Stability Tracking Detection of Moving Objects in Video Images Based on Computer Vision Technology

Ningning Wang[1(✉)] and Qiangjun Liu[2,3]

[1] Aba Teachers University, Wenchuan 623002, China
abtuwnn@163.com
[2] Krirk University, Bangkok 10220, Thailand
[3] Guangxi Vocational and Technical College, Nanning 530226, China

Abstract. In order to accurately collect images of moving targets and improve the accuracy of target tracking and detection, this paper proposes a new gray-scale image moving target stability tracking and detection method based on computer vision technology. Set camera parameters and light source parameters, and accurately collect moving target images through computer vision technology to improve the accuracy of moving target tracking and detection. The video image is preprocessed, and the specific preprocessing steps include image enhancement and edge detection. The random forest is used as the classifier to eliminate the background, generate a rough target ROI map, and implement the corresponding scale recognition on the ROI area to realize the recognition of moving objects in video images. Combining the Camshift algorithm and the Kalman filter algorithm, the existing moving target tracking method is improved, and the stable tracking of the moving target is implemented. The stability detection of moving targets is implemented by the background difference method, and the selected background difference method is the ViBe algorithm. The test results show that the design method correctly handles more frames and has a higher accuracy rate. It processes more frames per second and has lower tracking and detection errors. The average tracking detection time of this method is less than 3000 s.

Keywords: Computer Vision Technology · Video image · Edge Detection · Moving Target · Tracking Detection

1 Introduction

Human perception of the external world is mainly obtained through people's five senses, of which more than 80% of information is obtained through people's eyes, which is what we call vision. As we all know, the human eye is an important sensory organ, and it is one of the main sources of human information from the objective world [1]. It can be seen that image visual information plays a vital role in human life and work. For this reason, image processing has naturally become one of the most popular research fields. Among

B. Wang et al. (Eds.): ICMTEL 2023, LNICST 533, pp. 57–72, 2024.
https://doi.org/10.1007/978-3-031-50574-4_5

them, the application and research of high-tech image processing technology play a vital role in the realization of security detection technology, aerospace technology, robot vision technology, vehicle navigation, global positioning and even medical and military aspects [2]. In real life, we mainly divide images into moving and stationary ones. In most cases, people tend to be more interested in moving objects. Therefore, it is of great significance and broad application prospects to study the tracking and detection of moving objects. With the rapid development of digital signal detection and transmission processing technology and computer science and technology, as well as the continuous maturity of theory and the continuous improvement of hardware, people began to use cameras, video cameras, digital cameras and other high-precision image acquisition tools to obtain various image information and convert it into digital signals, then, computer algorithms and digital image processing methods are used to achieve the processing of image visual information, so as to meet people's various needs.

With the rapid development of research in recent years, intelligent video surveillance system is a new application direction rising in recent years. It has become a more advanced technology with high-tech content in the world. It is a frontier topic in the field of image research in the world today. It has the characteristics of interdisciplinary integration and technical integration. The intelligent video monitoring system is simply to understand, analyze and process the video signal, and further control the video monitoring system. When necessary, it can also alarm and other processing. Intelligent monitoring technology involves computer science and technology, psychology, physics, applied mathematics, etc., as well as sensor technology, detection technology, signal processing technology, image processing technology and other disciplines. The tracking and detection of moving objects in video is not only one of the core technologies of intelligent video surveillance system, but also a solid foundation of artificial intelligence research, and also one of the key technologies to realize intelligent robots and artificial intelligence. It has potential economic value and broad application prospects in intelligent monitoring systems, advanced human-machine interfaces, human motion analysis, virtual reality, and content-based image retrieval and storage. Therefore, the research focus is on the stability tracking and detection technology of moving objects in video images, and a stability tracking and detection technology for moving objects in video images based on computer vision technology is designed.

For the research of this problem, the current research on the detection and tracking methods of moving targets at home and abroad is still limited by the particularity of environmental factors. Most tracking algorithms or tracking systems still have tracking loss and drift problems [3] when the moving scene environment is complex or the target scale changes greatly, resulting in the reduction of the image acquisition accuracy of moving targets. Therefore, the detection and tracking of moving objects in complex environments is still a major problem being studied in this field.

