weight 53–99 kg) were invited to collect the data of human falls and daily activities. Each person performed 11 kinds of actions, and each action was repeated 3 times. As shown in Table 1, the actions include 5 types of falls, and 6 types of daily activity. 2 RGB cameras from the front and lateral perspectives, and six infrared sensors were deployed under indoor environment. 5 IMU sensors were deployed in different parts of the human body to collect the dynamic data of human activities.

The IMU sensors are distributed in the left wrist, chest (near the neck), right pocket, middle of waist and left ankle, where the left wrist sensor is used to simulating the status of wearing a smartwatch, the pocket sensor is used to simulating the sensor on the phone. For each sensor, 3 parameters from accelerometer and 3 parameters from gyroscope are received. Also, the timestamps are recorded for the multimodal data alignment. The sampling rate of is 18.4 Hz, signals are uploaded to the computer through Bluetooth. Two cameras record the scene from different perspectives at approximately 18 fps, signals and timestamps are saved to the computer through USB cables. After the collection process, the dataset is consolidated, signals from IMU sensors and RGB cameras are associated and aligned according to the saved timestamps.

Table 1 Activity types, durations, training samples

Id	Activities	Duration	Original	Augmented
1	Falling forward using hands	10	33	624
2	Falling forward using knees	10	30	576
3	Falling backwards	10	34	640
4	Falling sideways	10	33	624
5	Falling sitting in empty chair	10	33	624
6	Walking	60	658	658
7	Standing	60	882	882
8	Sitting	60	637	637
9	Picking up an object	10	33	624
10	Jumping	30	320	320
11	Laying	60	980	980

4.2 Implementations

RGB camera and wearable IMU modal data were selected to test the multi-modal fusion algorithm presented in this paper. Since the duration of human activity is usually less than 3 s, the width of the sliding window is set as 51 (about 2.77 s), each video clip with the same label is segmented by a sliding window. All the data in a sliding window are adopted as IMU modality data, while 11 frames with 5 interval length are taken to generate a 3channel trajectory map as visual modality data. Since clipped video may capture arbitrary segments of human action, starting and ending moments are uncertain, this method will increase the difficulty of classification, but it is closest to the real application scenario. In practical applications, human activity monitoring usually requires rapid response, sampling may be very random. From Table 1, we can see that the number of Falls and Picking up an object are much smaller than those of other activities. Because the UP-Fall dataset is a fall detection dataset and falls' duration is usually short, as a consequence, the split dataset has the problem of imbalanced samples' number. The number of original samples in each class is shown in the 4-th column of Table 1. The number of fall samples (class 1–5) is much smaller than that of daily activities (class 6–11). Therefore, in order to make the model learn more distribution of fall samples, the number of fall samples in the training stage is augmented as below.

On the training set, the frame-taking strategy of the visual modality of the fall sample is changed to random interval frame-taking, that is, one image is taken at every 5 ± 1 time points, and a total of 11 images are taken. Finally, stop taking frames randomly till the number of fall samples is approximately equal to the number of daily activities. Augmented samples number is shown in the 5-th column of Table 1. Such a frame-taking strategy increases the diversity of fall samples and the probability of model learning fall samples.

The test environment is a server with Ubuntu operating system (E5-2620 CPU, 128GB memory, TESLA M40 GPU), the program is implemented by Pytorch and Python. The training and testing sets were split according to the competition's setting (Ponce and Martínez-Villaseñor 2020).

In terms of visual modality processing, after evaluated the result, only lateral view camera's data are adopted. In each video clip, the size of the original image $480 \times 640 \times 3$ is scaled to $120 \times 160 \times 3$. In practice, human contour width is about 30 pixel after rescaling, thus, H in Eq. 3 is selected to be 30×30 to reflect the human contour's change. Subsequently, the generated human activity trajectory map with size $120 \times 160 \times 3$ is input into the Resnet101 network. The 1×40960 flattened feature map is reduced to 1×64 through an MLP which composed by 4096, 512 and 128 neurons hidden layer as the final feature of the visual modal of the video.

