Received September 10, 2020, accepted October 7, 2020, date of publication October 21, 2020, date of current version November 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3032699

Human Activity Recognition Based on Gramian Angular Field and Deep Convolutional Neural Network

HONGJI XU[1], **(Member, IEEE), JUAN LI**[1], **HUI YUAN**[2], **(Senior Member, IEEE), QIANG LIU**[1], **SHIDI FAN**[1], **TIANKUO LI**[1], **AND XIAOJIE SUN**[1]

[1]School of Information Science and Engineering, Shandong University, Qingdao 266237, China
[2]School of Control Science and Engineering, Shandong University, Jinan 250061, China

Corresponding authors: Hui Yuan (huiyuan@sdu.edu.cn) and Hongji Xu (hongjixu@sdu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0831001 and Grant 2017YFC0803400, in part by the National Natural Science Foundation of China under Grant 61771292, and in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2016FM29.

ABSTRACT With the development of the Internet of things (IoT) and wearable devices, the sensor-based human activity recognition (HAR) has attracted more and more attentions from researchers due to its outstanding characteristics of convenience and privacy. Meanwhile, deep learning algorithms can extract high-dimensional features automatically, which makes it possible to achieve the end-to-end learning. Especially the convolutional neural network (CNN) has been widely used in the field of computer vision, while the influence of environmental background, camera shielding, and other factors are the biggest challenges to it. However, the sensor-based HAR can circumvent these problems well. Two improved HAR methods based on Gramian angular field (GAF) and deep CNN are proposed in this paper. Firstly, the GAF algorithm is used to transform the one-dimensional sensor data into the two-dimensional images. Then, through the multi-dilated kernel residual (Mdk-Res) module, a new improved deep CNN network Mdk-ResNet is proposed, which extracts the features among sampling points with different intervals. Furthermore, the Fusion-Mdk-ResNet is adopted to process and fuse data collected by different sensors automatically. The comparative experiments are conducted on three public activity datasets, which are WISDM, UCI HAR and OPPORTUNITY. The optimal results are obtained by using the indexes such as accuracy, precision, recall and F-measure, which verifies the effectiveness of the proposed methods.

INDEX TERMS Deep convolutional neural network, Gramian angular field, human activity recognition, multi-source sensor data fusion.

I. INTRODUCTION

With the rapid development of the 5th generation (5G) mobile networks, Internet of things (IoT) and artificial intelligence (AI), the technology of human activity recognition (HAR) is becoming more and more important in people's daily lives because of its ability to analyze and recognize human activities by the raw sensor data. It has been widely used in many aspects, such as daily activity analysis [1], video surveillance [2], gait analysis [3] and gesture recognition [4]. At present, HAR is mainly divided into two categories: sensor-based activity recognition [5]–[7] and video-based activity recognition [8]–[10]. Video-based activity recognition mainly processes the video and image data collected by cameras, while sensor-based activity recognition is used to analyze and process the data collected by sensors such as accelerometers and gyroscopes. The sensor-based activity recognition has become the research focus due to its merits of good privacy and convenience.

The recognition models used in the HAR system can be roughly divided into two categories: one is based on classical machine learning algorithms; the other is based on deep learning algorithms. The commonly used classical machine learning algorithms include decision tree (DT) [19], [20], random forests (RF) [21], [22], and support vector machine (SVM) [23], [24]. Researchers all over the world have done a lot of studies on the sensor-based HAR using these algorithms. Zhou *et al.* proposed a two-layer classification

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenyu Zhou.

and recognition method based on DT [11]. Firstly, time-frequency domain features were extracted from the raw data of accelerometers and gyroscopes. After that, eight types of daily activities could be identified through the two-layer decision tree.

However, these classical algorithms all require complex and time-consuming feature engineering, which requires not only manual design of extracted features, but also feature selection or dimensionality reduction to screen out highly representative features. The method based on deep learning avoids the step of feature engineering, enabling researchers to pay more attention to other aspects of HAR, such as power consumption in practical applications. Chen *et al.* used an improved one-dimensional convolutional neural network (CNN) to classify the collected activity data [12], and compared with traditional methods such as DT and logistic regression, the recognition accuracies of various activities were improved, but the accuracies of going downstairs and going upstairs were still less than 70%. Kuang *et al.* compared the recognition performance of deep CNN and long short-term memory (LSTM) network on public datasets [13]. The results showed that the deep CNN using dropout got better recognition performance, and the training time of deep CNN was much less than that of LSTM. Deng *et al.* proposed two methods for activity images construction, transforming the sensor data into activity images by matrix rearrangement. And then, the CNN was applied to extract features and identify the types of activities [14]. Ravi *et al.* proposed an improved deep learning model, in which the deep features were extracted from the time-frequency graph of the original data by CNN. Meanwhile, the shallow features such as amplitude, mean, and variance of the original data were extracted in the time-frequency domain. After the combination of the deep features and the shallow features, the types of activities were identified [15]. Ordonez *et al.* proposed a new deep learning model that combined CNN and LSTM units, which was a significant improvement for traditional machine learning methods [16]. Xu *et al.* combined the CNN and the gated recurrent unit (GRU) to identify human activities, and verified the effectiveness of the proposed method on three public datasets [17]. Uddin *et al.* proposed a multi-sensors data fusion network based on recurrent neural network (RNN) [18]. Firstly, the time-frequency domain features were extracted from the original sensor data. Secondly, the effective features were selected by principal component analysis (PCA), and finally the human activities were recognized through RNN after processing the effective features.

Although deep learning method has many advantages, it also has some unavoidable problems and disadvantages. The recognition methods based on RNN [25], [26] can only be carried out in sequence because the calculation of the next step in the training process depends on the results of the previous step, and the training process consumes a long time. Recognition methods based on CNN mostly use one-dimensional convolution kernel [27], [28], which is difficult to fully exploit the rich high-dimensional data features. Many methods are used to convert the one-dimensional sensor data into the two-dimensional data through the matrix rearrangement [14], [29], [30], which is the simple listing and superposition of the data, but lacks interpretability. In [15], [31], [32], the one-dimensional time series were converted to the two-dimensional time-frequency images by Fourier transform, which led to a sharp increase in the amount of computations. Due to the requirements of portability and real-time property of wearable sensor devices, it is inevitable that sensor-based activity recognition methods require fewer computing resources and faster calculation speed.

To solve the above problems, a new HAR method based on Gramian angular field (GAF) [34] and deep CNN is proposed in this paper. Because of the weight sharing mechanism [33] of CNN, the training speed of CNN is much faster than the other networks. At the same time, the GAF algorithm increases the interpretability of transforming from one-dimensional time series to two-dimensional images and also lays a foundation for the effectiveness of feature extraction. The experimental results show that the new HAR method proposed in this paper can effectively improve the multi-scale feature extraction capability and the accuracy of activity recognition by combining the characteristics of GAF algorithm, the structure and advantages of CNN, residual learning and dilated convolution.

The main contributions of this paper are as follows:

(1) The GAF algorithm is used to quickly transform one-dimensional time series into two-dimensional images similar to the real image data, which makes it more effective to the application of two-dimensional CNN.
(2) A new improved deep CNN network Mdk-ResNet based on the multi-dilated kernel residual (Mdk-Res) modules is proposed in this paper, which makes it possible to extract rich features among the sampling points with different time intervals, thereby improving the recognition accuracy.
(3) The Fusion-Mdk-ResNet, a multi-source sensor data fusion network, can automatically fuse data collected by different sensors is proposed. The effectiveness of the proposed method is verified by comparison experiments on three public datasets, and better experimental results are obtained on both single sensor datasets and multi-sensors datasets.

