Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1440–1447, DOI 10.12694/scpe.v26i3.4171

THE CONSTRUCTION OF MATHEMATICAL MODEL OF SWIMMERS' TECHNICAL MOVEMENTS USING MULTIMODAL DEEP LEARNING FRAMEWORK

MENGMENG WANG*AND YANGWEN HE†

Abstract. This paper proposes a mathematical model construction method based on a multi-modal deep learning framework aiming at the accuracy and real-time requirements of swimmers' technical movement analysis. The model can extract the image features and timing information of athletes' movements from video sequences by integrating spatiotemporal modules. This paper introduces the translation partial channel strategy to overcome the limitation of spatiotemporal information separation in traditional methods, which can seamlessly integrate spatiotemporal features and enhance the recognition ability of complex action patterns. In addition, NetVLAD is used as the feature aggregation layer. This layer can capture and encode the global and local features of the athlete's movements, thereby improving the classifier's performance. In the experimental part, the model is strictly verified, and the results show that compared with the prior art, the model in this paper shows higher accuracy and faster processing speed in the swimmer's action classification task. This provides the possibility of immediate feedback for coaches and athletes and lays a solid foundation for further research in the field of sports science.

Key words: Multimodal deep learning; Space-time module; Translation part of the channel; NetVLAD; Classification of swimmers.

1. Introduction. In pursuing excellence in sports competition, technical improvement and scientific analysis are the driving forces to promote athletes to reach their peak state. Swimming, an ancient and vibrant water sport, attracts millions of fans and professionals worldwide with its unique charm. With the progress of science and technology and the continuous innovation of data analysis methods, applying deep learning technology in swimming has gradually become essential to improve the training effect and competition performance. In particular, the emergence of multimodal deep learning frameworks has provided unprecedented opportunities to capture and analyze the complex technical movements of swimmers. The technical movement analysis of swimmers is a highly specialized work that requires accurate capture and in-depth understanding of multi-dimensional information such as the athlete's posture, power distribution and movement rhythm in the water. Traditional methods often rely on intuitive observation by experienced trainers or use expensive and cumbersome motion-capture systems. Although these methods can provide valuable information to some extent, their subjectivity, limitation and lack of real-time limit their wide application in practical training.

The rise of multimodal deep learning frameworks has brought new hope to solve this problem. This framework can comprehensively utilize various sensor data, such as video streams captured by high-speed cameras and time series data recorded by motion tracking devices, and automatically extract key features through advanced algorithms to build mathematical models that reflect the nature of athletes' technical movements. The spatiotemporal module plays a core role in this framework, which can simultaneously process the spatial structure of visual images and the temporal evolution of temporal information, providing a comprehensive and detailed data basis for motion analysis.

Behavior recognition is essential to behavior prediction, posture analysis, etc. Its core purpose is to realize the accurate cognition of pedestrian behavior characteristics in the video, and the most critical problem is to fully dig out the adequate information of various parts in the video. One of the main differences is how the timing information is used and modeled. The previous research mainly used spatial-time descriptors to extract and recognize image features. Literature [1] applies it to dense moving trajectory images. Reference [2] optimizes IDT and improves feature regularization and feature coding. Remarkable results have been achieved

*School of Physical Education, Jiangsu University of Technology, Changzhou 213001, Jiangsu, China (Corresponding author, 2022500025@jsut.edu.cn)

†Department of Orthopedics, Changzhou West the Taihu Lake Hospital, Changzhou 213149, Jiangsu, China

in motion recognition. In recent years, with the rise of deep learning technology, more and more research has been done on image features. The 2-D convolution model only extracts a specific image from a single frame in video understanding, and it cannot be effectively modeled as a time series. In reference [3], for the problem of human behavior feature extraction, it is proposed to use a two-layer parallel convolutional network for image deep learning and a two-channel convolutional neural network for human behavior extraction to identify human behavior efficiently. In literature [4], a time-sequence segmented network was established based on the two-stream network, and a complete time-sequence data set was designed through segmented training of videos. The 3D convolution algorithm can better capture the temporal and spatial information, but it requires a large amount of computation. In literature [5], a 3D convolutional neural network based on time-domain information is applied to behavior recognition for the first time, and a 3D kernel function is used to conduct feature extraction on both space-time and behavior dimensions. Literature [6] pooled a 3D convolutional neural network called C3D. With the increasing demand for real-time image processing technology, algorithms based on lightweight models have gradually become a research hotspot. Reference [7] establishes MFNet, a mobile feature network containing action modules—the effective fusion of spatial and temporal information between frames within the same frame. In literature [8], a simple and effective STM module is designed to encode space and motion information using a two-dimensional network as the framework. Literature [9] analyzes the dynamic content in videos by using the fusion of slow and high-resolution CNN and fast CNN, respectively. Two parallel convolutional neural networks are applied to the same video sequence to improve the image quality. Literature [10] introduces RGB and optical flow into the two-in-one first-line network. The motion information of the streaming image is obtained in the moving state layer to realize the precise control of the underlying RGB signal.