2 Gray-Scale Image Moving Target Stability Tracking Detection

2.1 Computer Vision Acquisition of Moving Target Images

Design a computer vision detection system based on computer vision technology to implement image acquisition. The hardware of the designed computer vision inspection system consists of VT-LT2-HR1260W ring light source and DFK23G445 camera.

In order to meet the lighting needs of the CCD camera and obtain high-quality images, the light source is selected. In the selection, according to the characteristics of various lighting methods and the application of the subject, it is believed that the use of flat LED scattered light sources for lighting can improve most of the lighting. Display force for parts with irregular appearance. The illumination light source of the selected flat LED array is VT-LT2-HR1260W ring light source. In this light source, the light emitting diodes are closely arranged on the ring-shaped light source housing, which is divided into seven areas with a total of seven circles. The LEDs in any circle and any area can be controlled to turn on and off through the control buttons. If a single area is lit, its optical characteristics are similar to parallel light in a certain direction; When all lights up, it is uniformly scattered light vertically irradiated. It is helpful to avoid part shadow; The diameter of the inner circle of the ring shape is larger than the diameter of the CCD lens. The CCD lens can be moved up and down to expand and contract into the light source, so that the distance between the lens and the workpiece can be adjusted. When using, fix the light source on the bracket frame of the light source frame. The light source frame is mainly composed of a chassis, a bracket and a bracket frame. The bracket and the chassis form an angle of about 70 degrees. The chassis and the bracket frame are parallel to the horizontal plane. The light source is installed on the bracket frame and is vertical. The assembly line is flat and facing above the camera lens, the bracket frame moves up and down to adjust the distance between the light source and the camera lens, and the light source controller is connected to the light source line port.

The selected charge-coupled device CCD device is a DFK23G445 camera. When in use, support the camera on the camera bracket. The camera bracket is mainly composed of a chassis, a bracket and a runner. The bracket and the chassis form an angle of about 70°. The bracket chassis is parallel to the horizontal plane. After the CCD camera is installed on the bracket, it is perpendicular to the assembly line. It is flat and just above the parts flowing on the assembly line. The camera terminal mesh is connected to the PC mesh, and the camera power supply is connected to the power socket. The runner is used to adjust the focal length between the CCD camera lens and the parts to be inspected. In order to obtain better. For the image, the position of the light source and the CCD is very important. Therefore, the position of the CCD and the light source can be determined by visual inspection before the measurement, and then accurately adjusted according to the actual shooting situation. DFK23G445 camera volume is height: 29 mm, width: 29 mm, length: 57 mm.

2.2 Image Processing

The gray-scale image is preprocessed, and the specific preprocessing steps include image enhancement and edge detection [4].

The processing method used in image enhancement is gray-scale transformation. A gray-scale transformation function based on particle swarm optimization is designed to implement gray-scale transformation of gray-scale images.

1) Normalized original image

The image to be processed and the enhanced image are represented by d and r, and the image pixels with coordinates (X, Y) are represented by $d(X, Y)$. Normalize d to obtain image D:

$$D(a, b) = \frac{d(X, Y) - K_{\min}}{K_{\max} - K_{\min}} \tag{1}$$

In formula (1), $K_{\max}$ refers to the maximum gray value of the d image; $K_{\min}$ refers to the minimum gray value of the d image.

2) Using the particle swarm algorithm PSO, initialize the speed a_j and position b_j of each particle, and update the optimal position L_{best} and the optimal position U_{best} passed by the group according to the fitness function. At the same time, update the corresponding a_j and b_j until the optimal solution c and d [5] are found.

When contrast transformation is used for image enhancement, the measure of similar variance shown in the following formula is used as the fitness function of particle swarm optimization algorithm to determine the optimal position of particles.

$$E(D_a) = \frac{1}{m} * \sum_{X=1}^{Q} \sum_{Y=1}^{P} K^2(X, Y) - \left| 1/m^* \sum_{X=1}^{Q} \sum_{Y=1}^{P} K(X, Y) \right| \tag{2}$$

In formula (2), Q is the width of the image; P is the height of the image; m is the number of particles; $K(X, Y)$ is the gray value of the d image.

The larger $E(D_a)$, the higher the image quality and the better the visual effect of the image.