The rest of the paper is arranged as follows. In section 2, the principles of residual block structure in ResNet, the inception module in GoogLeNet and the dilated convolution are introduced. In section 3, the proposed models of HAR based on deep CNN and GAF algorithm is presented. The experimental environments and the results of the simulations are demonstrated in section 4, and the conclusion is drawn in section 5.

II. RELATED WORK

A. THE PRINCIPLE OF THE RESIDUAL BLOCK

He *et al.* proposed the ResNet, in which the residual learning suppressed the problem of gradient disappearance in deep

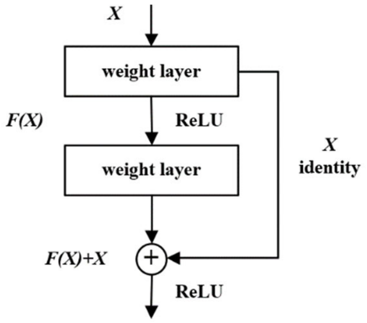


FIGURE 1. The residual block structure in ResNet [37].

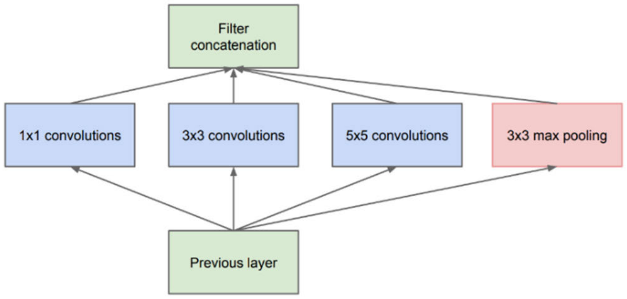


FIGURE 2. The inception module in GoogLeNet.

neural networks (DNN), improved its learning ability and increased the recognition accuracy significantly [35]. The residual block structure diagram is shown in Fig. 1. The input data X is the image data, $F(X)$ is the mapping of X obtained through multi-layer calculations, and the output of the residual block is $F(X) + X$.

B. THE PRINCIPLE OF THE INCEPTION MOUDULE

The GoogLeNet, a 22-layer neural network proposed by Szegedy *et al.*, won the first place in the ImageNet large scale visual recognition challenge 2014 (ILSVRC 2014) [36]. The inception module, as the core building block of the network, played an important role and its initial structure is shown in Fig. 2.

The inception module consists of four main components, namely 1 × 1 convolutions, 3 × 3 convolutions, 5 × 5 convolutions, and 3 × 3 max pooling. After being processed by these four parts, the feature maps of the upper layer will be combined into a new feature map to be transmitted to the next inception module. Such a structural design increases the width of the network, extracts information of different scales of images through multiple convolution kernels with different sizes, improves the feature extraction capability of the network, and can obtain better images representation.

C. THE PRINCIPLE OF THE DILATED CONVOLUTION

However, using large-scale convolution kernels greatly increases the amount of network parameters and the computation. In order to solve the above problems, Yu *et al.* proposed the dilated convolution, which increased the receptive field without losing information. Thus, the convolution output could contain a larger range of the information and the network can extract features from a larger scale without increasing the number of parameters of the convolution kernels [37]. The schematic diagram of dilated convolution is shown in Fig. 3. The basic idea is to add interspaces to the standard convolution operation so as to increase the receptive field. Compared with the original normal convolution process, the dilated convolution has a super parameter called the dilation rate, which refers to the pixel value of the interval on the feature map during convolution.

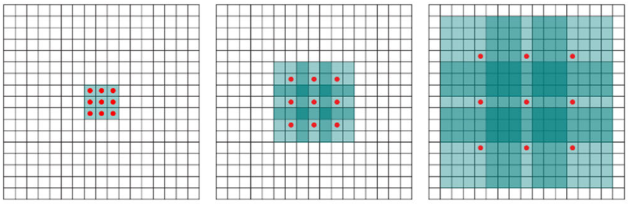


FIGURE 3. Dilated convolution of different dilation rates [37].

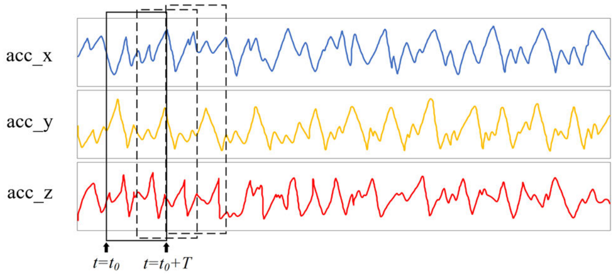


FIGURE 4. Sliding window segmentation of data collected by a triaxial accelerometer.

III. THE PROPOSED MODELS OF HAR BASED ON DEEP CNN

A. THE GAF ALGORITHM

In the sensor-based HAR, the accelerometer is one of the most commonly used sensors in current research, which can directly collect movement information by wearing it on body. The data collected by the sensor is continuous time series and can be segmented by the sliding window. In order to ensure a complete motion can be contained in a sliding window, the length of the sliding window is determined according to the sampling rate of the sensor and the type of human activity. In addition, the sliding window generally selects 50% overlap to ensure the integrity of the information. Fig. 4 is a diagram of the sliding window segmentation of the data collected by a triaxial accelerometer, in which a sliding window starts at $t = t_0$, the length of window is T and the window overlap is 50%. In Fig. 4, the graphs from top to bottom represent the x, y, and z axes of the raw data collected by the acceleration sensor, respectively.

Most of the original data collected by the sensor is one-dimensional time series. And it is usually necessary to convert one-dimensional time series into a format similar to two-dimensional images in the application of two-dimensional CNN.

Wang *et al.* proposed the GAF algorithm to convert the one-dimensional time series into the two-dimensional images which is one of the commonly used time series imaging algorithms [34]. The specific implementation steps are as follows:

Suppose a time series is $X = \{x_1, x_2, \ldots, x_i, \ldots, x_N\}$, containing N observations. Firstly, X is normalized so that all values of X can be in the range of $[-1, 1]$ or $[0, 1]$, which can be expressed as follows, respectively:

$$\tilde{x}^i_{-1} = \frac{(x_i - \max(X)) + (x_i - \min(X))}{\max(X) - \min(X)}, \qquad (1)$$

$$\tilde{x}^i_0 = \frac{x_i - \min(X)}{\max(X) - \min(X)}. \qquad (2)$$

Next, convert the one-dimensional time series from cartesian coordinate system to polar coordinate system, which can be expressed as

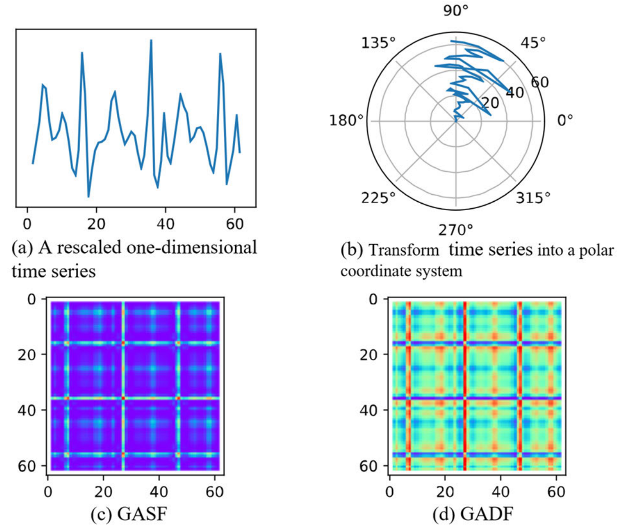
$$\begin{cases} \phi_i = \arccos(\tilde{x}_i), & -1 \le \tilde{x}_i \le 1, \tilde{x}_i \in \tilde{X} \\ r_i = \dfrac{i}{N}, & i \in N \end{cases} \qquad (3)$$

where the inverse cosine of the normalized observation $\tilde{x}_i$ is taken as the angle ϕ_i in the polar coordinate system, and the time label i/N is taken as the radius. The data processes by the two types of normalization operations has different angle ranges when converted to the polar coordinate system. The angle range of the cosine function corresponding to the data within the range $[0, 1]$ is $[0, \pi/2]$, and the angle corresponding to the data in the range of $[-1, 1]$ is $[0, \pi]$.

