Hindawi
Complexity
Volume 2021, Article ID 5515357, 13 pages
https://doi.org/10.1155/2021/5515357

Research Article

Automatic Construction and Extraction of Sports Moment Feature Variables Using Artificial Intelligence

Zhao Zhang,[1] Wang Li ,[2] and Yuyang Zhang[3]

[1]*Sichuan Normal University, Chengdu, Sichuan 610066, China*
[2]*Sichuan Tourism University, Chengdu, Sichuan 610100, China*
[3]*Sichuan Nursing Vocational College, Chengdu, Sichuan 610100, China*

Correspondence should be addressed to Wang Li; 0000164@sctu.edu.cn

Received 20 February 2021; Revised 23 March 2021; Accepted 2 April 2021; Published 24 July 2021

Academic Editor: Wei Wang

In this paper, we study the automatic construction and extraction of feature variables of sports moments and construct the extraction of the specific variables by artificial intelligence. In this paper, support vector machines, which have better performance in the case of small samples, are selected as classifiers, and multiclass classifiers are constructed in a one-to-one manner to achieve the classification and recognition of human sports postures. The classifier for a single decomposed action is constructed to transform the automatic description problem of free gymnastic movements into a multilabel classification problem. With the increase in the depth of the feature extraction network, the experimental effect is enhanced; however, the two-dimensional convolutional neural network loses temporal information when extracting features, so the three-dimensional convolutional network is used in this paper for spatial-temporal feature extraction of the video. The extracted features are binary classified several times to achieve the goal of multilabel classification. To form a comparison experiment, the results of the classification are randomly combined into a sentence and compared with the results of the automatic description method to verify the effectiveness of the method. The multiclass classifier constructed in this paper is used for human motion pose classification and recognition tests, and the experimental results show that the human motion pose recognition algorithm based on multifeature fusion can effectively improve the recognition accuracy and perform well in practical applications.

1. Introduction

In practice, human movement gestures generally include walking, running, jumping, squatting, falling, and other daily dynamic behavioural performance of human beings. These motor gestures not only express people's physical activity state in life, study, and work but also convey the information of human's behavioural purpose and emotional reaction during the activity [1]. By recognizing these motion gestures and mastering the spatial and temporal characteristics of the movement, we can effectively recognize and analyse the dynamic process of human motion and obtain the information conveyed by the human body, thus realizing intelligent analysis and detection and providing the basic basis for other intelligent applications [2]. The process of human motion gesture recognition is based on motion target detection, extracting motion gesture features and automatically classifying and recognizing human motion gestures by analysing the extracted human motion gesture feature information [3]. Traditional sports analysis is to obtain sports training data through actual observation by coaches and use it to analyse and formulate training and instruction plans. Therefore, the actual observation and experience analysis of coaches are the main basis for the development of athletes' training plans. However, facing a large amount of training day after day, it becomes difficult for coaches to collect and summarize and analyse training data in a comprehensive and detailed manner [4]. Video-based sports analysis is to take athletes' video information as the processing object and obtain sports data through computer vision intelligent analysis to achieve objective and efficient intelligent analysis, which helps athletes quickly find

out problems and expert action essentials in the training process. Accurate computer analysis can minimize the probability of injury to athletes, to achieve modern scientific training.

In the face of the problems that exist in current sports video analysis research, such as low-level video features cannot accurately reflect high-level human semantic concepts, high time complexity, and low recognition accuracy of action recognition algorithms in traditional RGB videos and the use of single features cannot meet the massive growth of existing video data and its recognition of complex actions, the research on automatic description of competitive sports, with free gymnastics as a typical representative, has theoretical research significance and practical application value. In terms of theoretical research, automatic sports description research is a cross-cutting topic that integrates machine learning, pattern recognition, video analysis, computer vision, and cognitive science, which provides a good research object for these fields and can promote the development of related disciplines [5]. The movement of the human body in sports videos is very complex and skilful, and the analysis of sports videos is more difficult and challenging compared with daily sports. The analysis of sports videos can not only bring more viewing effects to sports competitions but also help coaches to analyse the competitions and assist athletes in training [6]. Through the study of automatic understanding of free gymnastics, we can improve the accuracy of sports action recognition while conducting action data analysis, to explore the regular characteristics of the development of gymnastics technical innovation and realize the function of assisting training [7]. For example, with relevant competitors as the main research object, we analyse the gap between the difficulty, choreography, and quality of set movements between award-winning and ordinary competitors, study the trend of free gymnastics movement development and innovation, and adjust training countermeasures to improve the athletes' skill level.

In the research on automatic description of sports and more specifically the research on automatic description of human motion in sports videos, human motion analysis in videos refers to the intelligent representation and labelling of specific human motion present in video sequences through various technical means of computer vision and pattern recognition. Therefore, in this paper, the current research status is analysed from three aspects: automatic video description methods, sports video content research, and human motion classification in the video.

2. Related Studies

Prakash et al. proposed the idea of spatial-temporal interest points and proposed a 3D Harris detector. This method is mainly based on the principle of two-dimensional Harris corner point detection with the addition of temporal information of video sequences so that spatial-temporal interest points can be detected [8]. Wang et al. proposed a high-efficiency Dollar detector by using the Gaussian function and Gabor wavelet function to filter the video directly [9]. Although spatial-temporal interest points are a classical excellent local feature description algorithm, the complex and variable light and background, as well as camera motion, make the interesting point detection method less effective and difficult to be applied to real-life situations. The two-stream method proposed by Gültekin et al. uses a CNN network containing two branches [10]. The spatial branch takes the RGB image as input to extract the appearance features of the video, and the temporal branch takes the optical flow map as input to extract the motion features of the video, and then the two branches are used to classify the motion separately, and finally, the results are fused using the average or support vector machine (SVM) method [11]. Stachl et al. combined the traditional feature extraction DT method and the deep learning feature stream method and proposed a method to describe dense trajectories using CNN networks to improve the representation of action features [12]. The P-CNN method proposed by Ma et al. combines the human pose information and the 2-stream method; firstly, the human body is segmented into different limb parts using the human pose information, and then the RGB images corresponding to each limb part are fused [13]. The RGB images and optical flow maps corresponding to each limb part are then fed into the stream network for feature learning, and finally, the linear SVM is used for action classification; also, CNN features are combined with DT features to further improve the recognition effect [14].

On the one hand, an attention mechanism is introduced to improve the existing algorithm model to improve the accuracy of free gymnastics automatic description, and then different convolutional networks are tried as feature extractors to analyse the effect of feature extraction networks on free gymnastics automatic description [15]. On the other hand, the free gymnastics video automatic description problem is transformed into a multilabel classification problem, and to extract the temporal and spatial feature representations in the video, a 3D convolutional neural network is used as a feature extractor [16]. Then, a binary classifier for individual decomposed actions is constructed, and each video will perform binary classification computation for all categories to complete the multilabel classification process. Finally, the classification results are concatenated into a natural language description of the free.

