Received August 18, 2020, accepted September 7, 2020, date of current version November 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023755

Motion Recognition Algorithm in VR Video Based on Dual Feature Fusion and Adaptive Promotion

KUNNI HAN
School of Journalism and Communication, Qingdao University, Qingdao 266071, China
Institute for Research on Portuguese-Speaking Countries, City University of Macau, Macau 999078, China
e-mail: hkn0532@sina.com

ABSTRACT VR video recognition in complex environment, a motion recognition algorithm based on two-feature fusion and adaptive enhancement is proposed to solve the problems of inaccurate target position, target drift and recognition error caused by the vulnerability to light change, target rotation and occlusion. First, based on the spatio-temporal context (STC) mechanism, image sequence features are extracted through spatio-temporal context relationship and visual system characteristics to reduce the influence of light changes and occlusion on behaviors. Secondly, reliable feature point tracks are obtained through image feature point tracking and background track cutting, and a rich set of action descriptors (AD) are calculated from which local motion information, shape and static appearance information of the track are retained. After that, the principal component analysis operator is introduced to define the double feature fusion rules, and the STC feature and AD feature are combined to form a more accurate and complete feature representation. Finally, adaptive boosting algorithm (ABA) is used to train the classification through the new features obtained and complete the decision judgment of behavior and action. The experimental results show that the proposed algorithm has higher recognition accuracy and robustness compared with the current commonly used motion recognition methods, and can better adapt to complex background and motion changes.

INDEX TERMS Motion recognition algorithm, VR video, dual feature fusion, adaptive promotion.

I. INTRODUCTION

In recent years, behavior recognition in video has become an important task in the field of computer vision, and it is of great significance to video surveillance, video information retrieval, human-computer interaction and other work [1]–[4]. With the continuous development of various behavior recognition algorithms, behavior recognition tasks in videos have made great progress [5]. It is very important for behavior recognition and understanding to extract features that can represent human movement reasonably from image sequences.

Existing behavior recognition algorithms can be roughly divided into manney-based methods, global feature-based methods and local feature-based methods [6]–[8]. The behavior recognition algorithm based on local features has become a widely used and effective method [9], [10]. Common local descriptors include spatial temporal interest points (STIP), spatial temporal gradient (HOG) [11], dense trajectory characteristics [12] (DT), etc. The inference method based on graph model is widely used in the recognition method based on human body model. This method can be broadly divided into the method based on generative models (GM) [13] and the method based on discriminative models (DM) [14], [15]. Chao *et al.* [16] used conditional random field model (CRF) for behavior recognition. CRF does not need to model observations, so it avoids the assumption of independence and can satisfy the interaction between states and observations on a long time scale. Lazarowska *et al.* [17] proposed an action decision algorithm for feature point trajectory. Classification learning is carried out by combining different features, but the limited information of the action described by a single local feature makes the action difficult to understand. Mabrouk *et al.* [18] designed an effective behavior recognition method for dense trajectory characteristics. The curve is formed by the characteristic points of the optical flow field sampling, and the position vector of the curve and the three information of the neutron block of the curve are respectively measured as the action representation. However, the dense trajectory results in many redundant features and the calculation cost is too high. Ding *et al.* [19] proposed a posture sequence state maneuver discrimination technique. However, this method is easy to be disturbed by the environment and has low stability. Meng *et al.* [20] extracted both

The associate editor coordinating the review of this manuscript and approving it for publication was Zhihan Lv.

local action and static features from the video. Since these two types of original features contained noise, stable action features and static features without noise were obtained through statistical methods. This algorithm achieves good recognition effect on YouTube real behavior data set. Singh *et al.* [21] studied the recognition of real human behavior in feature films. Video clips containing actions to be recognized were collected from movie scripts to form a complex behavior data set, and a motion representation and behavior classification algorithm based on local spatio-temporal characteristics, spatio-temporal pyramid model and multi-channel nonlinear SVM was proposed. However, due to the complexity of the data set, the time is more serious. Xu *et al.* [22] proposed to model spatio-temporal context information in a hierarchical manner. There are also serious time consuming problems. Han *et al.* [23] proposed a motion representation method based on the neighborhood shape of spatio-temporal features, aiming at the defects of the existing Bag-of-Words (BOW) model in expressing the spatio-temporal relations of features. Hall *et al.* [24], based on the spatiotemporal context (STC) theory, formed a highly robust action expression. However, in complex scenes, the recognition effect based on STC is limited, and the recognition rate of occlusion and illumination change is not good.

The above methods all adopt the local image feature as the motion representation. These features can be extracted from the spatio-temporal interest points [25], [26], and can also be extracted from the feature point trajectory [27], [28]. However, the vast majority of current behavior recognition research only considers one of these characteristics, and there are few reports on the combination of these two characteristics. In addition, in a complex environment, motion recognition in VR video is vulnerable to illumination change, target rotation and occlusion, resulting in inaccurate target position, target drift and recognition error.

Based on this, a two-feature fusion and adaptive lifting motion recognition algorithm is proposed. The STC features are extracted by STC relation and visual system. A series of trajectory descriptors AD, which are invariant to scale, translation, rotation, etc., are calculated to effectively improve its robustness and enable it to realize recognition of occlusion, rotation and other targets. In order to obtain more accurate and comprehensive feature representation, principal component analysis (PCA) was used to effectively fuse STC features and CNN features. At the same time, according to the new features of fusion, the adaptive enhancement algorithm (ABA) is used for classification learning, so as to better recognize and understand the movement and accurately identify the movement.

II. BASIC THEORY OF HUMAN BEHAVIOR RECOGNITION

A. HUMAN MOTION ANALYSIS

The movement of the human body is a complex process, which is affected by many factors such as gender, age, health status, and natural environment. The differences in the data of different intuitive behaviors may not be obvious, such as standing, walking, running and other behaviors usually have large differences, but walking, going upstairs, going downstairs and other behaviors have great similarities. The same behavior has different motion states in different environments. For example, walking can be divided into fast walking, medium speed walking and slow walking according to speed; according to the change of speed, it can be divided into accelerated walking, constant speed walking and decelerated walking. In the face of complex and changeable human behaviors, it is impossible and unnecessary for us to recognize all behaviors, only the main behaviors that are closely related to life.

Human walking is a periodic and repetitive movement. When the human body is walking, the two feet are alternately coordinated to move forward, and the front foot is in contact with the ground at the same time, and the forward reaction force is generated by the kicking motion of the back foot, which pushes the human body forward and the center of gravity also moves forward. As the center of gravity of the person moves forward, after the current foot hits the ground, the contact between the ground and the sole of the foot will produce a backward force, and the human body will slow down. At the same time, the back foot retracts and moves forward. Alternate back feet to complete a walking process. During a gait cycle, the angle of the thighs changes periodically. With the bending and extension of the legs, the center of gravity of the human body moves forward in a spiral shape.

The walking process of a person is a rhythmic movement, and a complete walking cycle is defined as the step from the heel of one side to the heel of the same side. This process can be divided into the support phase and the swing phase. The support phase refers to the period when the foot touches the ground, and the swing phase refers to the period when the foot is in the air. The supporting phase and the swing phase can be divided into multiple sub-phases. Medically, a walking process is divided into eight phases. The support phase is specifically divided into the initial landing phase, load response phase, mid-support phase, and end-support phase. The swing phase is divided into pre-swing, early swing, mid-swing, and end swing.

B. HUMAN BEHAVIOR RECOGNITION SYSTEM STRUCTURE

The human behavior recognition system is essentially a system that classifies and predicts input information. The input information is usually sensor data related to human behavior, and the output is the behavior to be recognized. Human behavior recognition is usually based on pattern recognition classification models, so the method of human behavior recognition can be explored from pattern recognition. Pattern recognition can be understood as discovering and extracting features from data samples, and using the features for classification and recognition in the classification model. According to the different classification and recognition methods, pattern recognition can be further divided into supervised learning and unsupervised learning. Among them, supervised learning uses samples with category labels to train the model, finds the correlation between sample features and labels, and enables the model to obtain the ability to predict the category of unknown samples. The classification methods of supervised learning include support vector machine, neural network, and Bayesian network and so on. Unsupervised learning uses the structural relationship of the data itself to classify and identify unknown models, such as clustering, without any training samples. Human behavior classification and recognition models are usually implemented in a supervised manner, and the system structure is shown in Figure 1.