This representation method based on the polar coordinate system provides a new view for understanding time series. That is, as time goes by, the sequence value varies from the original amplitude change to the angular change in the polar coordinate system. By calculating the sum/difference of the trigonometric function among sampling points, the time correlation among them is identified from the perspective of angle. Gramian angular summation field (GASF) and Gramian angular difference field (GADF) are defined as follows, respectively:

$$\textbf{GASF} = \begin{bmatrix} \cos(\phi_1 + \phi_1) & \cdots & \cos(\phi_1 + \phi_n) \\ \cos(\phi_2 + \phi_1) & \cdots & \cos(\phi_2 + \phi_n) \\ \cdots & \ddots & \cdots \\ \cos(\phi_n + \phi_1) & \cdots & \cos(\phi_n + \phi_n) \end{bmatrix}, \qquad (4)$$

$$\textbf{GADF} = \begin{bmatrix} \sin(\phi_1 - \phi_1) & \cdots & \sin(\phi_1 - \phi_n) \\ \sin(\phi_2 - \phi_1) & \cdots & \sin(\phi_2 - \phi_n) \\ \cdots & \ddots & \cdots \\ \sin(\phi_n - \phi_1) & \cdots & \sin(\phi_n - \phi_n) \end{bmatrix}. \qquad (5)$$



(a) A rescaled one-dimensional time series

(b) Transform time series into a polar coordinate system

(c) GASF

(d) GADF

FIGURE 5. One-dimensional time series converted to two-dimensional images in polar coordinate system.

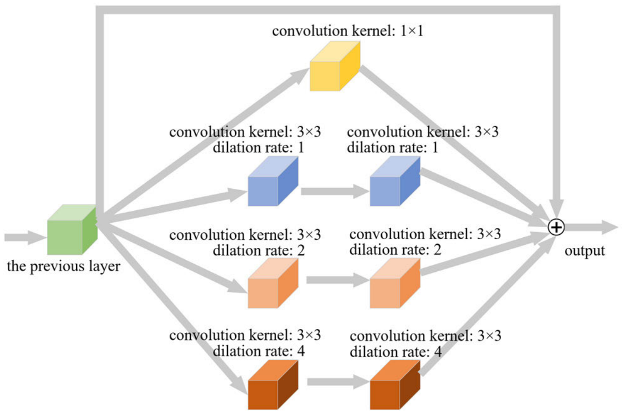


FIGURE 6. The structure diagram of Mdk-Res module.

The GAF algorithm is adopted to transform the one-dimensional time series into the two-dimensional images through three steps of scaling, coordinate axis transformation and trigonometric function, so as to apply the computer vision technology to the study of time. Fig. 5 shows the mapping relationship between one-dimensional time series and two-dimensional images. The time series is transformed into a polar coordinate system according to (3). The GASF and GADF images can be obtained by (4) and (5), respectively.

B. THE PROPOSED MDK-RES MODULE

Based on the advantages of the above networks, a multi-dilated kernel residual (Mdk-Res) module is proposed in this paper, and its structure is shown in Fig. 6.

It can be seen that the feature maps of the previous layer are input into the Mdk-Res model and processed through the four convolution channels. And then the results of the four parts and the input feature maps are added as the output. In the model, the convolution hyper-parameters for all four parts are set to "padding = same, stride = 1" to ensure that the dimensions of the output and input are the same.

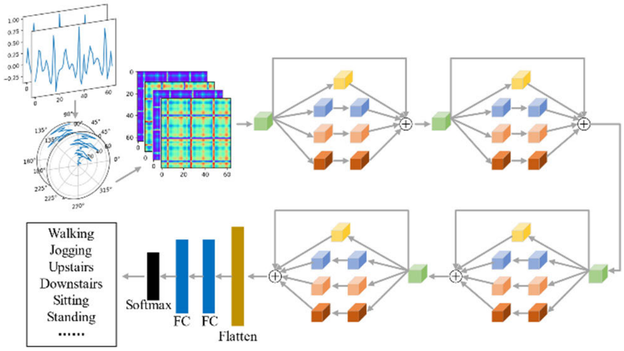


FIGURE 7. HAR network Mdk-ResNet for single sensor.

The green block represents the previous layer of the network, the yellow block represents the convolution whose kernel size is 1 × 1, the blue blocks represent the normal convolution whose kernel size is 3 × 3, the light orange blocks represent the dilated convolution whose kernel size is 3 × 3 and the dilation rate is 2, and the dark orange blocks represent the dilated convolution whose kernel size is 3 × 3 and the dilation rate is 4. The Mdk-Res module uses multiple normal convolution kernels and dilated convolution kernels at the same time, which improves the ability of the network to extract features of different scales. In addition, the residual learning is added into the Mdk-Res module, which suppresses the common gradient disappearance phenomenon in DNN and improves the fitting ability of the models.

The use of the dilated convolution allows the Mdk-Res module to extract the feature relationship among the sampling points with longer intervals. For example, the normal 3 × 3 convolution kernel extracts the feature relationships in a receptive field of 3 × 3. The 3 × 3 convolution kernel with a dilation rate of 2 extracts the feature relationships in a receptive field of 7 × 7, and the 3 × 3 convolution kernel with a dilation rate of 4 extracts the feature relationships in a receptive field of 15 × 15.

C. HAR NETWORK MDK-RESNET FOR SINGLE SENSOR

For the HAR with a single sensor, Mdk-ResNet is proposed in this paper, which is an improved HAR network based on Mdk-Res modules. The structure of Mdk-ResNet is shown in Fig. 7.

The time series data obtained by the sensor is converted from the rectangular coordinate axis system to the polar coordinate system, and then the two-dimensional images are obtained by the GAF algorithm. The GAF algorithm is used to convert each sliding window segmentation from the one-dimensional time series to the two-dimensional image format. Taking the data collected by triaxial acceleration as an example, assuming that the length of the sliding window is T, since the data contains three channels, the dimension of each sample is $T \times 3$. The GAF algorithm can transform the one-dimensional series of each channel into two-dimensional matrixes of GASF and GADF, and the dimension of each sample will become the form as same as the common RGB image. After that, the two-dimensional matrixes will be processed by the HAR network Mdk-ResNet which is the deep CNN network proposed in this section.

The two-dimensional matrixes are processed by multiple Mdk-Res modules and then transmitted to the fully connected (FC) layers after being flattened into one-dimensional data, and finally output by the Softmax layer. The corresponding highest probability of the labels is the result of the classification. In the model, the number of Mdk-Res modules can be selected according to the size of the dataset and the complexity of the activities to be identified, rather than being limited to the four modules listed in Fig. 7.

