J Inf Process Syst, Vol.19, No.4, pp.554~562, August 2023
https://doi.org/10.3745/JIPS.04.0285

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Action Recognition Method in Sports Video Shear Based on Fish Swarm Algorithm

Jie Sun* and Lin Lu

Abstract

This research offers a sports video action recognition approach based on the fish swarm algorithm in light of the low accuracy of existing sports video action recognition methods. A modified fish swarm algorithm is proposed to construct invariant features and decrease the dimension of features. Based on this algorithm, local features and global features can be classified. The experimental findings on the typical sports action data set demonstrate that the key details of sports action can be successfully retained by the dimensionality-reduced fusion invariant characteristics. According to this research, the average recognition time of the proposed method for walking, running, squatting, sitting, and bending is less than 326 seconds, and the average recognition rate is higher than 94%. This proves that this method can significantly improve the performance and efficiency of online sports video motion recognition.

Keywords

Action Recognition, Fish Swarm Algorithm, Image Features, Sports Video, Sports Video Shear

1. Introduction

Sports video motion recognition is a multi-category pattern recognition task and primarily meet two difficulties. One is to extract useful features from similar sports actions in various sports videos; the other is to build a machine learning model to complete the classification of action features. Silhouette and contour are the two primary motion feature types derived from sports video. With a high dimension and a lot of noise, silhouette action affects the performance of machine learning classification algorithms [1].

For sports video action analysis, as the traditional recognition methods are complex in computation, it is difficult to apply them to sports video analysis [2]. In addition, the fish swarm algorithm and support vector machine are used to recognize the sports action. Sports action feature modeling is first achieved using the fish swarm approach, and the results of action recognition are then obtained by completing the classification of the created model using a support vector machine. However, the feature dimension of fish swarm algorithm is very large, making it difficult to classify support vector machines and having a negative effect on sports action identification [3,4].

This paper proposes a motion video clipping motion recognition method based on fish swarm algorithm to address the issues in the motion identification. This method obtains the motion video clip motion features based on the fish swarm algorithm, and obtains the fused invariant features through feature

Manuscript received July 1, 2022; first revision November 7, 2022; second revision February 3, 2023; third revision April 5, 2023; accepted April 25, 2023.

* **Corresponding Author:** Jie Sun (pdvh42@163.com)
Institute of Physical Education, China University of Geoscience, Wuhan, China (pdvh42@163.com, 1547058490@qq.com)

dimensionality reduction. The mixed kernel method is used to classify the features and acquire the motion video motion recognition. The test results show that this method has better performance in the recognition of motion video clip motion, and this approach performs better in identifying motion in video clips and can serve as a benchmark for motion recognition research.

2. Related Works

Pourpanah et al. [5] provided a brief description of the continuous fish swarm technique that included the fundamental algorithm, its improvements, and hybrid models, as well as how they were used. Since 2013, publishing in high-caliber journals has drawn attention. The review offers insights into parameter modifications, the approach, and sub-functions of the fish swarm algorithm. The main factors that contributed to these gains are discussed along with comparison results with other hybrid approaches. The hybrid, multi-objective, and dynamic fish swarm algorithm models have been proposed to address continuous optimization issues. It also explored prospective upgrades and highlighted brand-new regions for the fish swarm algorithm-based model.

Elias et al. [6] investigated the potential applications of meta-heuristics in sports. In recent years, it has become clear that sophisticated algorithms are necessary to investigate the different NP-hard issues that arise in sports analytics. A promising remedy for these problems is the use of meta-heuristics. Increased knowledge and discussion of intelligent algorithms can be advantageous for future meta-heuristic application research in sports analytics. Human body is represented as a sequence of 2D or 3D joint locations, with joint motion trajectory used to indicate movement. These approaches can study human motion in depth, but they require a powerful computer. Based on the above analysis, few studies discuss the application of fish swarm algorithm to sports motion recognition, so this article chooses this aspect to explore.