In terms of IMU modality processing, due to the error of the dataset, part of the IMU data deployed on the right pocket are missing. Consequently, only the IMU sensor data of the remaining four parts are considered. The four body parts' IMU 6-axis data that contain 51 time points in a period, which are generated into a matrix of 24×51. Flattened feature with dimension 1×624 is fed into an MLP which composed by 256 and 128 neurons hidden layer. Finally, the feature vector is reduced to 1×64 as the final feature vector of IMU modality.

The feature vectors of the above two modalities are reduced to the same dimension by MLP so that the subsequent modality fusion stage is carried out.

In terms of modality fusion, concatenation fusion, and soft attention mechanism fusion are tested in this paper. After sum fusion and concatenation fusion, 1×64 and 1×128 dimensional fusion vectors are formed, respectively. For the attention fusion method, the feature vectors of two modalities are concatenated and input into an MLP composed by 128 and 64 neurons hidden layer, two corresponding scalar weight are then obtained. The 1×128 dimensional fused vector is then obtained by weighted concatenation of two modalities' feature.

In the classification stage, the fused vector is reduced to 1×32 through an MLP with two hidden layers, which contains 64, 32 neurons in the hidden layer. During training, the L-Softmax with cross entropy loss function is used for classification and punishment, after evaluating the results, the condition of margin = 2 is proved to be optimum. During testing, 7 times random frame-taking are conducted on the tested samples to comprehensively identify the predicted class by soft voting.

As for the training parameters, the batch size is set to 32, the learning rate starts from 0.001 with a decaying rate of 0.7 every 10 epochs. The Adam optimizer is employed for the gradient descent. The dropout rate for the last layer of the MLP is 0.5. Training lasts for 50 epochs.

4.3 Ablation and comparison experiments

The comparison results are shown in Tables 2 and 3, where Table 2 shows the 11-class recognition results and Table 3 shows the binary classification results. In Table 2, Gjoreski et al. (2020) introduces a multi-sensor data-fusion machine-learning algorithm using 5 wearable inertial sensor's data, Ponce and Martínez-Villaseñor (2020) employ MLP, RF, SVM, KNN as extractors and fuse the RGB visual modality, IMU modality, EEG modality data. Their per-class accuracy and F1-Macros are shown in the table. It can be seen that our proposed method has the highest F1-Macros of 0.92. The recognition accuracy on falls and daily activities are better than that of the Martinez-Villasenor's method, most of the per-class performance are higher than that of the Gjoreski et al. [40]'s method, except for fall classes 3–5, and daily activity class 6. Table 2 shows that the proposed method can distinguish the daily activities (class 6–11) as well as the first two fall activities effectively, the values displayed in bold represent the highest scores. All compared methods perform slightly worse on overall fall activities. The F1-Macro comparison shows that the proposed architecture outperforms other methods. Compared with Gjoreski et al. (2020), the performance distribution of the proposed method on each class is more balanced, and the accuracy of the first 2 classes are higher, which leads to a higher F1-Macro.

The UP-Fall dataset also provides a binary fall detection task, where 11 classes are summarized as fall and non-fall. The proposed method is trained for the binary classification task, the experimental results are shown in Table 3. The proposed architecture reaches 100% accuracy in binary classification task with either fused modalities and single

Table 2 The comparison of F1-Score for each class

	1	2	3	4	5	6	7	8	9	10	11	F1-Marco
proposed method	**0.83**	**0.83**	0.80	0.96	0.78	0.99	**0.99**	**0.99**	**1.0**	**1.0**	**0.98**	**0.92**
Gjoreski et al. (2020)	0.74	0.54	**0.94**	**0.97**	**0.84**	**1.0**	0.99	0.98	0.91	1.0	0.98	0.89
Ponce and Martínez-Villaseñor (2020)	0.40	0.70	0.53	0.33	0.39	0.93	0.93	0.98	0.57	0.99	0.98	0.70