D. HAR NETWORK FUSION-MDK-RESNET FOR MULTI-SOURCE SENSOR DATA FUSION

For the HAR system using multi-source sensor, a multi-source sensor fusion network Fusion-Mdk-ResNet is proposed in this paper, as shown in Fig. 8. For the proposed Fusion-Mdk-ResNet, firstly, the one-dimensional time series collected by multiple sensors is transformed into the two-dimensional images through the GAF algorithm, and then features are extracted by multiple Mdk-Res modules, respectively. After that, the obtained feature data is merged by the feature fusion layer and processed by multiple Mdk-Res modules again. Finally, the feature data is input into the FC layers and the Softmax layer to obtain the result of activity recognition.

The number of the Mdk-Res modules can be increased or decreased according to the application scenarios. The network structure shown in Fig. 8 is adopted here. The fusion processing of the input feature maps of each sensor can be expressed as

$$Y = \sum_{i=1}^{N} w_i \times X_i \tag{6}$$

where Y is the output of feature fusion layer, X_i is the feature maps input by the i-th sensor to the feature fusion layer, w_i is the weight or confidence coefficient of the i-th sensor in the feature fusion layer, and N is the number of sensors. The feature fusion layer has two working methods. One is to set the corresponding confidence or coefficient according to the type and deployment position of each sensor, and perform weighted stitching according to the specified coefficient during fusion. While for the other fusion method, the coefficients are not specified. The coefficients of each feature map are the same initially, and then learned by the Fusion-Mdk-ResNet automatically.

IV. EXPERIMENTS AND EVALUATIONS

A. EVALUATION INDEXES

For the binary classification problem, the samples can be classified into four categories based on the true category and the prediction results:

- true positive (TP): The true category is positive, and the predicted category is positive.

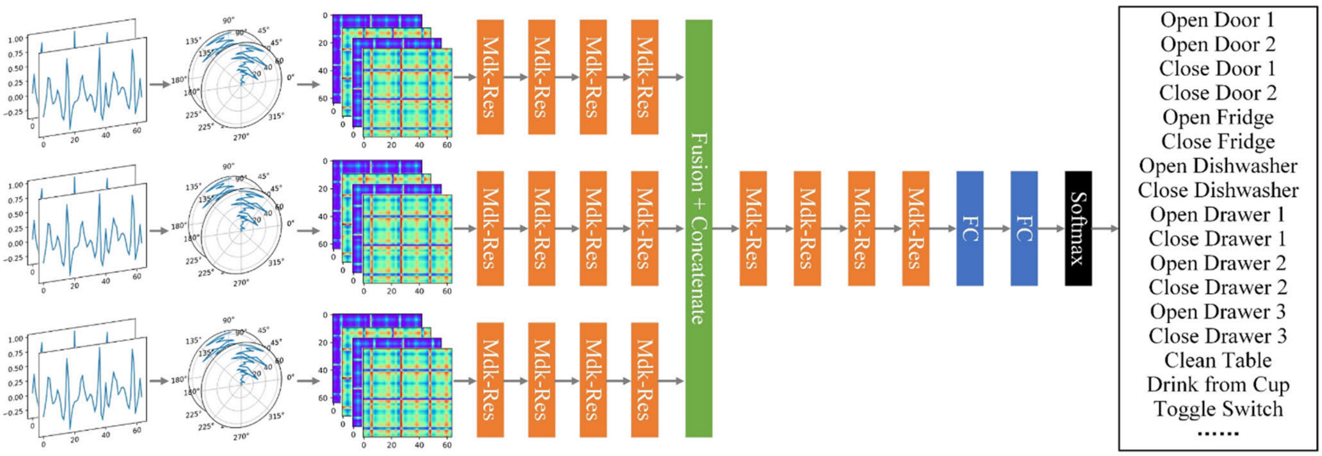


FIGURE 8. HAR network fusion-Mdk-ResNet for multi-source sensor data fusion.

- false positive (*FP*): The true category is negative, and the predicted category is positive.
- false negative (*FN*): The true category is positive, and the predicted category is negative.
- true negative (*TN*): The true category is negative, and the predicted category is negative.

The commonly used evaluation measures include accuracy, precision, recall, and F-measure. The accuracy can be expressed as

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{7}$$

The precision can be expressed as

$$precision = \frac{TP}{TP + FP}. \tag{8}$$

The recall can be expressed as

$$recall = \frac{TP}{TP + FN}. \tag{9}$$

The F-measure is the harmonic average of precision rate and recall rate, which can be expressed as

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{(\beta^2 \times precision) + recall}. \tag{10}$$

When $\beta = 1$, the commonly used variant F_1 can be expressed as

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}. \tag{11}$$

For multi-classification problems, we need to use macro average and micro average. The former one is to calculate the accuracy, recall rate and F-measure of each category firstly, and then, the arithmetic average is calculated, which is equivalent to giving each category the same weight, thus distribute more attention to categories of small sample size in the unbalanced dataset. The latter one is to count the *TP*, *FP*, *FN*, and *TN* of each category, and then calculate the corresponding the accuracy, recall, and F-measure, that is, giving each sample the same weight. For the unbalance dataset, the categories with a large number of samples will get more attentions. In particular, when the statistical range includes all categories, the precision, recall, and F-measure calculated by the micro average will be the same as the accuracy. Assuming that the number of categories is N, the macro average and micro average can be expressed as follows:

$$Macro\_Precision = \frac{1}{N}\sum_{i=1}^{N} precision_i, \tag{12}$$

$$Macro\_Recall = \frac{1}{N}\sum_{i=1}^{N} recall_i, \tag{13}$$

$$Macro\_F_1 = \frac{1}{N}\sum_{i=1}^{N} F_{1i}, \tag{14}$$

$$Micro\_Precision = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i}, \tag{15}$$

$$Micro\_Recall = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i}, \tag{16}$$

$$Micro\_F_1 = \frac{2 \times Micro\_Precision \times Micro\_Recall}{Micro\_Precision + Micro\_Recall}. \tag{17}$$

B. DATASETS

1) WISDM DATASET

The WISDM dataset [39] was released by the Laboratory for Wireless Sensor Data Mining at Fordham University in the United States. The data was collected from 36 volunteers who had smartphones in their front trouser pockets. The smartphone contains a three-axis accelerometer with a sampling rate of 20Hz and the dataset contains a total of 1098207 sampling points. The types of the collected activities

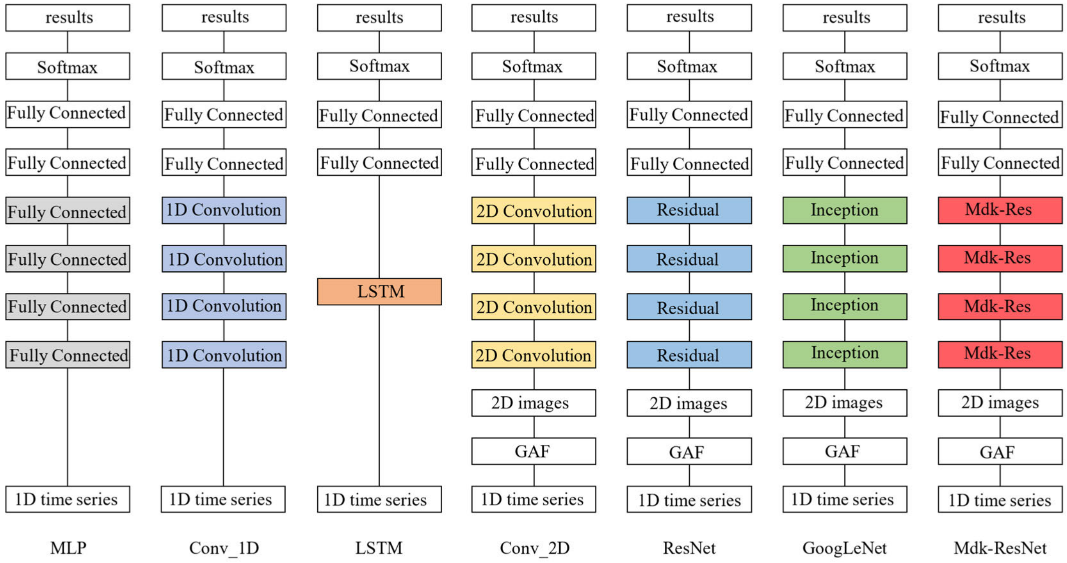


FIGURE 9. Seven network structures adopted in this paper.