3) Using particle swarm optimization algorithm to find

After the two optimal parameters c and d of the Bata function, use formula (3) to transform the image D to obtain O, and use formula (4) to inversely transform O back to the gray space to obtain the final enhanced image R.

$$O = O(c, d; D) \tag{3}$$

$$R(X, Y) = (K_{\max} - K_{\min}) * O + K_{\min} \tag{4}$$

Edge detection uses an improved edge detection method.

1) Extension of Sobel operator

In view of the defects of the conventional Sobel operator, the edge detection algorithm can be improved. First, Sobel operator lacks the correlation information of adjacent points outside the horizontal and vertical directions, resulting in the loss of

image edge information. In order to ensure the integrity of image edge information, Sobel operator should be extended to calculate image gradient information in multiple directions. Sobel operator is extended in 45° direction and 135° direction. There are 4 expanded operators in total, which increases the gradient information of adjacent points, making the gradient operation more complete. Since the gradient is a vector, it contains both magnitude and direction information. However, if it is synthesized strictly according to the vector synthesis rule, the amount of computation is extremely large, so the gradient synthesis is approximated as follows: in the amplitude, the maximum value in the four directions is selected as the gradient value of this point; in the direction, the direction with the largest amplitude is determined is the gradient direction of this point. The gradient direction matrix is D(x,y). D(x,y) = 1 means the gradient magnitude direction is 0°, D(x,y) = 2 means the gradient magnitude direction is 45°, D(x,y) = 3 means the gradient magnitude direction is 90°, D(x,y) = 4 indicates that the gradient magnitude direction is 135°.

Where the gradient is represented by a two-dimensional vector composed of amplitude and direction.

2) Edge refinement

In order to accurately locate the image edge information, non maximum suppression can be used. Because the gray value jump at the edge of the original image is not an ideal step signal, but there is a transition interval. When the Sobel operator is used to calculate the transition interval, all the pixels in the interval will be judged as edge parts, resulting in a wide edge detected in the image. Non maximum suppression can effectively solve this problem. Non maximum suppression is a process of finding the local maximum in the direction perpendicular to the image edge, that is, in the same direction as the gradient maximum.

3) Binarization threshold that conforms to the visual characteristics of the human eye

Aiming at the problem of adaptive binarization threshold setting, a fitting scheme that conforms to human vision is proposed on the basis of consulting relevant materials and conducting a large number of tests.

There are two aspects of human vision that need to be considered, one is the sensitivity of the human eye to the image structure, and the other is the sensitivity of the human eye to the gray value.

According to the noise structure mask theory, the human visual system has a low sensitivity to noise at the edge of the image or where the image structure is fine, and a high sensitivity to noise in the smooth area of the image. The gradient image can better reflect the structure of the image. If the gradient image has a large value in an area, the structure here is relatively fine, and the human eye has a low sensitivity to noise; While the area with small gradient value is the smooth area of the image, and the human eye has high sensitivity to noise [6]. Therefore, the gradient image of the image to be processed should be considered when setting the threshold. In view of the above problems, the mean value of the data in the 3×3 window of the gradient image can be judged. If the mean value in the window is large, it means that the original image has a large change in this area, which belongs to the structured area or the edge of the image. The human eye is not sensitive to noise in this area. To ensure the edge integrity of the image, a small binarization threshold is set. If the mean value in the window is small, it means that this area of the original image is a

smooth area, and the human eye is more sensitive to noise here. A larger threshold should be set. Because the sensitivity of human eyes to different gray values is not the same, even if two pixels have the same gradient, some pixels will be considered as image edges, while others will not. Therefore, in order to better fit the human visual characteristics, a smaller threshold should be used in the gray-scale interval with high visual resolution, and a larger threshold should be used in the gray-scale interval with low visual resolution.

2.3 Moving Target Recognition

The problem of target recognition can be regarded as the problem of target and background classification, which separates the target from the background. A random forest is used as a classifier to eliminate the background, so as to generate a rough target ROI map, and then carry out corresponding scale recognition on the ROI area to realize the recognition of moving targets in gray-scale images [7].

First, in the training process, a large number of positive and negative samples can be added to the random forest for training, and the random forest can be generated according to the FHOG-Lab features; Then, the FHOG-Lab features are extracted for the video frame images to be recognized. If the number is lower, the execution speed of the random forest is accelerated; Then the random forest target probability is obtained through the trained random forest, and the target probability map is generated; Finally, the target center ROI area is enhanced by threshold segmentation and morphological filtering.