3. Artificial Intelligence Feature Variable Design

In the process of computer vision recognition and artificial intelligence analysis to complete the preprocessing of video through motion target detection, its processing effect has a direct impact on much subsequent high-level vision processing, such as motion target tracking and recognition, video search, and action analysis. However, videos captured from real-life often contain interference from factors such as weather, lighting, and shadow changes, resulting in background images that also show dynamic changes [17]. Therefore, the selection of motion target detection algorithms should consider the actual application environment and the real-time nature of the algorithm. The detection

effect of such algorithms depends mainly on the establishment of the background model, and in practice, the background model is often interfered with by environmental factors, mainly the following interference effects: changes in lighting, such as outdoor daylight for a short period of time by dark clouds, indoor lighting equipment on and off, and light with the sun changes throughout the day; the relationship between the moving target and static background, such as the movement of the target at a location within the lens of the stationary or its stationary for a period of time after the start of movement again change; moving objects by shadows or other objects reflections caused by interference, such as light under the shadow of the moving target itself or other objects shadows and the floor or mirror objects reflections on the movement of the target interference; and background has been the existence of small changes, such as the natural environment caused by the breeze swaying branches, water fluctuations, and fires produce smoke. The sports video analysis system is the product of the application of image technology and sports mechanics in sports and has become an indispensable and important means for the development of sports. However, the motion video analysis system still has serious shortcomings, mainly due to slow information feedback and complicated operation. After sports video shooting, a lot of manual work is required, and it takes one or two months or even three or four months to obtain the analysis data, which makes the information feedback seriously lags behind, cannot realize real-time monitoring, and does not have the value of timely guidance.

The algorithm is applied to obtain static features, such as size, colour, contour, and shape; dynamic features, such as optical flow, speed, rate, and trajectory; spatial-temporal features, such as spatial-temporal interest points, directly from the video or image sequence; or hybrid features, which fuse several different features. Since it is difficult to classify an action by one feature, multiple features are usually extracted during feature extraction, and the format of these features is usually inconsistent, so the extracted features need to be normalized so that the features can follow a uniform form for action representation [18]. Feature extraction is the first processing stage in the action recognition process and plays a decisive role in the operation of the whole algorithm. The merit of feature selection determines the performance of the action recognition algorithm, and efficient features can improve the accuracy of action recognition. This section introduces the feature extraction methods in human action recognition, mainly from two aspects, global features as well as local features, and the common feature extraction methods are classified as shown in Table 1.

Global features are usually used to detect people or movements using human detection methods and then get some parameters of the human body, such as position or velocity or gradient, to describe human movements. The global features can be divided into apparent features, motion features, and hybrid features. The commonly used methods to represent these global features include motion energy image (MEI), motion history image (MHI), optical flow features, HOG, and HOF. Apparent features usually include the geometric structure and contour of the human body, through which the movement of the human body is described, and usually the background difference method can be used to obtain the contour of the human body. Here, we mainly introduce two classical methods of action recognition using apparent features.

Convolutional networks are neural networks that use at least one layer of convolutional (or intercorrelation) operations instead of the generic matrix computation. When constructing a model, cascading multiple layers of convolution allows the model to learn features with higher robustness and deeper levels. At the same time, the local receptive fields of convolutional kernels of different sizes are different, so the convolutional layers can output feature maps with different receptive fields and multiple scales. If a two-dimensional convolution kernel S is used to perform the convolution operation on the input image I, it can be described as follows:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i+m-n, j+n - m)K(m-n, m+n), \tag{1}$$

where K represents a constant coefficient, i and j represent the number of steps in the i-th and j-th steps, and m and n represent the values in the m and n states. In some deep learning software frameworks, mutual correlation is often used instead of convolution for computational convenience. The difference between mutual correlation and convolution as equivalent is that the convolution kernel does not need to be flipped counter clockwise during use. The mutual correlation operation can be described as follows:

$$S(i,j) = \sum_m \sum_n I(i+m+j, j+n-i)K(m-n+i, m+n-j). \tag{2}$$

The hybrid Gaussian background modelling approach is based on the dynamic updating of parameters so that the background model it builds is better adaptive. The basic principle is that each pixel point in the image is assumed to be independent and can be mixed by K Gaussian distributions, where K generally takes values from 3–5. A weighted sum of K Gaussian probability density functions can represent the probability of each pixel value.

$$p(x_t) = \sum_{i=1}^{M} w_{i,t} \times \chi\left(x_t, u_{it}, \gamma_{i,t}\right), \tag{3}$$

$$\chi\left(x_t, u_{it}, \gamma_{i,t}\right) = \frac{1}{(2\pi)^{n/2}\left|\sum i,t\right|^{1/2}} e^{-\chi\left(x_t, u_{it}, \gamma_{i,t}\right)}, \tag{4}$$

where p is the corresponding probability and t is the value of the corresponding time.

When a new video frame is an input, the pixel value x_t of each new input is compared in turn with the K models already established for the current pixel point at that location according to equation (3) where D is the confidence parameter, usually taken as 2.5.

Table 1: Commonly used feature extraction methods in the field of human action recognition.

Category	Overall model	Form	Representative method
Global characteristics	Human body part model	Apparent characteristics Movement characteristics Mixed features	MEI, MI, 3D space-time cube Light flow HOG, HOF
Local features	Significant location detection	Posture feature Points of interest	Deformation part model 3D SIFT, 3D Harris

$$\left|x_t - \mu_{i,t-1}\right| \le D \times \sigma_{it-1}, \tag{5}$$

where D represents a constant. After finding the extreme values of the pixel points by the Hessian matrix, the feature points are determined by the nonmaximum suppression method. By setting the threshold value, all feature points smaller than the threshold value are discarded, and the number of detected feature points is reduced by increasing the value of the threshold, and finally, a few of the strongest feature points are identified. In the process of feature detection, a filter of the size corresponding to the resolution of the scale layer image is used. For example, in a 3×3 filter, any point in the scale layer image is compared with 8 points around it and 18 points in the layer above and below it. If the feature value of the point is greater than the pixel value of the surrounding 26 points, the point is identified as the feature point in the region. This method is more accurate in the analysis of the prediction results of the changes in the characteristics of the human body at the moment of movement, so that less sports injuries and health care training can be analysed.

The Lansky function-based motion target detector is a vector image model that is commonly used to detect changes in pixels at the same position in two adjacent frames. In this model, it is assumed that each pixel in a frame is correlated with its neighbouring pixels [19]. Each pixel in an image frame can be represented by a vector, as shown in Figure 1; the components of this vector consist of the central pixel and its neighbouring pixels. To detect the change of pixel position in two image frames, this support region can be tested by performing a linear independent test. If the pixel at the same position is linearly independent of the corresponding pixel in the adjacent frame, it is the pixel that has changed. The Lagosian basis function can determine the linear correlation or independence of the vector.

It is shown experimentally that the determinant of the Lansky matrix is zero when two pixels are linearly correlated.

$$|W| = \begin{vmatrix} F_t(x,y) & F_{t+1}(x,y) \\ F_t'(x+1,y+1) & F_{t+1}'(x+1,y+1) \end{vmatrix}, \tag{6}$$

where (x, y) has grayscale values $F_t(x, y)$ and $F_{t-1}(x, y)$ at different moments, which can be simplified to the following equation:

$$|W| = \left| \left(\frac{F_t'(x+1,y+1)}{F_{t-1}'(x+1,y+1)} \right)^2 - \frac{F_t(x,y)}{F_{t+1}(x,y)} \right|, \tag{7}$$

and the Lansdowne-based motion target detector is applied to a pixel region with spatio-temporal properties, as shown in Figure 1 which represents the pixel point (x, y) support region, shown in an image frame with a window size of 3×3. At a specific location (x, y), a Lagosian basis function is used to detect the changing relationship between frame t and frame $y + 1$, as shown in the following equation:

$$|W| = \frac{1}{n} \sum_{i=1}^{n} \left| \left(\frac{F_t'(x+1,y+1)}{F_{t-1}{}'(x+1,y+1)} \right) - \frac{F_t(x,y)}{F_{t+1}(x,y)} \right|. \tag{8}$$

Through the LANSKY matrix correlation and pixel value judgment, the LANSKY matrix judgment interval is the pixel point change of five frames of the image, where the judgment interval is selected five frames from the test of different videos to get five frames that can produce ghost shadow because the ghost shadow region is generated fast and stationary in each frame; through the LANSKY function to judge the movement of pixel points while considering the pixel value, we can find the ghost shadow region. The pixel points in the ghost region are judged to be the background of the template update to suppress the effect of ghost shadow on the subsequent motion target detection; to get a clear motion target, reduce the noise and interference caused by the nonblurred edges and provide accurate information for the subsequent processing.