In this paper, an innovative translation partial channel method is proposed. Introducing translation operation in the feature extraction process makes the model more sensitive to capturing the subtle changes in the action process, thus enhancing the ability of spatiotemporal feature representation [11]. In addition, as an efficient feature aggregation technology, NetVLAD can reduce data redundancy while maintaining feature diversity, which provides strong support for action classification tasks. The research content of this paper focuses on the following key points. Firstly, this paper designs and implements a multi-modal deep learning framework with integrated spatiotemporal modules, which can automatically extract swimmers' technical motion features from continuous video streams. Secondly, utilizing the translation partial channel method, this paper optimizes the fusion process of spatiotemporal information and improves the feature representation ability of the model. The NetVLAD aggregation mechanism is used again to build a compact and efficient feature descriptor, which provides a solid foundation for the subsequent action classification [12]. Finally, through many experiments, this paper shows the superior performance of the proposed framework on the task of swimmer's movement classification.

2. Swimmers' technical movement recognition system.

2.1. System Architecture. This project begins with a systematic definition of swimming behavior, regarded as a set of actions on different time series. A complete human behavior database is formed through data collection, preprocessing and feature extraction. This project addresses global characteristics such as swimming speed, stroke frequency and lap time and specific characteristics such as intensity and duration of individual movements [13]. The overall design of the system is shown in Figure 2.1. A single pose sensor obtains the movement information of athletes. Using the swimming image collected by the high-speed camera, the corresponding actual behavior markers are extracted, and the swimming database is built by combining the feature quantity. The model is divided into two parts: the first is to classify and identify the motion behavior to maximize the utilization of the training sample; The second is to verify the model. A series of continuous behavior sequences are obtained by analyzing the motion information obtained by each sensor component.

2.2. Data acquisition device and experiment. In the test, a 36 mm× 51.3mm×21mm integrated pose sensor module was used to complete the measurement of swimming posture. The sensor has a three-way acceleration sensor, a three-dimensional gyroscope and a three-way geomagnetic field sensor. A high-performance microprocessor is used for acquisition. Kalman dynamic filter is used to obtain the real-time pose of the sensor component [14]. Through WIFI wireless transmission technology, realize the sensor components

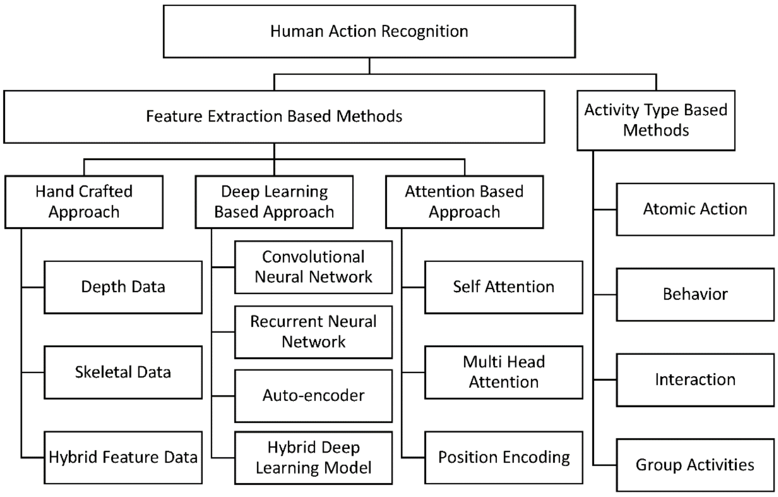

Fig. 2.1: Swim Action Recognizer system enclosure.