Table 3 Accuracy comparison on fall and non-fall binary classification

Precision	Recall	Specificity	F1-Score	Accuracy	
proposed method (multi-modal fusion)	1.0	1.0	1.0	1.0	100%
proposed method (visual modality)	1.0	1.0	1.0	1.0	100%
Ponce and Martínez-Villaseñor (2020), Galvão et al. (2021) (multi-modal fusion)	1.0	–	1.0	–	100%
Ponce and Martínez-Villaseñor (2020), Galvão et al. (2021) (visual modality)	0.999	–	1.0	–	99.99%
Galvão et al. (2021), Espinosa et al. (2019) (visual modality)	0.95	0.98	0.82	0.97	95.24%

Table 4 Performance with different backbone networks

	1	2	3	4	5	6	7	8	9	10	11	F1-Marco
Resnet101 (He et al. 2016)	**0.83**	**0.83**	**0.80**	**0.96**	**0.78**	**0.99**	**0.99**	**0.99**	**1.0**	**1.0**	0.98	**0.92**
Inception-V3 (Szegedy et al. 2016)	0.64	0.27	0.52	0.69	0.75	**0.99**	0.98	0.98	0.74	0.99	**0.99**	0.78
X3D (Feichtenhofer 2020)	0.57	0.32	0.58	0.60	0.72	0.93	0.90	0.44	0.85	0.97	**0.99**	0.72

Table 5 Ablation study on modalities

	1	2	3	4	5	6	7	8	9	10	11
Proposed global soft attention fusion	**0.83**	**0.83**	**0.80**	**0.96**	0.78	**0.99**	**0.99**	**0.99**	**1.0**	**1.0**	**0.98**
Multi-modal element-level fusion	0.74	0.70	0.80	0.72	0.80	0.99	0.99	0.98	0.91	1.0	0.98
Multi-modal local attention fusion	0.80	0.79	0.80	0.80	0.82	0.79	0.99	0.98	0.96	0.99	0.98
Multi-modal concatenation fusion	0.74	0.58	0.80	0.80	0.81	0.98	0.91	0.89	0.95	0.90	0.95
Only visual modality (speed-weighted)	0.77	0.75	0.75	0.74	**0.83**	0.98	0.94	0.97	0.89	0.90	0.98
Only visual modality (star RGB)	0.75	0.77	0.75	0.76	0.75	0.99	0.93	0.97	0.89	0.92	0.95
Only IMU modality	0.64	0.38	0.44	0.62	0.34	0.95	0.93	0.88	0.79	0.94	0.92

RGB modality. This result shows that the proposed method is effective against fall detection task.

The performance of different backbone networks for visual feature extraction is evaluated. Resnet101 (He et al. 2016), Inception-V3 (Szegedy et al. 2016) for 2D-CNN and X3D (Feichtenhofer 2020) for 3DCNN are selected. The comparison results are shown in Table 4. The result shows that Resnet101 produces better performances in 2D-CNN with 3-channel star RGB MHI, the values in bold represent the highest scores among different backbones.

In order to verify the difference between the multi-modal fusion model and the single-modal performance, the ablation experiments are carried out, and the differences between the modalities and different fusion methods are compared. The experimental results are shown in Table 5, where the values in bold represent highest scores among different fusion methods.

The ablation experiments show that the proposed fusion method is effective in balancing the weights between modalities, and gains higher accuracy than concatenation fusion, local attention fusion and element-level fusion. Also, the IMU modality failed to precisely distinguish the first 5 classes. This is because that in the dataset, the distribution of the wearable sensors are relatively sparse, the collected signals are similar for the classes 1–5. For example, falling forward using hands and using knees are hard to distinguish. This fact indicates that the IMU modality data are not suitable for classifying activities with similar motions, and it is difficult to extract the joint angle changes. On the other hand, the visual modality performs much better in the first 5 classes, indicates that the RGB data are more suitable for classifying activities with small motion differences. The proposed speed-weighted star RGB representation achieves further improvement over the original star RGB, the biggest improvement is 7% on the recognition rate of class 5 (falling, setting in empty chair). Last, the multi-modal approach that combines 2 modalities performs better than single modal approaches.