3.1. Experimental Design of Motion Moment Feature Construction and Extraction. Feature fusion refers to the fusion algorithm to obtain a fusion-optimized combination of feature vectors when facing multiple types and dimensions of feature vectors so that the fused feature set reflects the complementarity of information but also reduces the redundancy of information and ensures the real-time performance of the algorithm. The purpose of the search-based feature fusion method is to select a subset of feature information with strong distinguishing ability from the multisource feature information set, and compared with other feature fusion algorithms, this type of algorithm has strong classification while reducing dimensionality [20]. Among them, the genetic algorithm chromosome encoding is binary, simple, and easy to handle; it can obtain the global optimal solution of the optimization problem, and the optimization result is independent of the initial conditions, with good robustness; the problem is solved from the population and its inherent parallelism. From the above, the genetic algorithm can achieve the screening of effective features so that the feature vector dimensionality is reduced. Therefore, the new feature vector set after fusion helps to improve the motion pose recognition rate and the efficiency of the algorithm.

In the video, since the camera is in motion, there are also many trajectories in the background, and the trajectories of

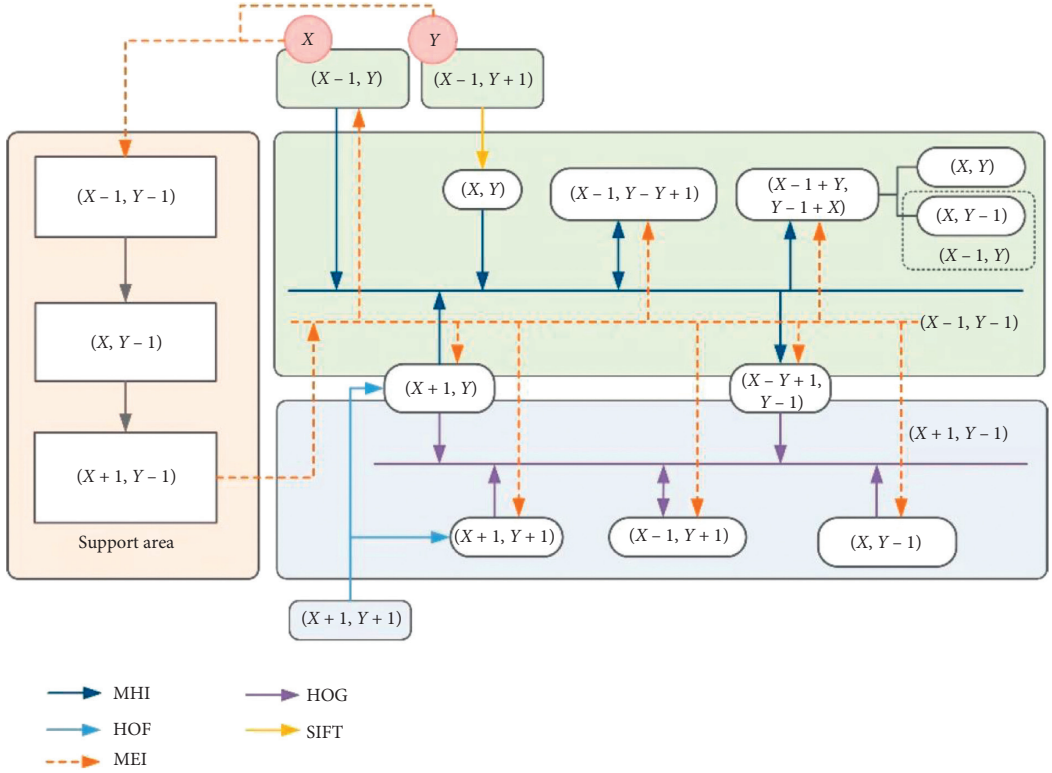

Figure 1: Support area of pixel point (x, y).

people are greatly influenced by the camera's motion. And this information is not very relevant to the action to be recognized and is interference information. So, it is desired to recognize and eliminate these trajectories. The motion of the trajectory is also calculated by computing the optical flow information. Therefore, after intensive sampling of the image, camera motion estimation is needed to eliminate the interference information in the background for better trajectory tracking and feature extraction.

The KTH dataset and HMDB51 dataset were used for the experiments. The KTH dataset contains 2391 video sequences, which are obtained from 25 individuals performing all actions in 4 relatively homogeneous background scenarios, involving 6 daily human behaviours, namely, walking, jogging, running, boxing, waving, and clapping. Each behaviour has at least 100 video samples. Since the data sources of this dataset are mainly movie clips and web videos, its intraclass variation is large, and it is one of the most challenging behaviour recognition datasets available. The homogeneity of the variance test is a method in mathematical statistics to check whether the population variances of different samples are the same. The basic principle is to first make a certain hypothesis about the characteristics of the population and then make inferences about whether this hypothesis should be rejected or accepted through statistical inference of sampling research.

The human brain pays different attention to different parts of the signal when processing, known as the visual attention mechanism. Human vision obtains the target area to be focused on, which is generally referred to as the focus of attention, by quickly scanning the global image, and then devotes more attentional resources to this area to obtain more detailed information about the target to be focused on and suppress other useless information. It is to test whether the difference between the mean of a sample and a known population mean is significant. When the population distribution is a normal distribution, if the population standard deviation is unknown and the sample size is less than 30, then the deviation between the sample mean and the population mean statistics is distributed in t. The reason for the need to utilize the attention mechanism in this paper is quite intuitive; the video frames that are decisive for the automatic description of the free gymnastic breakdown movements should be the method, direction, and angle of the athlete's body flip, and the weight of these video frames should be greater. In this paper, an attention mechanism is used, which allows the decoder to weigh each temporal feature vector of the free gymnastics video. Figure 2 illustrates the network structure after the introduction of the attention mechanism.

