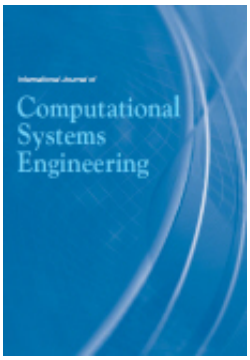


International Journal of Computational Systems Engineering

ISSN online: 2046-3405 - ISSN print: 2046-3391
https://www.inderscience.com/ijcsyse

A study on the application of 3DHOG-assisted technology in physical education movement recognition

Yu He, Na Chen

DOI: 10.1504/IJCSYSE.2025.10060558

Article History:

Received:	10 April 2023
Last revised:	07 June 2023
Accepted:	11 September 2023
Published online:	23 April 2025

A study on the application of 3DHOG-assisted technology in physical education movement recognition

Yu He

Department of Physical Education and Research,
Fuzhou University,
Fuzhou, 350108, China
Email: He036626@126.com

Na Chen*

Ministry of Sports,
Xiamen Institute of Technology,
Xiamen, 361021, China
Email: Na025Chen@outlook.com
*Corresponding author

Abstract: An image feature extraction technique based on histogram of oriented gradients (HOG) technology is proposed as a method for human body detection, while 3D convolutional neural networks (3D CNN) technology is combined as a key technology for action recognition. And the two are combined to generate 3DHOG assistive technology which is applied to the physical education video parsing. The results show that the false recognition rate of 3D CNN model in the training set is stable around 0.03, corresponding to a loss of 0.05. The average accuracy of each action of 3D HOG model is 96.25%, while the recall rate of the model is 81.2%. Its mean absolute error (MAE) value is 1.18% and root mean squared error (RMSE) value is 0.91%. The 3D HOG model has superior performance and has good application value for action detection and recognition of physical education videos.

Keywords: action recognition; HOG; human detection; physical education; 3D CNN.

Reference to this paper should be made as follows: He, Y. and Chen, N. (2025) 'A study on the application of 3DHOG-assisted technology in physical education movement recognition', *Int. J. Computational Systems Engineering*, Vol. 9, No. 8, pp.1–11.

Biographical notes: Yu He obtained his Bachelor's in Sports Training from Fujian Normal University in July 2010. In July 2012, he obtained his Master's in Physical Education from Fujian Normal University. Currently, he is a Lecturer at the Physical Education Teaching and Research Department of Fuzhou University, with a main research direction in physical education. His work experience: from 2013 to present, he has worked in the Sports Teaching and Research Department of Fuzhou University, as well as the Sports Training Teaching and Research Group and Deputy Team Leader.

Na Chen graduated from Jimei University with a bachelor's degree in Physical Education in 2010. In 2015, she graduated from Hubei University with a Master's in Physical Education and Sports Training. Since 2010, she works in Xiamen Institute of Technology as a physical education teacher. She is an editorial board member of the book *College Sports and Health*. In 2021, he published *Computer Software: Aerobics Music Production Software V1.0*; *Aesthetic Characteristics and Connotation of Cheerleading* was published in Neijiang Technology in 2019; *Investigation and Analysis of Aesthetic Penetration in Cheerleading Teaching in Xiamen Universities* published in 2022 and *A Preliminary study on Legal Liability of Sports Injuries in Uuniversities in Xiamen* was published in 2015.

This paper was originally accepted for a special issue on 'AI and Cognitive Computing for Next Generation Mobile Networks?' guest edited by Dr. Arvind Dhaka, Dr. Amita Nandal, Dr Edmar Candeia Gurjao and Dr. Dijana Capeska Bogatinoska.

1 Introduction

In the context of COVID.19, video teaching has become one of the main teaching methods in physical education, where teachers can record teaching videos and put them online for students to download and study (McDonough et al., 2022; Hawani 2021). The key problem that how to detect and identify the movements should be solved in the video teaching of physical education. The human action recognition technology needs to implement human body detection first. Then the action of the detection target is recognised and the next judgment and prediction are made (Qin and Liu, 2022; Singh et al., 2022). Human body detection technology and action recognition technology have been studied more deeply in the fields of human-computer interaction and unmanned vehicles. As a part of machine learning, deep learning (DL) has better applications and research in human detection and action recognition (HR-AR). In the current stage of DL research in HR-AR, feature extraction and classification can be achieved using neural networks and classifiers (Nolasco et al., 2022). In different environments, DL can be used as the characteristics extract method of human motion, so the detection and recognition of human motion can be realised. Then the classification model can be used to classify the extracted features, so as to improve the accuracy of human motion detection and recognition. Researches show that DL can be better applied to the detection and recognition of human motion in various fields (Stephen et al., 2022). HOG technology and 3D CNN technology have been well applied in the field of motion recognition. 3D CNN technology is an improvement based on 2D CNN technology, which has higher recognition accuracy than 2D CNN technology. Although these technologies can achieve better feature extraction and feature classification in various fields, there is less research material on techniques based on human detection and action recognition (HR-AR) in physical education videos. In this experiment, it is proposed to establish 3DHOG technology based on HOG technology and 3D CNN technology as an auxiliary technology for motion recognition in sports teaching videos. And it compares improved method and basic method to prove improved method's performance. In the experiment, 2D CNN and 3D CNN were compared, as well as the methods in the literature. The model's misidentification rate, loss, accuracy, recall, mean absolute error (MAE), root mean square error (RMSE) and other indicators were compared. It is hoped to improve the teaching efficiency of physical education teachers in daily teaching, while helping students identify and fill in gaps in daily learning, and engage in better physical education knowledge learning.

2 Review of the literature

Human action recognition has a wide range of practical applications, and it needs to be built on the basis of human body detection. In the research of sports action recognition, it is necessary to detect human action first, and then conduct action recognition to predict and analyse the next action. HOG technique is the main method of feature extraction in image processing and has been well studied in the fields of power detection, video analysis, etc. (Bai et al., 2022). Image feature extraction technology based on HOG is one of the main methods to realise human body detection in video. Hadjadji et al. (2022), through the improvement of HOG technology, can generate static features and dynamic features when detecting human falls as supplementary information for human fall detection. And the method has good performance in human fall detection. The HOG technique not only enables the extraction of static and dynamic features, but also feature identification and accurate classification based on micro features in the classification of pulmonary infection of COVID.19 (Dixit et al., 2023). This has implications for the improvement of human detection methods. It suggests that, when performing feature extraction, researchers can integrate microscopic features with hand-crafted features to improve the accuracy of human detection. Automated human detection system can make adjustments according to the changes of factors such as lighting conditions in real scenes. On the basis of HOG, the human detection system built by Konwar et al. (2021) can eliminate the effect of noise caused by image background in different environments. And it can perform human detection better. In addition, HOG technology has good robustness for detecting indoor objects affected by obstructions. In the experiment, HOG was used to extract features from images and establish image blocks. At the same time, SVMs were used in the experiment to classify image features and image blocks, which can effectively improve the accuracy of the detection model (Lee et al., 2021a).

