Multimodal Body Sensor for Recognizing the Human Activity Using DMOA Based FS with DL

M. Rudra Kumar[1], A. Likhitha[2(✉)], A. Komali[1,2], D. Keerthana[1,2], and G. Gowthami[1,2]

[1] Department of CSE, G. Pullaiah College of Engineering and Technology, Kurnool, India
[2] G. Pullaiah College of Engineering and Technology, Kurnool, India
ambalalikhitha11@gmail.com

Abstract. The relevance of automated recognition of human behaviors or actions stems from the breadth of its potential uses, which includes, but is not limited to, surveillance, robots, and personal health monitoring. Several computer vision-based approaches for identifying human activity in RGB and depth camera footage have emerged in recent years. Techniques including space-time trajectories, motion indoctrination, key pose extraction, tenancy patterns in 3D space, motion maps in depth, and skeleton joints are all part of the mix. These camera-based methods can only be used inside a constrained area and are vulnerable to changes in lighting and clutter in the backdrop. Although wearable inertial sensors offer a potential answer to these issues, they are not without drawbacks, including a reliance on the user's knowledge of their precise location and orientation. Several sensing modalities are being used for reliable human action detection due to the complimentary nature of the data acquired from the sensors. This research therefore introduces a two-tiered hierarchical approach to activity recognition by employing a variety of wearable sensors. Dwarf mongoose optimization process is used to extract the handmade features and pick the best features (DMOA). It predicts the composite's behavior by emulating how DMO searches for food. The DMO hive is divided into an alpha group, scouts, and babysitters. Every community has a different strategy to corner the food supply. In this study, we tested out a number of different methods for video categorization and action identification, including ConvLSTM, LRCN and C3D. The projected human action recognition (HAR) framework is evaluated using the UTD-MHAD dataset, which is a multimodal collection of 27 different human activities that is available to the public. The suggested feature selection model for HAR is trained and tested using a variety of classifiers. It has been shown experimentally that the suggested technique outperforms in terms of recognition accuracy.

Keywords: Human action recognition · Dwarf mongoose optimization algorithm · Camera-based approaches · Key poses extraction · Convolutional Neural Network

S. Kadry and R. Prasath (Eds.): MIKE 2023, LNAI 13924, pp. 1–12, 2023.
https://doi.org/10.1007/978-3-031-44084-7_1

1 Introduction

Ubiquitous sensing, which uses data collected by sensors placed strategically about a building, has become increasingly popular in recent years [1]. Wearable sensor research for (HAR) has exploded in recent years, thanks in large part to its widespread potential use in fields as diverse as sports, interactive gaming, healthcare, and other monitoring schemes. Multimodal HAR is best understood as a variation on the long-honoured series classification problem [2], wherein a sliding time window is used to partition incoming sensor data into discrete time intervals from which discriminative features may be extracted. Techniques, such as may be used to further distinguish each time frame [3]. In addition, it can be challenging for shallow learning to capture the essential elements of complicated actions, and feature selection is often a laborious process [5]. Studies into automatic feature extraction with little human effort are of paramount importance as a means of addressing the aforementioned issues. The field of multimodal HAR is shifting its focus from surface learning to deep learning at the moment [6].

To improve system performance and do away with the requirement for hand-crafted features, recent research in sensor-based HAR has focused heavily on deep learning, in which many layers are layered to build (DNNs) [7, 8]. In particular, the rich representation power of (CNNs) has substantially advanced the performance of HAR. DNNs will improve in performance as their model capacities for rich representation grow, but this will unavoidably increase the need for highly labelled data. Annotated or "ground truth labelled" training data is a source of difficulty for deep HAR identification [9, 10]. Annotating the ground truth requires the annotator to sift through raw sensor data and physically identify all activity instances. This is a time-consuming and costly process. As compared to data captured by other sensor modalities, such cameras, the time series data recorded by multimodal embedded sensors like accelerometers and gyroscopes is far more challenging to comprehend [11]. To effectively segment and classify a specific activity from a lengthy needs significant human work. Thus, while these DNN models can automatically extract relevant features for categorization, they still need precise truth, which would necessitate significantly more human work to provide an ideal training dataset for HAR in a supervised learning situation [12].