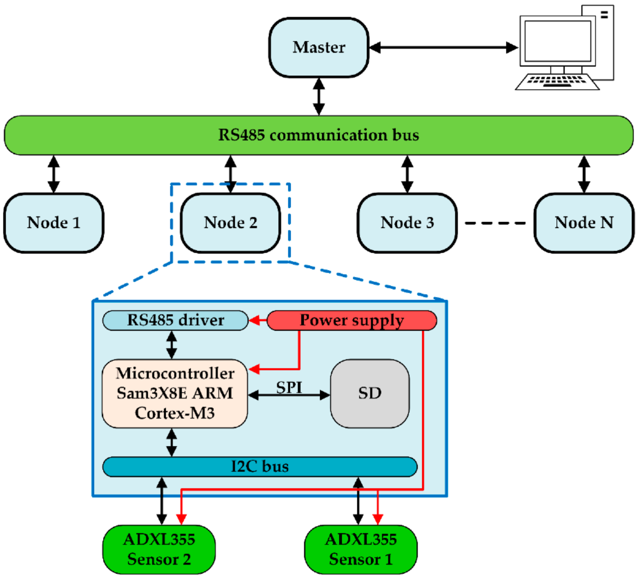

Fig. 2.2: Structure of swimmer condition monitoring sensor assembly.

of the pose and timely sequence information transmission. When the sensor component stops working for 5 minutes, it will automatically enter sleep mode, thus saving energy. Sleep automatically returns to its normal operating mode when the action is activated. The sensor is powered by 3.3V 5.0V. A USB charging port is installed externally. Figure 2.2 is the structure of swimmer condition monitoring sensor assembly. The sensor group's power supply is provided according to the swimming action characteristics. External USB charging port. Attach the sensor assembly to the waist's center using a strap (Figure 2.3). The position, acceleration, Angle and velocity sensor's arrow point to the square of each axis. Considering that different swimming conditions will lead to the change of geomagnetic field, this paper gives up the detection of geomagnetic field, and only uses two different motion states, such as acceleration and angular velocity [15]. The high-speed camera Q2m takes

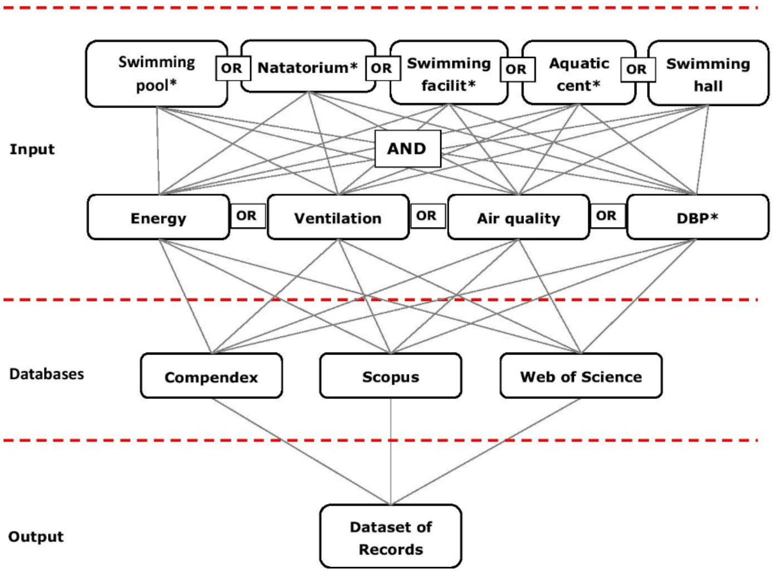

Fig. 2.3: Swimming data acquisition.

real-time photos of swimming motion marks at 5000 frames per second. The timing of an athlete's swimming movements can be watched frame by frame via video synced with sensor data.

2.3. Shift space-time module. An image sequence analysis method based on wavelet transform is proposed. In contrast to the convolution operation, the move operation does not require parameter values and floating-point operations but contains a set of operations with storage properties. Using 1x1 convolution for data fusion can reduce the computational cost. For example, in generic one-dimensional convolution, the prediction is expressed as a value derived from a weighted sum of the various inputs [16]. On the other hand, an input value is regarded as the input of the present moment and the adjacent moment. The input value is the input value of the three time points after displacement +1, 0, 1, multiplication, and addition. This kind of displacement convolution can be reduced to two processing methods: one is translation operation and the other is multiplication operation.

$$F_i = \lambda_1 T_{i-1} + \lambda_2 T_i + \lambda_3 T_{i+1} \tag{2.1}$$

$$T_i^{-1} = T_{i-1}, T_i^0 = T_i, T_i^{+1} = T_{i+1} \tag{2.2}$$

$$F = \lambda_1 T^{-1} + \lambda_2 T^0 + \lambda_3 T^{+1} \tag{2.3}$$

After sorting the Z-channel input in the K-frame picture, the tensor is shown in Figure 2.4.

