Received August 2, 2020, accepted August 21, 2020, date of publication August 26, 2020, date of current version September 11, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3019503*

# Video Key Frame Monitoring Algorithm and Virtual Reality Display Based on Motion Vector

ZHE WANG[1] AND YAN ZHU[2]

[1]Department of Physical Education, Qufu Normal University, Rizhao 276800, China
[2]College of Marxism, Qufu Normal University, Rizhao 276800, China

Corresponding author: Zhe Wang (wangzhe20011@126.com)

This work was supported by the Social Science Planning and Research Project of Shandong Province: research on collaborative supply and path optimization of social sports resources in Shandong Province from the perspective of national fitness under Grant 20CTYG05.

**ABSTRACT** In this article, a motion vector-based video key frame detection algorithm is proposed to solve the problem of miss election and missing selection caused by the difficulty in detecting the moving target characteristics of the video key frame. Firstly, the entropy of adjacent frame difference and the two-dimensional entropy of image are introduced, and the combination of the two is taken as the measurement of the difference between video frames. Secondly, outliers are detected by statistical tools to obtain the lens boundary, thus realizing the adaptive lens detection of video content. Then, ViBe algorithm is used to detect the foreground object in the video sequence and extract the scale-invariant feature transformation features of the foreground moving object. Finally, the motion vector is introduced, and the sum of the block matching results is motion vector by partitioning the two adjacent frames and performing block matching. The magnitude of motion vector reflects the intensity of motion in the video, so active and inactive motion regions are obtained, and the similarity of video frames is calculated according to the defined formula, and key frames are extracted in these regions respectively. The experimental results show that the detection algorithm proposed in this article improves the video results with rich motion information obviously, and the objective indexes and subjective scores are improved to some extent, which improves the universality of the algorithm. In addition, this article also studies the display mode of key frame extraction results in virtual reality environment. The key frame display mode of video in virtual reality is optimized mainly by changing the display mode of information and changing the scene and testing the result of user task execution.

**INDEX TERMS** Video, key frame, monitoring algorithm, virtual reality display, motion vector.

## I. INTRODUCTION

In recent years, with the continuous progress of Internet video technology and the continuous update of mobile video devices, browsing video has increasingly become an important way for people to understand the world and obtain information [1], [2]. The information content carried by video data is much larger than that of text and image data. Besides, video types tend to be diversified, and similar videos are also different in content. It is difficult for video users to quickly and accurately obtain the content they want from such a large amount of video information. For users, there is an urgent need for a way to quickly and accurately understand video content [3], [4].

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

A kind of method for text annotation of video content information is called text description-based retrieval method [5], [6]. This method has the disadvantage of losing video time information. Another kind of method is the video retrieval method for specific video content [7], [8].

The current VR technology can realize the functions of VR video playback, fast forward, pause and panoramic view of ordinary video, but compared with desktop and mobile platform, video display under virtual reality still has many shortcomings. Among them, video summary technology in VR environment is still in the blank field. This article not only introduces the video key frame extraction technology, but also studies and explores the key frame display mode and method in VR environment.

In this article, a motion vector-based video key frame detection algorithm is proposed to solve the problem of miss election and missing selection caused by the difficulty in detecting the moving target characteristics of the video key frame. Firstly, the entropy of adjacent frame difference and the two-dimensional entropy of image are introduced, and the combination of the two is taken as the measurement of the difference between video frames. Secondly, outliers are detected by statistical tools to obtain the lens boundary, thus realizing the adaptive lens detection of video content. Then, ViBe algorithm is used to detect the foreground object in the video sequence and extract the scale-invariant feature transformation features of the foreground moving object. Finally, the motion vector is introduced, and the sum of the block matching results is motion vector by partitioning the two adjacent frames and performing block matching. The magnitude of motion vector reflects the intensity of motion in the video, so active and inactive motion regions are obtained, and the similarity of video frames is calculated according to the defined formula, and key frames are extracted in these regions respectively. In this article, the key frame extraction effect is put into the virtual 3D scene for display, and the traditional display device is replaced by the virtual reality device, providing a new video summary browsing experience for users. The user experience under virtual reality environment is studied and the task execution ability of users under different display modes is explored.

