Hindawi
Complexity
Volume 2024, Article ID 9851756, 1 page
https://doi.org/10.1155/2024/9851756

Retraction

Retracted: Mean Shift Fusion Color Histogram Algorithm for Nonrigid Complex Target Tracking in Sports Video

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Liu and X. Wang, "Mean Shift Fusion Color Histogram Algorithm for Nonrigid Complex Target Tracking in Sports Video," *Complexity*, vol. 2021, Article ID 5569637, 11 pages, 2021.

Hindawi
Complexity
Volume 2021, Article ID 5569637, 11 pages
https://doi.org/10.1155/2021/5569637

Research Article

Mean Shift Fusion Color Histogram Algorithm for Nonrigid Complex Target Tracking in Sports Video

Yu Liu[1] **and Xiaoyan Wang**[2]

[1]*School of Physical Education, Qiqihar University, Qiqihar 161006, China*
[2]*School of Science, Qiqihar University, Qiqihar 161006, China*

Correspondence should be addressed to Yu Liu; 01407@qqhru.edu.cn

Received 3 March 2021; Revised 10 April 2021; Accepted 15 April 2021; Published 22 April 2021

Academic Editor: Zhihan Lv

We analyze and study the tracking of nonrigid complex targets of sports video based on mean shift fusion color histogram algorithm. A simple and controllable 3D template generation method based on monocular video sequences is constructed, which is used as a preprocessing stage of dynamic target 3D reconstruction algorithm to achieve the construction of templates for a variety of complex objects, such as human faces and human hands, broadening the use of the reconstruction method. This stage requires video sequences of rigid moving target objects or sets of target images taken from different angles as input. First, the standard rigid body method of Visuals is used to obtain the external camera parameters of the sequence frames as well as the sparse feature point reconstruction data, and the algorithm has high accuracy and robustness. Then, a dense depth map is computed for each input image frame by the Multi-View Stereo algorithm. The depth reconstruction with a too high resolution not only increases the processing time significantly but also generates more noise, so the resolution of the depth map is controlled by parameters. The multiple hypothesis target tracking algorithms are used to track multiple targets, while the chunking feature is used to solve the problem of mutual occlusion and adhesion between targets. After finishing the matching, the target and background models are updated online separately to ensure the validity of the target and background models. Our results of nonrigid complex target tracking by mean shift fusion color histogram algorithm for sports video improve the accuracy by about 8% compared to other studies. The proposed tracking method based on the mean shift algorithm and color histogram algorithm can not only estimate the position of the target effectively but also depict the shape of the target well, which solves the problem that the nonrigid targets in sports video have complicated shapes and are not easy to track. An example is given to demonstrate the effectiveness and adaptiveness of the applied method.

1. Introduction

Target tracking is a prerequisite for tasks such as action recognition and behavior analysis of targets [1]. The target tracking algorithm can automatically extract and analyze the action trajectory of the target, which can effectively compensate for the lack of target analysis through human eyes, thus achieving the function of finding abnormal information and tracking it more accurately and effectively [2]. Compared with manual detection, when the number of targets to be tracked increases, the video target tracking algorithm can be used for real-time feedback. This solves the problem of human negligence caused by subjective problems, and greatly reduces human resource investment while detecting leaks. In the process of target tracking, when the target deformation, occlusion, and light changes will have a certain impact on the tracking results, but the use of methods such as neural networks will have too long matching time, too high cost, it is difficult to achieve the real-time and economic requirements in the application. Thus, it is necessary to study an efficient and accurate feature matching algorithm.

For different scenes, the detection of moving objects has a variety of detection algorithms, the purpose of detection is to mark the region of moving objects of interest in the video sequence and to extract the region by certain methods to

further classify and analyze to understand the state of the region. When detecting a target, the most important thing is to build an accurate model of the target; usually, there are two main ways of modeling, target modeling and background modeling [3]. Background modeling is the most common modeling method in detecting moving objects, by creating an accurate background model, to obtain accurate parameters of the moving target; in this process, by the difference operation between the current frame image of the video and the background image, the result of the operation is the changing region of the object [4]. We must adopt the method of real-time background update, using different background update rates, which can reflect the external environment adaptively [5]. Accurate detection of the target is the basis for subsequent target tracking and accurate positioning, and the integrity of the detected target is related to whether the moving target can be tracked effectively.

The tracking of moving targets is mainly to judge the motion position of the detected targets from the previous moment to the next moment, for the moving tracking of targets in different scenes, which is generally divided into single target tracking and multitarget tracking in single scenes and overlapping and nonoverlapping tracking in multiple scenes [6]. The movement tracking of the target is mainly to locate the spatial position changes of the target and match the changes of state information such as posture, position, behavior, and action generated by the target due to different environments [7]. Currently, there are two main implementations of human target tracking, top-down and bottom-up, where the former refers to estimating the target state of the moving human body, learning a priori knowledge of the detection scene, and evaluating the assumptions of various aspects of the tracking process; the latter refers to extracting the salient human features directly from the image sequence without basing on various a priori knowledge, and then by matching the target information, again determine the moving position of the target, and then achieve the tracking effect of the human target and lay the foundation for the subsequent behavior recognition and analysis.

2. Related Works

The real-time nature of the video surveillance detection algorithm enables motion image segmentation of moving objects, rapid localization of the target human body, extraction of regions of interest, matching of location information, and also the identification of the behavioral dynamics of specific human bodies under complex conditions of multiple crowds, thus enabling object feature recognition [8]. Motion body segmentation technology is to segment the image sequence in the video according to specific criteria after reading the video file to form a certain region and then get the motion body region to be detected, as well as the human target features, shape, pose change state, and other pieces of information [9]. In contrast, motion body tracking is the statistics of the spatial information during the change of the moving body, such as the state of the human target, the height of the body, the width of the body, and other characteristics [10]. Through the main operations of these two techniques, video image data are analyzed and expressed to locate the position of the human target and use the data for subsequent classification and recognition [11]. There are many similarities between human segmentation and tracking techniques, one of which is that both methods operate with targets from video image capture, and both video images contain commonly desired target behavior segments; both techniques have high requirements in terms of real-time, accuracy, and robustness of image processing.