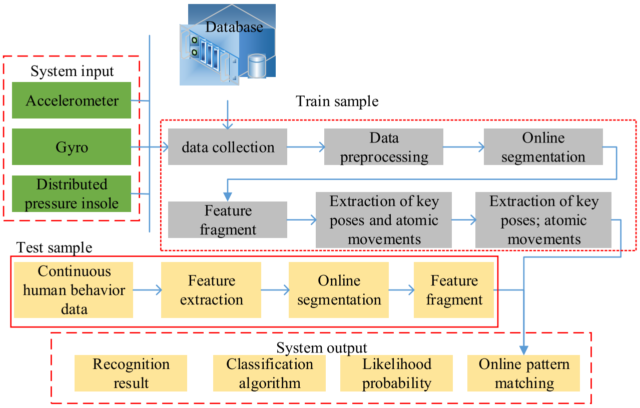


FIGURE 1. Human behavior recognition system structure.

The above figure is the overall structure diagram of human behavior recognition. Among them, the accelerometer, gyroscope, distributed plantar pressure insole and other sensors in the input layer of the system are responsible for sensing data related to human behavior and transmitting it to the data collection node. The raw data collected by the sensor cannot be directly used as the input of the classification model. It must go through the process of data preprocessing and feature extraction such as filtering and data segmentation before it can be input into the classification model. Since the structure map is based on the classification and recognition method of supervised learning, the features after feature extraction are divided into two parts: training samples and test samples. The training samples are labeled. After a lot of training, a trained classification model is generated. The characteristic data of the test samples are input into the trained classification model to directly identify the corresponding behavior. In addition, all input samples have corresponding category labels, so that the classification accuracy of the model can be evaluated to understand the classification effect of the classification method.

C. FILTERING AND DATA SEGMENTATION

Sensors such as accelerometers and gyroscopes will inevitably be interfered by noise signals in the process of data collection, including the measurement noise of the sensor itself and the noise generated by human body shaking. In order to remove the interference of noise, the original data is usually filtered. Commonly used filtering methods include moving average filtering, low-pass filtering. Data segmentation is an important part of the data preprocessing process. The data collected by the sensor is a data stream distributed according to time, which cannot be directly extracted and classified. The data usually needs to be segmented. In this way, data of different lengths can be intercepted to the same length, which is convenient for subsequent feature extraction and classification.

The data segmentation method used at this stage is mostly the sliding window segmentation method, which divides the data into a set of continuous data windows of the same length. The key issue of the sliding window segmentation method is how to choose the size of the sliding window. The size of the sliding window is related to the accuracy and effectiveness of the recognition of the entire system. If the window is too large, it will cause multiple actions in a window, the system needs too long waiting time, and the accuracy and effectiveness of the system are reduced; if the window is too small, each window cannot completely contain a motion cycle, and the extracted features cannot reflect the human behavior, which also affects the recognition accuracy.

Sliding windows are divided into overlapping windows and separate windows. Overlapping windows are also called window coverage, which refers to the overlapping part of the data between adjacent windows, while the separated windows do not overlap the data between adjacent windows. Window coverage usually reduces the impact of behavior transition on recognition, so 50% window coverage is used in most studies.

D. FEATURE EXTRACTION AND SELECTION

The pre-processed data cannot characterize human behavior and cannot be directly input into the classification model. It needs to be transformed into characteristic data that can characterize behavior through feature extraction. The feature extraction itself is not general. It is necessary to select appropriate features according to different recognition targets and research methods. In the field of behavior recognition, the extracted features mainly include time-domain features, frequency-domain features, and time-frequency features.

Time-domain features are usually directly calculated from time-domain signals. The amount of calculation is small and the solution process is simple. They are widely used in the study of human behavior recognition. Many time-domain features have clear physical meanings and can reflect the distribution characteristics of signals. Typical time-domain features include mean, root mean square, variance, standard deviation, skewness, kurtosis, interquartile range, and correlation coefficient between any two axes of the sensor. Frequency domain features are also a type of features that are frequently used in human behavior recognition. The frequency domain features of the signal are usually suitable for periodic motion, and the recognition accuracy can be improved by extracting the frequency domain features.

The time-frequency feature is proposed to make up for the lack of time-domain information in the frequency-domain feature. The time-frequency features currently used in the field of human behavior recognition are mainly based on the features obtained by wavelet analysis. Wavelet transform is a kind of time-frequency analysis. It inherits and develops the idea of short-time Fourier transform, and provides a window containing time-domain and frequency-domain information. In the field of behavior recognition, typical time-frequency features based on wavelet transform include wavelet energy, sum of squares of wavelet coefficients of different scales, and so on.

With the continuous introduction of new features, the dimensionality of feature vectors continues to increase. Although the introduction of new features will improve the recognition accuracy to a certain extent, the excessively high feature dimensionality greatly increases the amount of calculation and the time required for calculation becomes longer. In addition, too many redundant features will affect the construction of the system classification model and reduce the performance of the system. Therefore, feature selection has become a necessary step in the classification and recognition process. Feature selection can reduce feature dimensions, reduce computational complexity, and improve classification performance. Figure 2 shows the feature selection process. The original feature set is decomposed into different feature subsets, and then the feature subsets are evaluated using the evaluation function. Those that meet the evaluation criteria are output, and those that do not meet the evaluation criteria are returned, and finally selected out the required characteristics.

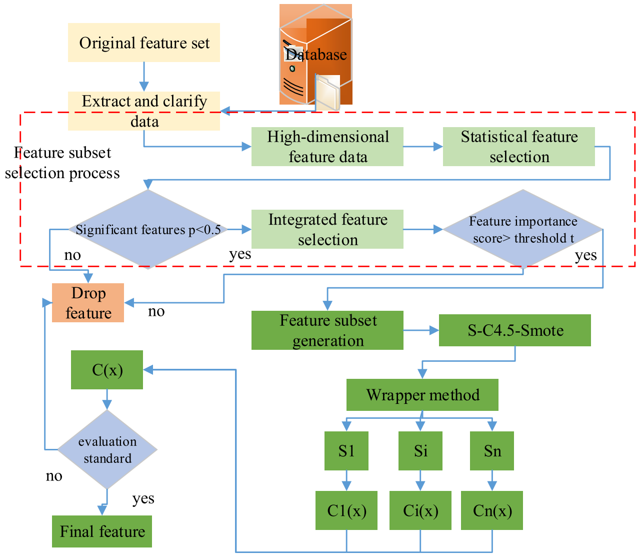

Original feature set
Database
Extract and clarify data
Feature subset selection process
High-dimensional feature data
Statistical feature selection
Significant features p<0.5
no
yes
Integrated feature selection
Feature importance score> threshold t
yes
Drop feature
no
Feature subset generation
S-C4.5-Smote
C(x)
Wrapper method
evaluation standard
no
yes
S1
Si
Sn
C1(x)
Ci(x)
Cn(x)
Final feature

FIGURE 2. **Feature selection process.**

III. MOTION RECOGNITION ALGORITHM BASED ON DUAL FEATURE FUSION AND ADAPTIVE LIFTING

Extracting the STC features of an image sequence through the temporal and spatial context and the characteristics of the visual system can reduce the impact of changes in illumination and occlusion on behavior. Through image feature point tracking and trajectory clipping, a reliable feature point trajectory set is obtained, and AD features are calculated based on these curves, which are used to capture the local motion information, shape and static appearance information of the trajectory. Therefore, these two features are adopted.

A. STC FEATURES

For visual tracking, the local context has a certain feature and its surrounding background area. Featured local backgrounds between consecutive frames have strong spatiotemporal connections. If a feature is blocked, it causes a big change in its appearance. However, since only a small part of the context area is blocked, the context areas before and after are similar, so the local context of this feature changes less. Therefore, the feature relationship of the next frame can be predicted through the local context of the current frame.