Due to its independence on the kind, distance, and arithmetical scale of distinct features derived from numerous sensory modalities [13], decision-level fusion has been the primary focus of existing research for multimodal HAR. Furthermore, the final for classification has fewer dimensions after decision-level fusion, and no post-processing of the retrieved features is required. Independent and stand-alone categorization choices pertaining to each sense modality, which are subsequently fused using some soft rule to generate the final conclusion, is the main shortcoming of the decision-level fusion [14]. On the other hand, feature-level fusion is useful for gathering features simultaneously from several sensors and integrating them to produce enough information for a sound judgement [15].

2 Related Works

Using machine and deep learning (DL) models, Pradhan and Srivastava [16] categorized multi-modal physiological inputs. Dahou et al. [17] provide a methodology to increase the performance of several applications using a wide variety of data kinds by addressing the large dimensionality of data transported via the SIoT system.

Islam et al. [18] recommend a fusion procedure for activity recognition using a multi-head (CNN) equipped with a Convolution Block Attention Module (CBAM) to process the visual data and a (ConvLSTM) to handle the time-sensitive multi-source sensor information. In order to evaluate and recover channel and spatial dimension attributes, the three CNN sub-architectures and CBAM for visual data are implemented.

Novel system architecture presented by Zhang et al. [19] consists of three parts: feature selection using an oppositional and chaos particle swarm optimization (OCPSO) algorithm, a multi-input (MI-1D-CNN) that takes advantage of signals, and deep decision fusion (DDF) that combines D-S evidence theory and entropy. Using the UCI HAR and WIDSM datasets, the suggested architecture is tested.

Using the combination of EEG and face video clips, Muhammad et al. [20] describe a multimodal emotion identification approach based on deep canonical correlation analysis (DCCA). We use a two-stage framework in which the first stage uses features extracted from a single modality to recognize emotions, and the second stage combines the highly correlated features from the two modalities and classifies the data. After fusing highly correlated data using a DCCA-based method, the SoftMax classifier was then used to categorize faces into one of three fundamental human emotion categories: joyful, neutral, or sad. The suggested method was explored using the MAHNOB-HCI and DEAP public datasets. The average accuracy of the experimental findings was 93.86% on the MAHNOB-HCI dataset and 91.54% on the DEAP dataset. By contrasting the proposed framework with other efforts, we were able to assess its competitiveness and provide justification for its exclusivity in the pursuit of this level of precision.

Human gait identification was the focus of Jahangir et al. [21].'s novel two-stream deep learning approach. In the first stage, we discussed a method for improving contrast by combining data from local and global filters. In the second stage, data augmentation is carried out to expand the dimensionality of the raw dataset (CASIA-B). Third, we use deep transfer learning using the supplemented dataset. Fourth, a serial-based method is utilized to combine the extracted features of the two streams; and fifth, an enhanced method is employed to further optimize the fusion. Eight different angles from the CASIA-B dataset were used in the experimentation procedure, with results of 97.3, 98.6, 97.7, 96.5, 92.9, 93.7, 94.7, and 91.2% accuracy. The results of head-to-head comparisons with SOTA methods revealed increased precision and decreased processing time.

3 Proposed System

We begin with a brief description of the experimental dataset, followed by a discussion of the methodology and metrics utilized in the experiments. We then detail how our suggested framework may be put into action. We conclude with a discussion of the qualitative results, which should give you some good ideas about the recommended approach.

3.1 Dataset and Implementation Details

The suggested technique was tested using the UTD-MHAD dataset, which is a publicly available multimodal HAR dataset consisting of 27 human activities performed by 8 people. Figure 1 shows a list of these activities along with several visual representations.

Fig. 1. Sample Images of the dataset [22]

Each participant performed each task four times. As a result, we have 8 participants $\times$ 4 trials per action $\times$ 27 actions totalling 864 trimmed data sequences. Three data sequences were corrupted during data recording, therefore after cleaning up the dataset, only 861 sequences remained. Both a Microsoft Kinect sensor (30 frames and a wearable inertial sensor (50 samples per second) were used to acquire the information in a controlled indoor environment. In order to record triaxial acceleration triaxial angular velocity, a Bluetooth-enabled hardware module was employed as a wearable inertial sensor (using a gyroscope). During activities 1–21, the participant wore the sensor on their right wrist; for actions 22–27, the sensor was attached to their right thigh. Each segmented action trial in the dataset is represented by four files, one for each of the four sensory modalities included in the dataset.

3.2 Feature Extraction

In particular, we made use of handmade features; the following sections outline each method in depth.