Each image channel represents the amount of image frame features captured at each moment. The characteristic quantity is simultaneously transversally shifted for multiple channels along the time direction. Some channel values are down one space, and some are shifted one space. The blank part is filled with 0, and the excess channel value of the feature image is transferred out, thus completing the translation of the two directions [17]. After the movement, the characteristic information of the neighboring frame is merged with the current frame. But more movement doesn't mean more exchange of information. When the displacement is too small, the function of the timing model cannot satisfy the correlation of complex timing. When the displacement is large, the learning effect of spatial characteristics will decrease. By adjusting only the local channel, the efficient union of multiple channels is realized. The displacement model is to be added to the residual blocks

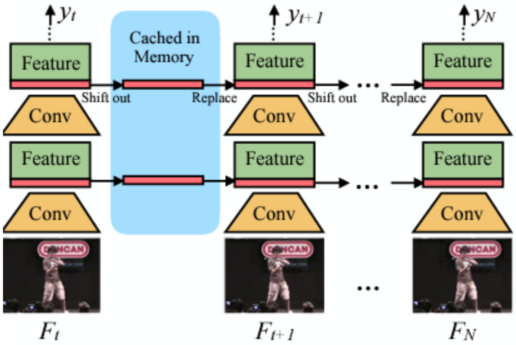

Fig. 2.4: Schematic diagram of the characteristics of the shift module.

of each branch of the residual network. Before the convolution operation, the displacement operation is carried out to complete the fusion of space-time spatial information without increasing the cost of three-dimensional computation. The time domain perception outfield value of each shift space unit is increased by 2 times to complete the construction of the time domain model.

2.3.1. Multi-mode. The multi-mode data processing of the image three-color difference is carried out based on fully mining the space and time information. In this way, the image is enhanced. Traditional feature-based image processing algorithms have a lot of operational overhead. This project intends to convert image color differences into RGB differences [18]. Then, the apparent changes and prominent moving areas are modeled to obtain the motion features of the image. Finally, the predicted value and the scores obtained from the spatial and temporal characteristics are added together to obtain the corresponding identification results.

2.3.2. NetVLAD Method. VLAD is based on locally aggregated information vectors, expressed through post-processing, and introduced into the terminal convolutional neural network to achieve image feature expression. In this paper, the NetVLAD algorithm is applied to the convolutional network and used as a pooling layer in the convolutional network to gather the characteristic information in the network. For A feature graph t, an S dimensional feature vector $t_i \in D^S$ must be obtained from the spatial position to represent the feature graph. First, J cluster centers z_j are given. The feature space D^s is divided into J units. Each feature vector t_i corresponds to a unit, and the residual vector $t_i - z_j$ is used to represent the difference between the feature vector and the cluster center. The resulting difference vector is expressed as:

The paper first needs to find its eigenvector A in the dimension S in space. First, J cluster centers B are given [19]. The model is divided into feature space C. Each feature vector D corresponds to the cell, and the residual vector E expresses the difference between the feature vector and the cluster center.

$$H(j,j) = \sum_{i=1}^{N} \frac{e^{-\beta\|t_i - z_j\|^2}}{\sum_{j'} e^{-\beta\|\|_i - z_j\|^2}} \left(t_i(j) - z_j(j)\right) \tag{2.4}$$

$t_i(j)$ and $z_j(j)$ represent the j component of the eigenvector t_i and cluster center z_j, respectively, where β is a trainable super parameter. The J column of the input matrix $h \in D^{JS}$ represents the eigenvectors gathered in the J cell, and then the matrix is normalized to a column and L_2 is normalized to a one-dimensional vector H to describe this property. Finally, the input data is fed into the fully connected layer for classification.

3. Simulation results and analysis.

3.1. Accurate rate and rate of image recognition. The efficiency and stability of the algorithm are verified by testing the accuracy and speed of swimming pose images of moving targets. In this experiment 1, the moving object is always within the shooting area of the camera. This paper counts the number of cooperative times and the time used [20]. The Kalman prediction and SIFT were combined to carry out the test, taking the swimmer's stroke as the research object (Figure 3.1). The calculation results of the optimal pairing rate