STC is mainly based on the Bayes theory to extract the temporal and spatial information of the object and its local context, and obtain the statistical properties of the object and its nearby low-level features. Using spatio-temporal information and the vision system to obtain the confidence map of the object, the point with the highest confidence is the coordinates of the object, and the feature with the highest likelihood is found. Regard the object tracking as the confidence map $Map(a)$ of the measured object, expressed as:

$$Map(a) = pro(a|b) \tag{1}$$

In the formula, variable a is the characteristic position and b is the tracking object. Let the current feature center point be a°, then the context feature is expressed as follows:

$$A^c = \{C(c) = (gray(c), c)|c \in \Psi_c(a^o)\} \tag{2}$$

In the formula: gray(c) is the gray value at c, and $\Psi_c(a^o)$ is the variable a^o of the feature center and the surrounding context area. By combining the context information with the tracking process, a further expression of the confidence map in equation (1) can be obtained:

$$\begin{aligned} Map(a) &= \sum_{C(c)\in A^c} pro(a, C(c)|b) \\ &= \sum_{C(c)\in A^c} pro(a|C(c)|b)pro(C(c)|b) \end{aligned} \tag{3}$$

In the formula, the variable $pro(C(c)|b)$ is the context prior probability, which represents the appearance feature. The variable $pro(a|C(c)|b)$ is the conditional probability, which represents the feature location and spatial connection.

In STC, it is mainly composed of prior probability, space model and spatiotemporal model. First, construct the prior probability through the local context:

$$pro(C(c)|b) = gray(c)G_\xi(c - a^o) \tag{4}$$

Among them, variable G_ξ is the Gaussian weighted operator. Generally speaking, the closer to a°, the more important the tracking feature and the greater the weight.

For the space model, it is represented by embedding a radial asymmetric operator in $pro(a|C(c), b)$:

$$pro(a|C(c), b) = \Gamma(a - c) \tag{5}$$

The variable $pro(a|C(c), b)$ reflects the distance and direction between the characteristic positions a and the local background c. Therefore, it represents the spatial relationship between the feature and the local area.

According to the calculated formulas (4) and (5), the confidence map $Map(a)$ can be evolved:

$$\begin{aligned} Map(a) &= \sum_{C(c)\in A^c} pro(a|C(c)|b)pro(C(c)|b) \\ &= \sum_{C(c)\in A^c} \Gamma(a-c)gray(c)G_\xi(c-a^o) \\ &= \Gamma(a) \otimes (gray(c)G_\xi(a-a^o)) \end{aligned} \tag{6}$$

Among them, symbol $\otimes$ is the convolution operation. Through STC feature description, it uses foreground motion features and morphological features to well characterize moving targets and effectively reflect various complex motion attributes. It contains object type, speed, and direction information and expands the confidence map $Map(a)$ obtained by STC into a matrix, expressed as:

$$Matrix = \begin{pmatrix} Map_{11} & Map_{12} & \dots & Map_{1n} \\ Map_{21} & Map_{22} & \dots & Map_{2n} \\ & \dots & & \\ Map_{m1} & Map_{m2} & \dots & Map_{mn} \end{pmatrix} \tag{7}$$

To avoid the influence of light, the image is de-averaged. And through the Hamming window to reduce the frequency interference caused by the edge to the fast Fourier transform.

B. FEATURE POINT TRAJECTORY DESCRIPTOR

Given a VR video sequence, first obtain a reliable feature point trajectory set through image feature point tracking and trajectory clipping, as a spatiotemporal curve. Then calculate the trajectory descriptor based on these curves, which is used to capture the local motion information, shape and static appearance information of the trajectory. The algorithm starts with the extraction of the salient points in space.

1) FEATURE POINT TRACKING

Feature point tracking is to extract feature points in the current frame, and then match these feature points in subsequent frames to predict and estimate the motion trajectory of the feature points. The process of trajectory extraction is based on the pairing of image feature points in consecutive frames. For a video sequence of m frames, mark it as $Video = \{V1, \dots, Vm\}$, establish the feature point matching relationship between V_i and V_{i+1}, and continue the pairing of multiple frames to form the action trajectory of the feature points. In order to reduce the generation of false trajectories caused by incorrect matching, the matching process is uniquely restricted, and the pairing with too large distance is discarded, because most realistic actions will not proceed very fast. Specifically, for any salient point $Point$ in the frame V_i, at most one candidate point $Point$‘ in the frame V_{i+1} can match it. And Point' must be located in the space window around Point. Under this constraint, when the trajectory reaches the camera boundary or is accompanied by a large occlusion, it will automatically terminate. Then the algorithm restarts to track a new trajectory. In addition, in order to further remove unreliable trajectories and reduce the possibility of confusion between long trajectories and continuous actions, such as people standing up and walking, during the tracking process, the length C of any effective trajectory is limited to $C_{min} < C < C_{max}$. In the experiment, set $C_{min} = 5$ and $C_{max} = 25$.

2) TRACK CUT

The trajectories extracted by the above methods are not always useful for behavior recognition. A large number of trajectories come from the background area, which must be removed in order to retain the most relevant trajectories describing human body movements. For this reason, consider the process of trajectory clipping. This paper adopts the method of trajectory pruning based on the trajectory dissimilarity measure and the region of interest detection in a specific time window. Details as follows.

Suppose there are Q tracks $Track = \{T1, \dots, Tm\}$ starting from frame V. For the displacement vectors $vector_i$ and $vector_j$ corresponding to any two trajectories in Track, we calculate the degree of dissimilarity to form a matrix M:

$$M = \sum_{f=1}^{4} ||vector_f^i - vector_f^j|| \tag{8}$$

Then the dissimilarity of trajectory T_i is calculated as $M_i = \sum_{f=1}^{n} vector_{ij}$. This value measures the degree of dissimilarity between this trajectory and all other trajectories starting from frame V in a time window of 5 frames. For video clips with small moving targets, when a trajectory is very similar to other trajectories, it is likely to be extracted from the dynamic background, which is unreliable and must be removed. Therefore, for frame V, calculate an adaptive threshold $M_{threshold}^V = \frac{\alpha}{n}\sum_{f=1}^{n} M_f$, where α is a constant and the value is 1.1. Then remove all trajectories whose dissimilarity is less than $M_{threshold}^V$.

After passing through the above constraints, assuming that there are n_V reliable trajectories crossing frame V, the center of ROI can be obtained by averaging the position coordinates of all reliable trajectories in frame V:

$$\vec{a} = \frac{1}{n}\sum_{f=1}^{n} a_f^i, \quad \vec{b} = \frac{1}{n}\sum_{f=1}^{n} b_f^i \tag{9}$$

In a relatively stable background, this method can obtain better motion positioning results. Then, remove all trajectories located outside of ROI.

The detection result of the region of interest can accurately locate the area where the obvious movement occurs. Different from the ROI detection algorithm based on two-dimensional interest point detection, the motion positioning method in this paper is based on the statistical analysis of the trajectory distribution of characteristic points. It does not require explicit target detection and tracking processes, and has a certain degree of performance for the video in the case of camera motion.

3) TRAJECTORY FEATURE EXTRACTION

Given any two consecutive points *Point* and *Point*‘ on the same trajectory, the displacement vector is *Vector* = PointPoint’. Consider quantizing the modulus length and direction of Vector respectively. For the modulus length, in order to make the quantization result scale-invariant, first use the largest displacement on the same track $||Vector||_{max}$ to normalize $||Vector||$, and then use 4 uniform quantization levels. For the direction, divide the upper semicircle and the lower semicircle into 8 sectors, each 22.5°, as shown in Figure 3. Combining the two quantifications of modulus length and direction, each trajectory can be described as a 32-bin histogram O.

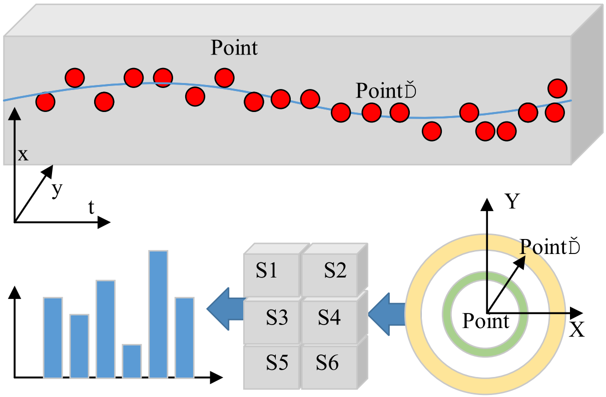


FIGURE 3. Quantify the trajectory and convert it into a histogram.