3.2.1 Handcrafted Features

Techniques for extracting features by hand are easy to implement and need less computing power. Simple statistical procedures more intricate frequency domain-based features, can be used to compute them on time series data. Table 1 summarizes the calculated characteristics and is followed by a detailed explanation of each. Each dimension of features received its own set of statistical calculations.

a) Extreme: Let X is the feature course. The $Max(X)$ function finds and revenues the largest feature value $x_i \in X$.

Table 1. List of Handcrafted Topographies

Skewness	Extreme
Norm of SOM	Percentile 50
Spectral energy	Percentile 80
Kurtosis	Minimum
Auto-correlation	Average
First-Order Mean (FOM)	Standard-deviation
Norm of FOM	Zero crossing
Second-order mean (SOM)	Percentile 20
Spectral entropy	Interquartile

b) Least: With X as input, the Min(X) function will locate the minimum story value (xi X) and return it.
c) Average: When there are N possible tale values, the average earnings are equal to the value in the middle of feature vector X. As in,

$$Average(X) = \mu = \frac{\sum_{i=1}^{n} x_i}{N} \tag{1}$$

d) Standard Deviation: It defines the amount of difference in feature vector $X = \{x1, x2 \ldots xN\}$ and can be calculated using the following preparation:

$$Stdev(X) = \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \tag{2}$$

e) The frequency with which the signal value passes zero in either direction is an indicator of how quickly or slowly the activity is changing.
f) The term "percentile" is used to describe a score where a specified fraction of all possible responses fall below that value. The pth percentile is defined as the number at which no more than (100 p) % of the capacities are lower than this value and no more than 100(1 p) % are higher than this value. The 25th percentile, for instance, indicates that the value is larger than 25 other values but lower than 75 other feature values.
g) To calculate the interquartile range, use the difference among the first and third quartiles.
h) The skewness of a distribution is a measure of how far off centre the data of relative to the mean:

$$Sk = \frac{1}{N\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^3 \tag{3}$$

i) Kurtosis: To what extent the distribution's tails deviate from the normal distribution's tails is measured by the statistic. A larger kurtosis number indicates that there are

more extreme deviations, or outliers, in the data. It may be calculated mathematically as:

$$Kr = \frac{1}{N\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^4 \tag{4}$$

j) Auto-correlation: It is a statistical method for determining how closely one set of the time series data is related to its own lagged version across a range of time periods and it may be calculated as:

$$r_k = \frac{\sum_{i=1}^{N-k} (x_i - \mu)(x_{i+k} - \mu)}{\sum_{i=1}^{N} (x_i - \mu)^2} \tag{5}$$

k) Order Mean Values: They are derived from the sorted list of numbers (in ascending order). That is, in an ordered collection of features X, the first ordered mean is just the smallest sample value X1, the second ordered mean is the value X2, and so on.

l) Norm Values: They help determine how far away from zero a feature vector actually is. There were two metrics we used: L1-norm.

m) Spectral Energy: We remember that in the recorded data numerous sensors are utilized to access the human actions; these sensors may be thought of as a function whose amplitude varies with time. The signal was changed from a time series to a frequency range using the Fourier transform, and the Spectral energy formulation was used to determine the energy levels at each frequency. The z value is the total amplitude squared of the frequencies present (n). As in,

$$S_E = \sum_{i=1}^{N} F(n)^2 \tag{6}$$

n) It is also calculable using normalized frequency spectra. As in,

$$\hat{F}(n) = \frac{F(n)}{\sum_{i=1}^{N} F(n)} \tag{7}$$

o) The normalized form of Eq. (6) is as follows:

$$NS_E = \sum_{i=1}^{N} \hat{F}(n)^2 \tag{8}$$

p) Spectral Entropy: It is a way to quantify the spectral distribution of a signal in terms of frequency, and it is entropy. One possible mathematical description of spectral entropy is as follows:

$$S_{EN} = -\sum_{i=1}^{N} \hat{F}(n) \times log\hat{F}(n) \tag{9}$$

These 18 features are the product of computations performed on each column in the features set for a single action and are joined together in a single row. Given that the input data is 61-dimensional, the resulting handmade feature will have a dimension of 1 (18 61) (or 1 1098). We tested the recognition accuracy of these generated features using a DL model.

