FFCI: A Camera and IMU Sensors Based Multi-modal Neural Network for Activity Recognition in Smart Factory

Yujue Wang, Xin Niu, Xianwei Lv, and Chen Yu *Member, IEEE*

***Abstract*—Worker activity recognition is an important aspect of the construction of smart factory. The development of deep neural networks and the widespread distribution of sensors in the smart factory have brought opportunities for the recognition of workers' activities. The existing methods based on camera and IMU sensors respectively have problems of visual occlusion and difficulty in recognition of similar activities, which result in the reduction of recognition accuracy. Therefore, we propose a feature-level fusion based on camera and IMU sensor (FFCI) for activity recognition in smart factory. The FFCI employs an optimized multi-modal fusion strategy to integrate the visual information and the action information of workers' activities. In order to verify the effectiveness of FFCI, we define a simple yet fine-grained assembly task and collect six common operations. Furthermore, we add visual occlusion and similar activity data to get closer to the smart factory. Finally, we evaluate the FFCI on the collected dataset and achieve a recognition accuracy of 98.17%, demonstrating its effectiveness in accurately classifying workers' activities.**

***Index Terms*—Camera sensor, IMU sensor, multi-modal fusion, feature-level fusion**

I. Introduction

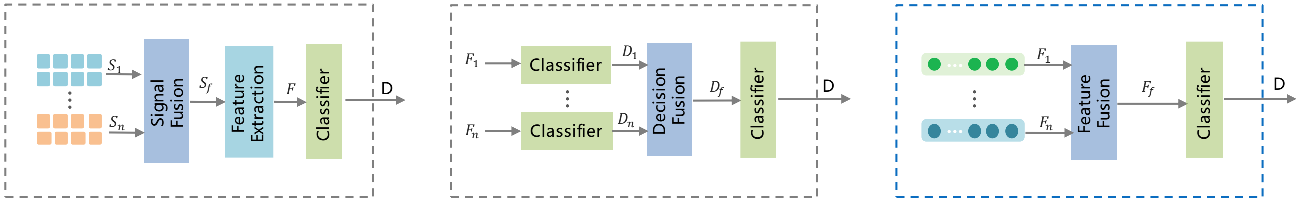
With the advent of the fourth industrial revolution [1], the smart factory has been transformed into automation. In the smart factory, the worker is indispensable. The main reason is that for small-batch personalized production orders, the cost of automated production lines is much higher than that of manual assembly [2]. In the manual assembly line, workers' negligence and unconscious operation may result in errors that seriously affect worker safety and factory productivity. Thanks to the development of deep learning [3], [4] and the widespread distribution of sensors in various fields, such as the smart factory [5], [6], smart healthcare [7], [8], smart manufacturing [9], [10], and smart transportation [11]. Especially in smart factory, it is possible to recognize and issue warnings for workers' activities that may cause errors in the smart factory.

At present, research on worker activity recognition typically uses only a single sensor, such as camera sensor [12] or IMU sensor [13]. While knowledge distillation effectively uses large models to guide the deployment of smaller models on sensor devices with limited computing capacity [14]–[16], enhancing recognition accuracy, relying on a single sensor often presents limitations [17]–[21]. For example, in the case of operational occlusion during production, the recognition method based on the camera sensor often fails to identify or misidentify. In addition, due to the different operating habits of workers, some similar operating actions hinder the accurate recognition of workers' activities by IMU sensors. Fortunately, the camera sensor and the IMU sensor are complementary [22], so we use multi-modal fusion technology to fuse the data of camera sensor and IMU sensor to improve the recognition accuracy of workers' activities.

The existing methods of multi-modal fusion can be classified into three categories [23]: signal-level fusion, decision-level fusion, and feature-level fusion. Signal-level fusion has high accuracy [24], however, it can only integrate the same type of data. Decision-level fusion inputs the extracted features into pre-trained models, then fuses the results of these models and outputs the final result [25]. Nevertheless, decision-level fusion cannot correlate information between multiple modalities. Feature-level fusion can extract the features of different modalities, and then use the correlation between different modalities to identify the target [26]. Considering the correlation between the visual data and motion data respectively collected by camera sensor and IMU sensor, we decide to use the feature-level fusion to recognize the workers' activities. When fusing the visual data and action data, the following issues need to be solved. First of all, the visual data also contains spatiotemporal features. In addition, there is a temporal correlation among the action data. Therefore, how to extract the features of visual data and motion data is the primary challenge need to be solved. Furthermore, how to establish the correlation between the visual data and the action data cannot be ignored.

To conquer the above challenges, we propose a feature-level fusion based on camera and IMU sensor (FFCI) for activity recognition in smart factory. FFCI mainly consists of three parts: visual feature extraction, action feature extraction, and multi-modal feature fusion. In order to extract the visual features of workers' activities, we use the C3D convolutional neural network (CNN) [27], which adds the temporal dimension analysis on the basis of the 2D convolutional neural network, so that the obtained visual features are more accurate. In addition, since the workers' activities change over time, the long short-term memory (LSTM) network can learn and store the characteristics of long-term dependencies [28], so we

This work was supported by NSFC under Grant 62272179. Yujue Wang, Xin Niu, Xianwei Lv, and Chen Yu (*corresponding author*) are with the National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Big Data Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China (e-mail: {yu_jue, xinniu, xwlv, yuchen}@hust.edu.cn.)



(a) Signal-level fusion unit (b) Decision-level fusion unit (c) Feature-level fusion unit

Fig. 1. Comparison of three multi-modal fusion strategies based on sensor systems.

use LSTM to extract the action features of workers' activities. After obtaining the above features, we optimize the feature-level multi-modal fusion strategy to fuse the visual feature and the action feature. Specially, we first perform channel stacking in the fusion layer to concatenate different features, and then use the convolutional kernel to learn the correlation between different activities. The main contributions of this paper are as follows:

- We propose a camera and IMU sensors based multi-modal neural network for activity recognition in smart factory.
- We design a feature-level fusion neural network, which fuses and optimizes the visual features and action features of workers' activities.
- We collect the data of six activities of numerous volunteers. On the collected data, we evaluate FFCI with an accuracy of 98.17%.

The remain of the paper is as follows. Section II reviews the previous research on multi-modal fusion. In the section III, we first describe the multi-modal data set we have built in detail. Then, we respectively use C3D and LSTM to extract the visual features and action features of workers' activities. Finally, we introduce the four feature fusion strategies. In the Section IV, we introduce the FFCI in detail. Section IV evaluates FFCI, including experimental evaluation, evaluation of different fusion strategies, and evaluation of different recognition methods. In the section V, we deploy FFCI to the monitoring system in smart factory and then test its accuracy and real-time of identifying workers' activities. The last section summarizes this work.

II. Related work

Multi-modal fusion is a type of machine learning technique [29], and it has achieved widespread application in various fields, such as audio-visual recognition [30], multi-modal sentiment analysis [31], pose estimation [32] and so on. All in all, most of the work of multi-modal fusion is based on the data collected by sensors. And the multi-modal fusion strategies can be divided into three [23]: signal-level fusion, decision-level fusion, and feature-level fusion.