In order to facilitate the effective representation and processing of video sequences, this paper adopts the bag-of-words representation method. Due to the large difference in the numerical range of these descriptors, in order to overcome the adverse effect of the difference on subsequent operations, consider using scores for data conversion, so that the average value becomes 0 and the standard deviation becomes 1. For each track, connect the standardized O to form the overall description vector of the track.

C. IMAGE ACTION RECOGNITION ALGORITHM

For a single feature, the ability to describe action features is insufficient, especially in complex scenes, which are susceptible to changes in lighting, target rotation, occlusion and other environments, making it difficult to accurately extract robust action features. This paper defines a dual feature fusion mechanism. In order to reduce the impact of lighting changes and occlusion on behaviors, STC technology is introduced to extract STC features of image sequences through temporal and spatial context and visual system characteristics. At the same time, in order to retain the local motion information, shape and static appearance information of the trajectory, the feature point trajectory descriptor AD is introduced. In order to fuse the two extracted features, PCA fusion technology is introduced. PCA forms a set of unrelated main features through linear transformation, which can effectively reduce redundant information while completing feature fusion. According to the obtained fusion dual features, the ABA algorithm is used for classification training to complete the action judgment. The entire action recognition process is shown in Figure 4.

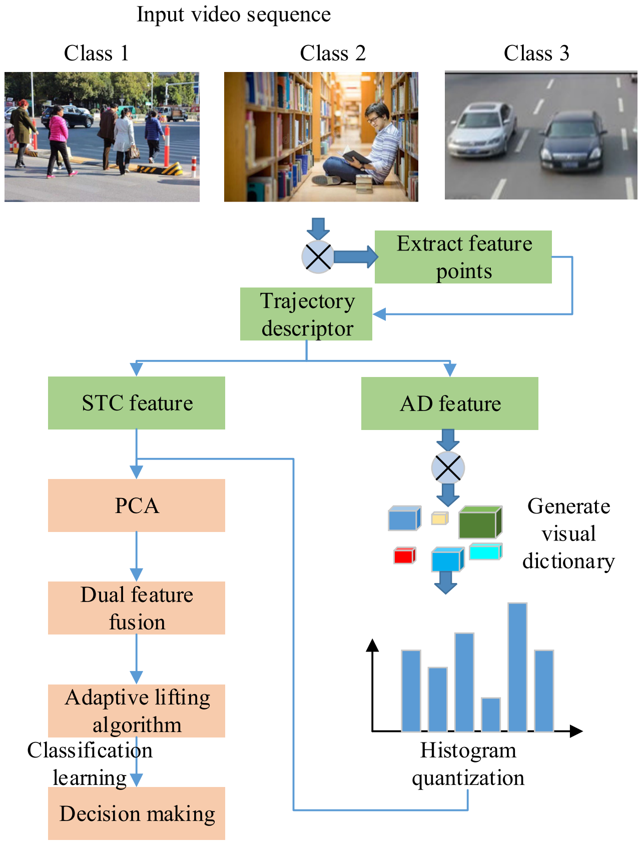


FIGURE 4. The algorithm flow of this article.

1) BASED ON PCA FEATURE FUSION

For the STC features and AD features extracted from the image sequence, they can well capture the local features of the action sequence at a certain point or scene, but they cannot express the action features completely and accurately. In order to obtain more accurate and comprehensive action features, this paper introduces a PCA-based feature fusion algorithm, which combines the extracted STC features with AD features to form a more effective feature representation. In the PCA algorithm, when using new components to reconstruct the original signal, the approximation effect is the best under the minimum mean square error, expressed as:

$$\lambda^2(m) = ||m - \sum_{i=1}^{n} (w_i^T m)w||^2 \tag{10}$$

$$v = M^T m = \{M_1, \ldots, M_n\}^T m \tag{11}$$

$$m' = \sum_{i=1}^{n} v_i M_i \tag{12}$$

Under this constraint, the condition to minimize λ is satisfied:

$$J = \sum_{i=k+1}^{n} m_i^T v_i M_i^T - \sum_{i=k+1}^{n} \lambda_i (m_i^T m_i - 1) \tag{13}$$

PCA is a linear transformation, which forms a set of new uncorrelated main features through linear transformation of information with a certain correlation to the initial state, effectively reducing redundant information. For example, for an image, the new component is used to reconstruct the original image, and the approximation effect is the best under the minimum mean square error using formula (10). Afterwards, the non-correlation between vectors is used to obtain the linear estimation under the minimum mean square error. Finally, formula (13) is used to obtain the final main feature.

2) BASED ON ADAPTIVE LIFTING CLASSIFICATION TRAINING

The ABA algorithm is an iterative algorithm. When the final classification error rate is less than a predetermined threshold, the algorithm stops working. ABA algorithm is designed to train the weak classification iteratively for the same training data set. In each iteration process, the weights of the overall data set are adjusted, and the weights of the weak classification are assigned by using errors. Initially, the weights of each training data are consistent. After each iteration, the data weights are allocated on the basis of reducing the weight for the correct type and increasing the weight for the wrong type. Therefore, after many iterations, the misclassified data is searched out and a new data distribution is obtained for the next weak classifier. Therefore, ABA algorithm is selected in this paper to improve our classification performance. Through the preset H iterations, U weak classifications are generated. Use each weak group to generate the final strong group according to the weight.

The ABA steps are as follows:

1) Input the training data set S, the number of iterations N.

2) Initialize the data weight distribution weight = 1/m, the initial classifier L.

3) Calculate the error λ_t of U_t through weight and weak grouping $U_t = L\,(S, W)$, expressed as:

$$\lambda_t = U_t(r_i(m_i)) \tag{14}$$

If $\lambda_t \geq 0.5$, then the iteration stops, otherwise the execution continues.

4) Calculate the weight of U_t, expressed as:

$$weight_t = (1/2)\ln((1 - \lambda_t)/\lambda_t) \tag{15}$$

5) Update the weight distribution of the data set, expressed as:

$$weight_{t+1}(i) = weight_t e^{-U_t \lambda_t}/\text{Normalized} \tag{16}$$

6) Through iterative operation, a strong grouping $Q(a)$ is obtained:

$$Q(a) = sign(\sum_{i=1}^{U} weight_i r_i(m)) \tag{17}$$

In order to evolve the two classifications of the ABA algorithm into multiple groups, different data category labels class are set in the initialization stage. Therefore, the weight update can be expressed as:

$$weight_{t+1}(i) = weight_t(i, \chi)e^{-U_t \lambda_t \chi}/\text{Normalized} \tag{18}$$

Therefore, the strong classifier Q(a,χ) can be expressed as:

$$\text{Q(a,}\chi\text{)=sign(}\sum_{i=1}^{T} weight_t \lambda_t(\text{a},\chi)) \tag{19}$$

IV. EXPERIMENTAL VERIFICATION

A. EXPERIMENTAL DATA DESCRIPTION

This section uses nine video sequences to evaluate the robustness and accuracy of the proposed recognition algorithm, and these nine video sequences include complex tracking scenes, such as scale change, posture change, lighting change, partial or total occlusion, background interference, etc. These sequence videos are datasets of car, girl, room, PETS, and pedestrian. The detailed information is shown in Table 1.

This section compares the recognition algorithm proposed in this article with seven other classic recognition algorithms. These recognition algorithms are literature [22], literature [23], literature [24], literature [29], literature [30], literature [31], and literature [32].

Table 2 lists the parameter settings of this algorithm. In addition, the difference thresholds and are also set empirically based on actual application scenarios. In this section, α is set to 0.3. β is the maximum duration parameter, which is associated with the target template. To get the best experimental results, the offline template update threshold is set based on the empirical values. The greater β, the greater the difference between the online and offline templates when updating. Conversely, the smaller β, the smaller the difference between the online and offline templates when updating.