The hardware environment of this chapter is Intel Core i5-4590 CPU, 3.30 GHz, and 4 GB memory, and the software environment is Matlab2015b under Ubuntu 14.04 system. The average processing speed of the KTH dataset is 37.2243 frames/second, while the HMDB51 dataset has more feature points due to more camera shaking, more interactive behaviours, and strong randomness of behaviours, and the average processing speed is 9.2665 frames/second [21]. The average training time of SVM parameters for each behaviour category is 6.23 s for the KTH dataset, and the average

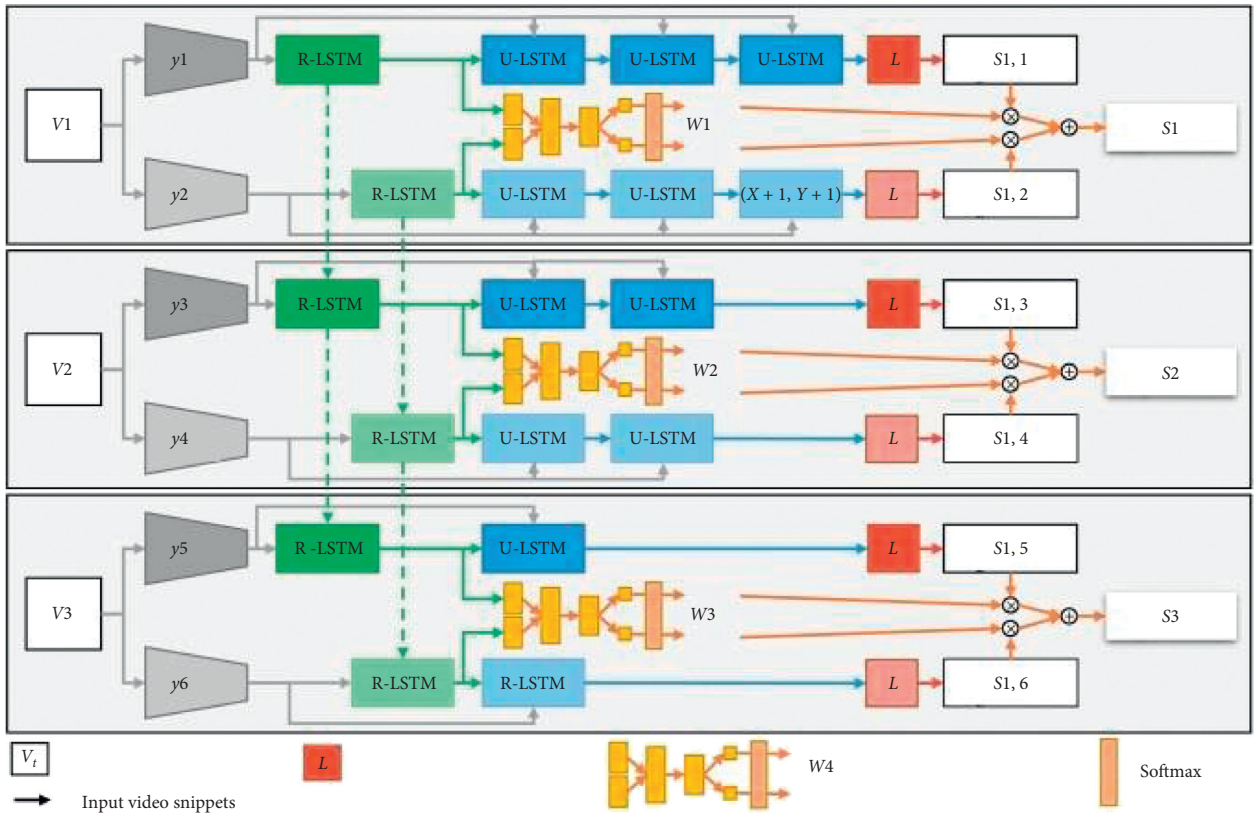

Figure 2: Illustration of the attentional mechanism.

adaptation time for 5-fold cross-validation is 31.15 s. The average training time of SVM for the HMDB51 dataset is 3.03 s, and the average calculation time off is 15.16 s. The computational complexity T of parameter search for a specific behavioural class depends on the population size N and the number of iterations n of the AMPSO algorithm. Since the SVM parameter search process for each behavioural class is independent of each other, a parallel computation strategy with multiple machines is used. In the test classification stage, the average time taken by a single model to discriminate the behavioural categories of a video segment is 6.83 ms. Since a one-to-many strategy is used to solve the multicategory problem, the overall computation time of the classification process is $6.83\,\text{ms} \times M$, where M denotes the number of behavioural categories.

The parameters of the RBF-SVM are optimized using the AMPSO algorithm on the KTH and HMDB51 datasets to verify the effectiveness of the proposed method. For the KTH dataset, the midlevel semantic expressions of all video segments are computed separately, and the global expressions of video segments of experimental bodies 2, 3, 5, 6, 7, 8, 9, 10, and 22 are used as the test sample set, while the rest are used as the training set [22]. In the optimization search process of model parameters γ and C, the current behaviour category in the training set is used as positive samples and the rest as negative samples, and the corresponding binary classification models are trained for each of the six behaviours using the LIBSVM toolbox. When the population size $N = 20$, γ is set to 3, and 5-fold cross-validation is used to obtain the classification accuracy under different parameters as the current particle fitness. Also, since the Fisher vector of each video segment is 41984 dimensions and the number of training set samples is large, the number of termination iterations is set to 50 in the parameter search process to reduce the time cost. We only change one of the same parameters. The change in this article is the number of corresponding hidden layers, while the other parameters remain unchanged. In this article, there is no formula or principle to observe the adjustment of parameters. We adjust based on our experience.

Applying this classification model to the process of human motion pose recognition, each motion pose sample feature information contains two independent feature information, and then the feature fusion multicategory classification model needs to be established. The training sample feature information is the feature data obtained from the videos in the standard video database, and two kinds of features can be extracted from each class of videos, i.e., fused features and SIFT features. In this paper, the two types of features of training samples are trained separately to generate two classification models, and the two types of features are extracted from test samples and input to the two classifiers, respectively, and the two multicategory classifiers get their respective voting results according to the test sample features, and the classification voting results are summed and fused to get the final cumulative value of classification votes, and the category corresponding to the largest cumulative value is chosen as the final decision category.

4. Analysis of Results

4.1. Technologies Feature Fusion and Behaviour Classification. Although the proposed HM-FT features can refine the densely sampled points to the motion boundary region of the actor, the proposed method lacks the necessary corrective measures for the foreground trajectory drift caused by camera motion. For this reason, we adopt the strategy of fusing HM-FT features with IDT features to weaken the above problem. IDT is a modified version of DT features, which can provide a reasonable estimation of camera motion information and correct the dense trajectory by this, weaken the effect of camera motion on feature performance, and make the trajectory features more focused on describing the actor from a global perspective. The object of the research in this article can only be humans and cannot be used for the research of other animals. Specifically, the IDT feature first assumes a single-strain transformation relationship since the image changes between adjacent frames are relatively weak. Subsequently, the camera motion estimation problem is solved by computing the single-response matrix between adjacent frames.

Figure 3 shows the test confusion matrix on the sub-JHMDB dataset; although the number of test samples for the same behaviour is different in the three subsets, the HM-FT features are consistently accurate for behaviours such as "playing golf" and "shooting a basketball." The accuracy of HM-FT features for behaviours such as "playing golf" and "shooting a basketball" is consistently good. Also, the two behaviours "climbing stairs" and "walking" can be easily confused with each other. Compared with the latter, the trajectory of "climbing stairs" tends to increase, but since the camera always adjusts the target to the centre of the lens, the difference becomes too weak to be used as effective discriminative information.

From the above confusion matrix, it can be inferred that although the HM-FT features effectively suppress the interference of background trajectories on behaviour recognition, however, the proposed method is not fully effective for two behaviours with highly similar motion patterns. Future work will focus on identifying motion-related interactive objects in the scene to provide more necessary semantic information for the behavioural process and effectively improve the robustness of the overall recognition framework.

As can be seen from Figure 4, the IDT features exhibit higher recognition accuracy than other trajectory features on the sub-JHMDB and Penn Action datasets, outperforming the HM-FT features by 2.3% and 3.4%, respectively. As an improved version of DT, the HM-FT feature improves the recognition rate of the original algorithm by 10.1% and 6.2% on the two datasets, respectively, which indicates that the recognition performance of the original DT significantly improved after effectively filtering out background trajectory interference. This chapter also uses two state-of-the-art saliency detection methods to generate separate masks and tests the recognition performance of the corresponding trajectory features on both datasets based on the above framework. However, the recognition accuracies of both methods are significantly lower than those of HM-FT using multiscale hybrid masks, which may be attributed to the lack of handling of failed saliency detection. In fact, due to the inherent challenges of saliency detection and the nature of the behavioural video, which does not necessarily contain significant moving subjects in the video frames, the saliency detection algorithm is insufficient to provide reliable a priori information for trajectory features without any auxiliary strategies.