The study of moving shadow detection techniques by Andres Sanin who used shadow removal as a key step to improve object detection and tracking and trackability as an unbiased method to determine the actual use of shadow detection methods summarized the different performances of all shadow detection methods. Masala proposed a method to extract unique invariant features from images method [12]. This method can be used to perform reliable matching between different views of an object or scene. These features are invariant to image scale and rotation and have been shown to provide robust matching over a wide range of affine distortions, variations in 3D viewpoints, increases in noise, and changes in illumination [13]. The method was experimentally shown to robustly identify objects in clutter and occlusion. Kitajima et al. proposed a contrast analysis-based method for nighttime visual surveillance that uses locally varying contrast to detect potentially moving objects, and the algorithm is effective for nighttime target detection [14]. Liao et al. proposed a real-time algorithm for foreground-background segmentation, where the sample background values at each image pixel are quantized into codebooks, and detection of moving targets is performed using the feature-layering idea with adaptive codebook background update, which can handle scenes containing moving backgrounds or illumination changes and enables reliable detection of different types of videos [15]. Ma et al. divided the human body into several parts, construct detectors for each part by SIFT features and AdaBoost, and finally detect travelers by fusing the results of each detection [16].

Severe degree of object deformation will directly lead to the efficiency of detection, the deformation is due to the object in motion, the camera is not stationary, the relative position between them is constantly changing, and the distance between the lens and the moving object is also changing, such as the human body from walking to jumping. Secondly, when the moving object itself undergoes nonrigid deformation, it will also fail to extract the contour of the moving target effectively, which seriously affects the operation efficiency and detection degree.

3. Mean Shift Fusion Color Histogram Algorithm for Video Nonrigid Complex Target Tracking

The mean shift fusion color histogram algorithm is a matching method based on kernel density nonparametric estimation.

3.1. Mean Shift Fusion Color Histogram Algorithm. This method uses the idea of local area template matching, which reduces the complexity caused by the global search for template matching [17]. It is calculated by iteratively obtaining the largest correlation coefficient in the target region and matching with that point and by continuously iterating the number of times and then achieving accurate matching of the features of the target and the candidate target, to quickly locate the target. This algorithm is widely valued and applied in the field of target tracking due to its good real-time performance. On the one hand, the algorithm is computationally small and easy to operate and has good real-time tracking performance, while it is not sensitive to information such as edge occlusion and nonrigid deformation of the target object; on the other hand, the algorithm itself has shortcomings, because the size of the target varies at different moments, and the mean shift algorithm's template is not updated in time, the tracking frame window is fixed, and it lacks a target color histogram. The algorithm of mean shift algorithm has its shortcomings due to the untimely update of the template, the fixed window of the tracking frame, the lack of target color histogram information, and so on, which can cause the low robustness of tracking targets. In the above shortcomings, the algorithm in this paper is improved by combining the mean shift algorithm with Kalman filtering and using an adaptive geometric feature fusion strategy.

Nonparametric density estimation, also colloquially called nonparametric estimation, is part of probability density estimation [18]. Parametric density estimation must determine a probability density function that follows a certain feature space, and this method is difficult to apply in practice. However, parametric-free density estimation does not require a priori knowledge, and the estimation is done based on training data only, which is not constrained in the target shape and does not require the initial setting of the probability density function, and several sample point estimates within the domain of a point represent the value of the density function at that point. Several methods are usually used: histogram method, nearest neighbor method, and kernel density estimation method, as shown in Figure 1.

When using the mean shift algorithm, no prior knowledge is required, and the associated density function values are obtained simply by taking the values of the sampled points in the feature space. When a set of sample data is available, the histogram method calculates parameter probability values by interval grouping, and the probability of each cell is represented by the ratio of the total amount of data in each group to the number of all parameters; kernel density estimation is similar to a histogram, except that kernel density uses a kernel function to smooth the data. Under good sampling conditions, it is possible to perform density estimation for any data that satisfy the distribution conditions.

$$M_h(x) = \frac{2}{k} \sum_{x_i \in s_h} (x_i + x)^2. \tag{1}$$

In the formula, k represents all sample points x, x denotes the offset vector of x, s denotes the high-dimensional spherical region of radius h, and $M_h(x)$ denotes the average of the sum values on sn. The sample points within sn, which are along the gradient of the probability density pointing in the direction of the increasing value, satisfy the following set at the point y.

$$s_h(x) = \left\{ y:\ (y + x)^T (y + x) \geq h^2 \right\}. \tag{2}$$

However, in real target tracking, the pixel values of the occluded part of the target differ significantly from the bare part due to the influence of the occluded objects, and the closer the pixel values to the target model are calculated, the more realistic results are obtained. Therefore, the importance of sample points at different locations varies in the actual calculation, and the closer the points to the outside edge are, the smaller their weights are. Therefore, adding kernel functions and weight coefficients to the algorithm can effectively improve the tracking performance.

$$H(k) = k\left(\left\| x + y - x^2 \right\|^2 \right). \tag{3}$$

The basic tracking steps of the mean shift algorithm are as follows: first, build a target region model, calculate the probability of each pixel feature value in the region, get the candidate model at the same time, compare the similarity between the object model and the current frame of the candidate model using the similarity function, select the maximum similarity function, and get the target position at the current moment after the motion change from the initial position. Then, according to the good convergence of the algorithm, the reliable position of the target is finally obtained by iterative calculation to achieve robust tracking.