Signal-level fusion. As shown in Fig. 1(a), it is also called data-level fusion, which is a low-information level fusion. The raw data is processed into the specified signal and then directly fused. Memmesheimer et al. [33] used reduction and enhancement techniques on signals, which improves accuracy on three datasets and achieves 96.11% accuracy on Simitate motion. Furthermore, Prabhakar [34] proposed a neural network architecture integrating phase and amplitude spectrum features, addressing phase ambiguity and the limited use of phase information in traditional speech emotion recognition, thereby improving accuracy. The signal-level fusion requires the data be the same type, and it is difficult to handle complex multi-modal data. To accurately identify users' exercise movements and provide guidance, Sun et al. [35] installed sensors on the equipment, classified and evaluated the movements using a neural network, achieving 89% recognition accuracy for distinguishing standard and non-standard movements, and generated personalized exercise recommendations.

Decision-level fusion. The Fig. 1(b) illustrates the process of decision-level fusion. the decision-level fusion is the highest level fusion. It uses averaging, weighting, voting schemes, and other mechanisms to fuse multiple classification results generated by single modality. Some decision-level fusion work [36]–[40] inputted the information of each modality into the CNN, they then fused the classification results for each modality to obtain the final decision. Among them, Zhang et al. [36] used the image and speed information of the vehicle, integrated the classification results of them, which could filter the low-information of the image and realized fast tracking of vehicles. Tao et al. [37] extracted information from the wearable sensor in both spatial and frequency domains, and video and image information from the camera sensor. Ren et al. [38] extracted three modalities of RGB, optical flow, and gray images from camera sensors. Al-Amin et al. [39] collected IMU, electromyography signals, and skeleton data from wearable sensors and camera sensors. Both Ren et al. [38] and Al-Amin et al. [39] fuse the resulting multi-modal inference results and then output the final result. Furthermore, Sharma et al. [40] proposed a decision-level fusion strategy that integrates three distinct emotional modalities (voice, facial expressions, and physiological signals), and developed a human-computer interaction system, achieving an 82.4% recognition accuracy for human emotions.

Feature-level fusion. The process of feature-level fusion is shown in Fig. 1(c). The main task of feature-level fusion is to connect feature vectors from multiple modalities and obtain a combined feature vector. Gao et al. [30] fused the high-level abstract features of single frame or audio. The fused features can assist the video-clip network to extract effective frames in the video and reduce the computing cost. Liu et al. [41] used the two stream CNNs to establish the connection between the spatial and temporal information, which efficiently exploits the attended spatial and temporal features. Experimental results

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2025.3534236

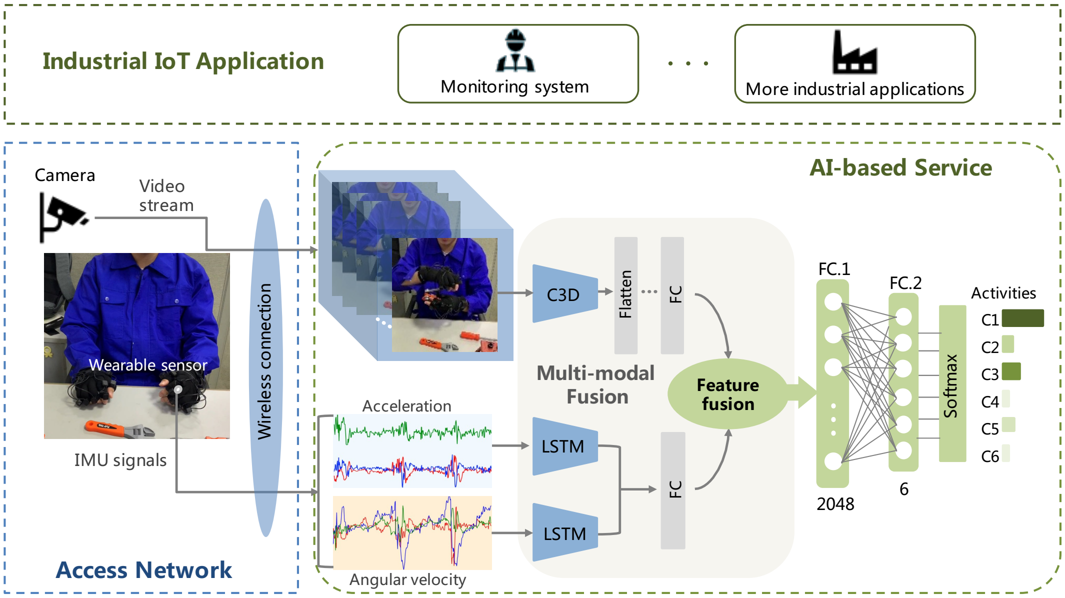


Fig. 2. A camera and IMU sensors based multi-modal neural network for activity recognition in smart factory.

proves that the fusion of spatial and temporal information can help to improve recognition accuracy. Zhao et al. [42] first obtained the shallow and deep modality of the image, and used deep belief networks to obtain the deep features of each modality. Then, they explored the correlation between the high-level features of the two modalities and achieved the improvement of recognition accuracy. Moreover, Ning et al. [43] integrated EEG and speech signals at the feature level and trained a classifier, the highest recognition accuracy of 87.44% was achieved on the healthy controls.

In the Fig. 1, we compare three multi-modal fusion strategies. The signal-level fusion first fuses the data and then classifies features from fusion data to complete the classification tasks, which requires the data to be consistent in format and dimension. And it is not easy to fuse different types of data. The decision-level fusion fuses the results of each classifier, which requires training multiple classifiers and is unable to obtain the association information between multiple modalities. The feature-level fusion can fuse features of different modalities, which can obtain the correlation between multiple modalities and only a single classifier needs to be trained [29].

III. Proposed system

In this paper, we propose the FFCI for activity recognition in smart factory. As shown in Fig. 2, we first input video streams and IMU signals into the FFCI. Then, we respectively use the C3D and the LSTM to process the video streams and IMU signals. Next, we use the multi-modal fusion strategy to fuse visual features and action features. The architecture of the multi-modal fusion model proposed in this paper is illustrated in Fig. 3. Finally, the FFCI outputs the recognition results of workers' activities. The meanings of the main symbols in the this paper are listed in Table I.

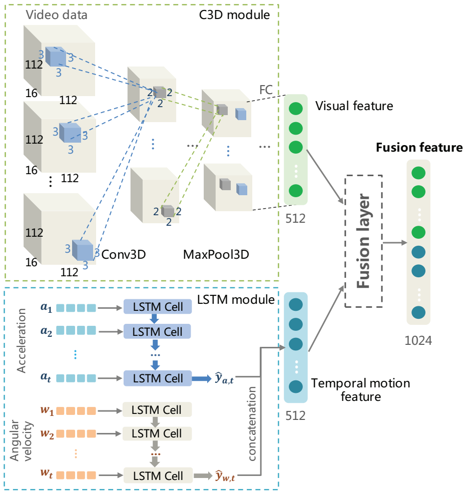


Fig. 3. Multi-modal fusion model. The visual features are obtained by C3D and temporal motion features are obtained by LSTM.

A. Data collection and preprocessing

We divide the assembly operation into fine-grained steps to make it easier for managers to get detailed information about each step. Specifically, we define a set of continuous assembly operations of workers as a dataset. In this paper, we mainly collect the following six operations of workers: Use a Screwdriver (US), Use a Wrench (UW), Use a Measuring tool (UM), Assembly Accessories (AA), Press Switch(PS), and Use a Power-screwdriver(UP). Table II details the six activities.