TABLE 1. Experimental hardware and software configurations.

Hardware configuration	
CPU	Intel i7-9700K
GPU	NVIDIA GeForce RTX 2080 Ti
RAM	Kingston 16G DDR4 3000
Software configuration	
Operating system	Windows 10 Pro
Programming language	Python
Deep learning framework	TensorFlow 2.0.0, Keras 2.2
IDE	PyCharm 2018.2

include walking, jogging, going upstairs, going downstairs, sitting, and standing. Among them, the walking accounts for 38.6%, the jogging accounts for 31.2%, the going upstairs accounts for 11.2%, the going downstairs accounts for 9.1%, the sitting accounts for 5.5%, and the standing accounts for 4.4%. It is observed that the number of each activity is quite different, therefore the WISDM dataset is an unbalanced dataset.

2) UCI HAR DATASET

The UCI HAR dataset [40] was released by the Laboratory for Nonlinear Complex Systems at the University of Genoa in Italy. The experiment involved 30 volunteers aged between 19 and 48. Each person attached the smartphone to the waist and used the embedded accelerometer and gyroscope in the smartphone to collect 6 kinds of activity data (walking, going upstairs, going downstairs, sitting, standing, and lying). The sampling rate is 50Hz, and the number of samples is 10929.

TABLE 2. Experimental parameter settings on the WISDM dataset.

Parameter	Experimental parameter setting
Dataset partitioning	Training set 60%, validation set 20%, test set 20%.
Loss function	Cross entropy
Optimizer	Adam (β_1=0.9, β_2 = 0.999)
Learning rate	0.001
Batch size	32
Epoch	100

3) OPPORTUNITY DATASET

The OPPORTUNITY dataset [41] was released by the Laboratory for Wearable Computing of the Federal Institute of Technology in Zurich, which mainly collected various activities in the kitchen environment. There are three types of sensors: body-worn sensors, object sensors, and ambient sensors. The experiment involved four volunteers, each of whom collected six types of data, five of which are activities of daily living (ADL) and the other is Drill. The ADL is a collection of daily activities that volunteers collected under natural conditions, including activities such as walking in the room, preparing coffee, cleaning and rest. Drill refers to a collection that contains a series of specific actions performed by volunteers in a preset order, including opening and closing the refrigerator, opening and closing the door, and turning on and off the lights.

C. EXPERIMENTAL MODELS AND ENVIRONMENT SETTING

GoogLeNet proposed by Google researchers and ResNet proposed by Microsoft researchers are aimed at the ImageNet dataset [38], which is an extremely large dataset with nearly 15 million images. In order to achieve accurate classification

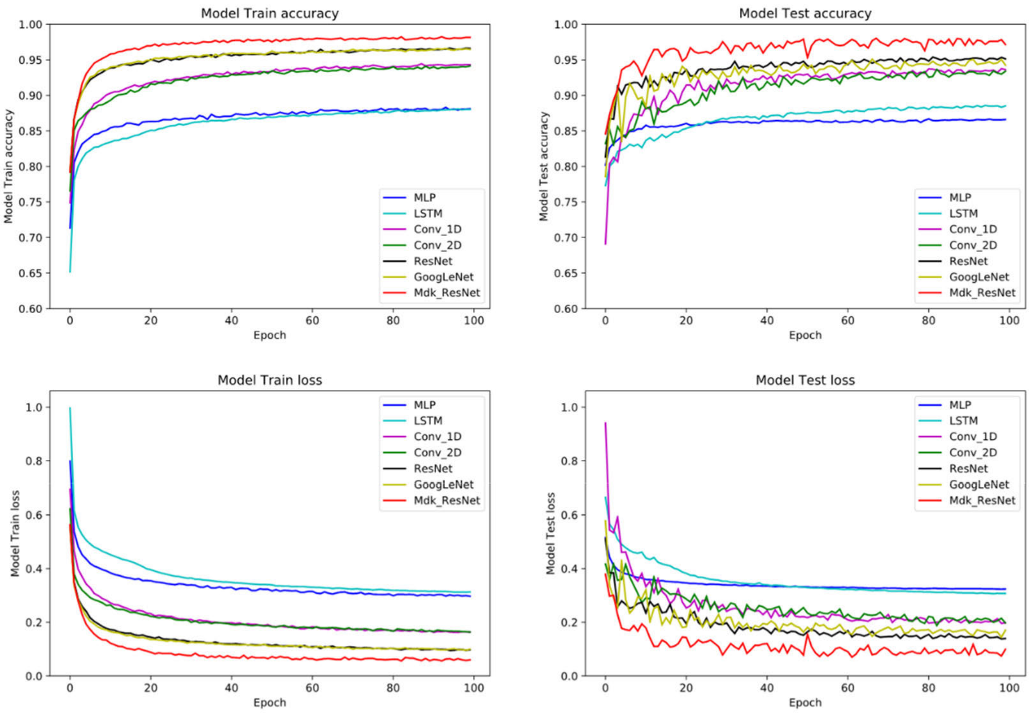


FIGURE 10. Accuracies and losses of training and testing on the WISDM dataset.

results, GoogLeNet and ResNet have been set to 22 layers [36] and 34 layers [35], respectively. The size and complexity of the public datasets used in this paper are smaller than the ImageNet dataset, so the original GoogLeNet and ResNet networks are not used in the comparative experiments. The GoogLeNet and ResNet have been retained in their original structural characteristics of the networks, but reduce the number of layers of the networks to avoid overfitting. As shown in Fig. 9, the Mdk-ResNet contains 4 Mdk-Res modules, the GoogLeNet and ResNet used in the comparative experiments also contain 4 inception modules and residual modules, respectively, and the other networks contain the same number of corresponding modules. Due to the training difficulties of LSTM, the number of LSTM layers is generally set to be 1 to 3. And the general optional parameter is the number of hidden units which is set to the length of the sliding window here.

The specific hardware and software configurations of the experiments are shown in Table 1.

D. EXPERIMENTAL RESULTS AND ANALYSES

1) EXPERIMENTAL RESULTS AND ANALYSES ON THE WISDM DATASET

The sampling rate of data in WISDM dataset is 20Hz, the length of sliding window is set to 64. The coverage time of such a sliding window is 3.2 seconds, which can meet the time length requirements of six behaviors in the dataset. The other parameter settings on the WISDM dataset are shown in Table 2.

TABLE 3. Experimental parameter settings on the UCI HAR dataset.

Parameter	Experimental parameter setting
Dataset partitioning	Training set 60%, validation set 10%, test set 30%.
Loss function	Cross entropy
Optimizer	Adam (β_1=0.9, β_2 =0.999)
Learning rate	0.001
Batch size	32
Epoch	100

TABLE 4. Experimental parameter settings on the OPPORTUNITY dataset.