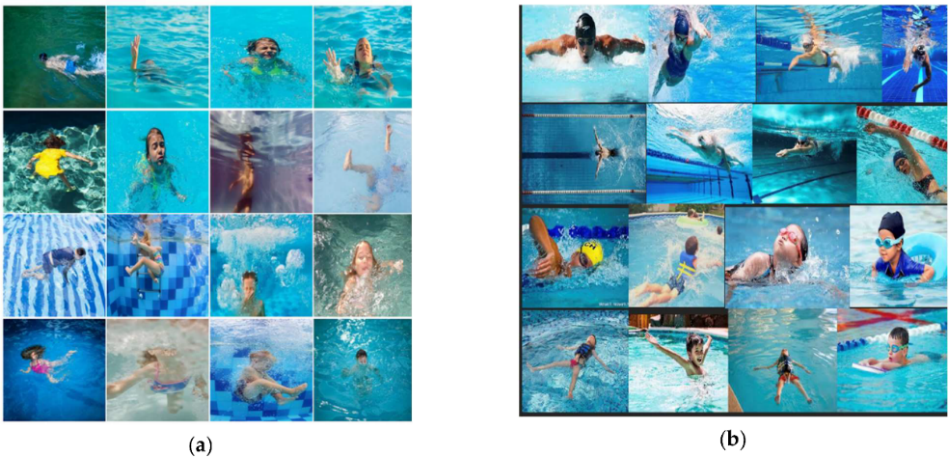

Fig. 3.1: Image recognition comparison diagram between traditional algorithm and the proposed algorithm.

Table 3.1: Comparison effect of moving objects.

Algorithm	Matches (N)	Successes (N)	Success rate (%)	Match time (S)	Average time per frame (S)
Textual algorithm	104	94	94	15	0.15
Traditional algorithm	104	68	68	197	1.97

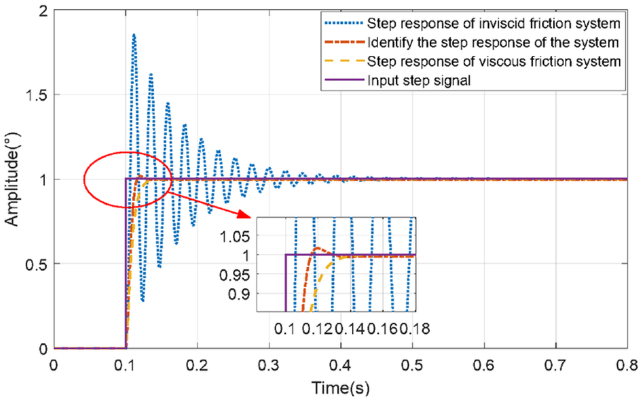

Fig. 3.2: Unit step response curve of the system.

and single frame time of the two algorithms are obtained (Table 3.1). This method can accurately classify the pose images of moving objects.

3.2. Image recognition effect. Experiments verify the effectiveness of this method, and it is tracked and identified. Experiments compare the performance of PID controller and fuzzy PID controller. The first is the Ziegler-Nichols algorithm. Then, the fuzzy PID controller is designed using the PID parameter set, and the step performance curve of the PID controller is recorded. The performance curve combining the above two controls uses the median filtering method (Figure 3.2).

The two systems' temperature rise and adjustment time are analyzed, and the results of related parameters

Table 3.2: Comparison table of step reaction diagram.

Controller	Overshoot	Rise time (t/s)	Adjustment time (t/s)
Conventional PID	0.25	2.08	4.17
Fuzzy PID	0.13	0.83	2.19

are obtained. The traditional PID method is used to adjust the attitude tracking of the moving object, and the lifting time is 2.08 seconds, the overshoot is 25%, and the adjustment time is 4.17 seconds. Using fuzzy PI D to adjust, the lifting time is reduced by 1.3 seconds. This reduces overshoot by 12%, reduces adjustment time by 1.9 seconds and improves the camera's performance in tracking moving objects. The algorithm in this paper completes the automatic recognition of moving objects and the tracking and retrieval of matching adjustment models (Table 3.2).

4. Conclusion. Using a multi-modal deep learning framework, this paper successfully constructs a mathematical model of swimmers' technical movements. By integrating a spatiotemporal module, this paper effectively extracts the image features and timing information of athletes' movements from video data, which provides a rich data basis for movement analysis. Applying the translation partial channel method further optimizes the fusion of spatio-temporal information, and enhances the recognition ability of the model for complex motion patterns. The introduction of the NetVLAD aggregation mechanism enables the model to process a large amount of feature information efficiently, significantly improving the accuracy of action classification. The experimental results show that this model performs excellently in classifying swimmers' movements, which provides a new technical analysis tool for coaches and athletes.