After obtaining the target region using the detection algorithm in this paper, assume that there are n pixels in the extracted region, $\{Z\}$ is the pixel location ($1 \leq i \leq n$), and the gray space in the region is uniformly divided into m to obtain the regional gray histogram, and then, the probability density of the model satisfies the following equation:

$$\begin{aligned} q_u &= C \sum_{i=1}^{n} k\left(\left\| x + z - x^2 \right\|^2 \delta(z_i - u) \right), \\ C &= \frac{2}{\sum_{i=1}^{n} k\left(\left\| x + z - x^2 \right\|^2 \right)}, \\ z_i &= \left(\frac{(x_i - x_0)^2 + (y_i + y_0)^2}{x_0^2 - y_0^2} \right). \end{aligned} \tag{4}$$

The iterative process is a gradual convergence from the target center of the candidate region to the actual target center.

$$f_{k+1} = f_k + \frac{\sum_{i=1}^{n} k\left(\left\|x + z - x^2\right\|^2\right) g\left(\left\|f_k + z_i/h\right\|\right)^2}{\sum_{i=1}^{n} wg\left(\left\|f_k + z_i/h\right\|\right)^2}, \quad (5)$$

$$g(x) = K'(x).$$

In the derivation of the above formula, the mean shift algorithm takes *f(k)* as the starting point and moves between the target model and the candidate model in the direction of the largest variable of color probability density all the time; when the moving distance is less than the threshold value, the target center position can be obtained by judgment, the point is taken as the center of the search box in the next frame, and the process is repeated until the end.

The main idea of the EKF algorithm is to transform the nonlinearly transformed Gaussian distribution into a Gaussian distribution by approximation and variation and to linearize the system around the operating point, after which the EKF has the same formulas and computational procedures as the KF. The main idea of processing nonlinear problems is to approximate them as linear problems [19]. Considering the mathematical calculation, we expand the nonlinear filter function into the form of Taylor's formula and keep only the first-order term, so that an approximate linear model is obtained with a certain error. When tracking a multiobjective problem, the target tracking algorithm can be divided into online tracking and offline tracking according to the sequence of the target tracking state and the target trajectory formation. The multihypothesis tracking algorithm (MHT) can be regarded as a probability-maximizing online tracking algorithm, which is essentially an extension of the Kalman filter-based tracking algorithm for multiobjective tracking problems. The MHT algorithm models the observed target states and the likelihood estimates separately, and the modeling approach generally chooses Gaussian and Poisson distributions. In the case where such assumptions hold, we can select the best association hypothesis by finding the maximum value of the sum of the observed likelihood and the likelihood of the association hypothesis.

In this paper, we combine Kalman filtering with the mean shift algorithm to obtain the optimal estimation of target information, while adding the occlusion judgment factor and focusing on the tracking effect of the objects in this. The video image sequence is read, the target information is initialized, the target area is obtained according to the target detection algorithm of this paper, the Kalman filter is initialized, and the search box is manually determined to lock the target range to be tracked; the path is predicted using the filter, the optimal estimate of the target position is obtained, the center position of the target and the color histogram are obtained by combining the mean shift method. The center position is used as the starting point of the target; meanwhile, the mean shift algorithm is used to iteratively calculate the candidate region of the next frame of the moving target, the obtained target position of the current frame is substituted into the Kalman filter, and this process is cycled to obtain the target position at the *k*th frame. The occlusion judgment factor ε is defined, and the threshold T is set to judge the target being occluded: using Kalman's predicted target result in the previous frame is iteratively calculated, the result of the previous frame is used as the starting point, and the target position in the next frame is predicted, while the target model is built using mean shift.

3.2. Experimental Design of Nonrigid Complex Target Tracking for Sports Video. Due to the influence of the imaging system sensor, external environment, and other factors, resulting in the uneven distribution of the read video image sequence of brightness, the degree of variation in certain areas is too large, these bright spots are also called noise, these noises seriously hinder the output image to be recognized and understood by people or machines, to remove these unnecessary noises, and the image smoothing operation is performed. Image smoothing is mainly to smooth out the regions with large variations in image brightness so that the overall image brightness remains the same [20]. Generally, low-frequency filtering is used to remove the noise in the target imaging and enhance the image.

However, during the image smoothing operation, the effect of smoothing varies due to the different degrees of image interference and the different filter sizes. Therefore, in the following subsections, the two most common filtering methods, mean filtering and median filtering, are introduced. The use of mean filtering is not unfamiliar in linear filtering algorithms, where a template with a good match to the target imaging is selected by certain methods, specifically by selecting eight-pixel points of the pixels close to the target imaging, excluding the target pixel itself, and filtering the template consisting of its neighboring pixels to calculate the average of all pixels in the template and overwrite the original image pixel values.

Image segmentation is an important step in the sequence of video images, from processing to analysis, and allows the process of dividing a target image into a certain number of specific regions and extracting the target regions with unique properties. In terms of mathematical analysis, the process of segmenting regions is like a marking process, where the image is divided into several disjoint parts and these different regions are marked with numbers. The commonly used segmentation methods are divided into the following categories: threshold-based segmentation, region-based segmentation, edge-based segmentation, and specific theory-based segmentation.

Under the given conditions, the pixels in the background and foreground of the target image differ greatly, and the selection of the threshold value determines whether the foreground and background can be accurately segmented, which requires the use of adaptive thresholding. The segmentation method used in this paper is also exactly the adaptive threshold segmentation method, which selects the threshold value according to the range of the neighborhood of each pixel point of the image to be segmented. The specific

method of region growth segmentation is to combine regions with similar features. In the first step, the central pixel is selected as the starting point for growth in the region of the image to be segmented; in the second part, pixels with the same, or similar, characteristics as the central pixel are combined into the neighborhood of the central pixel. Then, these new pixels are reselected and combined with the center pixel according to the above steps until all the pixels that meet the requirements have been absorbed, and the growth of an image region is completed, as shown in Figure 2.