Human action analysis based on visual information is a key direction of research in the field of computing. Vision is an important component to achieve human action recognition. Researchers extract features by processing RGB-D information and they build 3D human skeletal action recognition networks to improve performance of human action recognition models (Barkoky and Charkari, 2022). Temporal and spatial based skeleton recognition is a key direction of research (Kong et al., 2022). Through temporal measurements, Koch et al. (2022) detected skeletal data. They used search algorithms to combine time measurement data information with action recognition information. It can be used to measure action recognition in human-machine collaboration. CNN complexes have better recognition effects in human action recognition. But pure CNN complexes have limited recognition effects. So researchers improved and optimised the underlying neural networks. Long and short term memory networks can remember time-dependent sequences. By introducing short-term and short-term memory networks into CNN, the problem of its inability to effectively extract information containing temporal features in real-world scenarios can be improved (Senthilkumar et al., 2022). The introduction of spatio-temporal maps in graph convolutional networks can effectively recognise skeleton-based human actions. And

some scholars propose to use temporal dilation and spatial attention to expand graph convolutional networks. By this way, it can extract features in different time and space, reduce the influence of noise, etc. And the robustness of human action recognition models can be improved effectively (Zhang et al., 2022). Image recognition and feature extraction can be used for human action recognition in video. On this basis, incorporating temporal and spatial information for extracting human motion features can effectively improve the accuracy of motion recognition. Jiang and Zhang (2022) can obtain long-term and short-term temporal features by analysing the temporal dimension. Also considering the fine-grained spatial dimension, the representation of actions in videos can be enhanced.

From the above study, HOG technology in human detection can perform better image feature extraction, which lays a good foundation for image classification and action recognition. At the same time, the two-dimensional space action recognition technology has certain inapplicability at this stage. So it can improve the accuracy and effectiveness of human action recognition in realistic environment. And the effect of adding time dimension and space dimension on action recognition can be considered in the research. Meanwhile, CNN and image feature techniques are common methods and models for human body detection and action recognition. So in this experiment, we propose to establish 3DHOG technology based on HOG technology and 3D CNN technology as an auxiliary technology for action recognition in physical education teaching videos. It is hoped to improve the teaching efficiency of physical education teachers in daily teaching. At the same time, it is hoped to help students to check the gaps and make up for them in daily learning for better physical knowledge learning.

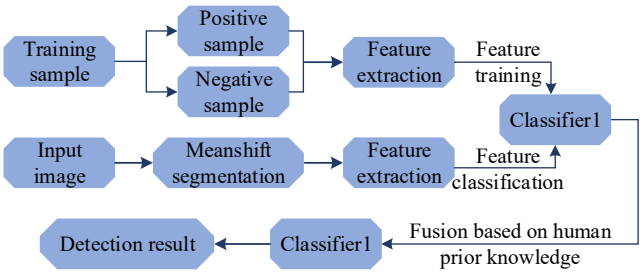
3 Research on sports teaching action recognition technology based on 3DHOG

3.1 Research on human detection techniques for physical education based on HOG and SVM

In the research of sports teaching action recognition technology based on 3DHOG auxiliary technology, it is necessary to use corresponding mathematical calculations for the research of technical methods. In Section 3, relevant methods and techniques are elaborated in detail, and corresponding calculation processes are listed for establishing the foundation of action recognition technology.

This chapter is based on existing HOG and SVM technologies. Due to the susceptibility of these technologies to environmental factors during application, the accuracy of detection is reduced (Sharma et al., 2022; Boudjit and Ramzan, 2022). Therefore, improvements have been made to these technologies in this chapter. In the traditional human detection technique of HOG and SVM, the whole technical process contains two steps. The extraction of HOG features from the image is the first step, which are used in the classifier for training. The second step is to extract the HOG features from the predicted image and input these features into a support vector machine (SVM) which has been pre-trained to do the classification process. Accuracy and effectiveness of human detection is a prerequisite for the application of human detection technology. And a single human detection method can lead to poor human detection in practice due to the influence of external conditions such as ambient light. A combination of multiple methods is often used in human detection technology. And prior knowledge and practical task characteristics are integrated to improve the performance of human detection technology. In Figure 1, HOG image features are extracted and coarsely scored. Then a priori knowledge aggregation is used to accurately identify the images to achieve accurate detection of the human body.

Figure 1 Human detector based on segmentation and clustering (see online version for colours)



Among them, the image segmentation method chooses Mean Shift algorithm, the feature extraction method chooses HOG, and the classifier chooses SVM. HOG technique is often used in computer vision technology for the description of local texture features of the image. It can be used to obtain the histogram of the local region by calculating the gradient information of the local region, which is the feature description of the region (Al-Obaidi et al., 2020). The feature descriptions of each region divided by image segmentation are concatenated to obtain the overall feature description of the target image. In the feature calculation process of HOG, the target image is first divided into regions, and each divided region becomes a cell. Then the gradient vector of each cell pixel point is calculated and the histogram is divided. Where the division is based on the gradient direction and the size of the weight value is decided according to the gradient size. Finally, the histogram obtained from the division is normalised, and the unit of normalisation is one Block. The specific image feature extraction steps of HOG are as follows. Firstly, the size of the target image is unified as 64*128, the pixel of size 8*8 is taken as a cell, and the cell of size 2*2 is taken as a block. In the cell grid, a rectangular HOG is used to all the blocks overlapped. Next, the colour image of size 8*8 is colour space normalised according to a certain ratio, see equation (1).

$$Gray = 0.3R + 0.59G + 0.11B \tag{1}$$

R, G, B denote 3 channels of red, green and blue in the image in equation (1). $Gray$ denotes the normalised

processed grayscale value. Also to reduce the effect of illumination on the blocks, the pixel point i of each cell is corrected for processing and the gamma correction in equation (2) is selected.

$$Y(x, y) = I(x, y)^{\gamma} \tag{2}$$

In equation (2), γ= 0.5, which denotes the correction factor. $I(x, y)$ denotes the size of the pixel value of the pixel point (x, y), and $Y(x, y)$ denotes the corrected value. After normalisation and correction, the direction and size of the pixel gradient for a cell of size 8*8 are calculated. The gradient in the horizontal direction $G_x(x, y)$ can be obtained by convolving the original image using the gradient operator [–1, 0, 1], see equation (3).

$$G_x(x, y) = I(x+1, y) - I(x-1, y) \tag{3}$$

The gradient in the horizontal vertical direction $G_y(x, y)$ can be obtained by convolving the original image using the gradient operator $[-1, 0, 1]^T$, see equation (4).

$$G_y(x, y) = I(x, y+1) - I(x, y-1) \tag{4}$$

Then the gradient size in equation (5) can be obtained.