## II. RELEVANT WORKS

Video key frame technology has been developed for many years since it was proposed. Histogram is a kind of statistical chart which is calculated and generated by counting the pixel value of the image. It is used by many algorithms because it can reflect the global color information of the image. Cui *et al.* [9] proposed to use the statistics of the gray values of the two images to obtain the two gray histogram, and calculate the difference between the two histograms to reflect the degree of difference between the two images. Griffiths and Boehm [10] improved the research results of the above algorithm, improved the histogram of gray value into the color histogram of RGB three-channel, and used the gap of the histogram of RGB three-channel results to reflect the degree of difference between color images. Goyal *et al.* [11] proposed an algorithm for automatic edge detection. Luo *et al.* [12] proposed a video key frame algorithm based on image segmentation. This method extracts the original video sequence with the same step length frame. After that, the extracted frames are divided into image blocks, each image block is 4 pixels in length and width. For each image block of each frame, it is matched with all image blocks of the next frame to find the most similar one, and the moving distance of this block is obtained, so as to obtain the gap between two consecutive frames. Song *et al.* [13] also proposed a very easy uniform sampling method based on video lens segmentation results. In this method, a certain number of key frames are extracted according to the user in advance, and the video content is extracted according to the original frame sequence every fixed frame to obtain uniform extraction results. Kim *et al.* [14] combined color distribution with the use of motion attributes. Since this method is not an adaptive algorithm, many parameters need to be set artificially. If the settings are unreasonable or the content of the video varies greatly, it may not be widely used, such as a lot of redundancy, or the general idea of the video cannot be fully expressed, etc. Li and Liu [15] proposed an algorithm to calculate the optical flow to reflect the degree of motion. Optical flow is a measure that can reflect the static degree of video information in a video. The optical flow value of a video frame is drawn as a curve according to the time series, and the minimum point in the curve is extracted. But some of these minima are not between two consecutive maxima, which will result in incorrect results. Aote and Potnurwar [16] set a number of clustering centers, that is, the video was divided into several parts, and finally obtained the key frame extraction results by extracting clustering centers. But there are some prominent problems, this algorithm first clustering way makes the time information is lost, something similar is not continuous time lens could be combined to cause information loss, second selection clustering center and clustering algorithm choice will affect the algorithm robustness, clustering result can't reflect video of content. Shabaninia *et al.* [17] used Shannon entropy in information theory to calculate the histogram divergence of color probability distribution between successive frames, which reflected the difference in color distribution between adjacent frames. This algorithm has the advantages of simple logic and fast calculation, but it is not good for the extraction effect. The extracted key frames sometimes have the disadvantage of not fully containing the video content, and a part of the original video sequence may not be accurately extracted. And the universality of the algorithm needs to be improved due to the artificial threshold. Yu and Principe [18] used mutual information in information theory to calculate the color probability distribution histogram divergence between successive frames, which also reflected the color distribution gap between two adjacent frames. Because different strategies are adopted for key frame extraction, the algorithm achieves excellent results. But at the same time, because the calculation of mutual information depends on the result of joint probability distribution between every two frames of histogram, the calculation amount is large, and the algorithm is easy to miss the drastic changes of video content.

Virtual reality (VR) [19], [21] is a hot computer graphics technology in recent years. At present, the frontier technology of virtual reality technology points to the direction of improving the user experience, such as improving the sense of immersion, improving the sense of reality and reducing vertigo.

## III. RELEVANT CONCEPTS

### A. VIDEO OVERVIEW

According to different shooting content and application fields, digital video can also be divided into different types, such as news video, game video, and surveillance video [22], [23]. Surveillance video as a kind of special form of digital video, is mainly used in areas such as security, road monitoring, record the contents of the mainly occur in a particular area of the events is given priority to, its means of object characteristics, record and final formation of the video formats such as with other forms of digital video there is a big difference.