We deploy cameras above the assembly line and collect video of works' activities. Then, we use the C3D network to extract the visual feature of workers' activities [27]. The IMU data is collected using a wearable device worn on the upper body of the volunteer. The device is Noitom Perception Neuron 2.0 [44], using upper body (all fingers) mode is shown in Fig. 4, sensors required both sides-upper arm, lower arm

TABLE I
IMPORTANT SYMBOLS INVOLVED

Symbols	Meaning
D_m	Worker activity data set collected by sensor equipment.
D_n	Worker activity data set after pre-processing.
$V_{n,T}$	The nth video sample under T window length.
$S_{n,T}$	The nth IMU sample under T window length.
$a^x_{n,t}$	Acceleration contained in $S_{n,T}$.
$w^x_{n,t}$	Angular velocity contained in $S_{n,T}$.
y	The ground truth of worker activity.
$\hat{y}$	Labels predicted by the model.
C_i	The ith 3D convolutional layer feature.
$f_{i,j}$	The ith 3D convolutional feature after flatten.
C_t	The cell state of LSTM at t time step.
h_t	The hidden state of LSTM at t time step.
f_t	The feature of the tth time step.
U, W, b	The parameters in different gates of LSTM.
P_y	The parameter that LSTM trained.
f_{avg}	Use average fused features.
f_{max}	Use maximum fused features.
f_{cat}	Use concatenation fused features.
f_{conv}	Use convolution fused features.

TABLE II
TASKS FOR COLLECTING WORKER ACTIVITY

No.	Tasks	Activities
1	Tighten/loosen 3 screws using a screwdriver	US
2	Tighten/loosen 3 screws using a wrench	UW
3	Measure accessories using the measurement tool	UM
4	Assembly 3 accessories	AA
5	Press switch using a single finger	PS
6	Tighten 5 screws using a power-screwdriver	UP

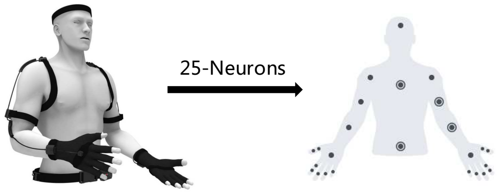


Fig. 4. Upper body mode of the Noitom Perception Neuron 2.0. There are 25 neurons in total.

and spine. We select 20 neurons (Table III) to collect data on arm and finger movements. Each neuron node obtains three-axis acceleration and three-axis angular velocity. Furthermore, the device returns data at a sampling rate of 60Hz. The video frame rate is 30fps. While collecting IMU data from the workers, the camera above the worker records his activities. Fig. 5 shows an example of capturing 6 activities.

The format of collected dataset is $\mathbb{D} = \{D_1, \ldots, D_m, \ldots, D_M\}$, where D_m is defined as follows:

$$D_m = \{[v_m, s_m], y_m\}, m \in [1, M]. \tag{1}$$

where v_m is the video clip sample corresponding to the IMU data, s_m is the sample set of the IMU data, and y_m is the ground truth of the activity class.

TABLE III
20 NODES SELECTED FROM WEARABLE SENSORS

Node no.	Node description	Node no.	Node description
1	LeftArm	11	RightArm
2	LeftForeArm	12	RightForeArm
3	LeftHand	13	RightHand
4	LeftHandThumb1	14	RightHandThumb1
5	LeftHandThumb2	15	RightHandThumb2
6	LeftHandIndex1	16	RightHandIndex1
7	LeftHandIndex2	17	RightHandIndex2
8	LeftHandMiddle2	18	RightHandMiddle2
9	LeftHandRing2	19	RightHandRing2
10	LeftHandPinky2	20	RightHandPinky2

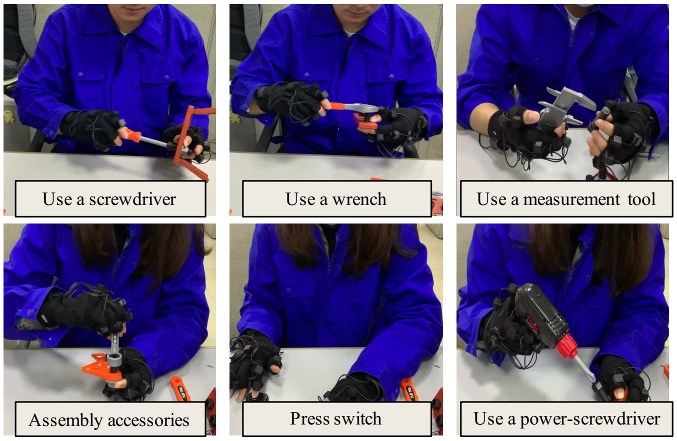


Fig. 5. The sample of 6 activities captured by the camera.

In general, the duration of assembly operations varies from seconds to minutes. Therefore, we use the sliding window mechanism to obtain fixed-length data. The sliding window mechanism is shown in Fig. 6, the window length is set to $T = 2s$, and there is 50% overlap between two adjacent windows. The IMU data has three acceleration channels and three angular velocity channels from 20 nodes. In addition, the timestamp of IMU data is synchronized with the video data.

In order to facilitate data processing, we normalize and standardize the collected data. Then, we will get N samples $\{X_1, \ldots, X_N\}$, and each sample contains two different inputs:

$$X_n = \{V_{n,T}, S_{n,T}\}, n \in [1, N]. \tag{2}$$

where T is the length of the window, $V_{n,T}$ and $S_{n,T}$ respectively are visual data and IMU data within T time period. More specifically, $S_{n,T} = \{S_{n,1}, \ldots, S_{n,t}, \ldots, S_{n,T}\}$, as follows:

$$S_{n,t} = [\underbrace{a^x_{n,t}, a^y_{n,t}, a^z_{n,t}}_{a_{n,t}\text{: acceleration}}, \underbrace{w^x_{n,t}, w^y_{n,t}, w^z_{n,t}}_{w_{n,t}\text{: angular velocity}}], \quad t \in [1, T]. \tag{3}$$

where t is timestamp, a^x, a^y, and a^z are the three-axis acceleration data. Moreover, w^x, w^y, and w^z are the three-axis angular velocity data. At the 60Hz sampling rate, the IMU data has 120 time steps. In the corresponding window, 16 frames are captured. There are a total of 9 volunteers complete each operation at different times, so each volunteer has a different

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2025.3534236

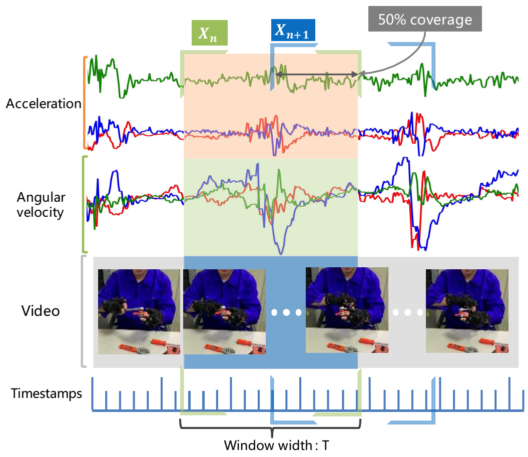


Fig. 6. Sliding window mechanism. Align the visual and IMU data, the window length is set to $T = 2s$, the time step is $1s$, the window overlap is 50%.