For different significant motion regions in the video, frames are clustered to form different motion frames, but since the size of the motion frames is not uniform, it is necessary to standardize the size and number of motion frames in each frame in the linked frame sequence [21]. If the number of motion frames in the video frames is not the same, then it may happen that some motion frames cannot form the motion tube, and it may cause the motion frames that are not part of the motion tube to become part of the motion tube, thus affecting the description of the motion. Therefore, to build a motion tube between consecutive frames by motion frames, the number of motion frames in each frame must be unified so that each frame can be connected between adjacent frames. Once the Euclidean distances of all motion frames between adjacent frames are found, all motion frames of adjacent frames can be connected.

In the modeling stage, the feature points of the target are selected at the location of the target region of the image frame, and the probability model of the feature points is established by training the samples. The traditional SIFT matching algorithm finds all the feature points of two images and matches them one by one. To reduce the computational effort, this paper divides the feature points that are physically close to each other into the same chunk and reduces the detection of feature points at the target location by predicting the target location.

In this paper, we first use the extended Kalman filter to predict the center of the target and determine the predicted position of each chunk based on the relative position of the chunk center and the target center; then, we set a certain scale at the chunk center and match the feature points in its range with the feature points in the original model center chunk and define the new center by the position of the successfully matched feature points [22]. The new central chunk position is defined by the position of the successfully matched feature points so that the new chunk center position can be obtained and the model can be modified. This matching strategy allows each chunk to be matched with only the feature points within the specified range, which greatly reduces the computational effort and can effectively improve the matching efficiency and at the same time greatly reduces the probability of mismatching.

The location of the target needs to be determined after the matching is completed for each chunk, and the location of the target is determined by the location of each chunk. The positioning strategy of the new chunk is to save the distance between each feature point and the chunk center as a star structure and determine the new position of the chunk according to the relative position of the chunk center and the successfully matched feature point. The use of the local chunking algorithm makes the target model highly robust and theoretically can effectively solve the problem of target deformation. However, when the target is occluded, a certain determination is needed to solve the occlusion problem of the target. Target occlusion can be divided into local occlusion and global occlusion, and the following is an example of how to determine the occlusion problem in this paper, as shown in Table 1.

The experiments for the 2-point grid were conducted using the points in the upper-left and lower-right corners of the actual target frame; the experiments for the 4-point grid added two points in the lower-left and upper-right corners to the 2-point grid, and the 9-point grid was in the same form as the $3 * 3$ grid mentioned in the paper. The experiments in the above table do not use the information fusion between the grid points. From the experimental results in the above table, the average detection accuracy achieved by the grid-directed localization is higher than that achieved by the regression model, and the more the grid points in the grid-directed localization method, the higher the average detection accuracy.

According to the above target occlusion process, we can see that when the target is partially occluded, there will be a partial chunk matching failure, but this partial chunk still exists as part of the target. When the object is occluded, if the target is still updated, the occluded object will be updated to the target template, and part of the target information will be lost at the same time, so it is very easy to lose the target. To avoid the situation of losing chunks, we adopt a model stop update strategy; that is, when the matching of chunks in the model fails, the feature information and structure information of the chunk are kept unchanged, and this part of the chunk is still matched with the target region in the next matching. In this case, to solve the value of the new target location, the EKF model of the center location is used to predict the center location of this chunk and adopt this location as the central location of this chunk at the next moment and update the relative location information of this location to the target center.

When all the chunks fail to match, that is, the target is globally occluded, only the center position of the target is predicted and the position of the target can be determined based on the structure information of each chunk. It is worth mentioning that when the target is globally occluded, a time threshold T_0 needs to be set, and if no target chunk is successfully matched after a time interval of T_0, the target is determined to be lost.

4. Analysis of Results

4.1. Algorithm Performance Results in Analysis. The experimental results of feature fusion on the UCF dataset and Hollywood dataset according to different dynamic and static feature contribution ratios are shown in Figure 3.

Through the experimental comparison, it can be found that the average recognition accuracy of these two datasets on 8 : 2 is the highest, so it can be considered that the optimal

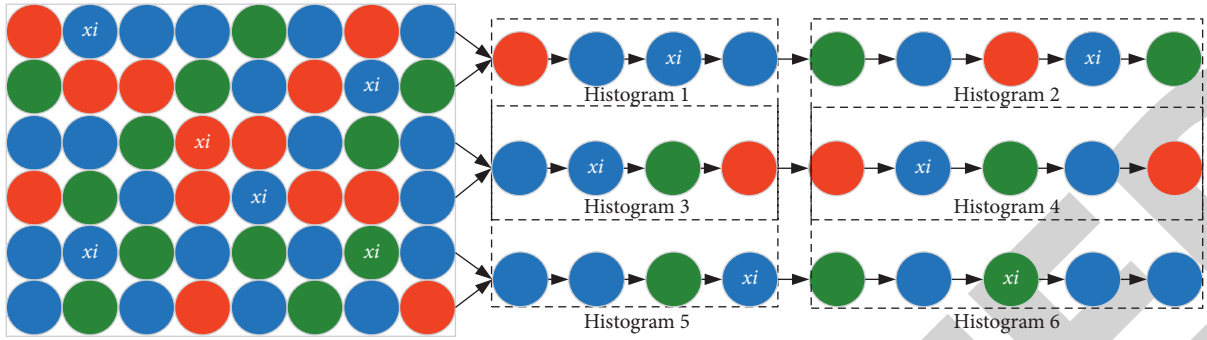

Figure 1: Mean shift fusion color histogram algorithm.

ratio of dynamic features and static features on these two datasets proposed in this paper should be close to this value.

However, the recognition accuracies for different motion categories with 8 : 2 and 6 : 4 contribution ratios of dynamic and static features are shown in Figure 4, and it can be found that the best contribution ratio of static and dynamic features is not 8 : 2 for all motion categories, for example, the best contribution ratio of standing and sitting is 6:4, which indicates dynamic function. This indicates that dynamic features are still important for motion recognition in these categories.