Also, the fusion strategy of HM-FT and IDT achieves better behaviour recognition accuracy than the other five methods on both datasets, reaching 68.3% and 93.3%, respectively. It is concluded that the proposed feature fusion framework can effectively exploit the complementarity between the two trajectory features to effectively improve the recognition performance of the overall framework.

The recognition results of different trajectory features were compared for each behaviour category on both datasets. For the sub-JHMDB dataset, to visualize the comparison results, the recognition accuracy of a behaviour category is defined as the quotient of the number of samples that have been correctly classified in the 3 subsets and the number of all tested samples. To visualize the computational complexity of the proposed HM-FT features, its performance is compared with three trajectory feature extraction methods in three aspects, including the time spent on processing video frames, the average number of trajectories per video clip, and the recognition accuracy. In the experiments, 12 video clips were randomly extracted from the sub-JHMDB dataset with the resolution of 320×240 and 14 video clips were extracted from the Penn Action dataset with the minimum resolution of 480×270 and the maximum resolution of 480×393. The test results are shown in Figure 5.

The proposed HM-FT feature further filters out invalid sampling points by a multiscale hybrid mask to produce a minimum number of dense trajectories, which further reduces the computational cost required to track the sampling points compared with DT-MB. However, the final computational consumption of HM-FT is larger than that of the original DT for both datasets due to the additional computational cost required for motion foreground detection in the proposed scheme. However, as can be seen in Figure 5, the disadvantage of HM-FT in terms of computational consumption is further weakened as the image resolution increases and more invalid sampling points are filtered out, and HM-FT significantly improves the recognition accuracy of DT. In terms of trade-offs, the effective choice tends to improve the recognition rate of behaviours at a smaller computational cost.

4.2. Parameter Sensitivity Evaluation Results. The relationship between the performance of HM-FT and the two parameters in the compensation scheme is shown in Figure 6. Overall, increasing the number of benchmarks of trajectories from 0 to 5 on both datasets improves the overall recognition performance, while increasing the number of benchmark trajectories from 5 to 10 leads to significant performance degradation. In the case of $p < 8$ and $d < 5$, increasing the

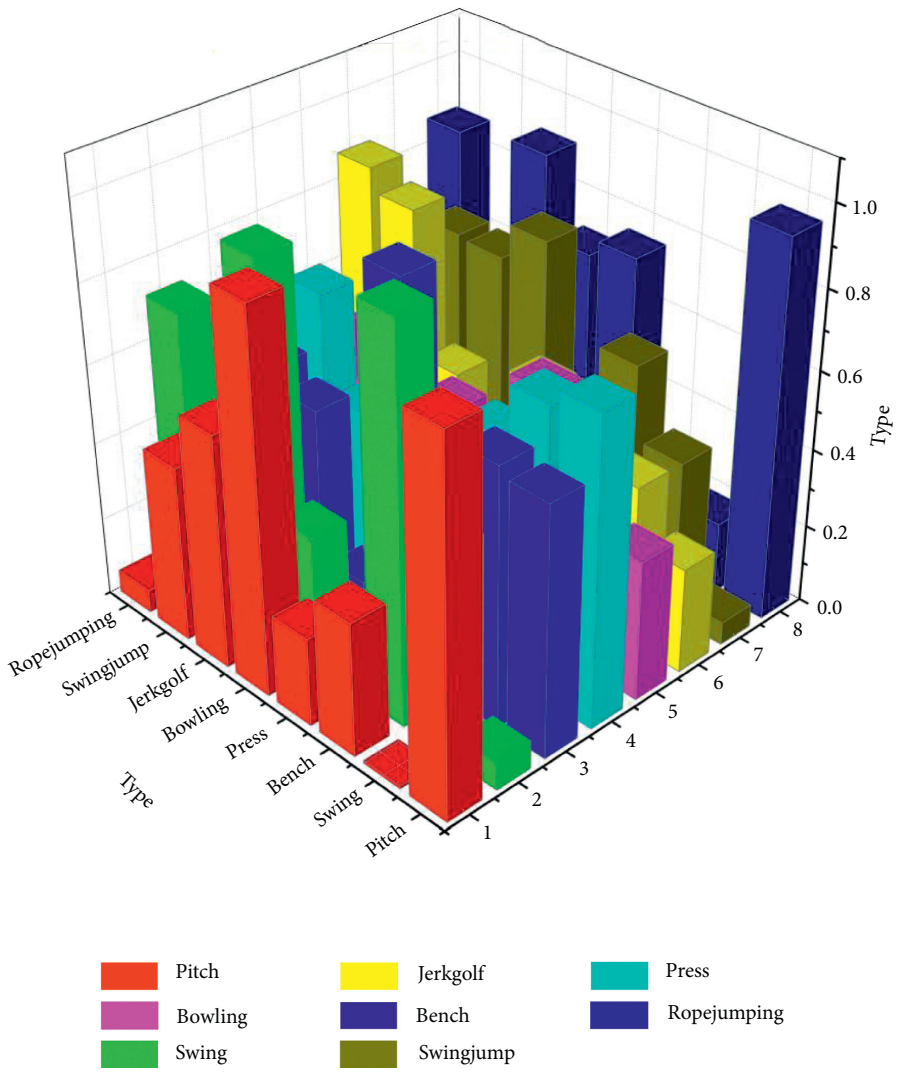

Figure 3: Test confusion matrix on two datasets.

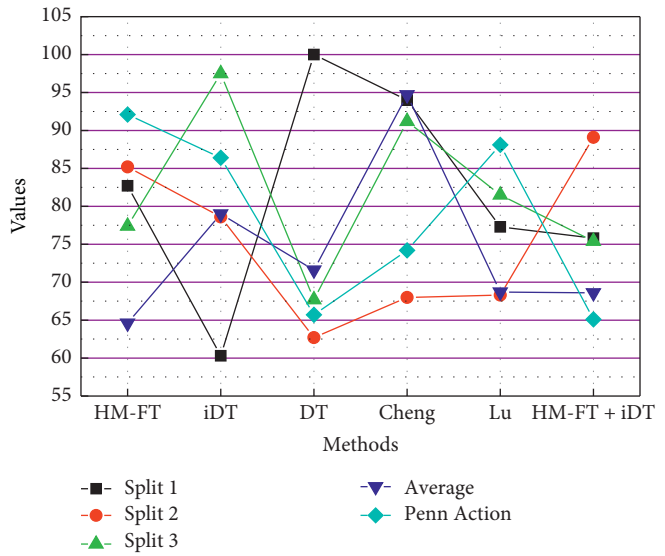

Figure 4: Comparison of the overall recognition performance of different algorithms.