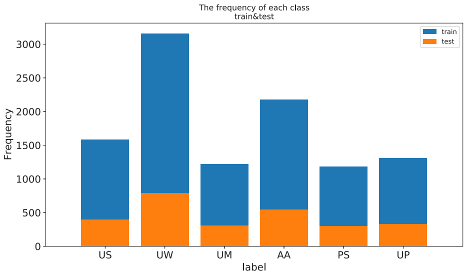


Fig. 7. The frequency of each action in the dataset.

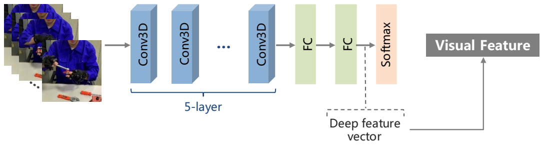


Fig. 8. The idea of extracting deep feature vector from C3D as visual features.

number of data samples. After sampling, we divide the test set with a ratio of 0.2, and the frequency of each action is listed in Fig. 7.

B. Visual feature extraction

The idea of extracting visual features from the video is shown in Fig. 8. For the input $V_{n,T} \in \mathbb{R}^{\mathrm{L\times C\times H\times W}}$, where $L = 16$ is the length of the video frame sequence, C is the image channel (3 channels), H and W respectively are the width and height of the video frame (112×112). Here, we use the pre-trained C3D [27] model to extract visual features.

The output feature $f_{i,j}$ of the j-th feature map after being convolved by the i-th 3D convolutional layer is expressed by the following formula:

$$f_{i,j}^{(x,y,z)_{i,j}} = \sigma\Big(\sum_{P_{i,j}=0}^{P_{i,j}-1}\sum_{q_{i,j}=0}^{Q_{i,j}-1}\sum_{r_{i,j}=0}^{R_{i,j}-1} w_{i,j}^{(p,q,r)} f_{(i-1)}^{(x+p)(y+q)(z+r)} + b_{i,j}\Big). \tag{4}$$

where $\sigma(\cdot)$ denotes the ReLu activation function, and $b_{i,j}$ is the bias term. $P_{i,j}$, $Q_{i,j}$, and $R_{i,j}$ are the three dimensions of the 3D kernel, where $R_{i,j}$ is the size of the temporal dimension. $(x,y,z)_{i,j}$ is the bit value of any pixel of the j-th feature map on the i-th convolutional layer, $w_{i,j}^{(p,q,r)}$ is the (p,q,r)-th value of the 3D kernel, which connected to the j-th feature map in the previous layer. After five layers of 3D convolutional layers, we extract feature vectors from the full connected layer as visual features:

$$f_{i,j} = \sigma\Big(\sum_{k=0}^{K_{(i-1)}-1} w_{(i,j)k} f_{(i-1,j)k} + b_{i,j}\Big). \tag{5}$$

where k is the index of the connection between the current feature map and the set of neurons in the $(i-1)$-th layer. $w_{(i,j)k}$ is the weight value that connects the j-th neuron in the i-th layer to the k-th neuron in the previous layer. And the $b_{i,j}$ is the bias term for this feature map.

C. Temporal action feature extraction

LSTM is a special recurrent neural network (RNN) that can learn and store long-term dependencies [28]. LSTM is suitable for processing data that changes with time, space, and other factors, such as IMU data. Furthermore, it is more in line with the process of human cognition in terms of representation. Compared with the traditional RNN, the LSTM introduces a gating mechanism. The definition of the LSTM gate is as follows:

$$\mathrm{gate}_{f,i,o}\left(h_{t-1}, x_t\right) = \sigma\left(Uh_{t-1} + Wx_t + b\right). \tag{6}$$

where f, i, o represent forget gate, input gate, output gate, U, W, b correspond to different parameters for different gates.

The LSTM brings in cell state C_t and hidden state h_t, corresponding to long-term and short-term information, where C_t changes slowly and h_t changes quickly. The transition of the two states is as follows:

$$\begin{aligned} C_t = \sigma\left(U_i h_{t-1} + W_i x_t + b_i\right) \odot \tanh\left(U_x h_{t-1} + W_x x_t + b_x\right) \\ + \sigma\left(U_f h_{t-1} + W_f x_t + b_f\right) \odot C_{t-1}. \end{aligned} \tag{7}$$

$$h_t = \tanh\left(C_t\right) \odot \sigma\left(U_o h_{t-1} + W_o x_t + b_o\right). \tag{8}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, $\odot$ is the element-wise product. The current cell state C_t is that the previous cell state being filtered by the forget gate and then superimposed with the information of the input gate. In the hidden state h_t, the output gate selectively retains and removes some data in the previous steps.

In the forward propagation, the feature of IMU data f_t that extracted by the LTSM can be defined as:

$$f_t = \sigma\left(P_y h_t + b_y\right). \tag{9}$$

where $\sigma(\cdot)$ denotes the activation function, P_y is the parameter that the network needs to learn, and b_y is the bias of this feature map. We use two LSTMs to deal with acceleration and angular velocity, respectively. The LSTM module is illustrated in the bottom left of Fig. 3, due to the application requirements of classification, we define LSTM as many to one, that is, output the h_t of the last time step.

D. Feature fusion strategies

The multi-modal fusion fuses the features of multiple modalities to provide supplementary information among different modalities. Mathematically, each model can extract useful features from the data of workers' activities, and we can use different strategies to fuse features from different models:

Average Fusion. In this method, we use the average value to fuse the features of multiple modalities and calculate the sum of the two feature maps in the feature dimension d:

$$f_{\text{avg}}^{d} = \frac{1}{2}\left(f_1^d + f_2^d\right). \tag{10}$$

where $1 \le d \le D$ and $f_1, f_2, f_{\text{avg}} \in \mathbb{R}^D$.

Maximum Fusion. This method compares the two feature maps bit by bit, and selects the larger one as the feature value of the corresponding position:

$$f_{\max}^{d} = \max\left\{f_1^d, f_2^d\right\}. \tag{11}$$

where $f_{\max}^{d}$ is the maximum value at position d in the two feature maps, and $f_{\max} \in \mathbb{R}^D$.

Concatenation Fusion. $f_{cat} = concat([f_1, f_2])$ stack two feature maps in the same dimension:

$$f_{\text{cat}}^{d} = f_1^d, f_{\text{cat}}^{D+d} = f_2^d, 1 \le d \le D. \tag{12}$$

where $f_{\text{cat}} \in \mathbb{R}^{2D}$. Concatenation fusion is the use of row or column lengthening for two features with the same dimensions. In this paper, we select the row lengthening method.

Convolution Fusion. This method first concatenates the two features using Eq.(12) and then convolves the concatenated features with the convolution kernel:

$$f_{\text{conv}} = f_{\text{cat}} * filter + b. \tag{13}$$

where $filter \in \mathbb{R}^{1\times D}$, $b \in \mathbb{R}^D$, and the dimension of the output is D. Here, the convolution kernel is used to reduce the dimension by half. It can model the weighted combination of f_1 and f_2 at the corresponding position.