Whether $S : M = 8 : 2$ or $S : M = 6 : 4$, static features occupy a larger proportion, a phenomenon that may be because the GRU features on the time series are fused again later in the whole framework, compensating for the fact that the CNN network used to identify single-frame images ignores the video on the time axis of feature extraction. For individual video frames, the output of the SoftMax layer is extracted as static features using the CNN model trained on the ImageNet training set, and for consecutive stacked frames, the trajectory is obtained by calculating the optical flow, the moving frame is constructed on each frame by the trajectory, the motion tube is proposed by connecting the motion frames on consecutive frame sequences, and the MBH features of the motion tube are extracted as dynamic features. The MBH features of the motion tube are extracted as dynamic features, then three mathematical models are used to fuse the static features and dynamic features, and the GRU-based video character classification model is proposed.

In the Cholesky variation-based feature fusion experiments, by setting four different sets of contribution ratios of dynamic and static features, we finally found that the recognition rate was the highest in most cases when the contribution ratio of static and dynamic features was 8 : 2, and the recognition rate of some categories of actions was highest when the contribution ratio of static and dynamic features was 6 : 4, which indicates that static features have a higher weight in feature fusion, which may be because the GRU captures the temporal features of the video sequences, and therefore the static features extracted by the CNN have a higher weight in the final motion feature fusion vector. However, dynamic features also play a significant role in feature fusion, because the clustered trajectory features can distinguish different motion salient regions on consecutive frames, making dynamic features more representative of motion information in different regions on the video.

TDB means To Be Detailed. The most widely used modeling approach for background modeling is a mixed Gaussian model- (GMM-) based modeling approach, which requires training with many samples to obtain a pixel-based sample statistic or a mixed Gaussian model histogram. In general, modeling the pixel points of the background should result in a distribution that approximates a Gaussian model, but it is possible to obtain a model with multiple peaks. If the obtained distribution model has multiple peaks, the distribution is superimposed on each distribution according to different weights to obtain a Gaussian distribution model of the pixel points, and the weights of each peak distribution are continuously updated according to the time series. The results of multitarget tracking using the algorithm in this paper are compared with other multitarget tracking algorithms by the above two evaluation metrics, and the results are shown in Figure 5. The preactivated ResNet of the P3DResNet algorithm is like the bottleneck ResNet architecture, but with differences in convolution, batch normalization, and ReLU order. In ResNet, each convolutional layer is followed by BN and ReLU, whereas in preactivated ResNet, each BN is followed by ReLU and convolutional layers. The shortcut pass connects the top of the block to the layer after the last convolutional layer in the block. The trace-before-detect (TBD) is proposed for detection tracking of weak targets. For detection tracking of weak targets, TBD, as a typical strongly nonlinear problem, can perform nonphase parameter accumulation and can achieve better results. For TBD, it has the following advantages: the original sensor data are processed directly, the information is obtained from the processing and accumulated over time, and the decision for the target is located in the last step of the whole processing chain. TBD does not need to carry out the problem of correlating the observation with the track data, which occurs in the traditional problem, because there is no threshold processing in the data processing, and there are no point-track data, so there is no need for the correlation of point-track data with the trajectory data which is performed.

According to Figure 5, the combination of the model construction algorithm used in this paper with the multiple hypothesis target tracking algorithms results in a small position error while maintaining a small missing rate, false-

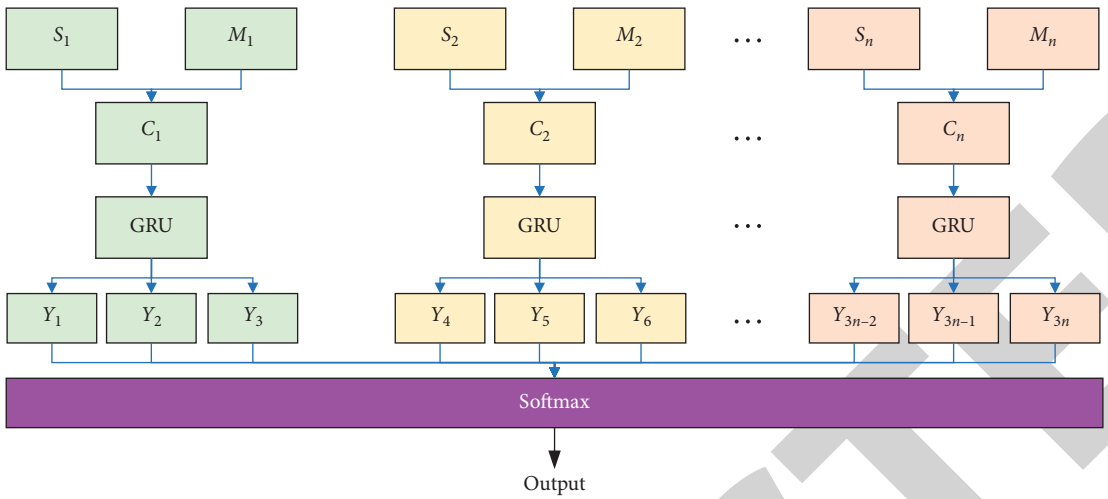

Figure 2: GRU-based video character classification model.

Table 1: Comparison of the accuracy of multipoint grid supervision.