3.3 Classification Using DL Models

3.3.1 Long-term Recurrent Convolutional Network (LRCN)

The goal of LRCN [23, 24] is to use convolution neural networks to extract spatial data from each frame. In order to categorise the data, the results from the convolutional networks are fed into a Bi-LSTM network, which combines the retrieved spatial characteristics with the temporal features. These models require an input size of 90 by 90 pixels. Convolutional filters are modelled as a matrix (in our example, 3×3 in size) with a random set of values that convolve across the picture and calculate the dot operation, and the output is then sent on to the next layer in the custom CNN model. Using convolution over k channels, the following Eqs. (10)–(13) summarise an input frame and provide a matrix as a result.

$$A_o^{(m)} = g_m(w_{ok}^{(m)} * A_k^{(m-1)} + b_o^{(m)}) \tag{10}$$

$$W - ok * A_k[s, t] = a_{p,q} * b \tag{11}$$

$$a = A_k\big[s + p, t + q\big] \tag{12}$$

$$b = w_{ok}[P - 1 - p, Q - 1 - q] \tag{13}$$

As per the above equations, max pooling is used to minimise the number of parameters after each convolutional layer in the network that lightens the convolutional burden. The rate of output is stable at this point. Our model used Rectified Linear Unit (ReLU), as seen in Eq. (16), and SoftMax, which converts a system's output into a probability distribution across projected classes. We imported an ImageNet-trained VGG-16 network and deleted the top layer to use its features in the VGG-LSTM model. Time-distributed layering was followed by a 256-filter bidirectional lstm, 256-filter ReLU-activated dense layering, and a final 2-neuron output layer.

$$y = A.x + b \tag{14}$$

$$y_i = \sum_{j=1}^{i} \big(A_{ij}, x_j\big) + b_i \tag{15}$$

$$y = \max(0, x) \tag{16}$$

$$y = A.x + b \tag{17}$$

$$y_i = \sum_{j=1}^{i} \big(A_{ij} x_j\big) + b_i \tag{18}$$

3.3.2 Convolutional Long Short-Term Memory (CLSTM) [25, 26]

While LSTM is limited to the temporal domain, we have also utilised ConvLSTM, which can be applied to the spatial domain. To do this, we employ ConvLSTM with spatially-oriented tensor inputs, cell outputs, hidden states, and gates. While both ConvLSTM and LSTM have a similar architecture, the two models diverge in how they handle transitions from input to state and from state to state. The activation function, convolution operator, and Hadamard product are respectively denoted by the symbols ““, “,” and “0” in the following Eqs. (19)–(24).

$$i_t = \sigma\left(w_{xi}x_t + w_{hiht-1} + w_{ci}{}^{\circ}c_{t-1} + b_i\right) \tag{19}$$

$$f_t = \sigma\left(w_{xf}x_t + w_{hfh}t - 1 + w_{cf}{}^{\circ}c_{t-1} + b_f\right) \tag{20}$$

$$\tilde{c}_t = \tanh(w_{xc}x_t + w_{hc}h_{t-1} + b_c) \tag{21}$$

$$c_t = f_t{}^{\circ}c_{t-1} + i_t{}^{\circ}c_t \tag{22}$$

$$o_t = \sigma(w_{xo}x_t + w_{h0}h_{t-1} + w_{co}\sigma c_t + b_o) \tag{23}$$

$$h_t = o_t \tanh(c_t) \tag{24}$$

The aforementioned formulas allow us to translate between the input value X_t and the output value Ht1 of the last neuron, where Ct1 is the current location. The convolution filter has a kernel size of k by k, where k is the dimension of kernel. In order to extract features from a movie, ConvLSTM reads in frames as input and performs a multidimensional convolution operation on each frame. To extract features more efficiently than the CNN model, ConvLSTM may transport and process input in both the inter-layer and the intro-layer.

3.3.3 3D Convolutional Neural Networks (C3D)

In contrast to 2D-CNNs, C3D [27–29] can extract both temporal and spatial data from videos. This is due to the fact that 2D convolution applied to a video section compresses the temporal features after convolving, leading to an overall feature map that fails to accurately reflect any motion. A 3D filter kernel is created by stacking many frames together to create the 3D cube needed for the 3D convolution. Frames × Height × Width × Channels in the following format: 25 × 90 × 90 × 3. A ReLU activation function follows the 64 filters in the first 3D convolutional layer. The next step is a max pooling, which takes the most notable features from each feature map patch and calculates their maximum value.