IV. Feature-level fusion based on camera and IMU sensors

In this section, we have a detailed introduction to FFCI. The pseudo-code of the FFCI is shown in the Algorithm 1. Before introducing algorithm 1, we need to make the following preparations.

After preprocessing, the video and IMU data collected from workers' activities are described in section III. There are N data samples $\{X_1, \ldots, X_N\}$, each sample contains two inputs, as shown in the Eq.(2).

$V_{n,T}$ is video-level input, the dimension of the data is $16 \times 3 \times 112 \times 112$, that is, the image pixels from the 3 channels are 112×112, and 16 frames of images are input as a video stream. For video stream data, we use the C3D network to extract vision features. C3D module includes five 3D convolutional layers, and the kernel size of each layer is $3 \times 3 \times 3$ (the first two 3×3 are in the spatial dimension and the last 3×3 is in the temporal dimension). The first 3D max-pooling layer connected to the convolutional layer uses a $2 \times 2 \times 1$ filter, and the remaining four layers use a $2 \times 2 \times 2$ filter. Finally, features are extracted in the fully connected layer. The dimensions of the features extracted by the two networks are 512.

Algorithm 1 Feature-level fusion based on camera and IMU sensors (FFCI).

```
Input: Training set $D = \{[V_{n,T}, S_{n,T}], y\}$, learning rate $\eta$,
    number of classes $M$.
Output: The predicted label vector $\hat{y}$ and parameters $\theta$.
for each $epoch = 1$ to $N$ do
    for all $[V_{n,T}, S_{n,T}], y \in D$ do
        // Obtain the Visual Features
        // Calculate the $i$-th 3D convolutional layer
        $C_i(V_{n,T}) \leftarrow$ Eq.(4)  for $i = 1$ to 5;
        $f_{i,j} \leftarrow$ flatten the outputs of $i$th 3D convolution ;
        $f_{\text{visual}} \leftarrow FC(f_{i,j})$ via Eq.(5) for $i = 1, 2$;
        // Obtain the Motion Features
        // Calculate the $C_t$, $h_t$ and $f_t$ of LSTM
        $C_t \leftarrow$ Eq.(7) for $t = 1$ to $T$;
        $h_t \leftarrow$ Eq.(8) for $t = 1$ to $T$;
        $f_t \leftarrow$ Eq.(9) for $t = 1$ to $T$;
        $f_{\text{motion}} \leftarrow f_T$;
        // Obtain the Fusion Features
        $f_{\text{Fusion}} \leftarrow Fusion(f_{\text{visual}}, f_{\text{motion}})$;
        $\hat{y} \leftarrow \text{Softmax}(f_{\text{Fusion}})$;
        Loss $\leftarrow \frac{1}{N}\sum_i -\sum_{c=1}^{M} y_c^i \log\left(\hat{y}_c^i\right)$;
        Update $\theta$ with the $\eta$ to minimize Loss;
    end for
end for
return  The $\hat{y}$ and $\theta$ after n epochs.
```

$S_{n,T}$ is IMU data, including acceleration and angular velocity (see Eq.(3)). There are three channels for each type of data, a total of 20 nodes, and 120 columns of data are generated each time. Here, we respectively set the window length and the acquisition frequency to $T = 2s$ and 60 Hz, each window has 120 rows of data. Therefore, the input dimension of IMU data is 120×120. In this paper, we select LSTM to process time-series data, and each LSTM contains 256 hidden neurons, and a temporal motion feature is the output of the LSTM at the last time step.

The purpose of training a deep learning model is to optimize the weight of the network, minimize the value of the loss function, that is, minimize the gap between the predicted value and actual value. We select the cross-entropy loss function to calculate loss during training. In order to obtain the smallest loss value, we choose to use the stochastic gradient descent (SGD) algorithm. Compared with non-random algorithms, SGD performs better in the initial stage, making the model converge faster. Aiming at overfitting, the dropout method is a better way to solve it. And the dropout rate is set to 0.5. We use L2 regularization to control the model complexity and improve the overfitting. Dropout and L2 regularization can reduce the dependence of the model on local features and

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2025.3534236

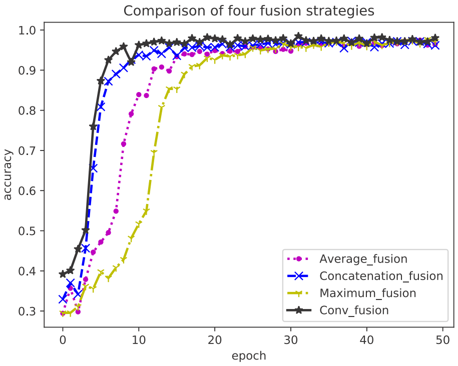


Fig. 9. Performance comparison of the four fusion strategies.

improve the generalization ability of the model.

Then, we will introduce Algorithm 1 in detail. We perform N rounds of training on the collected data. Firstly, we extracts the deep features of the video data through C3D. For i convolution layer in 3D, we use Eq.(4) to complete the corresponding computing task. Then, obtain the motion features by LSTM. Specially, input IMU data into the LSTM and calculate the $C(t)$, $h(t)$, and $f(t)$ to obtain the features of the last time step respectively according to Eq.(7), Eq.(8), and Eq.(9). Next, we fuse the two extracted features and classify the fusion features. Finally, we update the network parameters.

V. Performance Analysis

A. Experimental evaluation

In this section, we examine the accuracy and precision of FFCI, and we use the following metrics to evaluate the classification results:

- **Accuracy:** It is the ratio of the number of correct predictions to the total number of input samples.

$$\text{Accuracy } = \frac{\sum_{n}^{N} 1\left(\hat{y}_n = y_n\right)}{N}. \tag{14}$$

- **Precision:** It is the ratio of positive cases that were correctly identified.

$$\text{Precision } = \frac{TP}{TP + FP}. \tag{15}$$

where $1(\cdot)$ is an indicator function, $\hat{y}_n$ is the predicted value, y_n is the ground truth. For y_i, true positive (TP) is defined as the class y_i is correctly classified as y_i, false positive (FP) is defined as the non-y_i class is incorrectly classified as y_i.

B. Evaluation of different fusion strategies

Here, we compare the performance of four fusion strategies: 1) Average fusion, 2) Maximum fusion, 3) Concatenation fusion, 4) Convolution fusion.

Firstly, we test the convergence speed of the above four strategies. The convergence of different fusion strategies at 50 epoch is shown in Fig. 9, which shows that convolution fusion has the fastest convergence speed, concatenation fusion is slightly slower. And the convergence speeds of average fusion and maximum fusion are relatively slow.

TABLE IV
Comparison of Different Fusion Strategies

Fusion Strategies	Accuracy(%)	Precision(%)
Average fusion	96.88	97.30
Maximum fusion	96.69	97.47
Concatenation fusion	97.27	97.79
Convolution fusion	**98.17**	**98.39**

In terms of accuracy and precision, the performance of the four fusion strategies is shown in Table IV. The accuracy and precision of concatenation fusion are higher than that of average fusion and maximum fusion, but lower than that of convolution fusion. In the experiment, convolution fusion has the fastest convergence speed, the highest accuracy and precision. Therefore, we choose convolution fusion strategy to fuse the extracted visual features and action features in the following experiments.