Parameter	Experimental parameters setting
Dataset partitioning	Training set 60%, validation set 20%, test set 20%.
Loss function	Cross entropy
Optimizer	Adam (β_1=0.9, β_2 =0.999)
Learning rate	0.001
Batch size	16
Epoch	100

The comparative experiments are conducted on the WISDM dataset. Seven models are used for the comparative

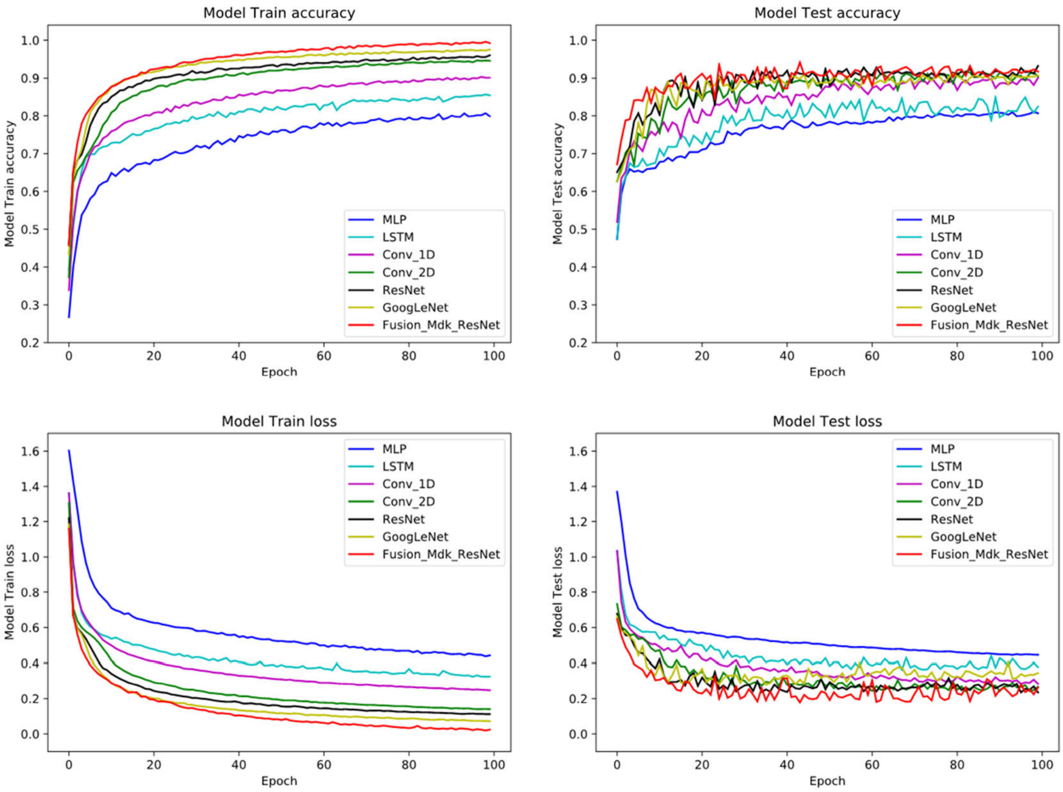


FIGURE 11. Accuracies and losses of training and testing on the UCI HAR dataset.

TABLE 5. Evaluation indexes of each model on the WISDM dataset.

Method	Accuracy (%)	Macro_Precision (%)	Macro_Recall (%)	Macro_F_1 (%)
MLP	86.95	82.00	80.62	81.15
Conv_1D	93.66	90.62	91.73	91.13
LSTM	87.53	83.26	80.97	81.79
GAF + Conv_2D	93.23	91.56	90.74	91.05
GAF + ResNet	96.08	93.68	95.10	94.33
GAF + GoogLeNet	94.27	90.96	92.83	91.60
GAF + Mdk-ResNet	**96.83**	**95.21**	**96.46**	**95.79**

experiments, which are MLP, Conv_1D, LSTM, Conv_2D, ResNet, GoogLeNet, and Mdk-ResNet. Fig. 10 shows the model training accuracy, model test accuracy, the value of model training loss function, and the value of model test loss function of different models, respectively. It can be seen that the proposed model is slightly improved in the accuracy and the convergence speed compared with the other models. Table 5 lists the evaluation indexes on the test set after each model is trained 100 times. Since all categories of activities to be identified are included, the precision, recall, and F-measure calculated by the micro average are the same as the accuracy, they will not be listed separately, and the accuracy is only listed here. As can be seen from the Table 5, the indexes of the Mdk-ResNet have been improved compared to the other methods. The accuracy of GAF + Mdk-ResNet is 96.83%, which is 9.88% higher than that of MLP and 2.56% higher than that of GAF + GoogLeNet. The accuracy of GAF + Mdk-ResNet is not much higher than GAF + ResNet, just 0.75%, but the other indexes are higher.

2) EXPERIMENTAL RESULTS AND ANALYSES ON THE UCI HAR DATASET

The sampling rate of the UCI HAR dataset is 50Hz, the sliding window size is 128, the window overlap is 50%, and the number of samples is 10929. The original dataset has been

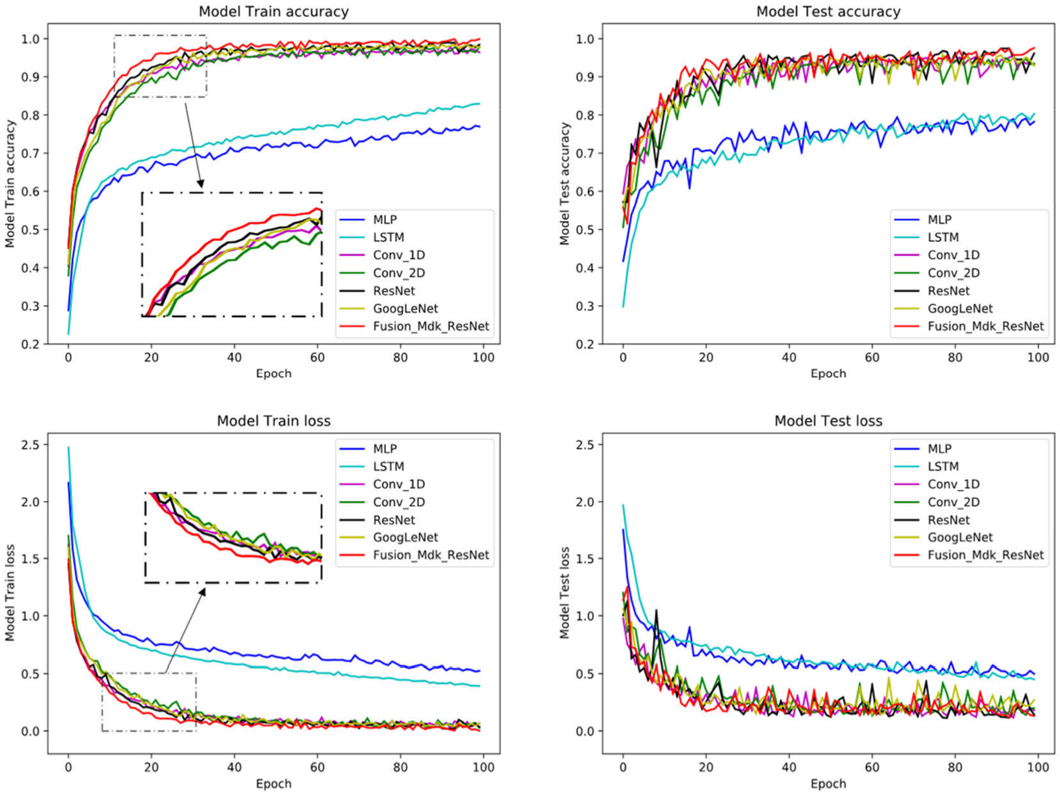


FIGURE 12. Accuracies and losses of training and testing on the OPPORTUNITY dataset.