4 Results and Discussion

Evaluation metrics including false positive rate (FPR), error rate (ER), accuracy (AUC), true positive rate (TPR), and precision (P) are used to make predictions about HAR detection (see Tables 2 and 3).

$$TPR = \frac{True\ Positive}{False\ Negative\ +\ True\ Positive} \tag{25}$$

$$FPR = \frac{False\ Positive}{True\ Negative\ +\ False\ Positive} \tag{26}$$

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \tag{27}$$

$$Error_{Rate} = \frac{False\ Positive\ +\ False\ Negative}{False\ Negative} \tag{28}$$

Table 2. Analysis of Various DL Classifiers without DMOA

Algorithms	TPR (%)	FPR (%)	Accuracy (%)	Error Rate
MLP	83.0	9.6	89.2	0.30
DBN	89.7	8.1	90.5	0.23
C3D	93.8	5.3	92.6	0.15
CLSTM	95.6	4.5	94.3	0.18
LRCN	96.8	2.5	96.4	0.16

In the analysis of TPR, the three proposed models achieved nearly 93% to 96%, DBN achieved 89.7% and MLP achieved 83%. When the models are tested with FPR, DBN and MLP achieved 9.6% and 8.1%, where the three models of proposed approach achieved 2.5% to 5.3%. The error rate is very low in C3D, CLSTM and LRCN, where DBN and MLP has high error rate. i.e., 0.30 and 0.23 (see Figs. 2 and 3).

Table 3. Analysis of Various DL Classifiers with DMOA

Algorithms	TPR (%)	FPR (%)	Accuracy (%)	Error Rate
MLP	89.7	8.9	91.2	0.28
DBN	91.8	7.2	92.1	0.20
C3D	94.4	4.6	95.2	0.13
CLSTM	96.1	3.5	96.4	0.15
LRCN	98.6	1.4	98.4	0.14

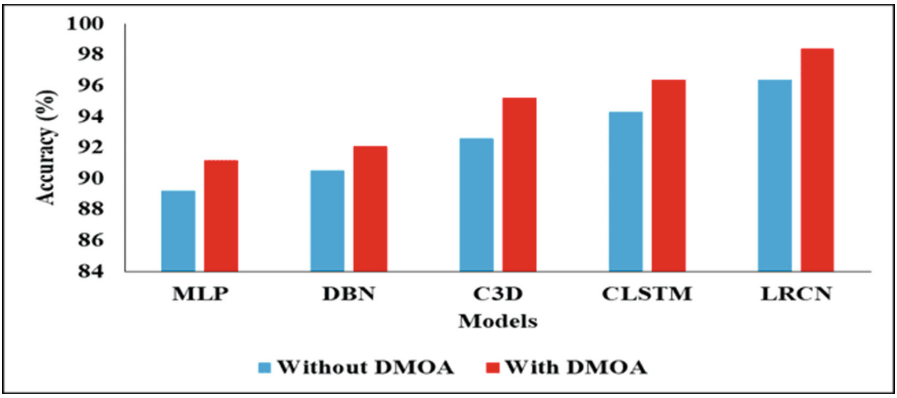

Fig. 2. Accuracy Validation

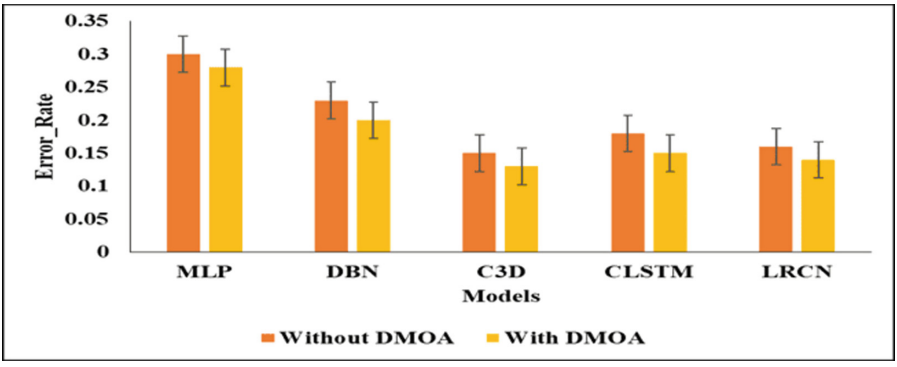

Fig. 3. Error_Rate Presentation

5 Conclusion

Dwarf Mongoose Optimization is a suggested optimization-based Feature selection method in this study for human action recognition. In order to detect an action, the proposed system combines the data derived from several sense modalities utilising a supervised trifecta of deep learning methods. The extensive experimental findings validate the validity of our proposed strategy for human action classification in comparison to standalone sensor modalities. Furthermore, as compared to state-of-the-art deep CNN approaches, the system's recognition accuracy is enhanced while computational cost is decreased by fusing time domain information calculated from inertial sensors with those from depth/RGB movies. In addition, it does not utilize Multi-view HAR, and the subject whose actions are being identified maintains their current orientation with relation to the camera. Further work will involve expanding the suggested HAR technique to compensate for these deficiencies. Moreover, we hope to explore the many uses for the suggested fusion architecture by utilizing an RGB-D camera and a set of wearable inertial sensors.