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{5}$$

The direction of the gradient in equation (6) can also be obtained.

$$\theta(x, y) = \arctan \frac{G_y(x, y)}{G_x(x, y)} \tag{6}$$

The gradient histograms are then obtained according to the gradient direction and the weight size. In order to reduce the influence of the light factor on the histogram, it is also necessary to normalise it. And the normalisation method used in this experiment is L2-norm, which is calculated in equation (7).

$$v = \frac{v}{\|v\|_2^2 + \varepsilon^2} \tag{7}$$

v denotes the feature vector before normalisation process in equation (7). $\|v\|$ is the k-parameter, which usually takes the value of 1 or 2. ε denotes a very small value, which is used to avoid the case of value 0. Finally, all the data after normalisation process are concatenated to form a large feature vector as image's HOG feature vector. SVM can be used for various recognition scenario classification of human detection techniques in the study. For linearly separable classes, SVM can find a plane to separate them while maximising the distance between other different classes and that plane. For a sample set $\{x, y\}$ containing samples A and B, there exists a hyperplane $w^T x + b = 0$. Where x_i denotes the set of feature vectors, y_i denotes the numerical labels of $\{x_i, y_i\}$, and $y_i \in \{1, -1\}$. This hyperplane needs to satisfy $y_i(w^T x_i + b) - 1 \geq 0$, where b is a constant and w is a weight vector. Then the classification interval M is denoted by $M = 2 / \|w\|^2$, and the minimum value of the formula $\varphi(w) = 2 / \|w\|^2$ can be found if the hyperplane formula $w^T x + b = 0$ is satisfied, and this result is the optimal classification surface. $\varphi(w)$ is defined as a Lagrange function whose expression is given in equation (8).

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left\{ y_i \left(w^T x + b \right) - 1 \right\} \tag{8}$$

α_i denotes the Lagrange coefficient in equation (8). The minimum of the Lagrange function is calculated by solving for the partial derivative of the weight vector w and the constant b. And the value of this partial derivative is taken to be zero, see equation (9).

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \dfrac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \Rightarrow \alpha_i \left\{ y_i \left(w^T x_i + b \right) - 1 \right\} = 0 \end{cases} \tag{9}$$

If equations (8) and (9) and $\alpha_i \geq 0$ are satisfied, the maximum value α_i is calculated in equation (10).

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \left(x_i^T x_j \right) \tag{10}$$

Assuming that the optimal solution in equation (10) is calculated as α_i^*, then equation (11) and equation (12) can be obtained.

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \tag{11}$$

$$b^* = \frac{1}{2}\left[w^* x^*(1) + w^* x^*(-1) \right] \tag{12}$$

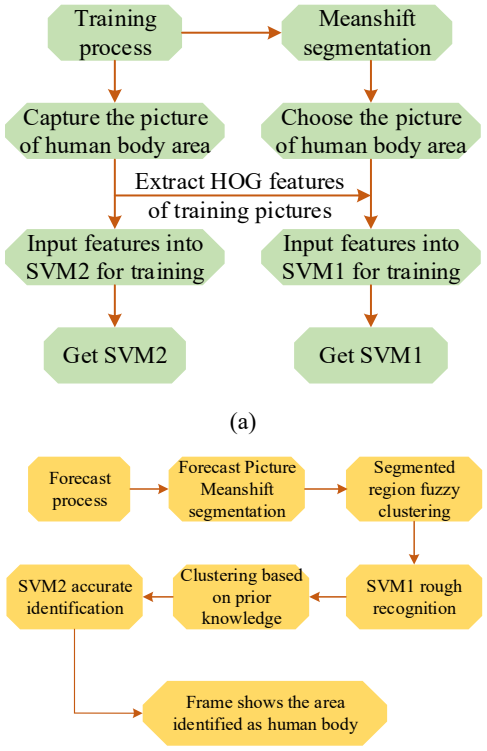
$x^*(1)$ and $x^*(-1)$ denote the support vectors that are not used by the category in equation (12). From the above derivation, the optimal classification function in equation (13) can be obtained.

$$f(x) = \operatorname{sgn}(w^* x + b^*) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i^n y_i x_i^* x + b^* \right) \tag{13}$$

sgn is a symbolic function in equation (13). In human detection technique, HOG technique can ensure the integrity of image boundary extraction while preserving the contour information better. And SVM technique is simple and easy to use and can model the data for classification. The combination of the two has better results for human detection. So the HOG-SVM algorithm is chosen for feature extraction and classification in this study. However, general human detection algorithms are prone to incomplete detection or missed detection. To improve the accuracy of the detection algorithm, the experiments are taken to first segment the image. And the HOG features of the suspected

human body are extracted. Then the feature data are inputted into the SVM1 model for coarse classification judgment. Next the secondary clustering is performed based on the human body prior knowledge, and the data is input to the SVM2 model for fine judgments. Figure 2 represents the training process and prediction process of this proposed method.

Figure 2 Training process and prediction process of the model (a) training process (b) forecast process (see online version for colours)



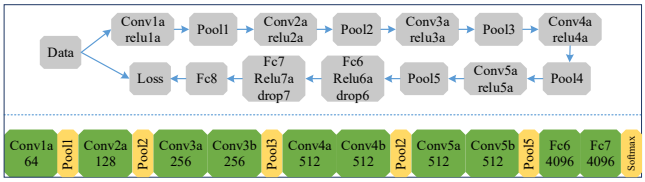
3.2 *Research on physical education action recognition technology based on HOG and 3D CNN technology*

For action recognition in physical education, the human action has a temporal extension, so 3D CNN is born (Zhao et al., 2020). 3D CNN technique can continuously collect the sequence information of images during model training. And by concatenating consecutive frames in other network layers for convolutional processing, feature maps in the convolutional layer are obtained. Both spatial and temporal dimensions are considered to finally obtain the action information in physical education (Banerjee et al., 2020). In the previous feature extraction process, the extraction of HOG features takes more time when the predicted image has a large size, so it is easy to miss or incomplete detection in the detection process (He et al., 2020). Therefore, the technique is optimised and improved by combining the above HOG features with 3D CNN technique in this experiment. 3D CNN technique can achieve feature extraction from both spatial and temporal dimensions. And it can be used for action recognition of video sequences, etc. Distinct from the 2D CNN complex, 3D CNN technique can be analogised to a cube formed by superimposing multiple images. And the cube is used to perform convolution operations and feature extraction. In 3D CNN, temporal dimension and spatial dimension are the factors that need to be considered to obtain motion information. Where the neurons at (x, y, z) in the feature map j on the i-th convolutional layer are calculated as shown in equation (14).