C. Evaluation of different recognition methods

In order to verify the effectiveness of FFCI, we compare FFCI to several popular recognition methods.

For the video recognition, we compare FFCI to the LRCN (CNN&LSTM) [45] and the C3D model. The LRCN combines a 2D convolutional neural network with LSTM, extracts image features through convolution and time relationship between images through LSTM. It is a typical video recognition model. The FFCI uses 3D convolution, which increases the convolution of time dimension, and extracts the spatial and temporal features of the video.

For the IMU data, we choose the LSTM to predict and classify. In addition, we compare FFCI to random forests (RF) and supported vector machine (SVMs). The traditional machine learning method uses feature engineering to extract 18 features(mean, variance, kurtosis, etc.) from the time and frequency domains. Therefore, we use the SVM and RF classifiers on the scikit-learn platform to complete the classification task. Since LSTM does not require feature engineering, we can directly input the raw data into the LSTM. Furthermore, we normalize and standardize the raw data to make the network convergence better.

The accuracy and precision of various methods are listed in Table V. For the recognition method using video data, the accuracy and precision of the C3D are higher than those of LRCN. Therefore, we use C3D to extract visual features of workers' activities. For the recognition method based on the IMU data, the accuracy and precision of LSTM are higher than those of RF and SVM, which indicates that LSTM learning the temporal motion features of workers' activities can help improve the recognition accuracy and precision. The accuracy and precision of FFCI are higher than those of other methods that use video data or IMU data separately. The experimental results show that the multi-modal method are effective in workers' activities recognition.

VI. Smart Factory Application

The working process of FFCI in smart factory is shown in Fig.10. The sensors in the monitoring system collect data

TABLE V
PERFORMANCE COMPARISON OF EXISTING POPULAR METHODS

Methods	Accuracy(%)	Precision(%)
RF	87.34	88.21
SVM	85.89	86.93
LSTM [28]	90.81	91.48
LRCN [45]	86.67	87.17
C3D [27]	91.34	92.93
FFCI	**98.17**	**98.39**

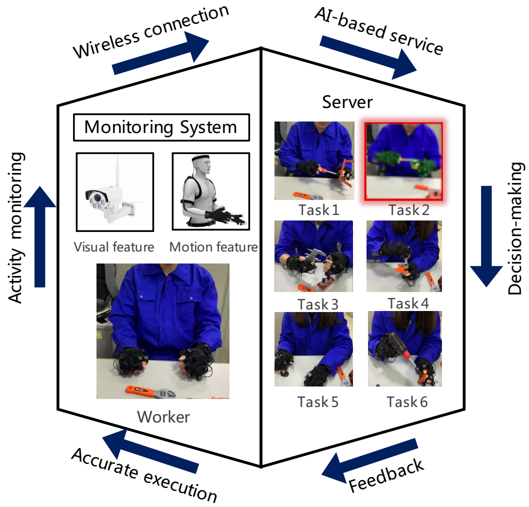
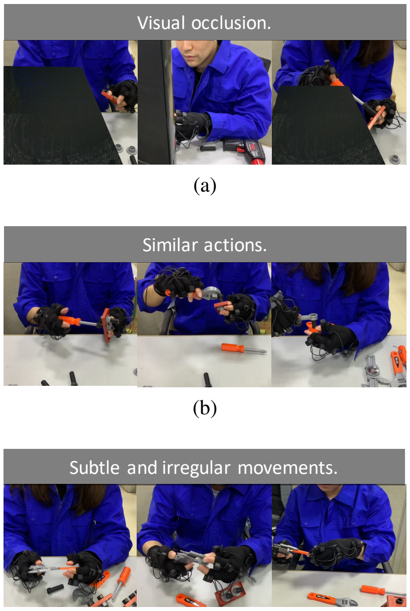


Fig. 10. Worker monitoring system using FFCI in smart factory.



(a)

(b)

(c)

Fig. 11. Three situations that affect activity recognition in smart factory.

and transmit it to the server. The FFCI will automatically analyze and make decisions. Then, the decisions will be sent to the workers and managers. In this paper, we evaluate the performance of the FFCI in terms of accuracy and real-time.

A. Accuracy analysis

In the data collection stage, we consider some of the following possible scenarios for smart factory.

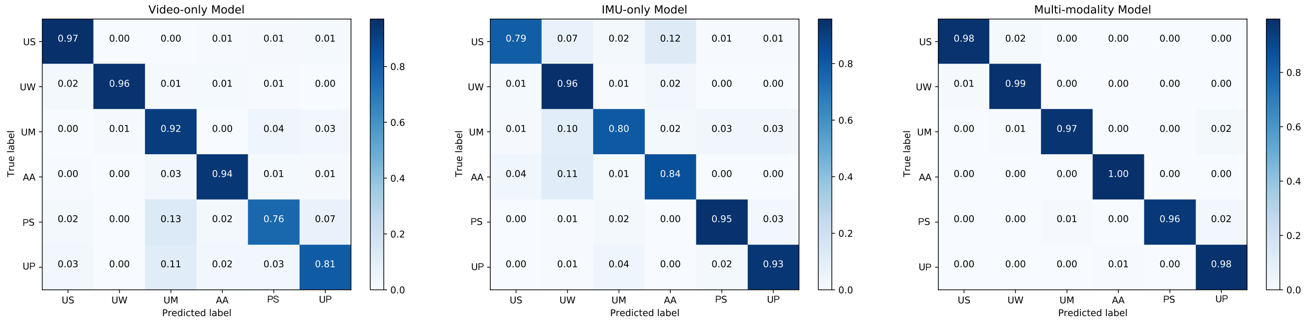
- **Visual occlusion**. As shown in Fig. 11(a), we collect a lot of visual occlusion data in the UM class, followed by PS and UP classes, and less in the rest.
- **Similar actions**. Using different tools to complete the task of tightening/loosening screws may have similar actions. As shown in Fig. 11(b), from left to right are the actions of using the screwdriver, using large wrench, and using small wrench. The similarity of these actions is high, it is difficult for the IMU sensor to distinguish.
- **Subtle and irregular movements**. The camera sensor has low accuracy in recognizing subtle movements, and it is difficult for the IMU sensor to recognize irregular movements. The operation in Fig. 11(c) is the use of measurement tool. The action is relatively subtle and varies with the operating habits of workers, it is difficult to identify these movements using camera or IMU sensor alone.

These situations are unfavorable to single sensor, and our proposed FFCI solves the problem of single sensor in smart factory. In Fig. 12, the accuracy of video recognition in the PS and UP class is low due to the UM data having the most occlusion among the UM, PS, and UP classes, the C3D network may classify the occluded data as the UM class, resulting in a significant decrease in the accuracy of PS and UP classes. In contrast, the UM class is less affected. Besides, the tasks of US, UW, and AA are all tightening/loosening screws. Although the tools are different, the action has a certain similarity, so the classification error of these three classes is relatively large. For example, based on the IMU method in Fig. 12(b), part of the US class is classified into the AA class, and part of the AA class is classified into the UW class, resulting in lower accuracy of these two classes. In addition, because the actions of UM class are subtle and irregular, the recognition accuracy is also low. As shown in Fig. 12(c), the method of fusing two sensors maintains a higher recognition accuracy in each class, while the method using single sensor has lower accuracy in some classes.