Let the training set be α and the total number of samples be M . extract V-dimensional FHOG-Lab features from each training sample, where the training data is represented by a matrix $\beta^{M\times V}$ with M rows and V columns.

A random forest is composed of multiple decision trees. Each decision tree can be viewed as a set of root nodes, branch nodes and leaf nodes. The root node can be regarded as a special branch node; Branch nodes can be regarded as a weak classification, mainly including feature number $v \in \{1, 2, \cdots, V\}$ and threshold information; The leaf node is used to calculate the target probability, mainly including information such as various samples of the node.

During training, the data in column V^* of V will be randomly selected, and the best column (with the smallest class purity) v^* will be selected according to the corresponding classification mechanism as the feature number of the current node and the threshold will be calculated. First, calculate the category purity of the current node. If the value is greater than zero, it means that the current node belongs to a branch node and can continue to split; Where samples smaller than the threshold are split to the left child node, otherwise they are classified to the right child node. When the class purity is close to zero or the sample is less than the minimum sample number, the termination condition is met, and the current node is the leaf node.

Generally, the expected value of information (entropy) is used to calculate the category impurity of branch points:

$$Z = -\sum_{j=1}^{l} q\left(\chi_j, \delta\right) \log_2 q\left(\chi_j, \delta\right) \tag{5}$$

In formula (5), l refers to the number of categories; $q(\chi_j, \delta)$ refers to the probability that δ belongs to class j.

Assuming that the trained random forest consists of m_f decision trees $\{R_f | f \in [1, m_f]\}$, and each detection window is represented by a vector feature t_i, the target probability generated by a leaf node passed to each tree is:

$$w_g(t_i) = q(\chi_{\text{target}}, \delta_{leaf}) \tag{6}$$

In formula (6), $q(\chi_{\text{target}}, \delta_{leaf})$ refers to the probability that the training set reaching the target node belongs to the leaf like node sample.

The final target probability obtained is:

$$w(t_i) = \sum_{g=1}^{m_f} w_g(t_i) \tag{7}$$

Among them, $w(t_i)$ is the probability that the recognition window belongs to the target, and thus the target probability map Y_w can be obtained. The threshold t is set, and the area larger than the threshold has a greater probability of belonging to the target, and the position smaller than the threshold will be regarded as the background. Then, the closing operation in morphological filtering is used to fill the small space in the target binary image object segmented by the threshold value, connect adjacent objects and smooth the target boundary, enhance the target ROI area, and ensure that the real target area is not missed.

Through the simple classification of random forest, the approximate center position of possible targets in the image can be quickly obtained. Therefore, the target ROI region can be used to provide the target central ROI region for subsequent target recognition, and the corresponding scale recognition of the classifier can be directly carried out in the target central ROI region, avoiding a large amount of time consumption required to search for targets in the whole image.

In recognition, the background is first separated by the trained random forest, and the ROI area of the target is obtained; Then the LIBLINEAR offline training of the linear SVM is used to generate the basic classifier. Set the window to browse the image sequence, calculate the corresponding scale on the ROI area, use the basic classifier trained in the learning stage to classify the obtained feature vector, and mark the target area with a rectangular frame [8].

Since the random forest can be trained in a large number of training samples, the problem of insufficient training samples for SVM is solved. When training the LIBLINEAR operator, the target samples and the difficult samples of random forest recognition errors can be used to complete the offline training of the classifier, and finally the best robust recognition operator can be obtained. The specific steps of offline training of classifier are as follows:

Step 1: Prepare the samples for training the linear classifier, including the positive sample set and the negative sample set;

Step 2: Extract the 34 channel FHOG-Lab features of the positive and negative samples used for training, and label the samples;

Step3: Train the features of the positive and negative samples according to the format required by the LIBLINEAR of the linear SVM, and generate the initial model φ_0 on the initial positive sample set and negative sample set ε_0 by means of Bootstrapping, and then execute φ_0 to negate all the sources. Scanning and detection of samples, and classify the detected wrong sub-images as difficult negative samples $J(\varphi_0, \varepsilon_0)$, and then perform secondary training on φ_0 with all positive samples, negative samples and difficult negative samples to obtain model φ_1, and then detect the source negative samples. The difficult negative sample $J(\varphi_1, \varepsilon_1)$ is obtained, and this cycle is repeated 3 times to reduce the false detection rate, and finally the final training result is saved.