$$v_{ij}^{xyz} = f\left[\sum_{k\in M}\sum_{p\in P_i}\sum_{q\in Q_i}\sum_{r\in R_i} w_{ijk}^{pqr} v_{(i-1)k}^{(x+p)(y+q)(z+r)} + b_{ij}\right] \tag{14}$$

k denotes the number of features, w_{ijk}^{pqr} denotes the weight value at the position of (p, q, r), P_i, Q_i denotes the convolution kernel size, R_i is the time dimension, and b_{ij} is the feature bias in equation (14). Also due to the weight sharing mechanism, the same convolution kernel will get the same features. To increase the representation of features, multiple types of convolutional kernels are generally chosen. In this experiment, the network structure in 3D CNN includes 1 input layer, 5 convolutional layers, 5 pooling layers, 2 fully connected layers and 1 layer of Sofmax layer. And the activation function is chosen as Relu function.

Figure 3 Network structure of 3d CNN (see online version for colours)



For the sampling of action image keyframes in sports instructional videos, first it is assumed that the video contains a total of N frames, each image sequence contains a starting frame, and the number of starting frames is the number of training samples that can be obtained. The interval size of the samples is set to U in the experiment, then the representation of the subscript of the starting frames is shown in equation (15).

$$S = \{1, 1+U, 1+2\times U, ..., 1+(T-1)\times U\} \tag{15}$$

The interval size is $U = N/T$, and T denotes the number of keyframe samples in equation (15). To preserve the sampling information as much as possible while ensuring no information redundancy, the interval of sampling frames is set to R. Sample numbers obtained in interval R is L frames. Then in equation (16), the subscript representation of the sampled frames in sample i can be obtained.

$$C_i = \{S_i, S_i + R, S_i + 2R, ..., S_i + (L-1)R\} \tag{16}$$

S_i denotes the starting frame subscript in equation (16). 3DHOG assisted technology is combined with HOG and 3D CNN technology. The action recognition algorithm based

on 3DHOG assisted technology is taken as the key technology for HR-AR. Where the action recognition method first needs to extract the image features from the image sequence. Then the difference between the features of two adjacent frames is calculated and this feature difference is treated as a class of HOG image features. The images of the same sequence are convolved and pooled in 3D CNN to generate the corresponding another class of 3D CNN image features. The obtained HOG image features and 3D CNN image features are learned in XGBoost, and the final result is the action recognition result of the above image sequence.

Based on the above research methods, further performance verification of these methods is needed. The commonly used datasets in the field of motion recognition include the KTH dataset, UCF101 dataset, and HMDB51 dataset. These datasets mostly contain videos related to sports, which are more relevant to the content of physical education teaching. Therefore, in the validation experiment in Section 4, the KTH dataset, UCF101 dataset, and HMDB51 dataset are selected as the training and prediction datasets for the model, and the performance of the model is tested and compared.

4 Simulation analysis of physical education action recognition technology based on 3DHOG

The increase in demand for human action recognition promotes the application of datasets in different scenarios, and the selection of action recognition datasets can be used to calibrate action recognition methods' performance metrics. The 3D CNN model was placed on UCF101 for training and optimised using the stochastic gradient descent method. With an initial value of 0.002, the learning rate decreased to 1/10 of the original value after proceeding to the 5th iteration, for a total of 14 training sessions. The loss value of the 3D CNN model reaches convergence by the 6th round of training, and its accuracy is highest at the 7th round, eventually stabilising at about 75.1%.

The 3D CNN model is put into XGBoost for training in the experiment. And the change of false recognition rate and loss curve in Fig. 5 can be obtained. With the increase of training times, 3D CNN's false recognition rate in the training set and the test set remains basically the same after about 60 times of training. 3D CNN's false recognition rate model in the training set is stable at about 0.03, and false recognition rate is about 0.21 in test set. The corresponding Loss of the 3D CNN model remains at about 0.05 in training set, and the corresponding Loss stays around 0.35 in test set. The Loss of the 3D CNN model gradually reaches convergence, indicating that the classification effect of the model is real and reliable.

The action recognition accuracy of 2D CNN model and 3D CNN model for KTH dataset, UCF101 dataset and HMDB51 dataset are given in Figure 6. The model's performance is tested in the KTH dataset, UCF101 dataset, and HMDB51 dataset for seven sports: jumping rope, playing football, boxing, skateboarding, weightlifting, playing basketball, and running. The recognition accuracy of each action of the 3D CNN model is higher than that of the 2D CNN model. It indicates that the influence of temporal and spatial dimensions on action recognition is considered in the 3D CNN model in this experiment, and it is better for all actions with continuity in sports.

The confusion matrix, also called the error matrix, is one of the measures that can be used for the classification accuracy of a classification model. To verify the classification effectiveness of the action recognition model in this experiment, the confusion matrix of action recognition is chosen to be used as a metric for performance evaluation. In the confusion matrix, each column of the matrix represents the predicted type of action and each row represents the true type of action. 3D CNN model's prediction accuracy is overall higher than 2D CNN model.

To compare the superiority of the 3D HOG model proposed in this experiment, a comparison of the different methods was performed in the UCF101 dataset. In Fig. 7, the action recognition accuracy of the 3D HOG method in this experiment is compared with the traditional CNN model, 3D CNN, C3D, and HOG+SVM methods (Lee et al., 2021b; Iftikhar S et al., 2022; Qu et al., 2022; Xu et al., 2022). For each action, 3D HOG model's recognition accuracy is higher, with an average accuracy of 96.25%. Traditional CNN model's average accuracy is 93.84%, 3D CNN model's average accuracy is 94.96%, C3D model's average accuracy is 95.68%, and HOG+SVM model's average accuracy is 95.53%.

For the performance evaluation of the different action recognition algorithms mentioned above, the more widely used evaluation metrics were chosen for this experiment, which include Precision (P) and Recall (R). Accuracy represents the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test dataset. The recall rate represents the ratio of the number of correctly classified relationship instances of a certain class to the total number of relationship instances of a certain class in the test set. The higher the values of quasi curvature and recall, the better the classification performance of the model. The differences of the metrics between the 3D HOG model, the traditional CNN model, 3D CNN, C3D, and HOG+SVM methods are compared in Figure 8. 3D HOG model's accuracy is 92.5%, and its recall is 81.2%, which are the highest among all models, indicating the superior performance of the 3D HOG model.