3. Movement Recognition in Sports Video Cut

3.1 Denoising Algorithm of Sports Video Action Recognition

The sports video analysis is dedicated to the identification of human motion in the footage [7]. A fish school algorithm is suggested and discussed in this study. In order to analyze sports video actions and extract sports video features, this method primarily makes use of the set features. The discipline of deep learning includes the classification of motion videos. In order to automatically extract characteristics and achieve the categorization of motion video actions and obtain the identification information, this approach primarily focuses on machine learning. The principle of this method is shown in Fig. 1.

Fig. 1 showed that it entails accurately engineering features from raw data in order to match a machine learning mode, which generally calls for in-depth domain knowledge and approaches from signal processing. The objective is to forecast a person's movement using sensor data. The feature extraction of motion video activities is completed by using motion trajectories to characterize and evaluate relevant motion types [8]. Fused invariant features are built using the fish swarm algorithm and feature dimensionality reduction techniques to produce reliable motion recognition results. It comprises the

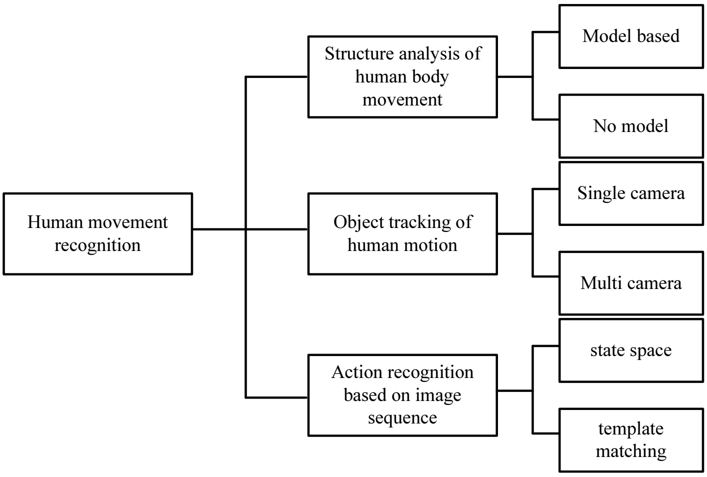

Human movement recognition
Structure analysis of human body movement
Model based
No model
Object tracking of human motion
Single camera
Multi camera
Action recognition based on image sequence
state space
template matching

Fig. 1. Main contents of human action recognition.

posture and body angle of an athlete in a certain action state when speaking of an invariant feature. Classifying and fusing invariant features with machine learning techniques produces results for motion video motion recognition. The sports elements that are derived during the analysis of sports footage vary between sports videos [9]. Therefore, a global fish school algorithm is needed for moving video sequences. Temporal, spatial, and motion boundary information are all considered global features. Then, the fish school algorithm is combined with this piece of information to obtain key features for motion recognition in moving video through dimensionality reduction. Suppose that $X = \{x_1, x_2, \dots, x_t\}$ represents the effective motion sequence of sports video, and$x_t \in R^d$. If it follows the independent distribution and d represents the feature dimension, then it conforms to the parameter set $\lambda = \{w_i, u_i, M_{p_i}\}, i = 1,2,..k$. The proposed fish swarm algorithm is composed of K fish swarm algorithms that can be given by the following formula. Assuming that w_i, u_i, M_{P_i} covariance matrix is composed of mixed weight, mean vector and covariance of fish swarm algorithm. In the above settings, several fish swarm algorithms are defined, among which the *i*-th fish swarm algorithm $p_i(x_i)$ can be defined as:

$$p_i(x_i) = \frac{exp\left\{-\frac{1}{2}(x_t - u_i)^T M_i^{-1}(x_t - u_i)\right\}}{(2\pi)^{d/2}|M_i|^{1/2} - p_\lambda(x_t)}. \tag{1}$$

The weight of each fish swarm algorithm in the hybrid model may be determined based on the fish swarm algorithm using the Bayesian formula, that is, the weight of the time-frequency frame t_x assigned to the *i*-th fish swarm algorithm is as follows:

$$r_t(i) = \frac{1}{\sum_{k=1}^{K} w_k p_k(x_t)} - t_x. \tag{2}$$

When extracting fusion invariant features, the logarithmic likelihood relationship between the sports video sequence x and its corresponding parameter set is as follows:

$$l_\lambda(X) = r_t(i)\log_2 p_\lambda(X) = r_t(i)\sum_{t=1}^{T}\log_2 p_\lambda(x_t). \tag{3}$$