References

1. Yadav, S.K., Tiwari, K., Pandey, H.M., Akbar, S.A.: A review of multimodal human activity recognition with special emphasis on classification, applications, challenges, and future directions. Knowl.-Based Syst. **223**, 106970 (2021)
2. Zhao, H., Miao, X., Liu, R., Fortin, G.: Multi-sensor information fusion based on machine learning for real applications in human activity recognition: state-of-the-art and research challenges. Inf. Fusion **80**, 241–265 (2022)
3. Ferrari, A., Mocci, D., Mobile, M., Napolitano, P.: Trends in human activity recognition using smartphones. J. Reliable Intell. Environ. **7**(3), 189–213 (2021)
4. Islam, M.M., Iqbal, T.: Multi-gat: a graphical attention-based hierarchical multimodal representation learning approach for human activity recognition. IEEE Robot. Autom. Lett. **6**(2), 1729–1736 (2021)
5. Rani, S., Babar, H., Coleman, S., Singh, A., Allandale, H.M.: An efficient and lightweight deep learning model for human activity recognition using smartphones. Sensors **21**(11), 3845 (2021)
6. Khan, I.U., Afzal, S., Lee, J.W.: Human activity recognition via hybrid deep learning based model. Sensors **22**(1), 323 (2022)
7. Challa, S.K., Kumar, A., Samwell, V.B.: A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data. Vis. Comput. **38**(12), 4095–4109 (2022)
8. Xiao, Z., Xu, X., Xing, H., Song, F., Wang, X., Zhao, B.: A federated learning system with enhanced feature extraction for human activity recognition. Knowl.-Based Syst. **229**, 107338 (2021)
9. Zhang, S., et al.: Deep learning in human activity recognition with wearable sensors: a review on advances. Sensors **22**(4), 1476 (2022)
10. Ramanujan, E., Perumal, T., Padmavathi, S.: Human activity recognition with smartphone and wearable sensors using deep learning techniques: a review. IEEE Sens. J. **21**(12), 13029–13040 (2021)
11. Wang, D., Yang, J., Cui, W., Xie, L., Sun, S.: Multimodal CSI-based human activity recognition using GANs. IEEE Internet Things J. **8**(24), 17345–17355 (2021)

12. Hamad, R.A., Kimura, M., Yang, L., Woo, W.L., Wei, B.: Dilated causal convolution with multi-head self-attention for sensor human activity recognition. Neural Comput. Appl. **33**, 13705–13722 (2021)
13. Gu, F., Chung, M.H., Chignell, M., Valaee, S., Zhou, B., Liu, X.: A survey on deep learning for human activity recognition. ACM Comput. Surv. (CSUR) **54**(8), 1–34 (2021)
14. Garcia, K.D., et al.: An ensemble of autonomous auto-encoders for human activity recognition. Neurocomputing **439**, 271–280 (2021)
15. Tasnim, N., Islam, M.K., Baek, J.H.: Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints. Appl. Sci. **11**(6), 2675 (2021)
16. Pradhan, A., Srivastava, S.: Hierarchical extreme puzzle learning machine-based emotion recognition using multimodal physiological signals. Biomed. Signal Process. Control **83**, 104624 (2023)
17. Dahou, A., Chelloug, S.A., Alduailij, M., Elaziz, M.A.: Improved feature selection based on chaos game optimization for social internet of things with a novel deep learning model. Mathematics **11**(4), 1032 (2023)
18. Islam, M.M., Nooruddin, S., Karray, F., Muhammad, G.: Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things. Inf. Fusion **94**, 17–31 (2023)
19. Zhang, Y., Yao, X., Fei, Q., Chen, Z.: Smartphone sensors-based human activity recognition using feature selection and deep decision fusion. IET Cyber-Phys. Syst.: Theory Appl. **8**, 76–90 (2023)
20. Muhammad, F., Hussain, M., Aboalsamh, H.: A bimodal emotion recognition approach through the fusion of electroencephalography and facial sequences. Diagnostics **13**(5), 977 (2023)