The hardware environment of this experiment is: PC, Intel Pentium G630, 2.70GHz, 2G memory; the operating system is Windows7, the development tool is VS2008, and OpenCV2.0 is configured. The programming languages are C language and C++ language.

B. QUALITATIVE ANALYSIS

1) POSTURE AND SCALE CHANGES

Intelligent-Room and Girl-1 video sequences are used to evaluate the recognition performance of the proposed algorithm in dealing with changes in pose and scale. As shown

TABLE 1. Video sequence description.

Video sequence	Video sequence	Resolution	Length (frame)	Disturbing factors	Complexity	Target
Car-scale	Car scale	768×576	114	Scale variation, background clutters.	High	White car
Girl-1	Girl 1	640×480	111	Scale variation, deformation.	Medium	girl
Intelligent-Room	Intelligent Room	320×240	210	Scale variation, deformation, and occlusion.	Medium	man
Girl-2	Girl 2	640×480	150	Scale variation, deformation, and occlusion.	High	girl
Car	Car	320×240	89	Illumination variation, deformation, scale variation.	High	car
Blur-car	Blur car	768×576	118	Blurred vision, scale variation.	Low	white car
PETS-2001	PETS 2001	384×288	149	Scale variation, deformation.	Medium	biker
Pedestrian-1	Pedestrian 1	720×576	104	Background clutters, deformation.	Low	red people
Pedestrian-2	Pedestrian 2	768×576	166	Background clutters, deformation, scale variation.	High	blue people

TABLE 2. Algorithm parameters.

Video sequence	α	β	Δ min-size
Car-scale	17	0.85	600
Girl-1	20	1.68	1200
Intelligent-Room	20	0.71	1100
Girl-2	26	1.33	1050
Car	19	0.45	600
Blur-car	19	0.58	1050
PETS-2001	19	0.82	600
Pedestrian-1	18	0.83	1100
Pedestrian-2	16	0.99	1200

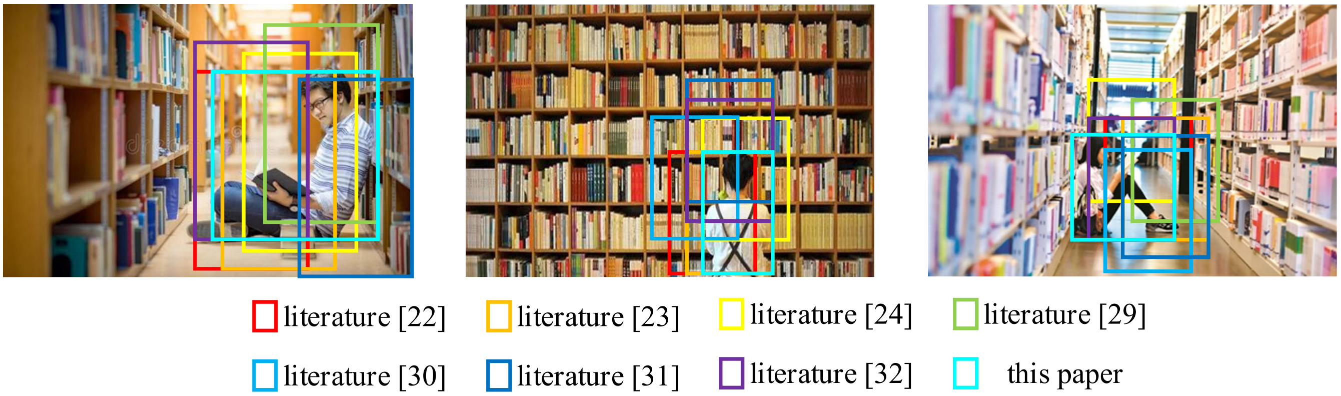


FIGURE 5. Intelligent-room video sequence recognition result.

in Figure 5, in the whole recognition process, only the recognition algorithm proposed in this paper successfully identifies the moving target, while other methods all have a certain degree of drift. Moreover, in the girl-1 video sequence, the moving target presents attitude changes and slight scale changes. The method in this paper is still very accurate in detecting the target, but the other methods all show serious deviation. As shown in FIG. 6, except for the algorithm in reference [24] and the algorithm in this paper, all other algorithms experience recognition drift rapidly. However, in this

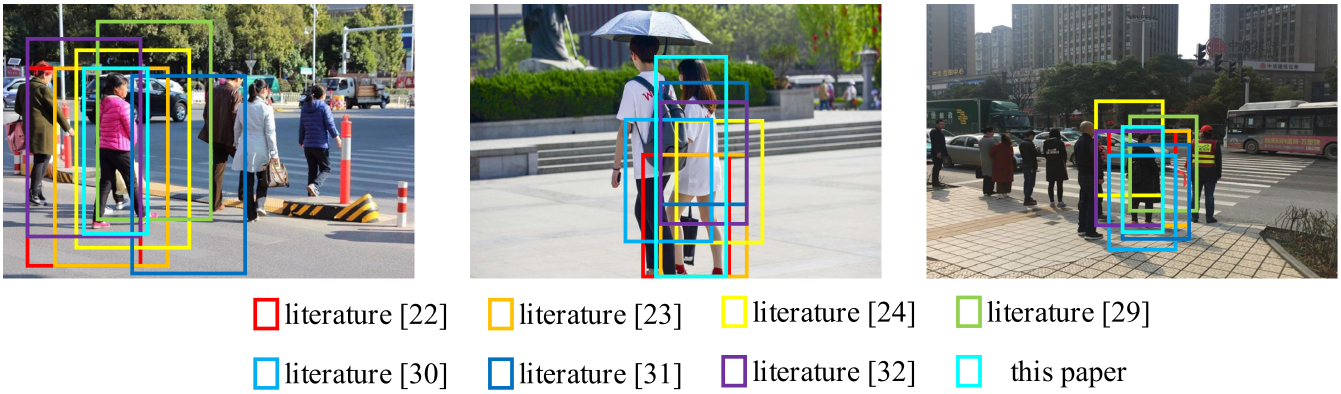


FIGURE 6. Girl-1 video sequence recognition results.

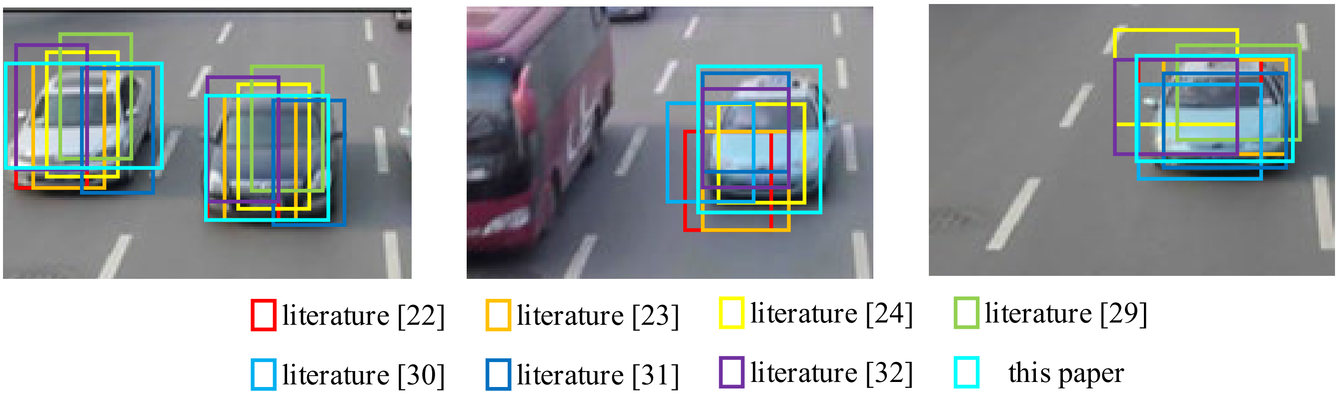


FIGURE 7. Car video sequence recognition results.