In order to speed up the removal of background by random forest, the image is processed by 4×4 interval sampling, and the target probability value of the unused point is obtained by taking the nearest neighbor among the 4 neighboring points of the interpolation point. This can reduce the operation time by 16 times and make the random forest classification operation consume a small amount of time, so that the recognizer can quickly identify the target.

2.4 Moving Target Stability Tracking Detection

Combining the Camshift algorithm and the Kalman filter algorithm, the existing moving target tracking method is improved, and the stable tracking of the moving target is implemented.

The Camshift algorithm obtains the color probability histogram of the target area by counting the color histogram of the target area, so as to realize the tracking of the target object. Camshift can solve the occlusion problem of the target object during the moving process, maintain continuous tracking of the object, and the tracking effect is better [9]. If the target to be tested moves too fast and the target position exceeds the Camshift search window, the target tracking will diverge. The idea of adding Kalman filter on the basis of Camshift is to predict the coordinates of the target in subsequent frames through the motion state of the target object. The tracking of the target through the Camshift tracking algorithm alone, or the tracking of the target through the Kalman filter alone cannot achieve the accurate tracking effect in practical applications. Therefore, the Camshift algorithm and the Kalman filter algorithm can be combined to improve the existing moving target. tracking method.

Suppose that in the K-th frame scene, the target object is located at the coordinate A, and then moves to the coordinate B in the K+1-th frame scene. When the Camshift tracking algorithm implements target tracking alone, it first creates a search window at the A coordinate, then iteratively searches along the path L, and finally obtains the actual coordinates of the target object in the next frame of image. In the process of Camshift processing, Kalman filter prediction is added, and the actual coordinates of the target in the K+1 frame image are predicted by the information of the moving target in the K-th frame. The actual location B is closer than the location A. The initial search window is first established in the range of point C, and the Camshift algorithm is performed to track, and the actual coordinate B can be found as long as the iterative search along the path G is performed.

The improved algorithm flow of Camshift fused with Kalman is as follows:

(1) First, read the recognized first frame image, frame select or automatically detect the target object in the image, and calculate the center coordinate point of the target object.
(2) The initial state color distribution histogram of the tracking target area is statistically calculated, and the color probability projection is obtained through formula calculation, and then the initial Kalman filter is established.
(3) Obtain the actual coordinates of the center point of the target object in the current frame image, estimate the coordinate position of the center point in the subsequent frame according to the Kalman filter prediction equation, then create a search window at the predicted coordinates, and iteratively search for the real coordinates of the moving target through the Camshift tracking algorithm until successful matching.
(4) The Kalman filter takes the calculated coordinate position as the observation value, updates the Kalman filter parameters, maintains the accuracy of the Kalman tracking algorithm, then extracts the subsequent frame images, and continues to track the target object according to the above steps.

The occlusion problem of the target is an unavoidable problem when dealing with target tracking in engineering applications, which directly affects the accuracy of moving target tracking. If the moving target encounters occlusion, the Kalman filter predicts its coordinates in the current frame through the target object position information of the previous frame image, and the target size calculated by the Camshift tracking algorithm at this time is very small. If the moving target is still in the occlusion area in the next frame, use the predicted value of the Kalman filter as the observation value, update the parameters of the Kalman filter, calculate the predicted center position as the initial coordinate of the Camshift iterative algorithm, and then use the target tracked by Camshift. The size is corrected to the original size of the moving target.

Then, the stability detection of moving targets is implemented by the background difference method, and the selected background difference method is the ViBe algorithm.

The focus of the algorithm mainly includes three aspects: model selection, model initialization, model update mechanism.

(1) Pixel background modeling

The basic idea of ViBe background modeling method is to model the background of each pixel, store a set of samples for each pixel, and the sample values in each sample set record a series of values displayed by the pixel at the same or adjacent positions in history. Read the current frame image and compare the similarity of all new pixel points with the sample set one by one. If the sampling value is relatively close to the sample set, it can be considered that it belongs to the background with a high probability. Otherwise, the observation value with a large difference is judged as a moving target.