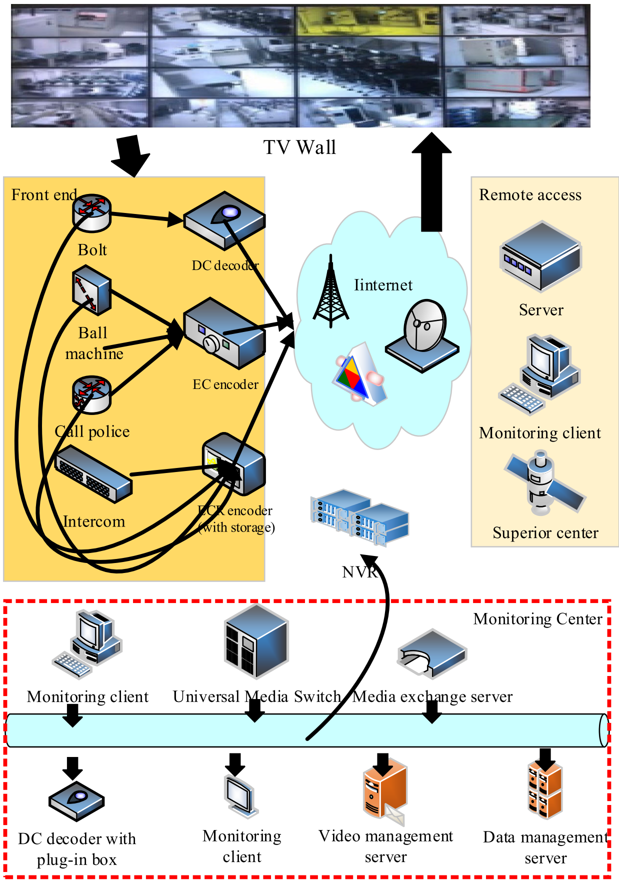

**FIGURE 1.** A typical network video surveillance system.

It can be seen from Figure 1 that the video file can be divided into the organizational structure of video stream scene, lens, and frame from top to bottom.

In the video hierarchy structure diagram shown in Figure 2, the frame is located at the bottom of the four-layer structure and is the smallest unit that constitutes video data. The essence of a frame is a still image, which has the general characteristics of an image. Lens refers to the video sequence shot by the camera in a continuous time period and spatial region. It consists of several adjacent frames and is the basic wood element of the video sequence. A scene is composed of shots with similar contents and a cluster of shots related to higher level contents, which describes the same event from different angles. Key frames are used to describe one or more frames of a shot, similar to the role of a summary. Using key frames to describe a shot can remove redundant information in the shot and reflect the main content of a shot while compressing the video data.

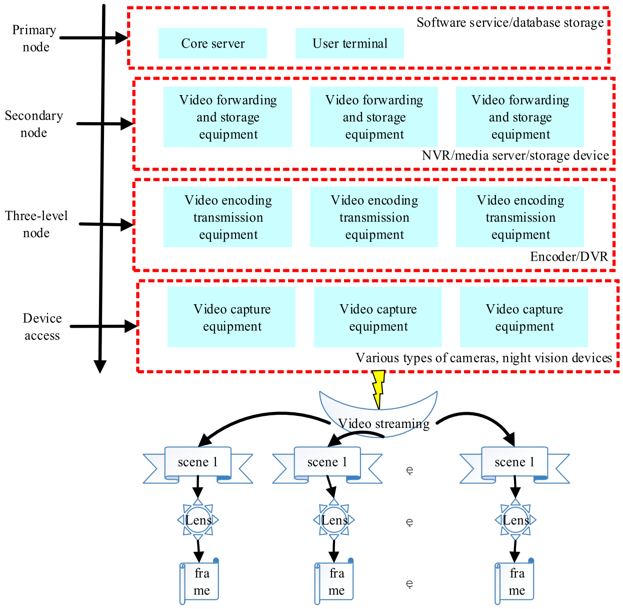
Primary node
Software service/database storage
Core server
User terminal
Secondary node
Video forwarding and storage equipment
Video forwarding and storage equipment
Video forwarding and storage equipment
NVR/media server/storage device
Three-level node
Video encoding transmission equipment
Video encoding transmission equipment
Video encoding transmission equipment
Encoder/DVR
Device access
Video capture equipment
Video capture equipment
Video capture equipment
Various types of cameras, night vision devices
Video streaming
scene 1
scene 1
scene 1
Lens
Lens
Lens
fra me
fra me
fra me

**FIGURE 2.** Hierarchical structure of video.

Frames, shots, scenes and video streams are defined as follows:

Frame: A static image located within a sequence of consecutive frames in time. It is the smallest unit of video composition.

Shot: A video sequence in which the camera presses start to stop. From the perspective of content, it refers to a series of frames with the same or similar content shot at adjacent locations in continuous time.

Scene: Consists of similar content and time continuous shots, usually with strong semantic correlation, describing a separate story unit, which is a semantic component unit of a segment of the eye cheek.

Video stream: The highest level of a Video, at which you can add a global explanation of the attributes and content of the Video file.

### B. ANALYSIS OF VIDEO KEY FRAME CHARACTERISTICS

Surveillance video is composed of many image frames connected in chronological order. Key frame a frame used to define a significant change in the content of an image sequence in a video is a frame containing key content. The key frame contains the motion and features of the object. Surveillance video can be regarded as a continuous sequence of images, and the general key frames of surveillance video can be described as follows [24]–[26]:

If video represents a surveillance video and $I_t$ represents an image frame in the video, the following form exists:

$$Video = f(I_t) \quad t = 1, 2, \ldots, T \tag{1}$$

where, variable T represents the maximum frame number of surveillance video. Surveillance video contains a lot of image frames, some of which have the same or similar content, or even meaningless, leading to a lot of redundant information in surveillance video. In order to distinguish from such image frames, key frames are usually used to define the image frames with obvious changes in the image sequence content in the surveillance video, that is, the image frames containing the key content.