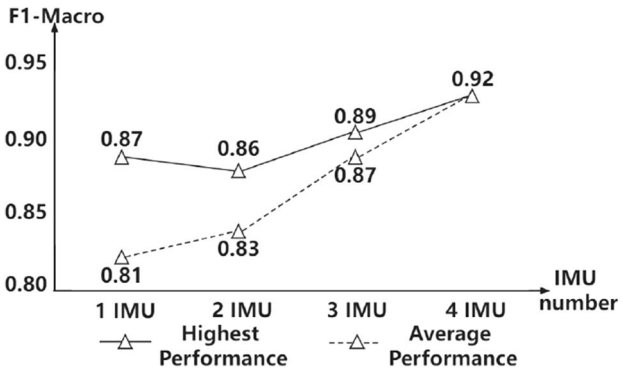

Fig. 5 Performances of different IMU combinations

The IMUs are deployed on 4 locations (neck, wrist, waist and ankle) in UP-Fall dataset. Hence, the performances on different deployment combinations of IMUs are evaluated based on the setting of UP-Fall dataset, the results are shown in Fig. 5. The solid line in the figure represents the highest F1-Macro of the IMU combination, and the dash line represents the average F1-Macro of the IMU combination. In the figure, 4 IMU combination gets the highest F1-Macro (0.92). Ankle produces the highest performance (0.87) in single-IMU test, ankle + neck produce the highest performance (0.86) in dual-IMU test, ankle+waist+neck produce the highest performance in triple-IMU test (0.89).

5 Conclusion

A multi-modal activity recognition method that fuses visual and wearable IMUs data modalities is proposed in our paper. The star RGB method is improved to strengthen the speed representations in the visual modality. In the IMU modality, FFT is used to aggregate temporal information and MLP is employed to further extract features. A global soft attention mechanism is introduced to adjust the weight between the visual features and dynamical features, and L-softmax combined with a soft voting algorithm is adopted to classify human activities. Experiments show that the proposed multi-modal fusion method is more effective on the recognition of human activities than single-modal methods, and performs better comparing with existing methods. The F1-Macro of 11-class classification is 0.92, and that of fall and non-fall binary classification reaches 1.0.

Funding This work was supported by the National Key Research and Development Plan under Grant 2020YFB2104400.

Data availability No data was used for the research described in the article.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abebe, G., Cavallaro, A.: Inertial-vision: cross-domain knowledge transfer for wearable sensors. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1392–1400 (2017)

Ahad, M., Rahman, A., Tan, J., Kim, H., Ishikawa, S.: Motion history image: its variants and applications. Mach. Vis. Appl. **23**(2), 255–281 (2012)

Balli, S., Sağbaş, E.A., Peker, M.: Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm. Meas. Control **52**(1–2), 37–45 (2019)

Barros, P., Parisi, G.I., Jirak, D., Wermter, S.: Real-time gesture recognition using a humanoid robot with a deep neural architecture. In: 2014 IEEE-RAS International Conference on Humanoid Robots. IEEE, pp. 646–651 (2014)

Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)

Brena, R.F., Aguileta, A.A., Trejo, L.A., Molino-Minero-Re, E., Mayora, O.: Choosing the best sensor fusion method: a machine-learning approach. Sensors **20**(8), 2350 (2020)

Chen, C., Jafari, R., Kehtarnavaz, N.: A survey of depth and inertial sensor fusion for human action recognition. Multim. Tools Appl. **76**(3), 4405–4425 (2017)