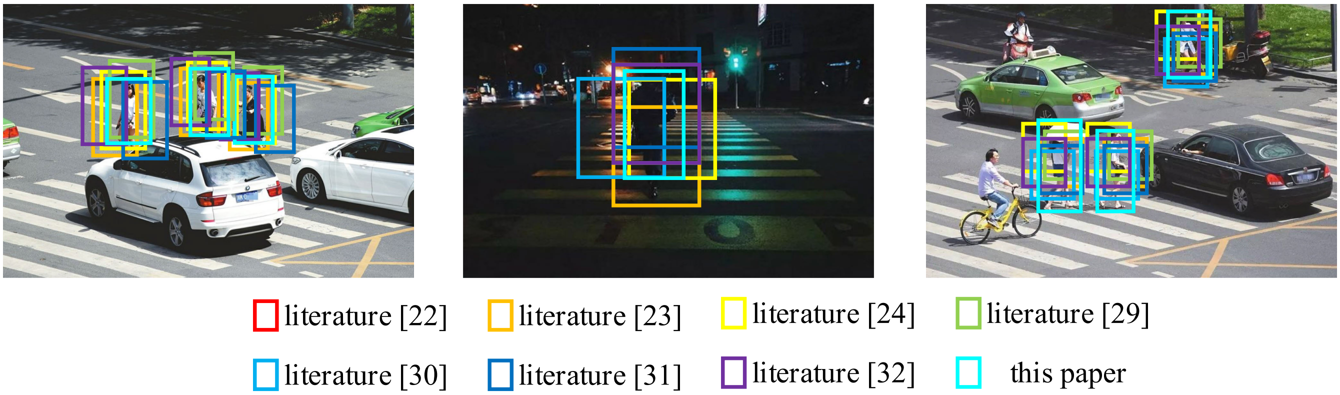


FIGURE 8. Girl-2 video sequence recognition results.

paper, the STC features of image sequences are extracted through spatiotemporal context and visual system characteristics, which can reduce the influence of light changes and occlusion on behaviors and actions. A reliable track set of feature points is obtained through image feature point tracking and track cutting. AD features are calculated based on these curves to capture the local motion information, shape and static appearance information of the track. The combination of multiple features overcomes the difficulty of changing the appearance of the moving target, thus obtaining accurate and stable recognition results.

2) LIGHT AND SCALE CHANGES

For the Car video sequence, the appearance of the moving target changes quickly when the moving target enters the shadow area from the illuminated area. As shown in Figure 7, in the whole video sequence, only the algorithm in this paper maintains good recognition results. When the illumination

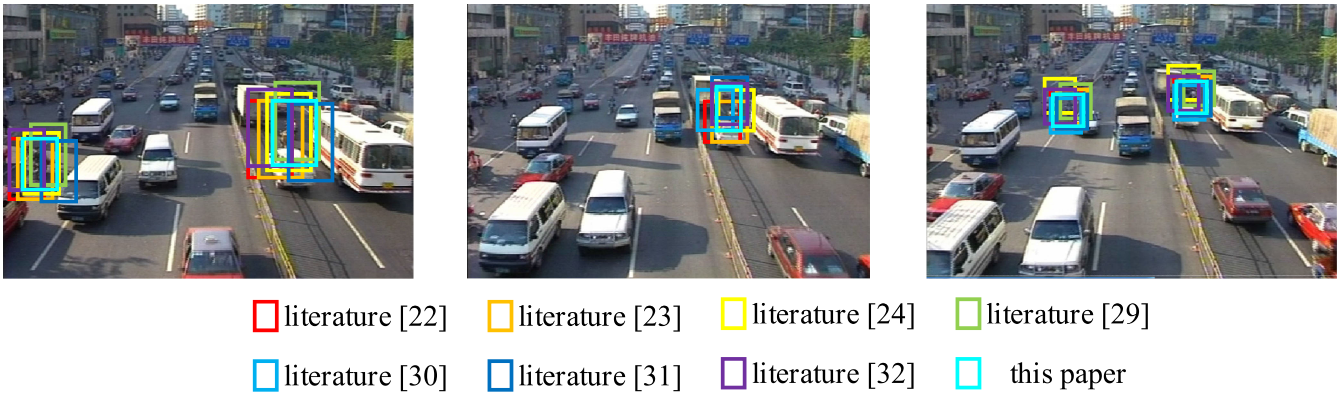


FIGURE 9. PEST-2001 video sequence recognition results.

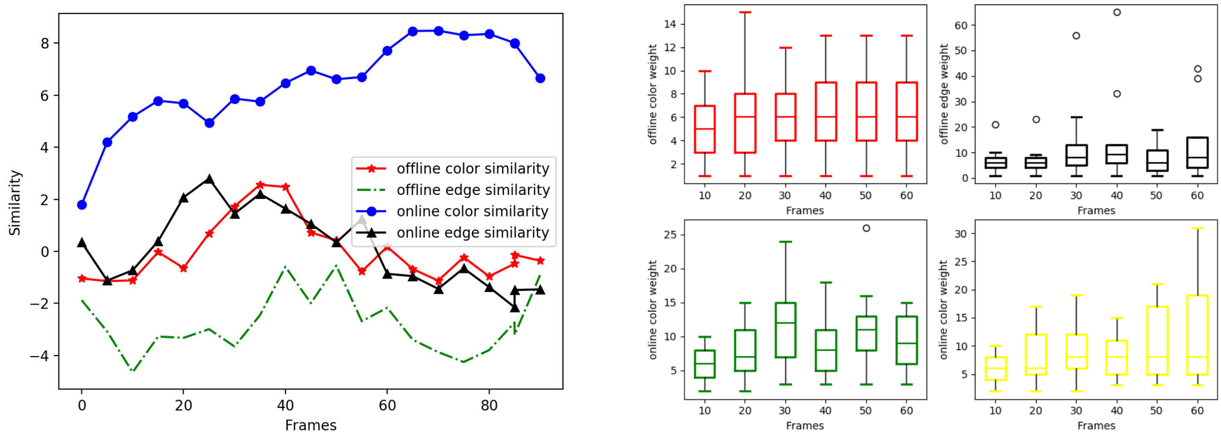


FIGURE 10. Similarity analysis of car video sequence.

changes dramatically, other recognition algorithms will drift to different degrees. In this paper, STC and AD features are adopted to reduce the influence of light changes and occlusion on behavior and action. This again proves the effectiveness of the two-feature fusion strategy proposed in this paper, which enables the proposed method to accurately locate the position of moving targets even under the condition of light changes.

3) BLOCKING

As shown in Figure 8, in the Girl-2 video sequence, there is a challenge of occlusion of all moving targets. From frame 87 to frame 104, at this time, the moving target is only partially occluded. In addition to the algorithm proposed in literature [22], other algorithms can better identify the target. However, starting from the 105th frame, when the target is completely occluded, other recognition algorithms lose the moving target, which eventually leads to serious drift. The algorithm in this paper accurately locates the target during the entire recognition process. In the PEST-2001 video sequence, from frame 39 to frame 46, the moving target part is occluded. And in the whole tracking process, there are still difficulties in scale change and deformation. As shown in Figure 9, in addition to the algorithms proposed in literature [29] and literature [32], other algorithms lose their targets in subsequent video frames. On the contrary, even in the case of partial occlusion or full occlusion, the algorithm

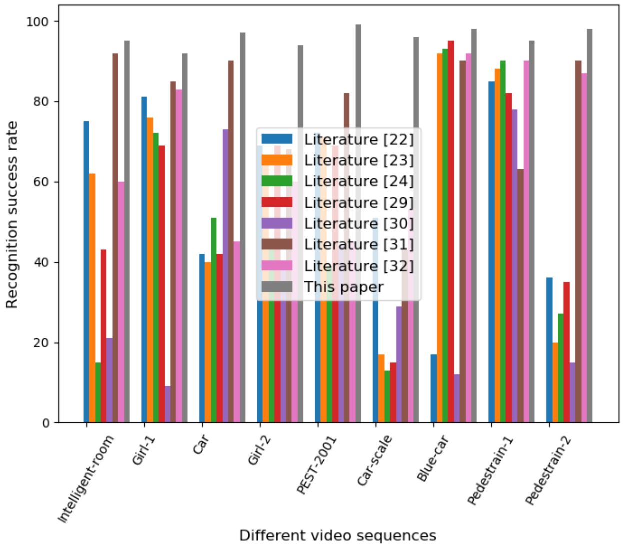


FIGURE 11. Recognition success rate results using different algorithms for different video sequences.