Method	Access point at 0(AP)	Access point at 0.25($AP_{0.25}$)	Access point at 0.0.5($AP_{0.5}$)	Access point at 0.75($AP_{0.75}$)
Regression	38.2	24.9	48.7	24.6
2 points	22.5	48.6	28.3	38.6
4-point grid	24.3	35	33.8	34.7
9-point grid	20.5	43.9	46.1	32.8

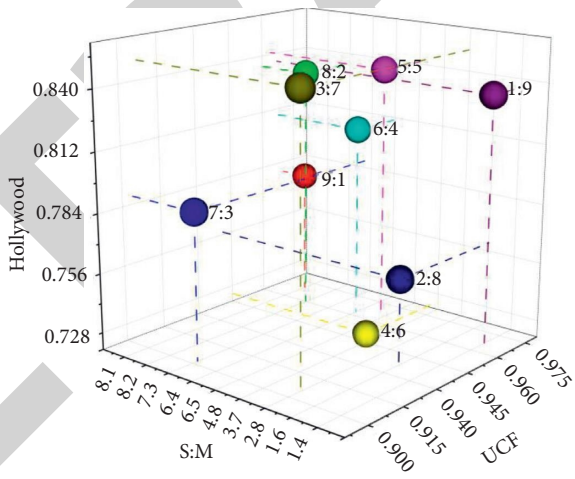

Figure 3: Results of different dynamic and static feature contribution ratios in two datasets.

positive rate, and mismatch rate for the target. Since this paper does not improve the MHT algorithm, we use its framework to improve the correct rate of multitarget tracking in video; that is, the range of areas matched to the correct target by improving the way the target is modeled and by subtracting the background. Therefore, in this paper, we demonstrate the superiority of this algorithm by using the visualization of the matching results when performing multitarget matching.

4.2. Analysis of Tracking Experimental Results. In this paper, we use the VGG-M-2048 model pretrained on the ImageNet dataset; however, the RGB image with channel number 3 in the first convolutional layer conv1 does not match the input data of the time-domain network, so we use the cross-modal cross pretraining method for this problem. The weights of conv1 are averaged and copied into 20 copies for the weights of conv1 of the time-domain network, and the other weights are kept the same to achieve the matching of the parameters of the pretrained network in the time-domain network.

Optical flow features are a set of dense optical flows, a set of displacement vector fields between adjacent frames, which can be applied to extract motion information and play an important role in video recognition. The network framework designed in this paper forms optical flow information encapsulating motion information of each static video frame

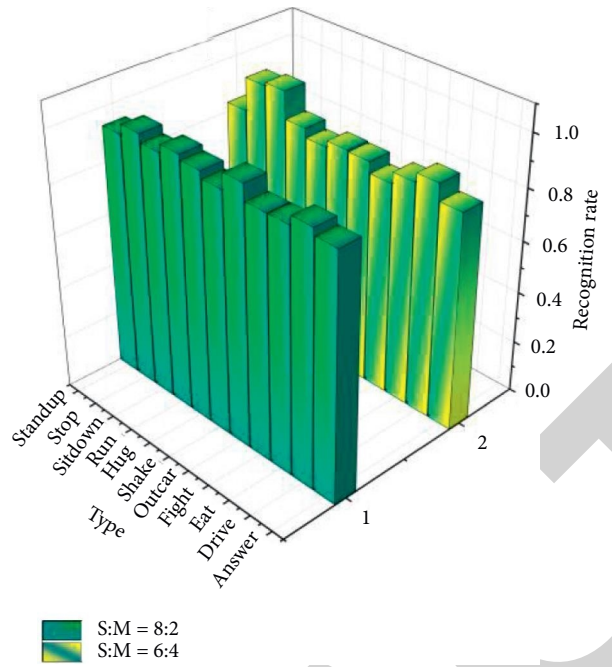

Figure 4: Hollywood for different classes of motion on $S:M$ of $8:2$ and $6:4$.

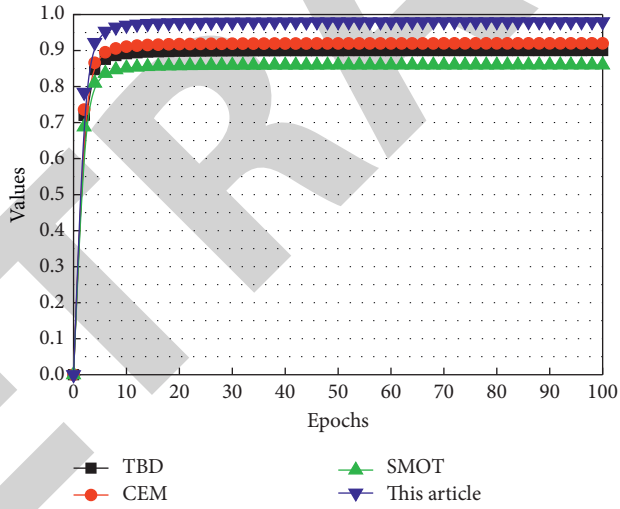

Figure 5: Comparison of the tracking results between this algorithm and other algorithms.

image in the input imported optical flow image, which improves the correlation of spatiotemporal features on pixel points and the robustness of processing video frame sampling. As shown in Figure 6, the optical flow map is a grayscale image calculated by decomposing the optical flow data into horizontal and hammer direction vectors, and there exist $2L$ image channels. In this paper, we set $L = 10$, and 10 consecutive frames of horizontal and vertical optical flow are stacked to form 20 dense optical flow images as the input to the time-domain network for the smaller deviations in model assumptions that can only have a small impact on algorithm performance.

In this paper, based on the improved dual-stream convolutional network algorithm, we achieved 92.1% and 66.1% recognition accuracy on UCF-101 and HMDB-51, respectively, for long-time video spatial-temporal modeling. Compared with the dual-stream method, the accuracy of this paper is improved by 4.1% and 6.7%, respectively, and the accuracy of this paper is also higher than other classical algorithms. Also, the algorithm in this paper implements an end-to-end network structure to achieve effectiveness on video-based behavior recognition tasks, as shown in Figure 7.