The pixel value of the marked point y is $\eta(y)$, and η_i is used to record the sample value of the background with the serial number i, and the background model formula is represented by the set of n background sample values:

$$\iota(y) = \{\eta_1, \eta_2, \cdots, \eta_n\} \tag{8}$$

Define a sphere with $\eta(y)$ as the center and r as the radius, and calculate the number of sample points in the union of the sphere and the background model formula, denoted by #. A minimum threshold #mim is set in advance, and then # and #min are compared. If #>#min, the pixel is judged as the background, otherwise it is judged as the foreground target. The radius r of the sphere and the threshold #in directly affect the accuracy of model detection.

(2) Model initialization

Many traditional background modeling methods require a long period of video image sequence in the model initialization phase, analyze a long period of data, and estimate the time distribution. The vibe modeling algorithm can use one image to complete the initialization of the background model. In the background model, each pixel contains λ sample values. These λ model sample values are obtained by sampling the neighboring pixels of this point in an image, and the neighboring pixels can be selected in a random manner, and these pixels have similar space-time distribution characteristics.

When selecting domain pixel values, the selection range should not be too large, which can reduce the statistical relationship between pixels, and also avoid the stored background model sample set being too large.

(3) Model update mechanism

The update strategy of ViBe modeling method is a combination of conservative update strategy and foreground point count method. Conservative update strategy: do not use foreground points to fill the background model, assuming that the stationary area is judged as the moving foreground, then this area is always regarded as the moving target foreground. Foreground point counting: Count and count the judgment results of each pixel point. If a pixel point is judged as a foreground point for ϖ consecutive times, the pixel point is judged as the moving target foreground, and if the condition is not established, it is classified as a background pixel point. Assuming that the judgment result of the current pixel belongs to the background point, for this pixel, the probability of updating its own background model is $1/\varsigma$, and the probability of updating the model sample value of its domain point is also $1/\varsigma$. Randomly select the replaced sample values to ensure that the smooth life cycle of the sample values is exponential decay. The detection of ViBe background modeling method is relatively stable, which can effectively adapt to interference such as changes in light brightness or camera rotation jitter. The algorithm is simple in calculation, low in space complexity, faster in processing than other background modeling methods, and takes up less system resources [10].

When using the ViBe algorithm for background modeling, if there is a moving target in the first frame image, the target is judged as a background point. Misidentified as foreground points of motion, a phenomenon commonly referred to as "ghosting". ViBe's update strategy can solve the "ghosting" problem caused by the first frame background image. However, the speed of this update is slow, and it takes a long period of time to eliminate the ghost area. If there are other moving targets in the process, the extracted foreground target will be wrong. At the same time, the method of vibe background modeling can not deal with the problem of false background introduced by the stop of moving objects, and it must be further studied and improved. For the first σ frames of the video image sequence, the moving area of the

foreground object can be obtained first by the inter frame difference method. When using vibe background modeling, reduce the update probability of vibe algorithm of pixels within the range of moving targets and increase the update probability of pixels outside the range. This method can speed up the elimination of ghost areas.

After the moving object detection process, the extracted binarized foreground image area often contains holes or narrow connecting lines, etc., and the image sequence obtained during the moving object detection process is morphologically processed by selecting the operation method of closing first and opening later. The opening operation is the process of first etching and then expanding the set ζ through the structural element ξ, which can break small discontinuities and eliminate small protrusions. The closing operation is the opposite of the opening operation, in which ζ is first expanded and then eroded through the structural element ξ, which can fill small gaps and voids. Assuming that there is a set ζ and a structural element ξ, the open operation of the set ζ through the structural element ξ can be expressed as $\zeta \circ \xi$, as follows:

$$\zeta \circ \xi = (\zeta \Theta \xi) \oplus \xi \tag{9}$$

The closing operation of ξ to ζ can be expressed as $\zeta \cdot \xi$, as shown in the following formula:

$$\zeta \cdot \xi = (\zeta \oplus \xi) \Theta \xi \tag{10}$$

In this way, the morphological processing of the moving target detection image is realized, and the stability tracking and detection of the moving target in the video image is completed.

3 Tracking Detection Performance Test

3.1 Experimental Platform and Experimental Data

The experimental platform is Intel Core i5 @ 2.4 GHz, 4 GB memory, Windows7 operating system, and the software environment is: Matlab2012b, Visual Studio2010, Opencv2.4.8. Before conducting the experiment, the target box initialization operation is performed on the target to be tracked in the first frame of the video sequence. The following tests are based on manually initializing the target area first.