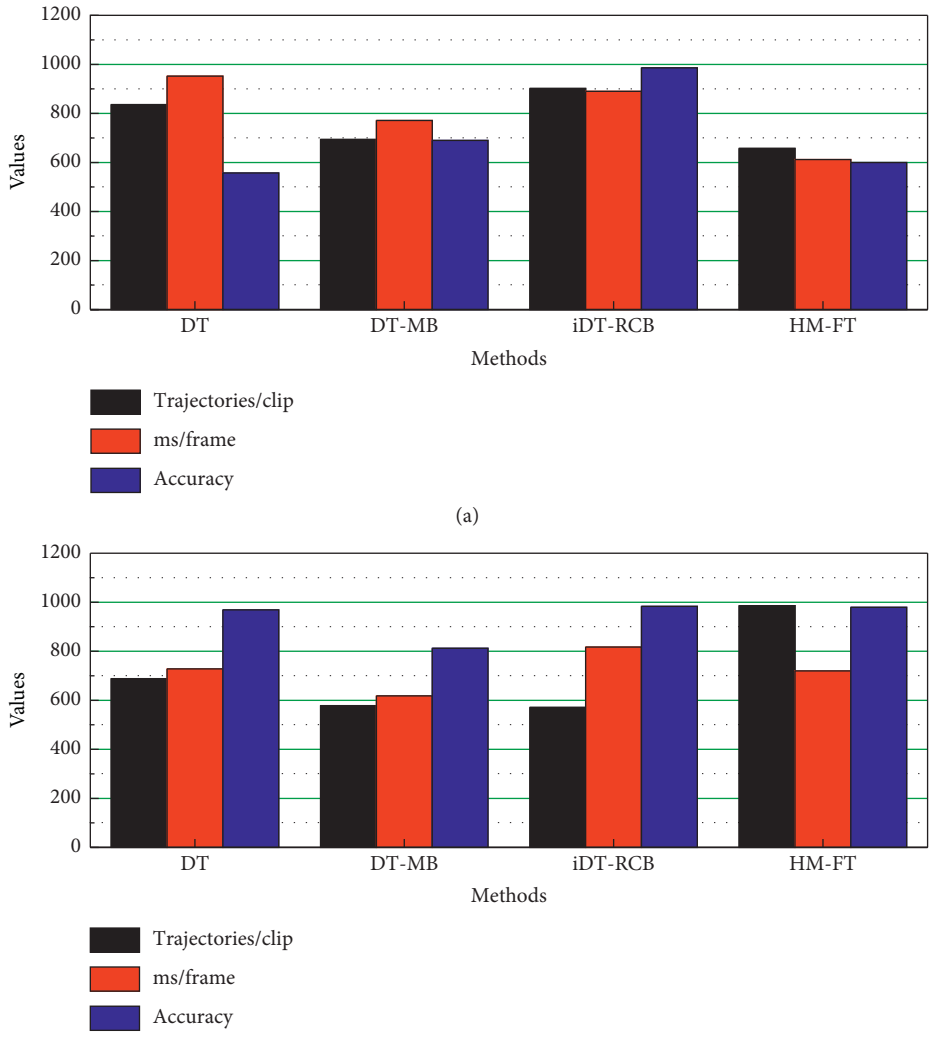

Figure 5: Comparison of the computational complexity of trajectory feature extraction methods: (a) Penn Action; (b) sub-JHMDB.

number of benchmarks of sampled points improves the performance, probably because samples with foreground detection bias are reasonably corrected. Ultimately, this chapter sets the two parameters to $p = 8$ and $d = 5$, respectively, to achieve a good compromise between performance and computation.

To obtain trajectories closely related to the actor and filter out many invalid background trajectories, a generated multiscale hybrid mask is used to refine the original densely sampled points. The hybrid mask is generated by a weighted sum of the weak saliency map optimized by the simultaneous update mechanism of the beta cellular automaton and the strong saliency map obtained using the MKB method. The co-optimization strategy is used to ensure that the foreground detection results are more reasonable and effective. Finally, the necessary compensation scheme is designed to improve the fault tolerance of the proposed features. The experimental results show that the HM-FT features effectively improve the recognition accuracy of the original DT. Also, the discriminative performance of the overall recognition framework can be significantly improved by using a feature fusion strategy. The accuracy of the algorithm in this paper has been improved by 5% compared with the previous research results, and the improvement is very large.

The proposals generated by the temporal behaviour nomination method are applied to temporal behaviour detection to further evaluate the performance of the proposed method. Behaviour detection not only locates the boundaries of actions but also identifies all action classes. Therefore, the proposals generated by the methods in this chapter are sent to the action classifier in the CBR network for behaviour detection experiments. mAP is evaluated by proposals using an average of 200 proposals generated per video with tIoU thresholds of {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7}. The results of the temporal behaviour detection are shown in Figure 7. From Figure 7, after adding the attention-guided network, the behaviour detection effect has a large improvement at low IoU requirements; at high tIoU requirements, the effect improvement is not obvious. At the

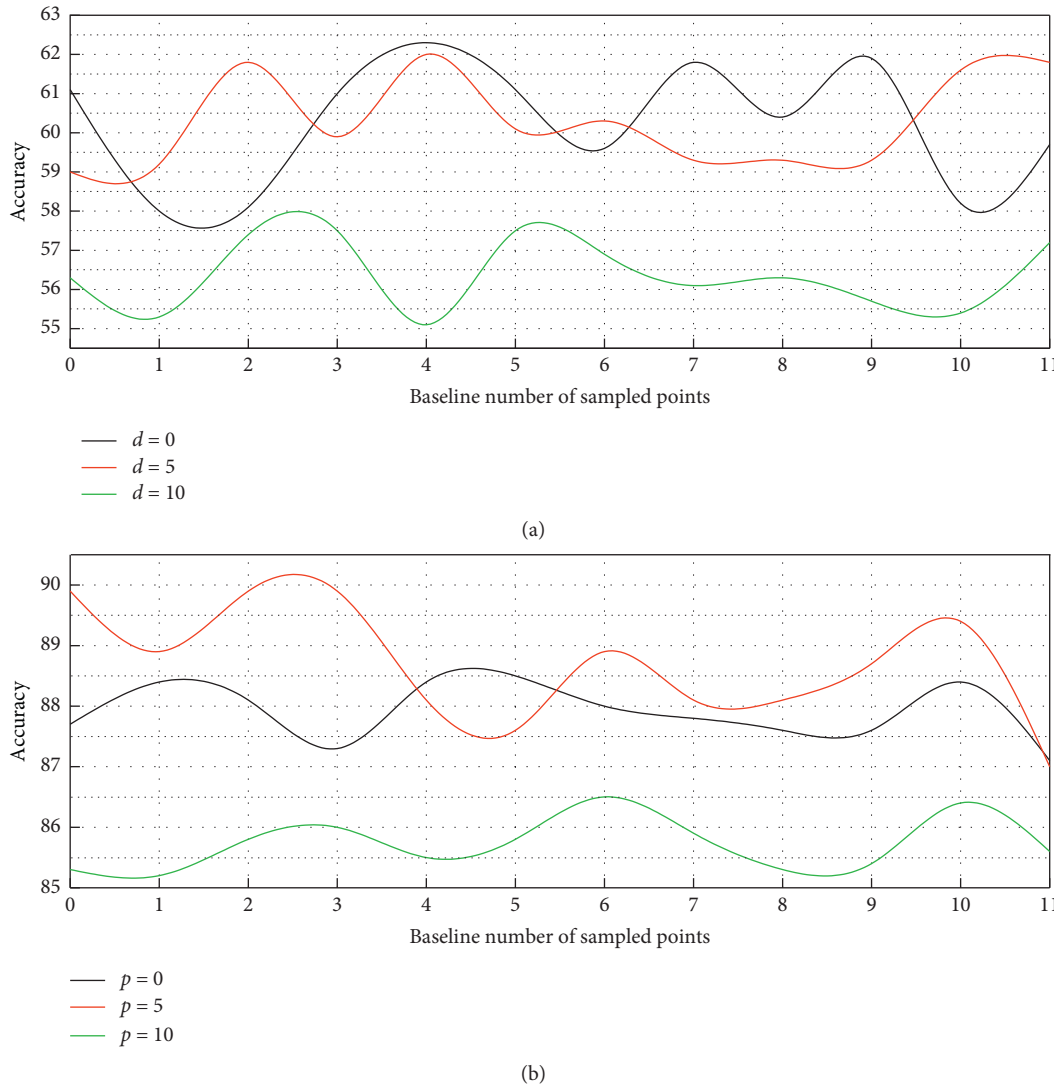

Figure 6: Performance as a function of parameters: (a) p and (b) d.

same time, it is compared with the current behavioural detection algorithms based on weakly supervised learning. From the results in Figure 7, the method in this chapter is superior to other weakly supervised learning methods at tIoU = 0.5. This is because the method in this chapter uses a fully supervised and attention mechanism strategy, which is superior to a certain extent to the weakly supervised learning methods.