After the above process, multidimensional invariant features are collected from each action in the video, and the gradient of the moving video sequence and the associated fish mixing function are determined [10]. The 2,010 dimension motion invariance features are made up of the 30 dimension motion position information, 480 dimension motion direction gradient information, 540 dimension motion optical flow information, and 960 dimension motion boundary information, as shown below:

$$X_1 = \left[x_t, t = 1,2, \dots, x_{T_1}\right] \tag{4}$$

where x_{T_1} is the feature of each sub dimension in the multi-dimensional invariance feature, and T_1 is the number of tracks in the sports video sequence. The feature dimension under multi-dimensional fusion feature is big, despite the fact that the feature derived by fish swarm technique has great invariance and good effect in sports video action recognition. Lack of training samples will result in issues with dimension functions and other issues when training the feature. The random projection feature reduction method is used to streamline the original fusion invariant features, which lowers the feature dimension and increases the classification and recognition efficiency of the fusion invariant features. This speeds up the classification of multi-dimensional fusion invariant features and enhances the effectiveness of sports video action recognition.

Fig. 2 demonstrated the identification difficulties vary considerably depending on application and data selection, proving that there is no one solution that addresses all challenges. It can be used in many different industries. Reducing the dimensions of feature vectors is the primary method for reducing the dimensionality of data. The covariance matrix is chosen for dimensionality reduction throughout the dimensionality reduction procedure. Assuming that an *n*-order symmetric matrix *A* has a diagonal matrix of diagonal elements *B*, its covariance matrix *P* is shown in Formula (5).

$$P^{-1}AP = B. \tag{5}$$

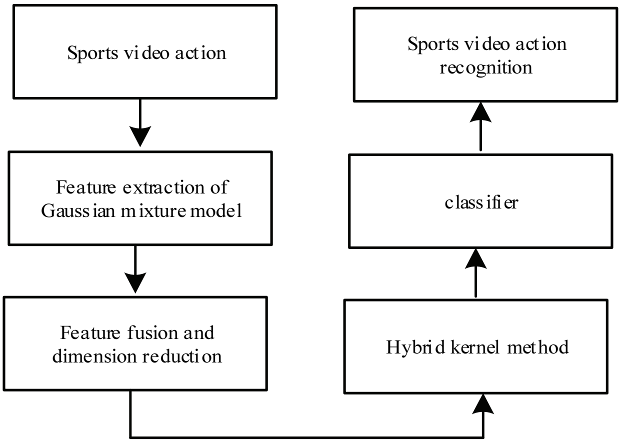

Sports video action
Feature extraction of Gaussian mixture model
Feature fusion and dimension reduction
Hybrid kernel method
classifier
Sports video action recognition

Fig. 2. Sports video action recognition process.

3.2 Realization of Video Shear Action Recognition

In sports videos with human motion objects, key poses must be extracted by choosing key frames. In a sports video, key frames are essentially one or more frames of pictures that can reflect the major storyline [10]. The movements of the people in the sports video can be incredibly succinctly expressed in this frame or numerous other frames. Additionally, most of the movements in each frame in these sports videos are repeated. Only the important frame content in sports videos is required when considering calculation efficiency and storage capacity, which can reduce storage needs and the workload placed on the CPU during redundant calculations and increase performance. The ratio of the horizontal to vertical dimensions of the moving object's contour region must be normalized in order to obtain the sequences of the contour region before computing the autocorrelation coefficient of the signal. Considering that the ratio of horizontal to vertical of the contour region of the corresponding action sequence in the sports video is $S = \{S_1, S_2, S_3, \dots, S_n\}$. Each S can be regarded as a signal, and the autocorrelation function is as follows:

$$R_{ss}(\tau) = \frac{1}{M}\sum_{i=1}^{M} s\,(i)s(i+\tau)D(x,y,\sigma). \tag{6}$$