In order to test the tracking detection effect of the design method, the performance of the design method is tested by selecting some moving target video sequences with relatively complex environments. The complex environment includes a series of situations, such as partial occlusion and total occlusion of moving objects, changes in the scale of moving objects, interference of light intensity on moving objects, and too fast moving speed of moving objects.

A total of 6 videos containing various situations are used for performance testing.

The video sequences (a) and (c) are the tracking of moving vehicles in environments with large changes in light intensity.

Video sequence (b) is motion tracking of a fast car in an environment where leaves are disturbed.

The content of video sequence (d) is the tracking of a fast moving ball.

Video sequences (e) and (f) track pedestrians on the road.

Video sequence (e) includes object occlusion and lighting changes.

In the video sequence (f), the target is occluded, but there is basically no obvious change in lighting.

The specific data of the experimental video sequences are shown in Table 1.

Table 1. Specific data of experimental video sequences

Experimental video sequence	(1)	(2)	(3)	(4)	(5)	(6)
Name	Car1	Car2	Car3	Ball	Manl	Man2
Total frames	542	635	852	185	365	755
Contains targets	452	512	632	154	326	701
Number of frames	have	nothing	have	nothing	have	nothing
Illumination change	nothing	have	nothing	have	have	have

The performance of the design method is evaluated from the following parameters: the correct number of frames processed, the correct rate of stable tracking detection, the number of frames processed per second, the tracking detection error, and the tracking detection time.

3.2 Test Results

The test results of the correct processing frame number of the design method are shown in Table 2.

Table 2. Correct processed motion video frame test results

Experimental video sequence	Number of correctly processed frames (frames)
(1)	448
(2)	509
(3)	630
(4)	152
(5)	325
(6)	700

According to the test results in Table 2, for the six experimental video sequences, the correct processing frames of the design method are high.

Table 3. Testing accuracy of video image moving target stability tracking and detection

Experimental video sequence	Accuracy rate of stability tracking detection (%)
(1)	86.32
(2)	84.63
(3)	87.54
(4)	86.32
(5)	81.20
(6)	79.32

The test results of the stability tracking detection accuracy of the design method are shown in Table 3.

The test results in Table 3 show that because the six video sequences contain a variety of situations, the accuracy of the stability tracking detection of the designed method is already high, indicating that its drift suppression ability is better.

The test results of the number of frames per second processed by the design method are shown in Fig. 1.

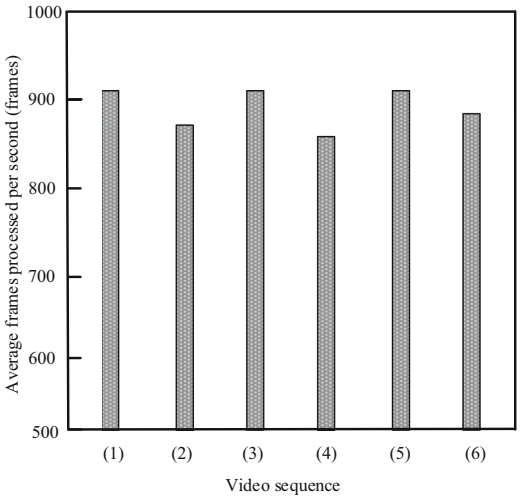

Fig. 1. Test results of the video frames processed per second

According to the test results in Fig. 1, the number of frames processed by the design method per second is higher than 800 frames.

The tracking detection error test results of the design method are shown in Table 4. The tracking detection error refers to the distance between the center coordinate of the tracking window and the actual center coordinate of the target in each frame image. The smaller the error, the higher the tracking accuracy.

Table 4. Tracking detection error test results of the design method

Experimental video sequence	Frame number	The coordinates of the real center point of the target	Tracking detection error of the design method (mm)
(1)	Frame 10	(563,48)	4.62
	Frame 20	(525,46)	4.20
(2)	Frame 10	(521,41)	4.75
	Frame 20	(504,43)	4.71
(3)	Frame 10	(574,35)	4.62
	Frame 20	(570,42)	3.85
(4)	Frame 10	(965,40)	3.95
	Frame 20	(658,41)	4.20
(5)	Frame 10	(704,51)	4.69
	Frame 20	(745,30)	4.71
(6)	Frame 10	(685,52)	3.52
	Frame 20	(685,62)	3.69

According to the error test results in Table 4, in the six experimental video sequences, the tracking and detection errors of the designed method are all lower, which proves that the tracking and detection performance of the designed method is better.