P3DResNet means Pseudo-3D Residual Networks. Several sets of comparison experiments are performed on UCF101 and HMDB51 datasets after invoking the pretrained model, and the experimental parameter settings and network structure are given in detail. The experimental

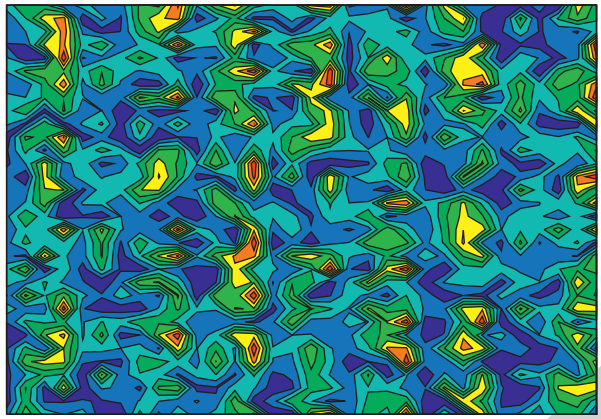

Figure 6: Optical flow diagram of continuous video frames.

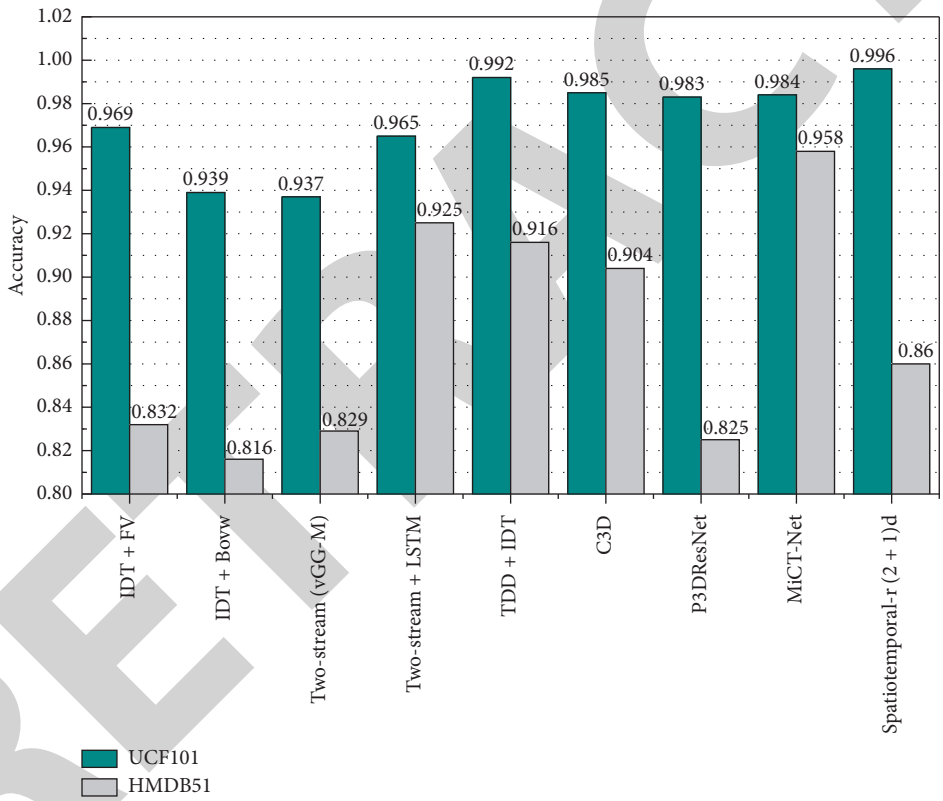

Figure 7: Comparison of video behavior recognition accuracy of different algorithms.

comparison of the number of video segments set in this paper is carried out, respectively, the weight ratio of dual-stream space-time fusion is compared experimentally, and finally the network is compared with other existing classic algorithms.

The evaluation strategy based on OTB50 initializes the algorithm with the state of the real target in the first frame of the video, and then the performance of all tracking algorithms is measured by two quantities: average accuracy and success rate. This straightforward approach is called one-pass evaluation (OPE). Precision is measured by the distance between the center of the bounding box of the tracking algorithm and the corresponding true bounding box. The accuracy graph shows the percentage of the tracking algorithm bounding box within a given threshold distance using the true value pixels. The threshold is typically set to 20 pixels to rank the tracking algorithm. The success rate is measured as the concatenation between the bounding box of the tracking algorithm and the true bounding box. The success graph represents the percentage of tracker bounding boxes with an overlap fraction greater than a given threshold OPE. Although this evaluation metric is simple, it also has two major drawbacks. First, a particular tracking algorithm may be more sensitive to using the first frame to initialize, but if

other different initial states or initial frames are used, the performance of that tracking algorithm may vary so much that it is uncertain whether it optimizes algorithm performance. Second, most algorithms do not have a mechanism for reinitialization, the tracking results obtained after a tracking failure do not provide meaningful information for algorithm optimization, and so on. The existing distance metrics for target feature selection are often only suitable for measuring the distance between two single-peaked distributions, while in practice both the target and the background are often multipeaked. The experimental results show that the distance metric between multipeak distributions proposed in this paper can select the best target features, thus improving the stability and accuracy of target tracking. In this paper, a fast and effective spatial color Gaussian mixture modeling method is proposed, and an approximate symmetric KL distance-based method is proposed to measure the similarity between the target model and the candidate model. Experimental results show that the modeling method in this paper can greatly save the time overhead in modeling and tracking, and the distance metric in this paper has stronger differentiation ability and better stability than existing metrics, thus greatly improving the robustness of the target tracking system.

Firstly, we introduce the commonly used tracking methods in the current tracking target scenario, analyze the problems and difficulties in tracking targets, and focus on the problem of target tracking by occlusion based on the existing algorithms. The two mainstream algorithms are investigated in-depth, the position of the target is predicted by establishing the Kalman prediction update mechanism, and the mean shift algorithm is incorporated into the algorithm to optimize the algorithm and improve the tracking robustness.