When choosing, it's important to take into account two fundamental criteria based on the traits of crucial actions in various poses: As a starting point, the chosen key action frame needs to precisely and thoroughly capture the action while reflecting the pose's primary qualities. Second, the amount of data processing should be as minimal as feasible, and the calculation should be made as simple as possible, to increase the speed of data processing. As a result, key frame extraction ought to be kept to a minimum. Image feature matching is an important research direction in image recognition matching and computer vision applications. First, local features that are independent of scale and rotation are chosen from among local features other than global features. These landmark characteristics, such as key points, intersections, and other notable elements in local imagery, are what make this feature point interesting. In addition, compared to the method based on image region matching, it compared to the method based on image region matching, it simply needs a few points or blocks instead of having to determine the local region's pixel gray value., thereby reducing the calculation of subsequent matching work. A significant area of research in image matching is picture feature matching. It possesses properties that are immediately apparent, such great precision, low processing complexity, and good robustness. Currently, local invariant feature extraction is divided into two categories: feature point extraction and feature region extraction. In the field of computer vision, image normalization is an important step in preprocessing. The image normalization here is to convert the image into a more standard form. This can prevent the impact of affine changes and geometric changes, and accelerate the convergence of the network. The purpose of linear function normalization is to linearize the original data and convert the data to a standard range. As shown in the formula, *X* is the original pixel data, Max and Min are the maximum and minimum values of image pixels in the dataset:

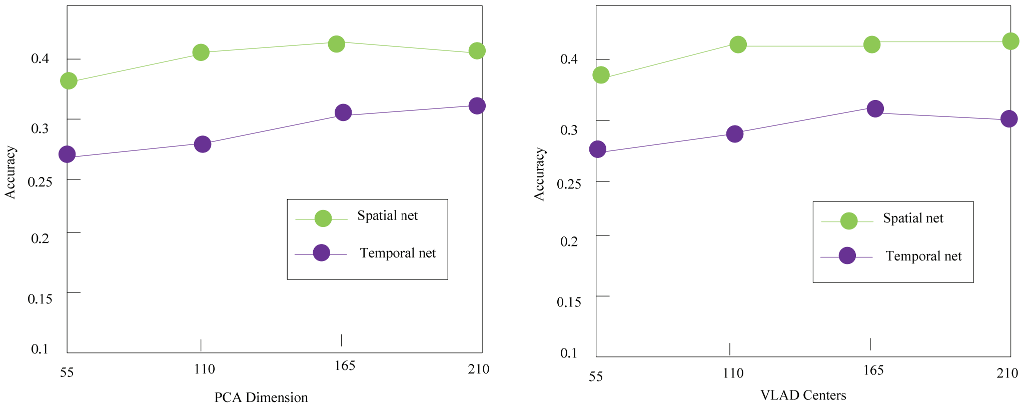
$$norm = \frac{R_{ss}(\tau)x - Min}{Max - Min}. \tag{7}$$

The data is split into two states throughout the processing stage. When extracting the features of the moving objects in sports video, the data is transformed from the sports video frame sequence to the spatial time series data. Before the transformation, the amount of data and information was large, and there was a lot of redundancy. First, the data capacity is reduced; second, the representation of motion features in time series and the action rules for classification can be easily obtained. This pre-processing phase is followed by the optimization of the human action recognition system. Traditional support vector machines are employed for classification after preprocessing is finished.

4. Analysis of Experimental Results

4.1 Extraction and Identification of Space Video Recognition

A comparative simulation experiment is built to finish the performance study of sports video action recognition in order to confirm the validity and efficacy of the suggested method. The training network model is accelerated in the experiment using the GPU and the PyTorch deep learning toolkit. Two of the most popular large-scale datasets—hmdb51 and ucf101ohmdb51—which are real-world datasets with complicated sports films compiled from movies and online sports videos, are chosen for the experiment. It contains 6,766 sports videos in 51 categories, with at least 100 sports videos in each category. It is divided into three overlapping subsets for evaluation. Each subset includes 70 training sports videos and 30 test sports videos. ucf101 dataset contains 101 sports video actions, and each action has at least 100 sports video segments. It is divided into 25 groups according to the executors. A total of 600 sports videos have been cut and processed. Similar to hmdb51, ucf101 is also divided into three subsets for evaluation. On ucf101 and hmdb51, the default training and test subset assignment is adopted, and then the average accuracy of the three subsets is reported. The test results are shown in Fig. 3.