We generally let the algorithm automatically modulate the parameters. To verify the recognition effect of the algorithm in this paper, experiments are conducted on KTH, UCF Sports, and Hollywood datasets with dense trajectory methods in two main aspects; finally, the experimental results are obtained by comparing the algorithm in this paper with different literatures. The improved dense trajectory action recognition method in this paper improves the SURF algorithm by introducing dynamized constructive pyramids and rBRIEF feature descriptors, uses the improved SURF algorithm to optimize the optical flow to remove the effect of camera motion, and then introduces an improved feature fusion method to improve the accuracy of recognition. The original dense trajectory uses the coding method of BOW, so to compare the effect of the improved SURF and the improved feature fusion method on the recognition rate when using the same coding method, the coding method in the improved method of this paper is set to BOW, and other parameters are set unchanged; then, Figure 8 shows the accuracy rate of the improved dense trajectory action recognition method of this paper and the original dense trajectory comparison.

From Figure 8, we can see that the improved dense trajectory motion recognition method using BOW coding does not improve the accuracy much for the KTH dataset, mainly because the background in the KTH dataset is simple and the camera is fixed, so the accuracy improvement is not obvious. For the UCF Sports and Hollywood2 datasets, the accuracy is improved by 1.3% and 2.5%, respectively, mainly because there are more camera motions in these two datasets, and the improved dense trajectory motion recognition algorithm can effectively suppress the effect of camera motions.

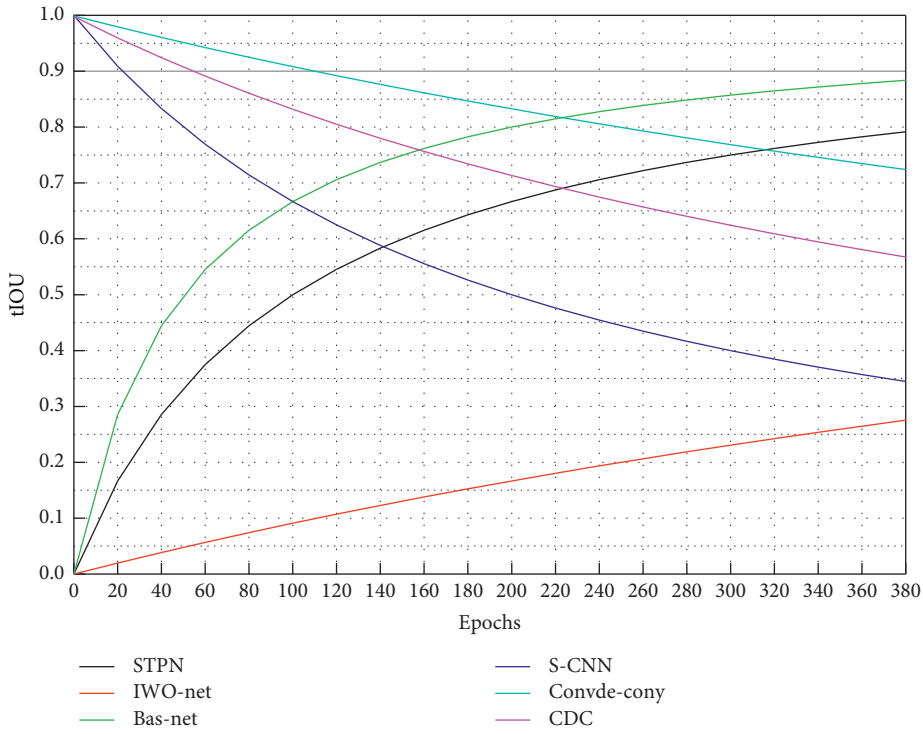

Figure 7: Behaviour detection results.

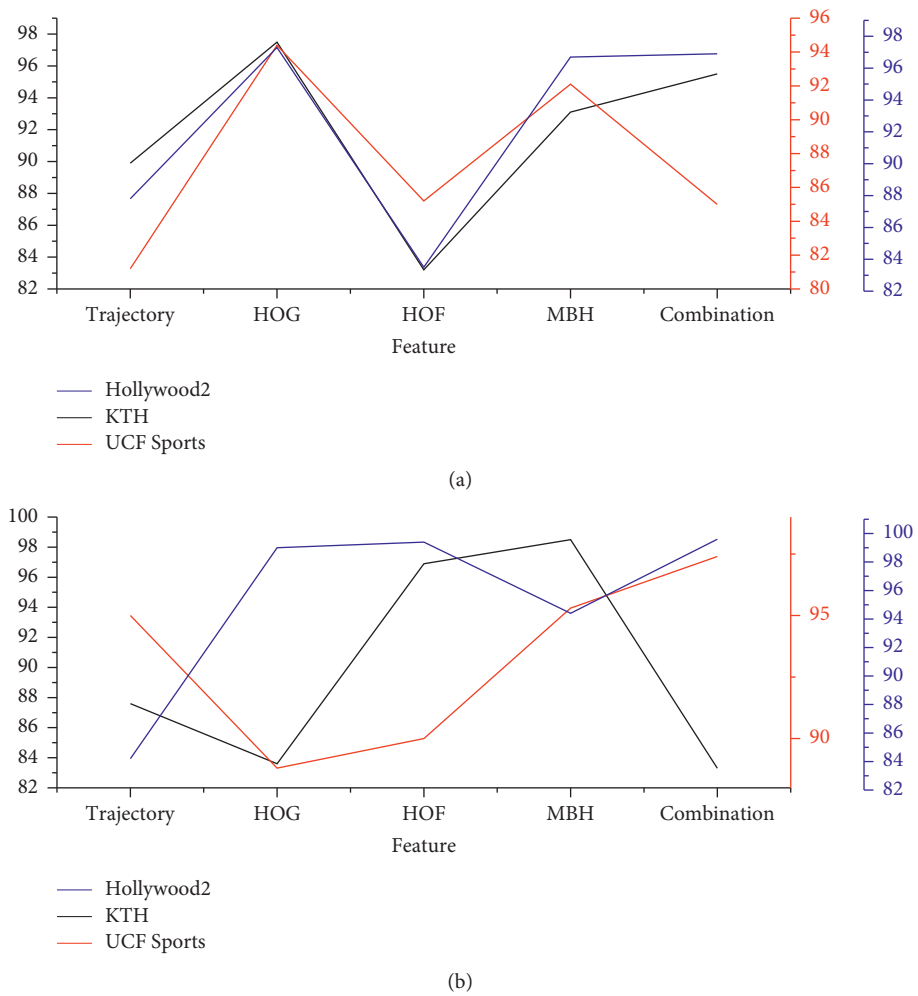

Figure 8: Comparison of improved dense trajectory and dense trajectory: (a) dense track; (b) improved dense trajectory.