The Receiver Operating Characteristic (ROC) curve is a curve drawn based on a series of different binary classification methods (boundary values or determination thresholds), with true positive rate (sensitivity) as the vertical axis and false positive rate (1-specificity) as the horizontal axis. The ROC curve is a comprehensive evaluation indicator of accuracy and recall, which can more intuitively reflect the performance of the model. The closer the ROC curve is to the upper left corner, the larger the area under the curve, and the higher the accuracy of the experiment. The ROC curves of 3D HOG model, traditional CNN model, 3D CNN, C3D, and HOG+SVM methods are compared in Fig. 9. Fig. 9(a) shows ROC curve' validation

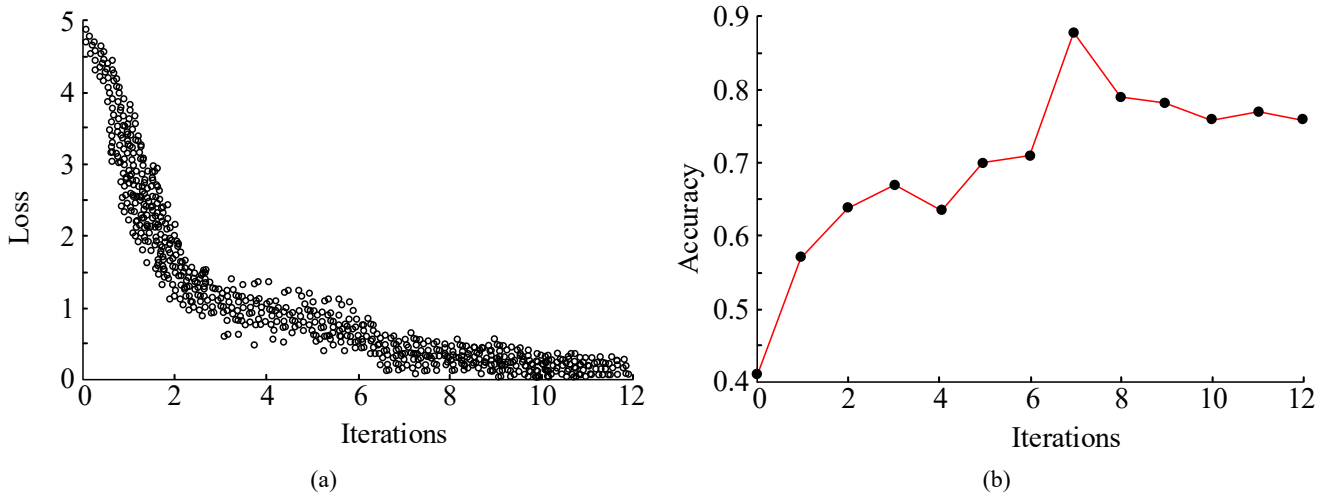
results and Fig. 9(b) shows the ROC curve's detection and identification results. The area under both the validation, detection and identification of the 3D HOG model is higher than that of the traditional CNN model, 3D CNN, C3D, and HOG+SVM methods, indicating that the 3D HOG model proposed in this experiment has better results.

Average absolute error (MAE) and root mean square error (RMSE) indicators are commonly used in the evaluation of machine learning models. RMSE represents the sample standard deviation of the difference (referred to as residual) between predicted and observed values, indicating the degree of dispersion of the sample. MAE is the average of the absolute error between the predicted value and the observed value. The smaller the values of MAE and RMSE, the smaller the difference between the predicted and true values, indicating higher accuracy of the model. The results of MAE and RMSE are shown in Figure 10. The average MAE of the 3D HOG model is 1.18% and average RMSE is 0.91%, lower than the traditional CNN model, 3D CNN, C3D, and HOG+SVM methods, proving that the results are accurate.

Table 1 Action recognition confusion matrix of different algorithms in UCF101 dataset

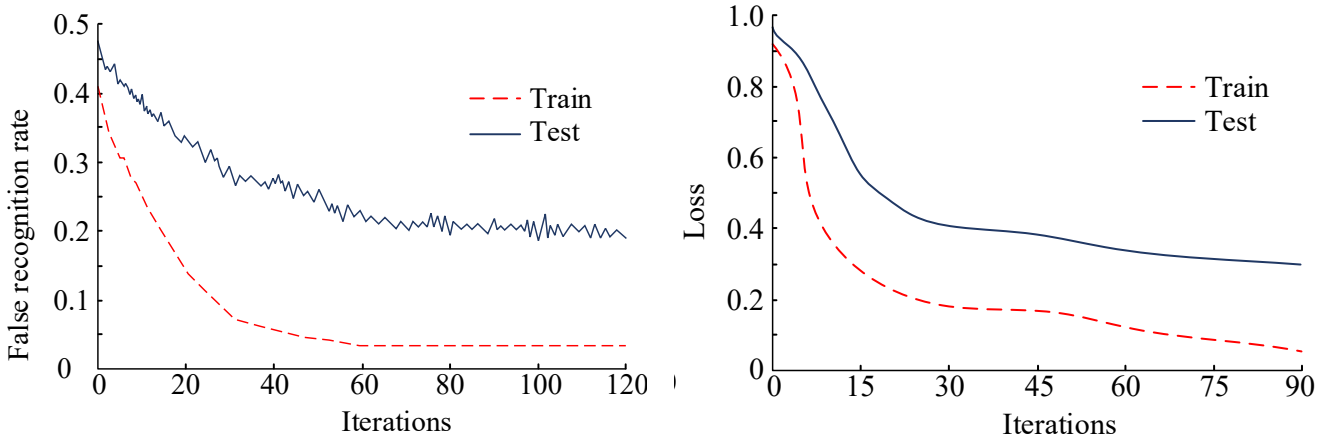
2D CNN							
Real type	*Forecast type*						
	Sports	*Skipping rope*	*Play football*	*Boxing*	*Skate*	*Weightlifting*	*Play basketball*	*Run*
Skipping rope	0.98	0	0	0	0	0	0.02
Play football	0.02	0.96	0	0	0	0	0.02
Boxing	0	0	0.98	0	0.02	0	0
Skate	0	0	0	0.97	0	0.01	0.02
Weightlifting	0	0	0	0	0.98	0.02	0
Play basketball	0	0	0	0	0	0.98	0.02
Run	0	0	0	0.02	0	0	0.98
3D CNN							
Real type	*Forecast type*						
	Sports	*Skipping rope*	*Play football*	*Boxing*	*Skate*	*Weightlifting*	*Play basketball*	*Run*
Skipping rope	1	0	0	0	0	0	0
Play football	0	0.98	0	0	0	0	0.02
Boxing	0	0	1	0	0	0	0
Skate	0	0	0	0.98	0	0	0.02
Weightlifting	0	0	0	0	1	0	0
Play basketball	0	0	0	0	0	0.99	0.01
Run	0	0.01	0	0.01	0	0	0.98