Table VI intuitively shows the classification accuracy of each activity class under different methods. The performance of the first five methods under different classes is not stable. For example, the first three methods based on IMU data have the low classification accuracy of the US, UW, UM, and AA classes. Besides, the classification accuracy of the PS and UP classes is also low in the video-based recognition method. However, multi-modal fusion can complement the two modal information and stably maintain high recognition accuracy for each class. In the smart factory, fusing the advantages of the two sensors can improve the recognition accuracy of workers' activities.

B. Real-time analysis

When worker activity recognition is applied to the monitoring system of smart factory, in addition to the accuracy, response speed is also not negligible. As mentioned above,



(a) Activity recognition model using video

(b) Activity recognition model using IMU data

(c) FFCI

Fig. 12. Classification accuracy of different mechanisms.

TABLE VI
CLASSIFICATION ACCURACY(%) OF EACH CLASS

Methods	Accuracy (%)					
	US	UW	UM	AA	PS	UP
RF	85.63	91.10	82.11	84.58	83.18	97.76
SVM	83.90	84.67	75.88	80.62	95.69	94.84
LSTM [28]	78.87	95.74	80.42	84.39	94.62	93.38
LRCN [45]	91.61	89.46	86.63	90.19	75.45	77.81
C3D [27]	96.98	95.78	92.17	94.12	76.23	80.57
FFCI	**98.24**	**98.60**	**97.39**	**99.63**	**96.31**	**98.47**

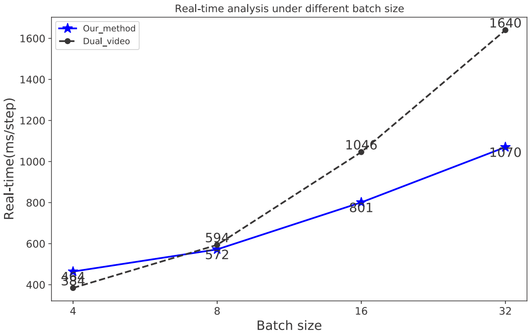


Fig. 13. The real-time comparison of 1 step between the two methods under different batch sizes.

video recognition is easily affected by visual occlusion. The solution of improving recognition accuracy is to add at least one other angle of video. Due to the low network complexity of the IMU sensor recognition method, although the model responds quickly, the accuracy is low. Therefore, we mainly compare the performance of real-time of the dual video streams recognition model and the FFCI. As shown in Fig. 14, train the model on the GPU, we compare the training time for 1 step of the two models under the different batch sizes. It turns out that the time cost of the proposed multi-modal fusion method increases slowly. Moreover, the larger the data scale, the more apparent the time advantage.

In general, our proposed FFCI maintains a high level of accuracy in each activity and has the advantage of real-time in smart factory.

VII. CONCLUSION

The activities of workers affect the safety and efficiency of factory production. With the development of smart factory, worker activity monitoring has become more automated and effective. However, the existing methods based on camera and IMU sensors respectively have problems of visual occlusion and similar operation recognition difficulties, which result in reduced recognition accuracy. Therefore, we propose a camera and IMU sensors based multi-modal neural network called FFCI for activity recognition in smart factory. To verify the effectiveness of FFCI, we establish a multi-modal dataset of worker activities, including six frequent operations generated by nine subjects. Simulation experiments and FFCI deployed on the smart factory prove that FFCI can significantly improve the accuracy of worker activities recognition.

REFERENCES

[1] W. S. Alaloul, M. Liew, N. A. W. A. Zawawi, and I. B. Kennedy, "Industrial revolution 4.0 in the construction industry: Challenges and opportunities for stakeholders," *Ain shams engineering journal*, vol. 11, no. 1, pp. 225–230, 2020.

[2] X. Qingxin, A. Wada, J. Korpela, T. Maekawa, and Y. Namioka, "Unsupervised factory activity recognition with wearable sensors using process instruction information," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–23, 2019.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] X. Zhou, W. Liang, K. I.-K. Wang, and L. T. Yang, "Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 171–178, 2021.

[5] X. Niu, C. Yu, and H. Jin, "Crsm: Computation reloading driven by spatial-temporal mobility in edge-assisted automated industrial cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 9283–9291, 2022.

[6] Z. Luo, C. Jiang, L. Liu, X. Zheng, and H. Ma, "Research on deep reinforcement learning based intelligent shop scheduling method," *Chinese Journal on Internet of Things*, vol. 6, no. 1, pp. 53–64, 2022.

[7] X. Zhou, X. Ye, K. I.-K. Wang, W. Liang, N. K. C. Nair, S. Shimizu, Z. Yan, and Q. Jin, "Hierarchical federated learning with social context clustering-based participant selection for internet of medical things applications," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 1742–1751, 2023.

[8] J.-H. Syu, J. C.-W. Lin, G. Srivastava, and K. Yu, "A comprehensive survey on artificial intelligence empowered edge computing on consumer electronics," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 4, pp. 1023–1034, 2023.

[9] X. Zhou, X. Zheng, T. Shu, W. Liang, I. Kevin, K. Wang, L. Qi, S. Shimizu, and Q. Jin, "Information theoretic learning-enhanced dual-generative adversarial networks with causal representation for robust ood generalization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[10] X. Zhou, W. Liang, K. I.-K. Wang, Z. Yan, L. T. Yang, W. Wei, J. Ma, and Q. Jin, "Decentralized p2p federated learning for privacy-preserving and resilient mobile robotic systems," *IEEE Wireless Communications*, vol. 30, no. 2, pp. 82–89, 2023.

[11] X. Zhou, W. Liang, A. Kawai, K. Fueda, J. She, and K. I.-K. Wang, "Adaptive segmentation enhanced asynchronous federated learning for sustainable intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 6658–6666, 2024.

[12] L. Malburg, M.-P. Rieder, R. Seiger, P. Klein, and R. Bergmann, "Object detection for smart factory processes by machine learning," *Procedia Computer Science*, vol. 184, pp. 581–588, 2021.

[13] X. Qingxin, A. Wada, J. Korpela, T. Maekawa, and Y. Namioka, "Unsupervised factory activity recognition with wearable sensors using process instruction information," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–23, 2019.

[14] X. Zhou, X. Zheng, X. Cui, J. Shi, W. Liang, Z. Yan, L. T. Yang, S. Shimizu, I. Kevin, and K. Wang, "Digital twin enhanced federated reinforcement learning with lightweight knowledge distillation in mobile networks," *IEEE Journal on Selected Areas in Communications*, 2023.

[15] X. Zhou, W. Huang, W. Liang, Z. Yan, J. Ma, Y. Pan, I. Kevin, and K. Wang, "Federated distillation and blockchain empowered secure knowledge sharing for internet of medical things," *Information Sciences*, vol. 662, p. 120217, 2024.

[16] X. Zhou, J. Wu, W. Liang, I. Kevin, K. Wang, Z. Yan, L. T. Yang, and Q. Jin, "Reconstructed graph neural network with knowledge distillation for lightweight anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[17] Y. Celik, S. Stuart, W. L. Woo, E. Sejdic, and A. Godfrey, "Multi-modal gait: A wearable, algorithm and data fusion approach for clinical and free-living assessment," *Information Fusion*, vol. 78, pp. 57–70, 2022.