The improved feature fusion method is proposed, followed by several common feature reduction and clustering methods, as well as the encoding methods of the BOW model and Fisher Vector, and finally, the human motion is classified and recognized. The improved algorithm in this chapter experiments on KTH, UCF Sports, and Hollywood datasets. Through the experiments, the improved camera motion estimation is shown to improve the recognition results, and the encoding methods of the BOW model and Fisher Vector are shown to affect the action recognition results. The efficiency of the action recognition algorithm based on the improved feature fusion in this chapter is verified. From the experimental results, the algorithm in this chapter can better remove the interference information of camera motion and improve the accuracy rate of action recognition.

5. Conclusion

Due to the spatial arbitrariness of human motion posture, it is difficult to comprehensively describe the motion posture by single feature information, while the complex information of multiple features although comprehensive in the description but computationally intensive and redundant. To describe human motion posture accurately and effectively, this paper proposes a multifeature fusion human motion posture feature model based on four features as human motion posture feature descriptors, and the extraction of four features can obtain 39-dimensional feature information, which is used for fusion optimization. The genetic algorithm is used as a feature vector optimization algorithm based on the advantages of efficient searchability and robustness, and its binary initial coding method is convenient for feature selection and expression, and the mean-variance ratio is used as the basis for constructing a differentiable fitness function among multiple categories based on the characteristics of similar and intercategory feature data, to achieve large intraclass stability and interclass variation of the optimally selected features in multiple categories. In turn, the independence of each motion pose between classes is ensured. We only change one of the same parameters, and the other parameters remain the same. The optimized features based on the genetic algorithm can help to improve the recognition efficiency and reduce the computational complexity. In terms of feature fusion and action recognition, we propose a weighted feature fusion method to express the features and recognize the actions, mainly for the problem of low recognition rate of extracted single features and directly combined features. The traditional dense trajectory-based action recognition method is also studied, and the experimental results are compared and analysed with the improved dense trajectory-based action recognition method proposed in this paper, and the results and experimental statistics show that the improved dense trajectory-based action recognition method proposed in this paper has certain advantages.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Consent

Informed consent was obtained from all individual participants included in the study references

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] F. Kong and Y. Wang, "Design of computer interactive system for sports training based on artificial intelligence and improved support vector," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 5, pp. 6165–6175, 2019.

[2] B. Xue and T. Liu, "Research on emotional model of sports arena based on artificial intelligence emotion calculation," *Cluster Computing*, vol. 22, no. 6, pp. 14927–14933, 2019.

[3] D. C. Angus, "Randomized clinical trials of artificial intelligence," *Jama*, vol. 323, no. 11, pp. 1043–1045, 2020.

[4] Y. Lu, "Artificial intelligence: a survey on evolution, models, applications and future trends," *Journal of Management Analytics*, vol. 6, no. 1, pp. 1–29, 2019.

[5] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, and K.-I. Shimizu, "Machine learning for catalysis informatics: recent applications and prospects," *ACS Catalysis*, vol. 10, no. 3, pp. 2260–2297, 2019.

[6] J. Odry, M. A. Boucher, P. Cantet, S. Lachance-Cloutier, R. Turcotte, and P. Y. St-Louis, "Using artificial neural networks to estimate snow water equivalent from snow depth," *Canadian Water Resources Journal/Revue Canadienne Des Ressources Hydriques*, vol. 45, no. 3, pp. 252–268, 2020.

[7] G. Nagarajan, R. I. Minu, and A. Jayanthila Devi, "Optimal nonparametric bayesian model-based multimodal BoVW creation using multilayer pLSA," *Circuits, Systems, and Signal Processing*, vol. 39, no. 2, pp. 1123–1132, 2020.

[8] C. Prakash, R. Kumar, and N. Mittal, "Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges," *Artificial Intelligence Review*, vol. 49, no. 1, pp. 1–40, 2018.

[9] C. Wang, Z. Li, N. Dey et al., "Histogram of oriented gradient based plantar pressure image feature extraction and classification employing fuzzy support vector machine," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 4, pp. 842–854, 2018.

[10] E. Gültekin, H. İ. Çelik, S. Nohut, and S. K. Elma, "Predicting air permeability and porosity of nonwovens with image processing and artificial intelligence methods," *The Journal of The Textile Institute*, vol. 111, no. 11, pp. 1641–1651, 2020.

[11] H. J. Escalante, S. Rodríguez-Sánchez, M. Jiménez-Lizárraga, A. Morales-Reyes, J. De La Calleja, and R. Vazquez, "Barley yield and fertilization analysis from UAV imagery: a deep learning approach," *International Journal of Remote Sensing*, vol. 40, no. 7, pp. 2493–2516, 2019.

[12] C. Stachl, F. Pargent, S. Hilbert et al., "Personality research and assessment in the era of machine learning," *European Journal of Personality*, vol. 34, no. 5, pp. 613–631, 2020.

[13] L. Ma and B. Sun, "Machine learning and AI in marketing-connecting computing power to human insights,"

International Journal of Research in Marketing, vol. 37, no. 3, pp. 481–504, 2020.

[14] H. Babajanian Bisheh, G. Ghodrati Amiri, M. Nekooei, and E. Darvishan, "Damage detection of a cable-stayed bridge using feature extraction and selection methods," *Structure and Infrastructure Engineering*, vol. 15, no. 9, pp. 1165–1177, 2019.

[15] P. Vračar, E. Štrumbelj, and I. Kononenko, "Automatic attribute construction for basketball modelling," *Knowledge and Information Systems*, vol. 62, no. 2, pp. 541–570, 2020.

[16] N. D. Schilaty, N. A. Bates, S. Kruisselbrink, A. J. Krych, and T. E. Hewett, "Linear discriminant analysis successfully predicts knee injury outcome from biomechanical variables," *The American Journal of Sports Medicine*, vol. 48, no. 10, pp. 2447–2455, 2020.

[17] F. De Grove, K. Boghe, and L. De Marez, "(What) can journalism studies learn from supervised machine learning?" *Journalism Studies*, vol. 21, no. 7, pp. 912–927, 2020.

[18] M. Inoue, S. Inoue, and T. Nishida, "Deep recurrent neural network for mobile human activity recognition with high throughput," *Artificial Life and Robotics*, vol. 23, no. 2, pp. 173–185, 2018.

[19] M. A. K. Quaid and A. Jalal, "Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm," *Multimedia Tools and Applications*, vol. 79, no. 9-10, pp. 6061–6083, 2020.

[20] H. Salah, I. Al-Omari, J. Alwidian, R. Al-Hamadin, and T. Tawalbeh, "Data streams curation for better machine learning functionality and result to serve IoT and other applications: a survey," *Journal of Computer Science*, vol. 15, no. 10, pp. 1572–1584, 2019.

[21] G. Rolan, G. Humphries, L. Jeffrey, E. Samaras, T. Antsoupova, and K. Stuart, "More human than human? Artificial intelligence in the archive," *Archives and Manuscripts*, vol. 47, no. 2, pp. 179–203, 2019.

[22] T. Schaffter, D. S. M. Buist, C. I. Lee et al., "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms," *JAMA Network Open*, vol. 3, no. 3, p. e200265, 2020.