(a) (b)

Fig. 3. (a) PCA dimension and (b) cluster center number exploration experiment.

In the simulation experiment, 10 athletes were selected, and each of them was asked to do a variety of simple sports actions. There were 600 different sports videos in all. Each sports video action included a random selection of five fixed movements, including walking, running, bending, squatting, and sitting down. The subjects' sports video action sequences were given, and five action sequences were given from six key frames. In the experiment, 200 of the remaining 400 sports action sports movies are utilized as test sets, and the training sets consist of the sports action sports videos. Fig. 4 shows the basic action sequence (including five kinds of actions) demonstrated by athletes in the experiment. The popular sports video motion recognition algorithm is chosen to finish the comparison of motion recognition algorithms, and the particle swarm optimization BP neural network, proposed fish swarm algorithm, and support vector machine are each chosen in turn. Based on Table 1, the three comparison algorithms' average recognition times for five basic sports actions- including walking, running, squatting, sitting, and bending- are provided.

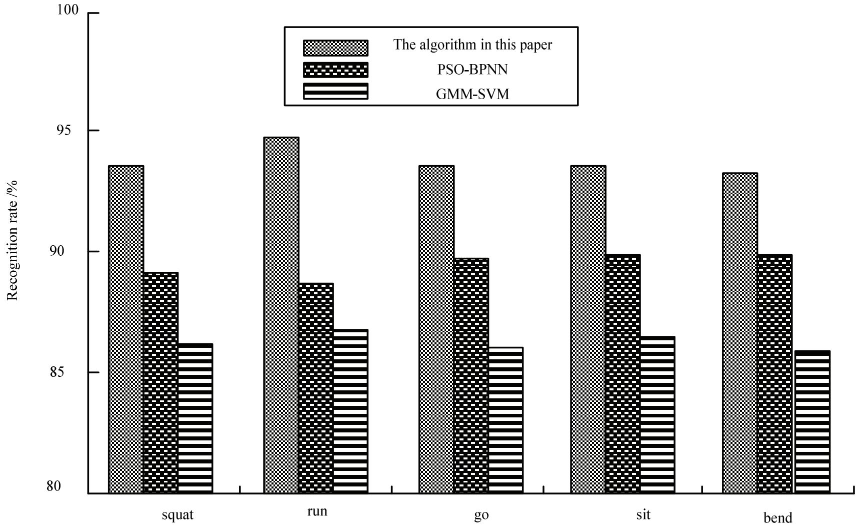


Fig. 4. Comparison of average recognition rate of sports video action recognition.

According to Table 1, the proposed method uses less than 326 seconds to recognize walking, running, squatting, sitting, and bending. It takes 288–423 seconds for GMM-SVM method to recognize walking, running, squatting, sitting, and bending. The recognition time of PSO-BPNN method for walking, running, squatting, sitting, and bending is 273–369 seconds. The proposed approach is more effective in identifying the five typical movements. The proposed method's usage of the fish swarm algorithm to build invariant features, reduce their dimension. Furthermore, the comparison results of the average recognition rate of sports video action recognition are given, as shown in Fig. 4.

Table 1. Average recognition time comparing of the three recognition algorithms for five common sports actions

Types of sports action	Recognition time (s)		
	Proposed algorithm	GMM-SVM	PSO-BPNN
Run	266	313	298
Walk	298	368	334
Sit	278	306	294
Squat	253	288	273
Bend	326	423	369

GMM=Gaussian mixture model, SVM=support vector machine, PSO=particle swarm optimization, BPNN=backpropagation neural network.

Based on Fig. 4, the average recognition accuracy of the proposed method in the five common sports actions of walking, running, squatting, sitting, and bending is higher than 94%. The average recognition accuracy of GMM-SVM method for walking, running, squatting, sitting, and bending is 87%–89%. The average recognition rate of PSO-BPNN in walking, running, squatting, sitting, and bending is 85%–86%. Therefore, the recognition rate of the proposed method is higher. The reason is that the proposed method completes the classification of motion features after dimensionality reduction, which improves the effect of video motion recognition. Therefore, the classification efficiency of this method is significantly higher than the other two methods.