[18] L. Wang, X. Zhang, Z. Song, J. Bi, G. Zhang, H. Wei, L. Tang, L. Yang, J. Li, C. Jia *et al.*, "Multi-modal 3d object detection in autonomous driving: A survey and taxonomy," *IEEE Transactions on Intelligent Vehicles*, 2023.

[19] X. Zhou, Q. Yang, Q. Liu, W. Liang, K. Wang, Z. Liu, J. Ma, and Q. Jin, "Spatial–temporal federated transfer learning with multi-sensor data fusion for cooperative positioning," *Information Fusion*, vol. 105, p. 102182, 2024.

[20] X. Zhou, Q. Yang, X. Zheng, W. Liang, I. Kevin, K. Wang, J. Ma, Y. Pan, and Q. Jin, "Personalized federated learning with model-contrastive learning for multi-modal user modeling in human-centric metaverse," *IEEE Journal on Selected Areas in Communications*, 2024.

[21] X. Sun, Y. Tian, W. Lu, P. Wang, R. Niu, H. Yu, and K. Fu, "From single-to multi-modal remote sensing imagery interpretation: A survey and taxonomy," *Science China Information Sciences*, vol. 66, no. 4, p. 140301, 2023.

[22] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3200–3225, 2023.

[23] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino, "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Information Fusion*, vol. 80, pp. 241–265, 2022.

[24] X. Lai, Q. Liu, X. Wei, W. Wang, G. Zhou, and G. Han, "A survey of body sensor networks," *Sensors*, vol. 13, no. 5, pp. 5406–5447, 2013.

[25] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, pp. 1–33, 2014.

[26] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, pp. 4405–4425, 2017.

[27] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[28] K. Muhammad, Mustaqeem, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, and V. H. C. D. Albuquerque, "Human action recognition using attention based lstm network with dilated cnn features," *Future Generation Computer Systems*, vol. 125, pp. 820–830, 2021.

[29] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[30] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[31] S. A. Abdu, A. H. Yousef, and A. Salem, "Multimodal video sentiment analysis using deep learning approaches, a survey," *Information Fusion*, 2021.

[32] S. Chen, F. Hu, Z. Chen, and H. Wu, "Correction method for uav pose estimation with dynamic compensation and noise reduction using multi-sensor fusion," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 980–989, 2024.

[33] R. Memmesheimer, N. Theisen, and D. Paulus, "Gimme signals: Discriminative signal encoding for multimodal activity recognition," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 394–10 401.

[34] G. A. Prabhakar, B. Basel, A. Dutta, and C. V. Rama Rao, "Multichannel cnn-blstm architecture for speech emotion recognition system by fusion of magnitude and phase spectral features using dcca for consumer applications," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 2, pp. 226–235, 2023.

[35] S. Sun, T. Peng, H. Huang, Y. Wang, X. Zhang, and Y. Zhou, "Iot motion tracking system for workout performance evaluation: A case study on dumbbell," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 4, pp. 798–808, 2023.

[36] Y. Zhang, B. Song, X. Du, and M. Guizani, "Vehicle tracking using surveillance with multimodal data fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 7, pp. 2353–2361, 2018.

[37] W. Tao, M. C. Leu, and Z. Yin, "Multi-modal recognition of worker activity for human-centered intelligent manufacturing," *Engineering Applications of Artificial Intelligence*, vol. 95, p. 103868, 2020.

[38] Y. Ren and G.-P. Li, "An interactive and adaptive learning cyber physical human system for manufacturing with a case study in worker machine interactions," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6723–6732, 2022.

[39] M. Al-Amin, R. Qin, M. Moniruzzaman, Z. Yin, and M. C. Leu, "An individualized system of skeletal data-based cnn classifiers for action recognition in manufacturing assembly," *Journal of Intelligent Manufacturing*, no. 2, pp. 633–649, 2021.

[40] A. Sharma and A. Kumar, "Dream: Deep learning-based recognition of emotions from multiple affective modalities using consumer-grade body sensors and video cameras," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1434–1442, 2024.

[41] T. Liu, Y. Ma, W. Yang, W. Ji, R. Wang, and P. Jiang, "Spatial-temporal interaction learning based two-stream network for action recognition," *Information Sciences*, vol. 606, pp. 864–876, 2022.

[42] G. Zhao, Z. Gao, C. Tsai, and J. Lu, "CGPM: poverty mapping framework based on multi-modal geographic knowledge integration and macroscopic social network mining," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings,Part V*, vol. 13717. Springer, 2022, pp. 549–564.

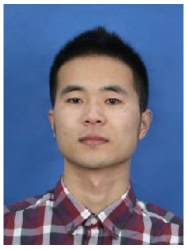
[43] Z. Ning, H. Hu, L. Yi, Z. Qie, A. Tolba, and X. Wang, "A depression detection auxiliary decision system based on multi-modal feature-level fusion of eeg and speech," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 3392–3402, 2024.

[44] R. Sers, S. Forrester, E. Moss, S. Ward, J. Ma, and M. Zecca, "Validity of the perception neuron inertial motion capture system for upper body motion analysis," *Measurement*, vol. 149, p. 107024, 2020.

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

Yujue Wang received the B.S. degree in computer science from Hunan University, Changsha, China, in 2019. She is currently working toward the M.S. degree in the Cluster and Grid Computing Lab, National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Big Data Technology and System Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, under the guidance of Pro. Chen Yu. Her research interests include multimodal interaction, the Industrial Internet of Things, and ubiquitous computing.

Xin Niu received the B.S. degree in Internet of Things engineering and the M.S. degree in computer science and technology from the College of Computer Science, Chongqing University, Chongqing, China, in 2017 and 2020, respectively. He is currently pursuing his Ph.D. degree in the Services Computing Technology and System Laboratory, Big Data Technology and System Laboratory, Cluster and Grid Computing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, under the guidance of Prof. Chen Yu. His research interests include edge intelligence, edge computing, and ubiquitous computing.



Xianwei Lv received the B.S. and M.S. degree in computer science from Huazhong University of Science and Technology, Wuhan, China, in 2016 and 2019, respectively. He is currently working toward the Ph.D. degree in the National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Big Data Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, under the guidance of Pro. Chen Yu. His research interests include ubiquitous computing, edge computing and deep learning.

Chen Yu received the B.S. degree in mathematics and the M.S. degree in computer science from Wuhan University, Wuhan, China, in 1998 and 2002, respectively, and the Ph.D. degree in information science from Tohoku University, Sendai, Japan, in 2005. From 2005 to 2006, he was a Japan Science and Technology Agency postdoctoral researcher in the Japan Advanced Institute of Science and Technology. In 2006, he was a Japan Society for the Promotion of Science Postdoctoral Fellow in the Japan Advanced Institute of Science and Technology. Since 2008, he has been in the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, where he is currently a professor and a special research fellow, working in the areas of ubiquitous computing, edge computing and Industrial Internet. He was a recipient of the Best Paper Award in the 2005 IEEE International Conference on Communication and the nominated Best Paper Award in the Proceedings of the 11th IEEE International Symposium on Distributed Simulation and Real-Time Application in 2007. He is a member of the IEEE.