TABLE 6. Evaluation indexes of each model on the UCI HAR dataset.

Method	Accuracy (%)	Macro_Precision (%)	Macro_Recall (%)	Macro_F$_1$ (%)
MLP	80.79	81.50	81.48	81.38
Conv_1D	85.41	86.07	85.89	85.90
LSTM	80.90	82.04	81.43	81.53
GAF + Conv_2D	88.19	88.49	88.35	88.29
GAF + ResNet	87.75	88.31	87.95	87.85
GAF + GoogLeNet	87.61	87.89	87.96	87.78
GAF + Fusion-Mdk-ResNet	**89.48**	**89.83**	**89.81**	**89.63**

randomly divided, 60% as training data, 10% as validation set and 30% as test data. The other parameter settings on the UCI HAR dataset are shown in Table 3.

The UCI HAR dataset contains data from two types of sensors, accelerometer and gyroscope. The UCI HAR dataset are classified and identified by using the Fusion-Mdk-ResNet, which is proposed in this paper for multi-source sensor of HAR. The other comparative experimental models include MLP, Conv_1D, LSTM, Conv_2D, ResNet, and GoogLeNet. Fig. 11 shows the model training accuracy, model test accuracy, the value of model training loss function, and the value of model test loss function of different models, respectively. It can be seen that the train accuracy and test accuracy of the proposed Fusion-Mdk-ResNet are higher than that of the other methods. At the same time, Table 6 lists the evaluation indexes on the test set after each model is trained 100 times. It can be seen that the results obtained by the Fusion-Mdk-ResNet are superior to that of the other methods. The accuracy, Macro_Precision, Macro_Recall and Macro_F1 of GAF + ResNet and GAF + GoogLeNet are ranges from 87.61% to 88.31%, while the indexes of GAF + Fusion-Mdk-ResNet are all higher, which are more than 89%.

TABLE 7. Evaluation indexes of each model on the OPPORTUNITY dataset.

Method	Accuracy (%)	Macro_Precision (%)	Macro_Recall (%)	Macro_F_1 (%)
MLP	80.31	75.59	72.41	70.74
Conv_1D	95.97	94.31	94.07	94.11
LSTM	80.74	74.02	73.19	72.84
GAF + Conv_2D	96.27	94.95	94.34	94.58
GAF + ResNet	96.87	96.02	95.49	95.73
GAF + GoogLeNet	95.03	93.11	93.37	93.04
GAF + Fusion-Mdk-ResNet	**97.27**	**96.09**	**96.07**	**96.04**

3) EXPERIMENTAL RESULTS AND ANALYSES ON THE OPPORTUNITY DATASET

The sampling rate of the OPPORTUNITY dataset is 30Hz, and the length of the sliding window selected is 90. One window contains data of 3 seconds, and the other parameters are shown in Table 4. Considering that there are more types of sensor data in the OPPORTUNITY dataset, which is slightly different from the previous experiments, the batch size of this experiment is changed from 32 to 16.

The OPPORTUNITY dataset contains data collected by multiple sensors, which are classified by using the Fusion-Mdk-ResNet. The models for comparison include MLP, Conv_1D, LSTM, Conv_2D, ResNet, and GoogLeNet. The accuracy curves and the loss function curves of different methods on the train set and the test set are shown in Fig. 12. It can be seen from Fig. 12 that although the proposed Fusion-Mdk-ResNet is not much different from the other methods in the accuracy, the convergence speed of the proposed method is faster than the other ones. At the same time, the evaluation indexes on the test set after each model is trained 100 times are listed in Table 7. As can be seen from Table 7, the results obtained by the Fusion-Mdk-ResNet are also superior to that of the other methods. The accuracy of GAF + Fusion-Mdk-ResNet is 2.24% higher than that of GAF + GoogLeNet and the other indexes are also higher.

V. CONCLUSION

In this paper, we propose two improved deep CNN models for sensor-based HAR and use the GAF algorithm to process time series. Firstly, the one-dimensional time series collected by the sensor is converted into two-dimensional images by the GAF algorithm, and then the improved deep CNN is adopted to identify the types of human activities. The Mdk-ResNet extracts the features of sampling points with different time intervals to form a more representative feature map. At the same time, a multi-sensor data fusion network Fusion-Mdk-ResNet is proposed, which can process data collected by different sensors and fuse data automatically. The proposed methods are validated on three public activity datasets, and the comparative experiments with seven different models are conducted under the same experimental conditions on each dataset. The experimental results show that the proposed methods are superior to the other methods in accuracy and convergence speed. The accuracy of GAF + Mdk-ResNet is 96.83% on the WISDM dataset, which is 9.88% higher than that of MLP and 2.56% higher than that of GAF + GoogLeNet. The indexes of GAF + Fusion-Mdk-ResNet are all more than 89% and 96% on the UCI HAR dataset and the OPPORTUNITY dataset, respectively, which are higher than the other methods.

The research in this paper is currently carried out on public activity datasets, and the follow-up plan is to perform real-time HAR by using wearable devices such as smart bracelets. The research directions of future research include model compression and miniaturization, so that the model can be transplanted to wearable devices and becomes more practical. In addition, reduce the computation, save hardware resources and increase equipment stand-by time are also the directions of future research.

REFERENCES

[1] P. Vepakomma, D. De, S. K. Das, and S. Bhansali, "A-wristocracy: Deep learning on wrist-worn sensing for recognition of user complex activities," in *Proc. IEEE 12th Int. Conf. Wearable Implant. Body Sensor Netw. (BSN)*, Cambridge, MA, USA, Jun. 2015, pp. 1–6.

[2] J. Qin, L. Liu, Z. Zhang, Y. Wang, and L. Shao, "Compressive sequential learning for action similarity labeling," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 756–769, Feb. 2016.

[3] L. Wang, Y. Sun, Q. Li, T. Liu, and J. Yi, "Two shank-mounted IMUs-based gait analysis and classification for neurological disease patients," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1970–1976, Apr. 2020.

[4] C. Tan, Y. Sun, G. Li, and G. Jiang, "Research on gesture recognition of smart data fusion features in the IoT," *Neural. Comput. Appl.*, vol. 5, no. 3, pp. 1–13, Jan. 2019.

[5] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[6] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016.

[7] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie, and J. Zhou, "TSE-CNN: A two-stage End-to-End CNN for human activity recognition," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 292–299, Jan. 2020.

[8] K. Kim, A. Jalal, and M. Mahmood, "Vision-based human activity recognition system using depth silhouettes: A smart home system for monitoring the residents," *J. Electr. Eng. Technol.*, vol. 14, no. 6, pp. 2567–2573, Sep. 2019.

[9] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.

[10] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatiotemporal VLAD for video action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2799–2812, Jun. 2019.

[11] L. Zhou, L. Lei, and L. Yang, "Human behavior recognition system based on multi-sensor," *Transducer Microsyst. Technol.*, vol. 35, no. 3, pp. 89–91, 95, May 2016.

[12] B. Chen, Q. Yu, and T. Chen, "On deep-learning-model-based sensor activity recognition," *J. Zhejiang Univ. Technol.*, vol. 46, no. 4, pp. 375–381, Aug. 2018.

[13] X. Kuang, J. He, Z. Hu, and Y. Zhou, "Comparison of deep feature learning methods for human activity recognition," *Appl. Res. Comput.*, vol. 35, no. 9, pp. 2815–2817, 2822, Sep. 2018.

[14] S. Deng, B. Wang, C. Yang, and G. Wang, "Convolutional neural networks for human activity recognition using multi-location wearable sensors," *J. Softw.*, vol. 30, no. 3, pp. 718–737, Mar. 2019.