5. Conclusion

In recent years, the detection and tracking of motion targets have remained a challenging research topic in the field of computer vision. It has been a target of interest for many researchers because of its wide range and its ability to achieve target recognition in numerous complex environments in reality. However, the detection and tracking of moving targets are still challenging due to various complex factors in the actual research environment, weather changes, target occlusion, noise interference, and so on. When tracking targets, the present algorithm has extremely high reliability in tracking single targets, however, the tracking performance of the algorithm is not stable for multiscale information fusion of targets in multidimensional space to establish tracking mechanisms. Three commonly used mainstream detection algorithms are compared and analyzed, and performance analysis is done for their advantages and disadvantages. Based on this, the improved T-GMM algorithm is proposed, and the model of the algorithm is built based on the improved hybrid Gaussian and combined with the three-frame difference method at the same time to achieve accurate detection of the target. The experimental simulation shows that the algorithm has greatly improved in removing "ghost images" and reducing noise. After accurately extracting the moving human region, we further realize the effective tracking of the moving human body, and firstly, we describe several common tracking methods, introduce the Kalman filtering principle and mean shift algorithm, respectively, propose the improved KMT algorithm, and simultaneously perform the occlusion judgment. The experimental results show that the tracking algorithm can achieve a robust tracking effect when the human body is blocked.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Q. Hui, "Motion video tracking technology in sports training based on Mean-Shift algorithm," *The Journal of Supercomputing*, vol. 75, no. 9, pp. 6021–6037, 2019.

[2] Y. Zhang, S. Feng, X. Sun, and H. Yang, "Research on tracking algorithm for fast-moving target in sport video," *Journal of Computational and Theoretical Nanoscience*, vol. 14, no. 1, pp. 230–236, 2017.

[3] F. Zhou, "Dynamic tracking algorithm suitable for intelligent video surveillance," *Journal of Computational Methods in Sciences and Engineering*, vol. 19, no. S1, pp. 157–164, 2019.

[4] O. Sliti, H. Hamam, and H. Amiri, "CLBP for scale and orientation adaptive mean shift tracking," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 3, pp. 416–429, 2018.

[5] N. Pourmomtaz and M. Nahvi, "Multispectral particle filter tracking using adaptive decision-based fusion of visible and thermal sequences," *Multimedia Tools and Applications*, vol. 79, no. 25-26, pp. 18405–18434, 2020.

[6] P. R. Kamble, A. G. Keskar, and K. M. Bhurchandi, "Ball tracking in sports: a survey," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1655–1705, 2019.

[7] H. E. Rouabhia, B. Farou, Z. E. Kouahla, H. Seridi, and H. Akdag, "Cooperative processing based on posture change detection and trajectory estimation for unknown multi-object tracking," *International Journal of Systems Science*, vol. 50, no. 13, pp. 2539–2551, 2019.

[8] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Unsupervised classification of erroneous video object trajectories," *Soft Computing*, vol. 22, no. 14, pp. 4703–4721, 2018.

[9] C. Jiang, H. Yin, F. Yang, and X. Jiang, "Application of 3-D sensor tracking imaging in detailed feature extraction of motion damage action," *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 4, pp. 842–846, 2020.

[10] M. Takahashi, S. Yokozawa, H. Mitsumine, and T. Mishina, "Real-time ball-position measurement using multi-view cameras for live football broadcast," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 23729–23750, 2018.

[11] G. Awad, W. Kraaij, P. Over, and S. i. Satoh, "Instance search retrospective with focus on TRECVID," *International Journal*

of Multimedia Information Retrieval, vol. 6, no. 1, pp. 1–29, 2017.

[12] G. Masala, F. Casu, B. Golosio, and E. Grosso, "2D recurrent neural networks: a high-performance tool for robust visual tracking in dynamic scenes," *Neural Computing and Applications*, vol. 29, no. 7, pp. 329–341, 2018.

[13] M. A. Rafique, M. Jeon, and M. T. Hassan, "Deformable object tracking using clustering and particle filter," *Computing and Informatics*, vol. 37, no. 3, pp. 717–736, 2018.

[14] T. Kitajima, E. A. Y. Murakami, S. Yoshimoto, Y. Kuroda, and O. Oshiro, "Privacy-aware human-detection and tracking system using biological signals," *IEICE Transactions on Communications*, vol. E102.B, no. 4, pp. 708–721, 2019.

[15] X. L. Liao and C. Zhang, "Toward situation awareness: a survey on adaptive learning for model-free tracking," *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21073–21115, 2017.

[16] S. Ma, J. Zhang, S. Sclaroff, N. Ikizler-Cinbis, and L. Sigal, "Space-time tree ensemble for action recognition and localization," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 314–332, 2018.

[17] Q. Wu, G. Xu, Y. Cheng, W. Dong, L. Ma, and Z. Li, "Histogram of maximal point-edge orientation for multi-source image matching," *International Journal of Remote Sensing*, vol. 41, no. 14, pp. 5166–5185, 2020.

[18] Z. Zhu, W. Wu, W. Zou et al., "End-to-end flow correlation tracking with spatial-temporal attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 548–557, Salt Lake City, UT, USA, June 2018.

[19] J. Yang, M. Xi, B. Jiang et al., "FADN: fully connected attitude detection network based on industrial video," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2011–2020, 2020.

[20] F. Terroso-Saenz, A. González-Vidal, A. P. Ramallo-González, and A. F. Skarmeta, "An open IoT platform for the management and analysis of energy data," *Future Generation Computer Systems*, vol. 92, pp. 1066–1079, 2019.

[21] M. Khare, R. K. Srivastava, and A. Khare, "Object tracking using combination of daubechies complex wavelet transform and Zernike moment," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 1247–1290, 2017.

[22] M. R. Keyvanpour, S. Vahidian, and M. Ramezani, "HMR-vid: a comparative analytical survey on human motion recognition in video data," *Multimedia Tools and Applications*, vol. 79, no. 43-44, pp. 31819–31863, 2020.