Figure 4 Loss function value and test accuracy results of 3d CNN model (a) loss (b) accuracy (see online version for colours)



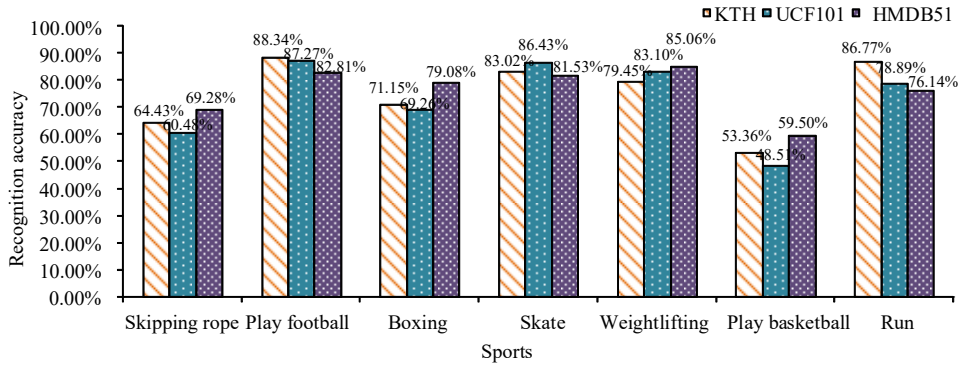
(a) (b)

Figure 5 Change of false recognition rate and loss curve (a) false recognition rate (b) loss (see online version for colours)

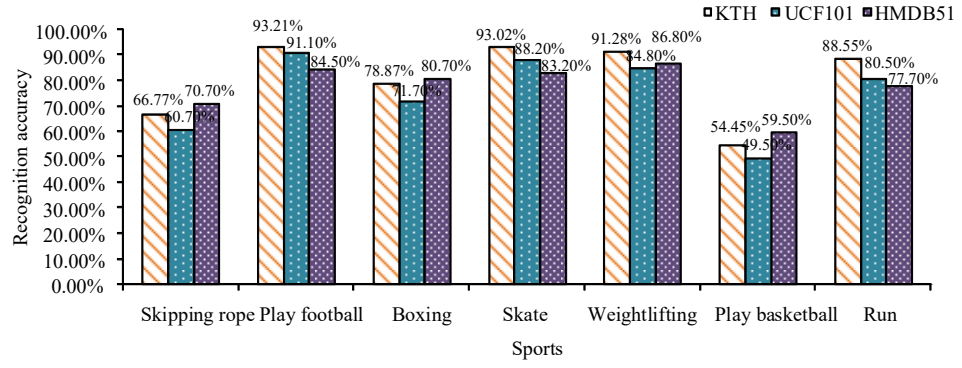


(a) (b)

Figure 6 Action recognition accuracy of 3d CNN in different datasets (a) 2D CNN (b) 3D CNN (see online version for colours)



(a)



(b)

Figure 7 Comparison results of action recognition accuracy of different methods (see online version for colours)

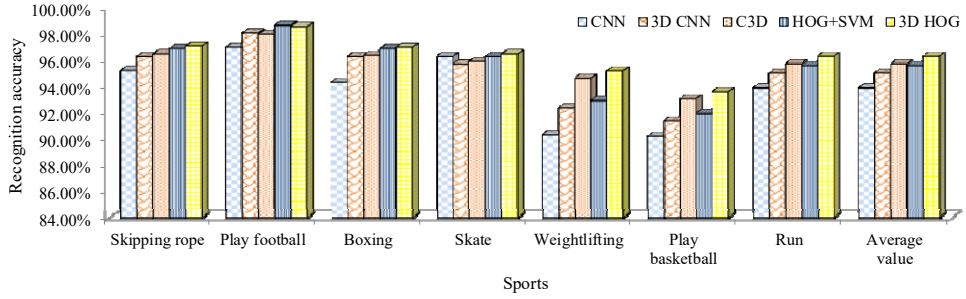
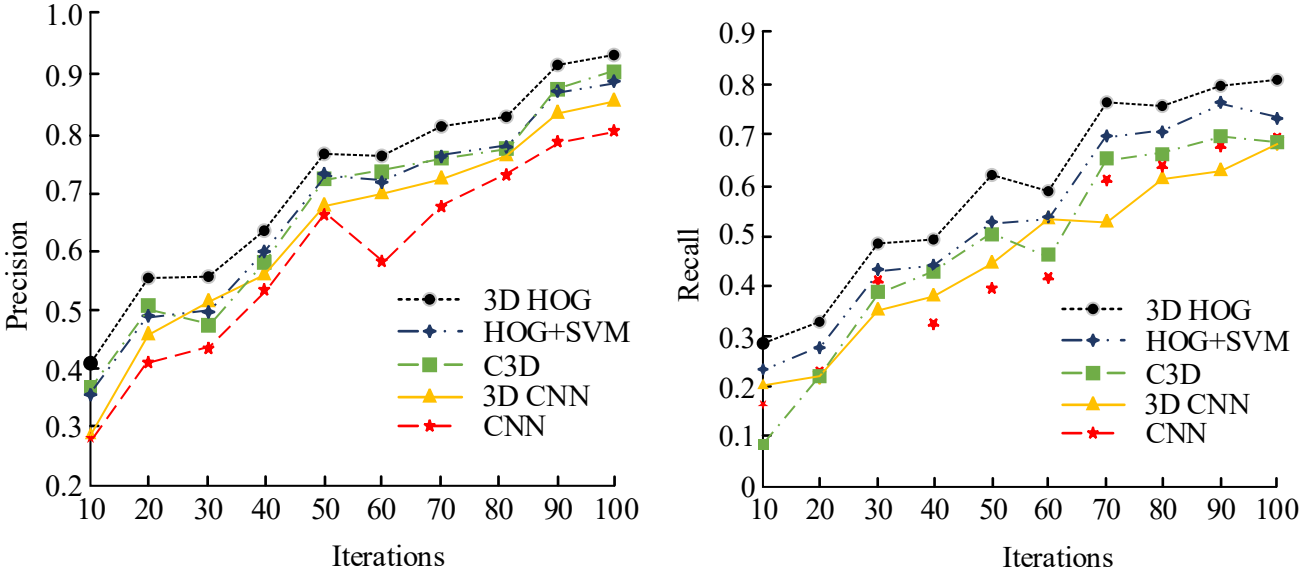
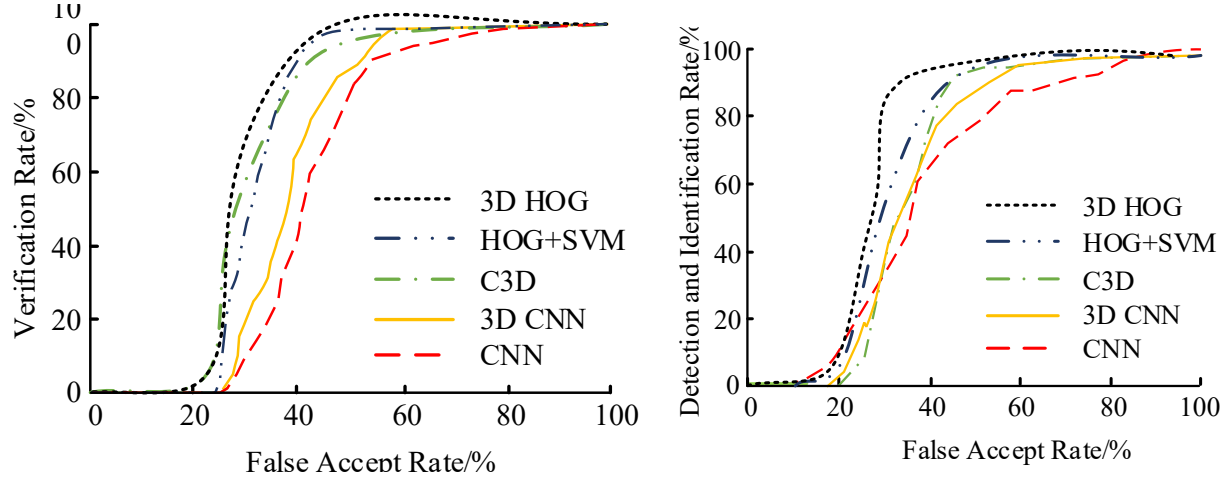