The average tracking and detection time-consuming test results for six experimental video sequences are shown in Fig. 2.

The test results in Fig. 2 show that, for the six experimental video sequences, the average tracking and detection time of the design method is less than 3000s, indicating that the design method has low time consumption.

In summary, because the method in this paper preprocesses the edge of the video image first, the image quality is enhanced. This saves time for image background culling mentioned below. After generating a rough target ROI map, the ROI area is identified with a corresponding scale, thereby realizing the stable tracking of moving objects in the image.

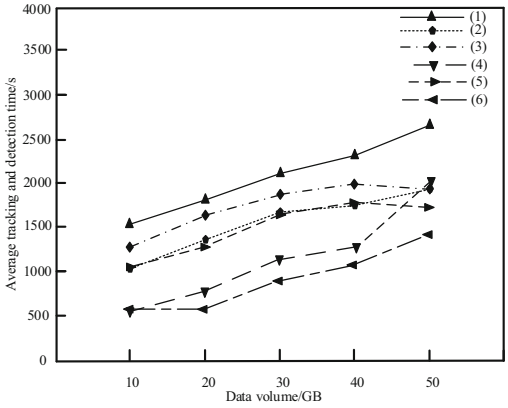

Fig. 2. Average tracking detection time-consuming test results

4 Conclusion

Moving target tracking and detection in grayscale images is a very hot topic at present. It has very broad application prospects in national defense, military, health, industrial intelligence, intelligent control, video surveillance, etc. Researching this technology has high practical application value. However, due to the complexity of the actual situation, there is still a big gap between the actual result and the ideal situation. Based on the existing results, this topic proposes a more effective method. On the basis of preprocessing the video image, the image background is proposed by using the random forest algorithm. And corresponding scale identification is performed on the ROI area. At the same time, it combines the Camshift algorithm and the Kalman filter algorithm to perform stable tracking on moving targets. It can track and detect moving targets more accurately and efficiently in complex environments. According to the experimental results, this method can correctly process more frames and lower tracking and detection errors.

References

1. Yoo, K., Chun, J.: Analysis of optimal range sensor placement for tracking a moving target. IEEE Commun. Lett. **24**, 1700–1704 (2020)
2. Andriyanov, N.A.: Application of computer vision systems for monitoring the condition of drivers based on facial image analysis. Pattern Recogn. Image Anal. Adv. Math. Theory Appl. USSR **31**(3), 489–495 (2021)
3. Xu, S., Wang, J., Shou, W., et al.: Computer vision techniques in construction: a critical review. Arch. Comput. Methods Eng. State Art Rev. **28**(5), 3383–3397 (2021)
4. He, Z., Li, H., Wang, Z., et al.: Adaptive compression for online computer vision: an edge reinforcement learning approach. ACM Trans. Multimedia Comput. Commun. Appl. **17**(4), 118.1–118.23 (2021)
5. Xu, X., Yuan, Z., Wang, Y.: Multi-target tracking and detection based on hybrid filter algorithm. IEEE Access **8**, 209528 (2020)
6. Zheng, Y., Li, Q., Wang, C., et al.: Magnetic-based positioning system for moving target with feature vector. IEEE Access **8**, 105472–105483 (2020)

7. Wang, W., Pei, Y., Wang, S.H., et al.: PSTCNN: explainable COVID-19 diagnosis using PSO-guided self-tuning CNN. Biocell **47**(2), 373–384 (2023)
8. Wang, W., Zhang, X., Wang, S.H., et al.: Covid-19 diagnosis by WE-SAJ. Syst. Sci. Control Eng. **10**(1), 325–335 (2022)
9. Xiong, Y., Liu, Y., Cheng, W., et al.: Research on vision system of grasshopper bionic robot. Comput. Simul. **38**(12), 345–348, 361 (2021)
10. Cao, J., Sun, Y., Zhang, G., et al.: Target tracking control of underactuated autonomous underwater vehicle based on adaptive nonsingular terminal sliding mode control. Int. J. Adv. Rob. Syst. **17**(2), 451–468 (2020)