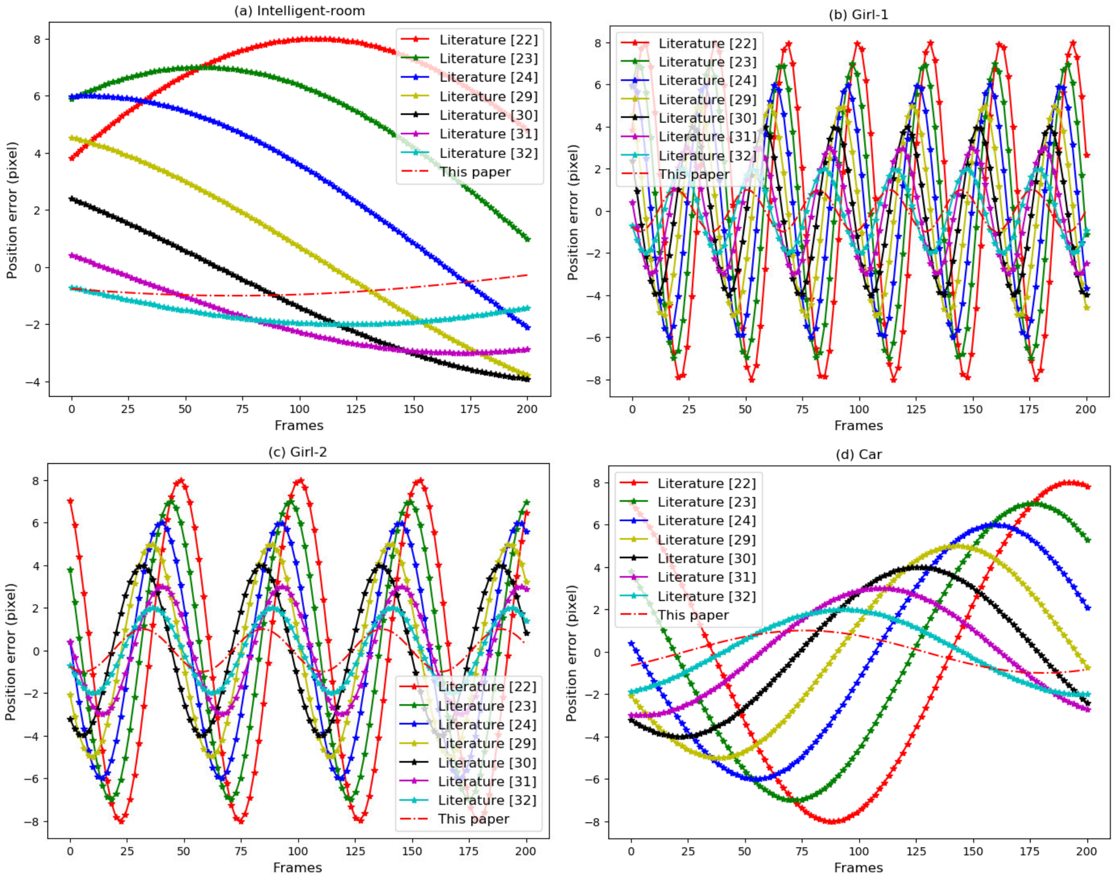


FIGURE 12. **The result of center position error CLE using different algorithms for different video sequences.**

in this paper can still accurately locate the position of the moving target. Since multiple features are used to represent the appearance model of the target, the algorithm proposed in this article can still maintain stable recognition performance even in the case of partial or full occlusion.

C. QUANTITATIVE ANALYSIS

To illustrate the rationality of the algorithm proposed in the article, this section analyzes the similarity of Car video sequences. Figure 10(a) shows the single feature similarity between the best candidate template and the online or offline template, and Figure 10(b) shows the normalized adaptive feature weights of the corresponding features. In the Car video sequence, from frame 37 to frame 49, the target happened to enter the shadow area from the illuminated area. During this process, the appearance of the moving target changed significantly. During this process, the appearance of the moving target changed obviously. As can be seen from Figure 10(a) and Figure 10(b), the offline template based on color features is very different from the target template, and cannot provide much useful information, resulting in the gradual decline of offline color similarity. Due to the illumination invariability of the edge direction feature in describing the spatial structure information of the moving object, the off-line edge similarity and on-line edge similarity basically maintain a stable value even if the moving object exists under the interference of the drastic light change.

Obviously, the adaptive feature weight reflects the importance of the corresponding feature in describing the target object, which is consistent with the analysis in the article, and also proves the rationality of the algorithm in the article.

In order to verify the effectiveness of the algorithm proposed in this paper, the algorithm in this paper is compared with literature [22], literature [23], literature [24], literature [29], literature [30], literature [31] and literature [32]. All these literatures are good methods at present. In addition, this section uses two evaluation criteria to quantitatively evaluate the performance of all tracking algorithms. The first evaluation criterion is the success rate. Figure 11 lists the experimental results of the recognition success rate. It can be seen that among all the recognition algorithms, the algorithm proposed in the article has the highest recognition success rate for all video sequences.

The second evaluation criterion is the center position error CLE. CLE refers to the Euclidean distance between the center position of the target obtained by the recognition result and the accurate center position manually marked. Figure 12 lists the experimental results of the center position error CLE.

In general, in the three video sequences, the CLE curve of the algorithm in this paper is lower than the CLE curves of the other seven algorithms, which again proves that the method proposed in the paper has better recognition performance than other algorithms.

The method of dual feature fusion and adaptive enhancement brings robustness to the algorithm proposed in this paper, but also increases the time complexity to a certain extent, and has an impact on the real-time performance. In fact, the time complexity of the dual feature fusion method is related to the number of selected features. The more features selected, the greater the time complexity. Table 3 lists the average recognition speed of all comparison algorithms. It can be seen from Table 3 that the algorithm in this chapter has a relative advantage in real-time compared to other algorithms. Our method gets 8FPS, which is 1/2 of the other algorithms.

TABLE 3. Average recognition speed of different algorithms.

Methods	Average FPS
Literature [22]	31
Literature [23]	12
Literature [24]	23
Literature [29]	28
Literature [30]	22
Literature [31]	19
Literature [32]	28
This paper	8

D. ANALYSIS OF ADAPTIVE LIFTING ALGORITHM

In order to better illustrate the robustness and accuracy of the proposed algorithm, this section analyzes the recognition success rate and center position error of the Pedestrian-2 video sequence.

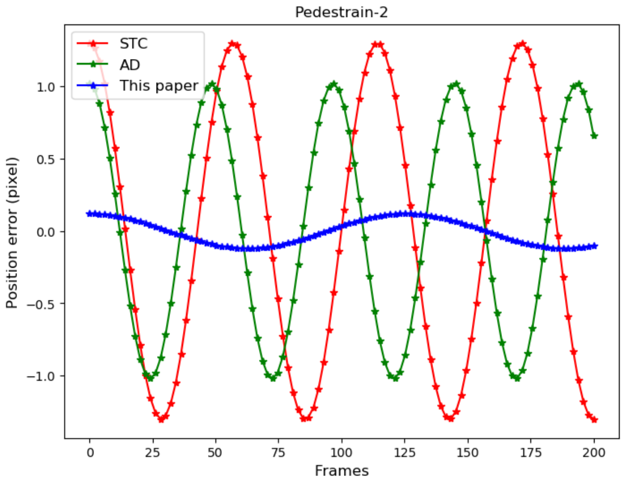


FIGURE 13. The CLE curve of center position error of Pedestrian-2 video sequence.

Figure 13 shows the center position error CLE curve of the Pedestrian-2 video sequence. Compared with other single action features, the method in this paper extracts the action features of the image sequence through STC and AD, merges the two action features through PCA technology to form a new feature, and defines an adaptive lifting algorithm to classify the action features training.