Acknowledgement. This work has been supported by a Social Science Project of Talent Introduction of Jiangsu University of Technology in 2022. The subject is Research on the Integration Development of National Traditional Sports and Early Childhood Physical Education Under the Background of Constructing a Sports Power, No.KYY22525.

REFERENCES

[1] Li, Z., Ye, X., & Liang, H. (2023). Sports video analysis system based on dynamic image analysis. Neural Computing and Applications, 35(6), 4409-4420.

[2] Strömbäck, D., Huang, S., & Radu, V. (2020). Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(4), 1-22.

[3] Chen, L., & Hu, D. (2023). An effective swimming stroke recognition system utilizing deep learning based on inertial measurement units. Advanced Robotics, 37(7), 467-479.

[4] Xia, H., Khan, M. A., Li, Z., & Zhou, M. (2022). Wearable robots for human underwater movement ability enhancement: A survey. IEEE/CAA Journal of Automatica Sinica, 9(6), 967-977.

[5] Zhang, X. (2021). Application of human motion recognition utilizing deep learning and smart wearable device in sports. International Journal of System Assurance Engineering and Management, 12(4), 835-843.

[6] Zhang, Y. (2023). Track and field training state analysis based on acceleration sensor and deep learning. Evolutionary Intelligence, 16(5), 1627-1636.

[7] Yang, M., & Zhang, S. (2023). Analysis of sports psychological obstacles based on mobile intelligent information system in the era of wireless communication. Wireless Networks, 29(8), 3599-3615.

[8] Amsaprabhaa, M. (2024). Hybrid optimized multimodal spatiotemporal feature fusion for vision-based sports activity recognition. Journal of Intelligent & Fuzzy Systems, 46(1), 1481-1501.

[9] Chinchilla Gutierrez, S., Salazar, J., & Hirata, Y. (2022). Mixed-reality human-machine-interface for motor learning of physical activities. Advanced Robotics, 36(12), 583-599.

[10] Kaseris, M., Kostavelis, I., & Malassiotis, S. (2024). A Comprehensive Survey on Deep Learning Methods in Human Activity Recognition. Machine Learning and Knowledge Extraction, 6(2), 842-876.

[11] Matsuyama, H., Aoki, S., Yonezawa, T., Hiroi, K., Kaji, K., & Kawaguchi, N. (2021). Deep learning for ballroom dance recognition: A temporal and trajectory-aware classification model with three-dimensional pose estimation and wearable sensing. IEEE sensors journal, 21(22), 25437-25448.

[12] McGrath, J., Neville, J., Stewart, T., & Cronin, J. (2021). Upper body activity classification using an inertial measurement unit in court and field-based sports: A systematic review. Proceedings of the institution of mechanical engineers, Part P: Journal of sports engineering and technology, 235(2), 83-95.

[13] Talha, M. (2022). Research on the use of 3D modeling and motion capture technologies for making sports training easier. Revista de Psicología del Deporte (Journal of Sport Psychology), 31(3), 1-10.

[14] Chen, G. (2024). An interpretable composite CNN and GRU for fine-grained martial arts motion modeling using big data analytics and machine learning. Soft Computing, 28(3), 2223-2243.

[15] Ramesh, M., & Mahesh, K. (2023). Efficient key frame extraction and hybrid wavelet convolutional manta ray foraging for sports video classification. The Imaging Science Journal, 71(8), 691-714.

[16] Van Leeuwen, T. (2021). The semiotics of movement and mobility. Multimodality & Society, 1(1), 97-118.

[17] Siddiqi, M. H., Alshammari, H., Ali, A., Alruwaili, M., Alhwaiti, Y., Alanazi, S., & Kamruzzaman, M. M. (2022). A template matching based feature extraction for activity recognition. CMC-COMPUTERS MATERIALS & CONTINUA, 72(1), 611-634.

[18] Kanwal, S., Khan, F., & Alamri, S. (2022). A multimodal deep learning infused with artificial algae algorithm–An architecture of advanced E-health system for cancer prognosis prediction. Journal of King Saud University-Computer and Information Sciences, 34(6), 2707-2719.

[19] Akila, K. (2022). Recognition of inter-class variation of human actions in sports video. Journal of Intelligent & Fuzzy Systems, 43(4), 5251-5262.

[20] Turmo Vidal, L., Márquez Segura, E., & Waern, A. (2023). Intercorporeal Biofeedback for Movement Learning. ACM Transactions on Computer-Human Interaction, 30(3), 1-40.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 25, 2024
Accepted: Jul 30, 2024