Figure 8 Precision and recall results of 3D hog (a) precision (recall) (see online version for colours)



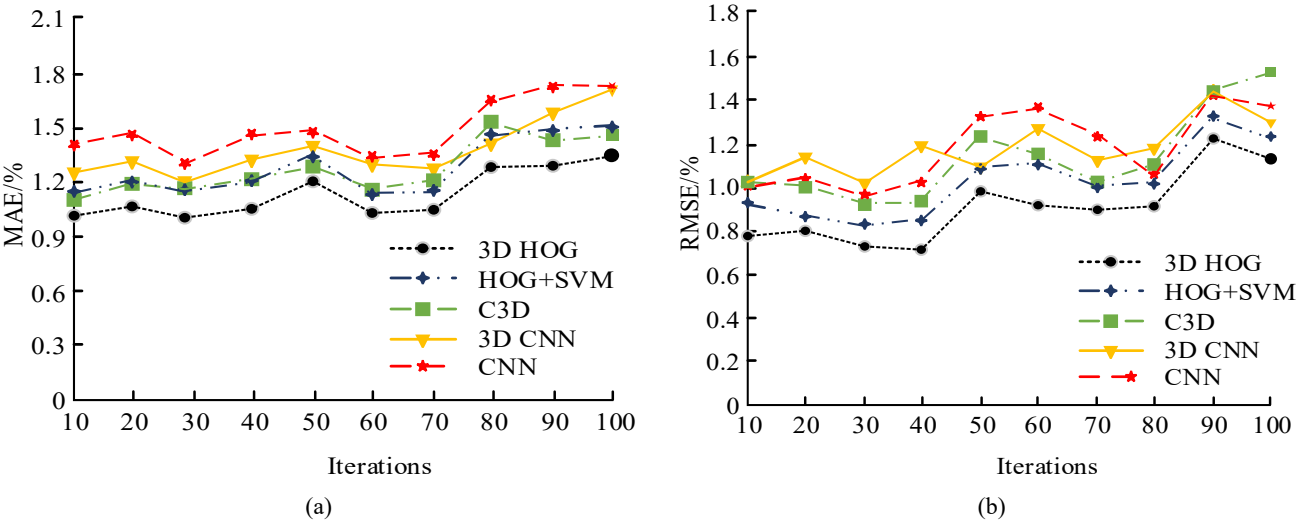
(a) (b)

Figure 9 Roc curve (a) verification (b) detection and identification (see online version for colours)



(a) (b)

Figure 10 Comparison of MAE and RMSE (a) MAE/% (b) RMSE/% (see online version for colours)



(a) (b)

5 Conclusions

In existing research, HOG technology can perform better image feature extraction in human detection, laying a solid foundation for image classification and action recognition (Hadjadji et al., 2022; Dixit et al., 2023). At the same time, the motion recognition technology in two-dimensional space has certain inapplicability at the current stage (Koch et al., 2022; Jiang et al., 2022). In order to improve the accuracy and effectiveness of human action recognition in the real environment, the influence of time and space dimensions on action recognition can be considered in the research. In the parsing of physical education videos, the actions need to be detected and recognised. And the combination of HOG-based technique and 3D CNN technique is proposed for the detection and recognition of physical education actions in this experiment. 3D CNN model's false recognition rate is stable around 0.03 in training set and 3D CNN model's false recognition rate is stable around 0.21 in test set. The corresponding Loss is kept around 0.05 in training set and the corresponding Loss is kept around 0.35 in test set. 3D CNN model takes into account the influence of time and space dimensions on action recognition, which has a good recognition effect on continuous actions in sports. Compared with 2D CNN model, this model's prediction accuracy is higher. Each action's average accuracy rate of 3D HOG model is 96.25%, and its recall rate is 81.2%. 3D HOG model's average MAE value is 1.18%, and its average RMSE value is 0.91%, indicating that 3D HOG model's performance is superior. Compared with existing research (Lee et al., 2021B; Iftikhar et al., 2022; Qu et al., 2022; Xu et al., 2022), the method proposed in this experiment has better accuracy in motion recognition by combining HOG technology and 3D CNN technology. The experimental results confirm the superiority of the improved method. However, in this study, we only discussed the characteristics of actions, and it is needed to judge the start and end time of actions, which needs to be further improved in the future work. Compared to image recognition, the proposed action recognition method in this experiment considers spatiotemporal information. However, the difficulty of action recognition is much greater than that of image recognition. The corresponding action recognition not only needs to consider the characteristics of the action, but also needs to accurately recognise the start and end times of the action, which significantly increases the difficulty of action recognition. Therefore, it is necessary to conduct in-depth research using appropriate datasets.