V. CONCLUSION

In order to reduce the interference of illumination changes and occlusion on the action, this paper extracts the image sequence features through the STC relationship and the visual system characteristics, and obtains the STC features. At the same time, considering the descriptor based on spatiotemporal interest points and the descriptor based on trajectory, the combination of the two forms a set of information-rich features AD, which can effectively capture the global and instantaneous movement information of the movement in time. In order to better fuse the two features obtained, the PCA operator is used to effectively combine the STC features and AD features to form a more accurate and complete feature representation. According to the new features, ABA is used for classification training to complete the decision-making and judgment of behavior. Experiments show that the method in this paper can effectively recognize and understand various actions, reduce the interference of light changes and occlusion, and still perform well in complex scenes with better stability.

REFERENCES

[1] S. Bench, C. Winter, and G. Francis, ''Use of a virtual reality device for basic life support training: Prototype testing and an exploration of users' views and experience,'' *Simul. Healthcare*, vol. 14, no. 5, pp. 287–292, 2019.

[2] S. Kavanagh, A. Luxton-Reilly, B. Wuensche, and B. Plimmer, ''A systematic review of Virtual Reality in education,'' *Themes Sci. Technol. Educ.*, vol. 10, no. 2, pp. 85–119, 2017.

[3] D. Velev and P. Zlateva, ''Virtual reality challenges in education and training,'' *Int. J. Learn. Teach.*, vol. 3, no. 1, pp. 33–37, 2017.

[4] L. Jensen and F. Konradsen, ''A review of the use of virtual reality head-mounted displays in education and training,'' *Edu. Inf. Technol.*, vol. 23, no. 4, pp. 1515–1529, Jul. 2018.

[5] R. Santos Silva, A. M. Mol, and L. Ishitani, ''Virtual reality for older users: A systematic literature review,'' *Int. J. Virtual Reality*, vol. 19, no. 1, pp. 11–25, Jan. 2019.

[6] C. C. T. Clark, C. M. Barnes, N. J. Swindell, M. D. Holton, D. D. Bingham, P. J. Collings, S. E. Barber, H. D. Summers, K. A. Mackintosh, and G. Stratton, ''Profiling movement and gait quality characteristics in pre-school children,'' *J. Motor Behav.*, vol. 50, no. 5, pp. 557–565, Sep. 2018.

[7] F. E. Costa, J. D. Pupo, J. Barth, E. D. S. Bezerra, and A. F. Salvador, ''The prevalence of injuries and its association with the characteristics of training in American football players in Brazil,'' *Hum. Movement*, vol. 20, no. 1, pp. 31–37, 2019.

[8] F. Luo, G. Cao, K. Mulligan, and X. Li, ''Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of chicago,'' *Appl. Geography*, vol. 70, pp. 11–25, May 2016.

[9] Q. Yang, D. Xiao, and S. Lin, ''Feeding behavior recognition for group-housed pigs with the faster R-CNN,'' *Comput. Electron. Agricult.*, vol. 155, pp. 453–460, Dec. 2018.

[10] S. Nazir, M. H. Yousaf, and S. A. Velastin, ''Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition,'' *Comput. Electr. Eng.*, vol. 72, pp. 660–669, Nov. 2018.

[11] C. Wang, Z. Li, N. Dey, Z. Li, A. S. Ashour, S. J. Fong, R. S. Sherratt, L. Wu, and F. Shi, "Histogram of oriented gradient based plantar pressure image feature extraction and classification employing fuzzy support vector machine," *J. Med. Imag. Health Informat.*, vol. 8, no. 4, pp. 842–854, May 2018.

[12] T. T. Nguyen, T. P. Nguyen, and F. Bouchara, "Directional dense-trajectory-based patterns for dynamic texture recognition," *IET Comput. Vis.*, vol. 14, no. 4, pp. 162–176, Jun. 2020.

[13] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360–365, Jul. 2018.

[14] M. A. Kiasari, G.-J. Jang, and M. Lee, "Novel iterative approach using generative and discriminative models for classification with missing features," *Neurocomputing*, vol. 225, pp. 23–30, Feb. 2017.

[15] M. Kang, J. Ahn, and K. Lee, "Opinion mining using ensemble text hidden Markov models for text classification," *Expert Syst. Appl.*, vol. 94, pp. 218–227, Mar. 2018.

[16] H. Chao and Y. Liu, "Emotion recognition from multi-channel EEG signals by exploiting the deep belief-conditional random field framework," *IEEE Access*, vol. 8, pp. 33002–33012, 2020.

[17] A. Lazarowska, "A new deterministic approach in a decision support system for ship's trajectory planning," *Expert Syst. Appl.*, vol. 71, pp. 469–478, Apr. 2017.

[18] A. Ben Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Syst. Appl.*, vol. 91, pp. 480–491, Jan. 2018.

[19] W. Ding, K. Liu, H. Chen, and F. Tang, "Human action recognition using similarity degree between postures and spectral learning," *IET Comput. Vis.*, vol. 12, no. 1, pp. 110–117, Feb. 2018.

[20] Q. Meng, H. Zhu, W. Zhang, X. Piao, and A. Zhang, "Action recognition using form and motion modalities," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1s, pp. 1–16, Apr. 2020.

[21] J. Singh and G. Goyal, "Anticipating movie success through crowdsourced social media videos," *Comput. Hum. Behav.*, vol. 101, pp. 484–494, Dec. 2019.

[22] W. Xu, Z. Miao, X.-P. Zhang, and Y. Tian, "A hierarchical spatio-temporal model for human activity recognition," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1494–1509, Jul. 2017.

[23] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Understand.*, vol. 158, pp. 85–105, May 2017.

[24] A. Hall, P. Ahonen-Rainio, and K. Virrantaus, "Insight provenance for spatiotemporal visual analytics: Theory, review, and guidelines," *J. Spatial Inf. Sci.*, no. 15, pp. 65–88, Dec. 2017.

[25] A. Ben Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," *Pattern Recognit. Lett.*, vol. 92, pp. 62–67, Jun. 2017.

[26] S. K. A. Kamarol, J. Parkkinen, M. H. Jaward, and R. Parthiban, "Spatiotemporal feature extraction for facial expression recognition," *IET Image Process.*, vol. 10, no. 7, pp. 534–541, Jul. 2016.

[27] S. Misra and R. H. Laskar, "Development of a hierarchical dynamic keyboard character recognition system using trajectory features and scale-invariant holistic modeling of characters," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 12, pp. 4901–4923, Dec. 2019.

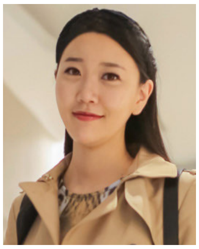
[28] J. M. Carmona and J. Climent, "Human action recognition by means of subtensor projections and dense trajectories," *Pattern Recognit.*, vol. 81, pp. 443–455, Sep. 2018.

[29] B. Shin, C. Kim, J. Kim, S. Lee, C. Kee, H. Seok Kim, and T. Lee, "Motion recognition-based 3D pedestrian navigation system using smartphone," *IEEE Sensors J.*, vol. 16, no. 18, pp. 6977–6989, Jun. 2016.

[30] Y. Zhou and Z. Gao, "Intelligent recognition of medical motion image combining convolutional neural network with Internet of Things," *IEEE Access*, vol. 7, pp. 145462–145476, 2019.

[31] J. Kim, H. Jung, M. Kang, and K. Chung, "3D human-gesture interface for fighting games using motion recognition sensor," *Wireless Pers. Commun.*, vol. 89, no. 3, pp. 927–940, Aug. 2016.

[32] J. Camargo and A. Young, "Feature selection and non-linear classifiers: Effects on simultaneous motion recognition in upper limb," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 743–750, Apr. 2019.



KUNNI HAN was born in Shandong, China, in 1986. She received the B.A. degree from Qingdao University, in 2007, and the M.F.A. degree in film, television, and digital media from the Academy of Film, Hong Kong Baptist University, in 2011. She is currently pursuing the Ph.D. degree with the City University of Macau, researching on digital media, smart city development, and cultural communication. She is also a Lecturer with the School of Journalism and Communication, Qingdao University.