Cippitelli, E., Gasparrini, S., Gambi, E., Spinsante, S.: A human activity recognition system using skeleton data from rgbd sensors. Comput. Intell. Neurosci. **2016** (2016)

Demrozi, F., Pravadelli, G., Bihorac, A., Rashidi, P.: Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey. IEEE Access **8**, 210-210 836 (2010). (**836**)

dos Santos, C.C., Samatelo, J.L.A., Vassallo, R.F.: Dynamic gesture recognition by using cnns and star rgb: a temporal information condensation. Neurocomputing **400**, 238–254 (2020)

Espinosa, R., Ponce, H., Gutiérrez, S., Martínez-Villaseñor, L., Brieva, J., Moya-Albor, E.: A vision-based approach for fall detection using multiple cameras and convolutional neural networks: a case study using the up-fall detection dataset. Comput. Biol. Med. **115**, 103520 (2019)

Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016)

Feichtenhofer, C.: X3d: expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 203–213 (2020)

Fortun, D., Bouthemy, P., Kervrann, C.: Optical flow modeling and computation: A survey. Comput. Vis. Image Underst. **134**, 1–21 (2015)

Galvão, Y.M., Ferreira, J., Albuquerque, V.A., Barros, P., Fernandes, B.J.: A multimodal approach using deep learning for fall detection. Expert Syst. Appl. **168**, 114226 (2021)

Gjoreski, H., Stankoski, S., Kiprijanovska, I., Nikolovska, A., Mladenovska, N., Trajanoska, M., Velichkovska, B., Gjoreski, M., Luštrek, M., Gams, M.: Wearable sensors data-fusion and machine-learning method for fall detection and activity recognition. In: Challenges and Trends in Multimodal Fall Detection for Healthcare. Springer, pp. 81–96 (2020)

Han, J., Bhanu, B.: Human activity recognition in thermal infrared imagery. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops. IEEE, pp. 17 (2005)

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

He, J., Zhang, Z., Wang, X., Yang, S.: A low power fall sensing technology based on fd-cnn. IEEE Sens. J. **19**(13), 5110–5118 (2019)

He, J., Zhang, C., He, X., Dong, R.: Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features. Neurocomputing **390**, 248–259 (2020)

Horn, B.K., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)

Hwang, I., Cha, G., Oh, S.: Multi-modal human action recognition using deep neural networks fusing image and inertial sensor data. In: 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). IEEE, pp. 278–283 (2017)

Li, Z., Wu, H.: A survey of maneuvering target tracking using Kalman filter. In: 2015 4th International Conference on Mechatronics, Materials, Chemistry and Computer Engineering. Atlantis Press, pp. 542–545 (2015)

Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 507–516 (2016)

Lu, Y., Velipasalar, S.: Autonomous human activity classification from wearable multi-modal sensors. IEEE Sens. J. **19**(23), 11 403-11 412 (2019)

Lucas, B.D., Kanade, T. et al.: An iterative image registration technique with an application to stereo vision. Vancouver **81** (1981)

Luo, F., Poslad, S., Bodanese, E.: Temporal convolutional networks for multiperson activity recognition using a 2-d lidar. IEEE Internet Things J. **7**(8), 7432–7442 (2020)

Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv:1508.04025 (2015)
Mallat, R., Bonnet, V., Khalil, M., Mohammed, S.: Toward an affordable multi-modal motion capture system framework for human kinematics and kinetics assessment. In: International Symposium on Wearable Robotics. Springer, pp. 65–69 (2018)
Mao, A., Ma, X., He, Y., Luo, J.: Highly portable, sensor-based system for human fall monitoring. Sensors **17**(9), 2096 (2017)
Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., Peñafort-Asturiano, C.: Up-fall detection dataset: a multimodal approach. Sensors **19**(9), 1988 (2019)
Ometov, A., Shubina, V., Klus, L., Skibińska, J., Saafi, S., Pascacio, P., Flueratoru, L., Gaibor, D.Q., Chukhno, N., Chukhno, O., et al.: A survey on wearable technology: history, state-of-the-art and current challenges. Comput. Netw. **193**, 108074 (2021)
Ponce, H., Martínez-Villaseñor, L.: Approaching fall classification using the up-fall detection dataset: Analysis and results from an international competition. In: Challenges and Trends in Multimodal Fall Detection for Healthcare. Springer, pp. 121–133 (2020)
Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: Aaai, vol. 5, no. 2005. Pittsburgh, PA, pp. 1541–1546 (2005)
Rivera, P., Valarezo, E., Choi, M.-T., Kim, T.-S.: Recognition of human hand activities based on a single wrist imu using recurrent neural networks. Int. J. Pharma Med. Biol. Sci. **6**(4), 114–118 (2017)
Salehzadeh, A., Calitz, A.P., Greyling, J.: Human activity recognition using deep electroencephalography learning. Biomed. Signal Process. Control **62**, 102094 (2020)
Steven Eyobu, O., Han, D.S.: Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network. Sensors **18**(9), 2892 (2018)
Stoeve, M., Schuldhaus, D., Gamp, A., Zwick, C., Eskofier, B.M.: From the laboratory to the field: Imu-based shot and pass detection in football training and game scenarios using deep learning. Sensors **21**(9), 3071 (2021)
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
Tsai, D.-M., Chiu, W.-Y., Lee, M.-H.: Optical flow-motion history image (of-mhi) for action recognition. Signal Image Video Process. **9**(8), 1897–1906 (2015)
Wan, S., Qi, L., Xu, X., Tong, C., Gu, Z.: Deep learning models for real-time human activity recognition with smartphones. Mob. Netw. Appl. **25**(2), 743–755 (2020)
Zhu, Y., Yu, J., Hu, F., Li, Z., Ling, Z.: Human activity recognition via smart-belt in wireless body area networks. Int. J. Distrib. Sens. Netw. **15**(5), 1550147719849357 (2019)
Zimmermann, T., Taetz, B., Bleser, G.: Imu-to-segment assignment and orientation alignment for the lower body using deep learning. Sensors **18**(1), 302 (2018)

Cheng Zhang received the M.S. degree in Software Engineering from Beijing University of Technology in 2019. He is a Ph.D. student at the Faculty of Information technology, Beijing University of Technology. His research interests include Human-Computer Interaction, Computer Vision and Deep Learning.

Tianqi Zu received the B.S. degree in software engineering from Beijing University of Technology, Beijing, China in 2018 and received the M.S. degree with the Faculty of Information Technology, Beijing University of Technology in 2021. His research interests include software engineering, computer vision, wearable computing and deep learning.

Yibin Hou received the Ph.D. degree in Electrical Engineering from Eindhoven University of Technology. He is a professor at the Faculty of Information Technology, Beijing University of Technology, China. He is also an adjunct researcher with Beijing Advanced Innovation Center for Future Internet Technology, China. His research interests include Human-Computer Interaction, Embedded Software and Systems, Internet Theory and Technology.

Jian He received the M.S. degree in Computer Software from Northwest University, Xi'an, China, in 2000, and received the Ph.D. degree in Computer Software from Xi'an Jiaotong University, Xi'an, China, in 2005. He is an associate professor at the Faculty of Information Technology, Beijing University of Technology. His research interests include Ubiquitous Computing, Embedded System, and HCI.

Shengqi Yang received the double B.S. degree in mechanical engineering and economics and the M.S. degree in electrical engineering from Peking University, Beijing, China, in 2000 and 2002, respectively, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2006. He is an Adjunct Professor with the Beijing Engineering Research Center for IoT Software and Systems, Beijing University of Technology, China. His research interests include IoT, embedded system design, and big data in digital health.

Ruihai Dong is an Assistant Professor with the School of Computer Science in University College Dublin. His research interests lie broadly in Machine Learning and Deep Learning, and their applications in recommender systems and finance.