References

Al-Obaidi, S., Al-Khafaji, H. and Abhayaratne, C. (2020) 'Modeling temporal visual salience for human action recognition enabled visual anonymity preservation', *IEEE Access*, Vol. 8, No. 1, pp.213806–213824.

Bai, K., Zhou, Y., Cui, Z., Bao, W., Zhang, N. and Zhai, Y. (2022) 'HOG – SVM – based image feature classification method for sound recognition of power equipments', *Energies*, Vol. 15, No. 12, pp.4449–4460.

Banerjee, A., Singh, P.K. and Sarkar, R. (2020) 'Fuzzy integral – based CNN classifier fusion for 3D Skeleton action recognition', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 31, No. 6, pp.2206–2216.

Barkoky, A. and Charkari, N.M. (2022) 'Complex network – based features extraction in RGB - D human action recognition', *Journal of Visual Communication and Image Representation*, January, Vol. 82, pp.1–9.

Boudjit, K. and Ramzan, N. (2022) 'Human detection based on deep learning YOLO-v2 for real-time UAV applications', *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 34, No. 3, pp.527–544.

Dixit, U.D., Shirdhonkar, M.S. and Sinha, G.R. (2023) 'Automatic logo detection from document image using HOG features', *Multimedia Tools and Applications*, Vol. 82, No. 1, pp.863–878.

Hadjadji, B., Saumard, M. and Aron, M. (2022) 'Multi – oriented run length based static and dynamic features fused with Choquet fuzzy integral for human fall detection in videos', *Journal of Visual Communication and Image Representation*, January, Vol. 82, pp.1–14.

Hawani A. (2021) 'The Practice of Tunisian physical education teachers at the end of initial training during the preparation for professional life', *The Journal of Quality in Education*, Vol. 11, No. 17, pp.99–130.

He, M., Song, G. and Wei, Z. (2020) 'Human behavior feature representation and recognition based on depth video', *Journal of Web Engineering (JWE)*, Vol. 19, Nos. 5–6, pp.883–902.

Iftikhar, S., Asim, M., Zhang, Z. and El-Latif (2022) 'Advance generalization technique through 3D CNN to overcome the false positives pedestrian in autonomous vehicles', *Telecommunication Systems*, Vol. 80, No. 4, pp.545–557.

Jiang, J. and Zhang, Y. (2022) 'An improved action recognition network with temporal extraction and feature enhancement', *IEEE Access*, Vol. 10, No. 1, pp.13926–13935.

Koch, J., Büsch, L., Gomse, M. and Schüppstuhl, T. (2022) 'A methods – time – measurement based approach to enable action recognition for multi – variant assembly in human – robot collaboration', *Procedia CIRP*, Vol. 106, No. 1, pp.233–238.

Kong, J., Bian, Y. and Jiang, M. (2022) 'MTT: multi – scale temporal transformer for skeleton - based action recognition', *IEEE Signal Processing Letters*, Vol. 29, No. 1, pp.528–532.

Konwar, L., Talukdar, A.K. and Sarma, K.K. (2021) 'Robust real time multiple human detection and tracking for automatic visual surveillance system', *WSEAS Transactions on Signal Processing*, Vol. 17, No. 1, pp.93–98.

Lee, H., Kim, Y.S., Kim, M. and Leeb Y. (2021) 'Low – cost network scheduling of 3D - CNN processing for embedded action recognition', *IEEE Access*, Vol. 9, No. 1, pp.83901– 83912.

Lee, S.J., Kim, B.H. and Min, Y.K. (2021) 'Multi – saliency map and machine learning based human detection for the embedded top – view imaging system', *IEEE Access*, Vol. 9, No. 1, pp.70671–70682.

McDonough, D.J., Helgeson, M.A., Liu, W. and Gao, Z. (2022) 'Effects of a remote, YouTube - delivered exercise intervention on young adults' physical activity, sedentary behavior, and sleep during the COVID - 19 pandemic: randomized controlled trial', *Journal of Sport and Health Science,* Vol. 11, No. 2, pp.145–156.

Nolasco, L., Lazzaretti, A.E. and Mulinari, B.M. (2022) 'DeepDFML-NILM: a new CNN-based architecture for detection, feature extraction and multi-label classification in NILM signals', *IEEE Sensors Journal,* Vol. 22, No. 1, pp.501–509.

Qin, Y. and Liu, B. (2022) 'KDM: a knowledge – guided and data - driven method for few – shot video action recognition', *Neurocomputing*, October, Vol. 510, pp.69–78.

Qu, W., Zhu, T., Liu, J. and Li, J. (2022) 'A time sequence location method of long video violence based on improved C3D network', *The Journal of Supercomputing*, Vol. 78, No. 18, pp.19545–19565.

Senthilkumar, N., Manimegalai, M., Karpakam, S., Ashokkumar, S.R and Premkumar, M. (2022) 'Human action recognition based on spatial–temporal relational model and LSTM - CNN framework', *Materials Today: Proceedings*, Vol. 57, No. 5, pp.2087–2091.

Sharma, S., Raja, L., Bhatnagar, V., Sharma, D., Bhagirath, S.N. and Poonia, R.C. (2022) 'Hybrid HOG-SVM encrypted face detection and recognition model', *Journal of Discrete Mathematical Sciences and Cryptography*, Vol. 25, No. 1, pp.205–218.

Singh, P.K., Kundu, S., Adhikary, T., Sarkar, R. and Bhattacharjee, D. (2022) 'Progress of human action recognition research in the last ten years: a comprehensive survey', *Archives of Computational Methods in Engineering: State of the Art Reviews*, Vol. 29, No. 4, pp.2309–2349.

Stephen, O., Maduh, U.J. and Sain, M. (2022) 'A machine learning method for detection of surface defects on ceramic tiles using convolutional neural networks', *Electronics,* Vol. 11, No. 1, pp.55–76.

Xu, P., Huang, L. and Song, Y. (2022) 'An optimal method based on HOG-SVM for fault detection', *Multimedia Tools and Applications*, Vol. 81, No. 5, pp.6995–7010.

Zhang, J., Ye, G., Tu, Z., Qin, Y., Qin, Q., Zhang, J. and Liu, J. (2022) 'A spatial attentive and temporal dilated (SATD). GCN for skeleton – based action recognition', *CAAI Transactions on Intelligence Technology*, Vol. 7, No. 1, pp.46–55.

Zhao, H., Xue, W., Li, X., Gu, Z. and Zhang, L. (2020) 'Multi - mode neural network for human action recognition', *IET Computer Vision*, Vol. 14, No. 8, pp.587–596.