5. Conclusion

There is significant scientific value in identifying and comprehending video data using the robust storage and computational capabilities of computers. This paper suggests a motion recognition model based on the fish swarm algorithm to improve the performance of current sports video motion detection algorithms. Experiments have been conducted to confirm its efficacy. This approach employs a dimensionality reduction technique while building mixed invariant features using a fish swarm algorithm. According to experimental findings on a common sports video motion dataset, the algorithm's effectiveness and performance have greatly increased. Future work needs to focus on developing stable features for more complicated motion motions and constructing real-time motion video motion recognition through models with improved classification performance and efficiency.

References

[1] H. Wei and N. Kehtarnavaz, "Simultaneous utilization of inertial and video sensing for action detection and recognition in continuous action streams," *IEEE Sensors Journal*, vol. 20, no. 11, pp. 6055-6063, 2020. https://doi.org/10.1109/JSEN.2020.2973361

[2] J. Xiong, L. Lu, H. Wang, J. Yang, and G. Gui, "Object-level trajectories based fine-grained action recognition in visual IoT applications," *IEEE Access*, vol. 7, pp. 103629-103638, 2019. https://doi.org/10.1109/ACCESS.2019.2931471

[3] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Applied Intelligence*, vol. 51, pp. 690-712, 2021. https://doi.org/10.1007/s10489-020-01823-z

[4] F. Liu, X. Xu, T. Zhang, K. Guo, and L. Wang, "Exploring privileged information from simple actions for complex action recognition," *Neurocomputing*, vol. 380, pp. 236-245, 2020. https://doi.org/10.1016/j.neucom.2019.11.020

[5] F. Pourpanah, C. P. Lim, and Q. Hao, "A reinforced fuzzy ARTMAP model for data classification," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 1643-1655, 2019. https://doi.org/10.1007/s13042-018-0843-4

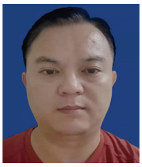
[6] P. Elias, J. Sedmidubsky, and P. Zezula, "Understanding the limits of 2D skeletons for action recognition," *Multimedia Systems*, vol. 27, pp. 547-561, 2021. https://doi.org/10.1007/s00530-021-00754-0

[7] Y. Y. Joefrie and M. Aono, "Multi-label multi-class Action recognition with deep spatio-temporal layers based on temporal Gaussian mixtures," *IEEE Access*, vol. 8, pp. 173566-173575, 2020. https://doi.org/10.1109/ACCESS.2020.3025931

[8] J. Xie, Q. Miao, R. Liu, W. Xin, L. Tang, S. Zhong, and X. Gao, "Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition," *Neurocomputing*, vol. 440, pp. 230-239, 2021. https://doi.org/10.1016/j.neucom.2021.02.001

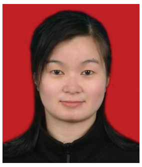
[9] J. H. Kim and C. S. Won, "Action recognition in videos using pre-trained 2D convolutional neural networks," *IEEE Access*, vol. 8, pp. 60179-60188, 2020. https://doi.org/10.1109/ACCESS.2020.2983427

[10] D. Ludl, T. Gulde, and C. Curio, "Enhancing data-driven algorithms for human pose estimation and action recognition through simulation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3990-3999, 2020. https://doi.org/10.1109/TITS.2020.2988504



Jie Sun **https://orcid.org/0000-0001-8904-1770**

He received a master's degree from the School of Physical Education, China University of Geosciences (Wuhan) in 2013. Since 2013, he has been engaged in the teaching and research of Taekwondo public physical education in the School of Physical Education, China University of Geosciences (Wuhan). His current research interests include sports event organization and management, taekwondo competitive training, etc.



Liu Lu **https://orcid.org/0000-0003-3839-3594**

She received her master's degree from the School of Sports Training of Wuhan Institute of Physical Education in 2012. Since 2013, she has been engaged in the teaching and research of Taekwondo public physical education in Wuhan Institute of Engineering and Technology. Her current research interests include sports core training theory, sports event programming, etc.