[15] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, "A deep learning approach to on-node sensor data analytics for mobile or wearable devices," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 56–64, Jan. 2017.

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

[17] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "InnoHAR: A deep neural network for complex human activity recognition," *IEEE Access*, vol. 7, pp. 9893–9902, Jan. 2019.

[18] M. Z. Uddin, M. M. Hassan, A. Alsanad, and C. Savaglio, "A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare," *Inf. Fusion*, vol. 55, pp. 105–115, Mar. 2020.

[19] L. Fan, Z. Wang, and H. Wang, "Human activity recognition model based on decision tree," in *Proc. 1st Int. Conf. Adv. Cloud Big Data (CBD)*, Nanjing, China, Dec. 2013, pp. 64–68.

[20] J. Pärkkä, L. Cluitmans, and M. Ermes, "Personalization algorithm for real-time activity recognition using PDA, wireless motion bands, and binary decision tree," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 5, pp. 1211–1215, Sep. 2010.

[21] Z. Feng, L. Mo, and M. Li, "A random forest-based ensemble method for activity recognition," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Milan, Italy, Aug. 2015, pp. 5074–5077.

[22] M. T. Uddin and M. A. Uddiny, "A guided random forest based feature selection approach for activity recognition," in *Proc. Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, Savar, Bangladesh, May 2015, pp. 1–6.

[23] Z. Chen, Q. Zhu, Y. C. Soh, and L. Zhang, "Robust human activity recognition using smartphone sensors via CT-PCA and online SVM," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3070–3080, Dec. 2017.

[24] M. B. Abidine, L. Fergani, B. Fergani, and M. Oussalah, "The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition," *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 19–138, Aug. 2016.

[25] L. Wang and R. Liu, "Human activity recognition based on wearable sensor using hierarchical deep LSTM networks," *Circuits, Syst. Signal Process.*, vol. 39, no. 11, pp. 1–20, Apr. 2019.

[26] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang, "Deep residual bidir-LSTM for human activity recognition using wearable sensors," *Math. Problems Eng.*, vol. 2018, pp. 1–13, Dec. 2018.

[27] H. Cho and S. Yoon, "Divide and conquer-based 1D CNN human activity recognition using test data sharpening," *Sensors*, vol. 18, no. 4, p. 1055, Apr. 2018.

[28] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th Int. Conf. Mob. Comput., Appl. Serv. (MobiCASE)*, Austin, TX, USA, Nov. 2014, pp. 197–205.

[29] S. Chen, W. Wei, B. He, S. Chen, and J. Liu, "Action recognition based on improved deep convolution neural network," *Appl. Res. Comput.*, vol. 36, no. 3, pp. 945–949, 953, Mar. 2019.

[30] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2015, pp. 1307–1310.

[31] X. Zheng, M. Wang, and J. Ordieres-Meré, "Comparison of data preprocessing approaches for applying deep learning to human activity recognition in the context of industry 4.0," *Sensors*, vol. 18, no. 7, p. 2146, Jul. 2018.

[32] C. Ito, X. Cao, M. Shuzo, and E. Maeda, "Application of CNN for human activity recognition with FFT spectrogram of acceleration and gyro sensors," in *Proc. ACM Int. Jt. Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Singapore, Oct. 2018, pp. 1503–1510.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[34] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, Buenos Aires, Argentina, Jul. 2015, pp. 3939–3945.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 34th IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, Las Vegas, NV, USA, Jun/Jul. 2016, pp. 770–778.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[37] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[39] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newslett.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.

[40] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. 21st Eur. Symposium Artif. Neural Netw., Comput. Intell. Mach. Learn. (ESANN)*, Bruges, Belgium, Apr. 2013, pp. 437–442.

[41] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Bayati, M. Creatura, and J. R. Millàn, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Netw. Sens. Syst. (INSS)*, Kassel, Germany, Jun. 2010, pp. 233–240.

HONGJI XU (Member, IEEE) received the B.E. degree in electronics engineering from the Shandong University of Technology, Jinan, China, in 1999, and the M.S. degree in signal and information processing and the Ph.D. degree in communication and information systems from Shandong University (SDU), Jinan, in 2001 and 2005, respectively. From 2004 to 2005, he was a Visiting Ph.D. Candidate with the Telecommunications Technological Center of Catalonia (CTTC) and the Department of Signal Theory and Communication, Polytechnic University of Catalonia (UPC), Barcelona, Spain, and did research in the areas of wireless communications and signal processing. From 2010 to 2015, he was a Postdoctoral Researcher with Tsinghua University—Inspur Group Postdoctoral Scientific Research Station, China, and focused on research in multimedia information processing for smart homes and cloud computing. From December 2014 to December 2015, he was a Visiting Scholar with the Department of Cognitive Science, University of California, San Diego (UCSD), USA, and did research in ubiquitous computing and human–computer interaction. From January 2018 to April 2018, he was a Visiting Scholar with the Virginia Polytechnic Institute and State University (Virginia Tech), USA, and did research in the interdisciplinary fields related to information science and computer science. He is currently an Associate Professor with the School of Information Science and Engineering, SDU. His research interests include wireless communications, ubiquitous computing, blind signal processing, human–computer interaction, and artificial intelligence.

JUAN LI received the B.E. degree in electronic information engineering from Shandong University (SDU), Jinan, China, in 2018. She is currently pursuing the M.S. degree with the School of Information Science and Engineering, SDU, Qingdao, China. Her research interests include human activity recognition, data fusion, and ubiquitous computing.

HUI YUAN (Senior Member, IEEE) received the B.E. and Ph.D. degrees in telecommunication engineering from Xidian University, Xi'an, China, in 2006 and 2011, respectively. He was a Lecturer and an Associate Professor at Shandong University (SDU), Jinan, China, from 2011 to 2014 and from 2015 to 2016, respectively. He worked as a Postdoctoral Fellow (Granted by the Hong Kong Scholar Project) and a Research Fellow of the Department of Computer Science, City University of Hong Kong (CityU), from 2013 to 2014 and from 2017 to 2018, respectively. He has been a Full Professor with the SDU since 2016. His current research interests include video/image/immersive media processing, compression, adaptive streaming, and computer vision.

QIANG LIU received the B.E. degree in electronic information engineering from the Harbin University of Science and Technology, Harbin, China, in 2019. He is currently pursuing the M.S. degree with the School of Information Science and Engineering, Shandong University, Qingdao, China. His research interests include ubiquitous computing, data fusion, and human activity recognition.

SHIDI FAN received the B.E. degree in electronic information engineering from Shandong University (SDU), Jinan, China, in 2018. She is currently pursuing the M.S. degree with the School of Information Science and Engineering, SDU, Qingdao, China. Her research interests include artificial intelligence, ubiquitous computing, context awareness, data fusion, and the quality of context.

TIANKUO LI received the B.E. degree in light chemical engineering from Northeast Forestry University, Harbin, China, in 2015. He is currently pursuing the M.S. degree with the School of Information Science and Engineering, Shandong University, Qingdao, China. From April 2014 to March 2015, he was a Visiting Undergraduate Student with the Kitami Institute of Technology, Hokkaido, Japan. His research interests include smart home systems, deep learning, and natural language processing.

XIAOJIE SUN received the B.E. degree in communication engineering from Shandong Normal University, Jinan, China, in 2019. She is currently pursuing the M.S. degree with the School of Information Science and Engineering, Shandong University, Qingdao, China. Her research interests include machine learning, artificial intelligence, activity recognition, and context-aware computing.