### C. ANALYSIS OF STATISTICAL CHARACTERISTICS OF VIDEO KEY FRAMES

Figure 3 shows the key frame, background frame and their corresponding gray histogram in a surveillance video. Gray histogram is an effective tool for statistical analysis of the gray distribution of images [27]. Generally, the horizontal axis represents the distribution range of gray level, while the vertical axis represents the number of pixels in the image with a certain gray level. The histogram shows the distribution of tones in the image, revealing the number of pixels at each brightness level in the image. Most of the spikes are distributed on the left, indicating that the details of the image are concentrated in the shadow, and the whole is dark, lacking in bright parts. The fact that most of the spikes are on the right indicates that the detail of the image is concentrated in the highlight area and lacks dark detail. The fact that most of the spikes are distributed in the middle indicates that the details of the image are concentrated in the middle tone. When the spike undulation is small and evenly distributed, it indicates that the details of the image are evenly distributed in shadows, middle tones and highlights, and the color transition is smooth.

Background frames and histograms of surveillance video are shown in Figure 3(c) and Figure 3(d). According to the statistical distribution characteristics of image gray scale or color histogram, the content of background frames in surveillance video is relatively stable and generally lacks large changes, and their histograms will not change greatly. At the same time, compared with the frame with the boundary of the shot, the overall visual content of the continuous frame in the same shot is similar, and its histogram changes are small. As shown in Figure 3(a) and Figure 3(b), for video frames containing moving objects that suddenly intrude into the scope of surveillance lens, their histogram will change significantly and usually show significant differences, which can be identified as key frames.

## IV. VIDEO KEY FRAME DETECTION ALGORITHM BASED ON MOTION VECTOR

In this article, a video key frame extraction algorithm based on motion vector is proposed, and the algorithm structure is shown in Figure 4. Firstly, the entropy of adjacent frame difference and the two-dimensional entropy of image are introduced, and the combination of the two is taken as the measurement of the difference between video frames. Secondly, outliers are detected by statistical tools to obtain the lens boundary and realize the adaptive lens detection of video content. Then, ViBe algorithm is used to detect the foreground object in the video sequence and extract the scale-invariant feature transformation features of the foreground moving object. Then, the motion vector is introduced, and the active region and inactive region are judged by calculating the magnitude of the motion vector. Finally, the similarity of video frames is calculated according to the defined formula, and the key frame extraction results are obtained for active and inactive regions respectively.

### A. VIDEO LENS DETECTION METHOD

As the basis of video shot segmentation, the calculation of the difference between successive video frames is the first step of the algorithm. The accuracy and complexity of the algorithm are directly affected by which measure is used to calculate the difference between frames. In this article, the entropy of adjacent frames and the two-dimensional entropy of image are selected as the measurement of the difference between video frames.

Image entropy is a statistical form of image features, which reflects the average amount of information in an image [28], [29]. The one-dimensional entropy of the image represents the information contained in the aggregation feature of the gray distribution in the image. Let gray represent the proportion of pixels whose gray value is I in the image, then the one-dimensional entropy S of the gray image can be defined as follows:

$$S = \sum_{i=0}^{RGB} gray_i \log gray_i \tag{2}$$

The one-dimensional entropy of the image can represent the aggregation characteristics of the image gray distribution, but it cannot reflect the spatial characteristics of the image gray distribution, especially the morphological characteristics of the object in the image. In order to better describe the spatial characteristics of image gray distribution, we can introduce feature quantities that can reflect the spatial characteristics of image gray distribution on the basis of one-dimensional entropy to approximately estimate the two-dimensional entropy of the image.

Under normal circumstances, the gray mean of image neighborhood is selected as the spatial characteristic quantity of gray distribution, and the gray scale of image pixel constitutes a feature binary group, denoted as (x,y), where x∈$(0 \le x \le 255)$ represents the gray value of pixel, and y ∈$(0 \le y \le 255)$ represents the gray value of neighborhood.

$$gray_{xy} = P(x, y)/H^*W \tag{3}$$

where, P(x,y) is the frequency of occurrence of characteristic binary group (x,y) in the image that conforms to the gray distribution characteristics of a certain neighborhood, and H × W is the size of the image.

The above formula can reflect the joint feature of the gray value at a certain pixel position in the image and its surrounding pixel gray distribution. In particular, when there are some meaningful objects in the image, the joint feature of this pixel gray distribution is relatively obvious. Therefore, the joint feature of pixel gray distribution is of great value for object detection in images. On the basis of formula (2)

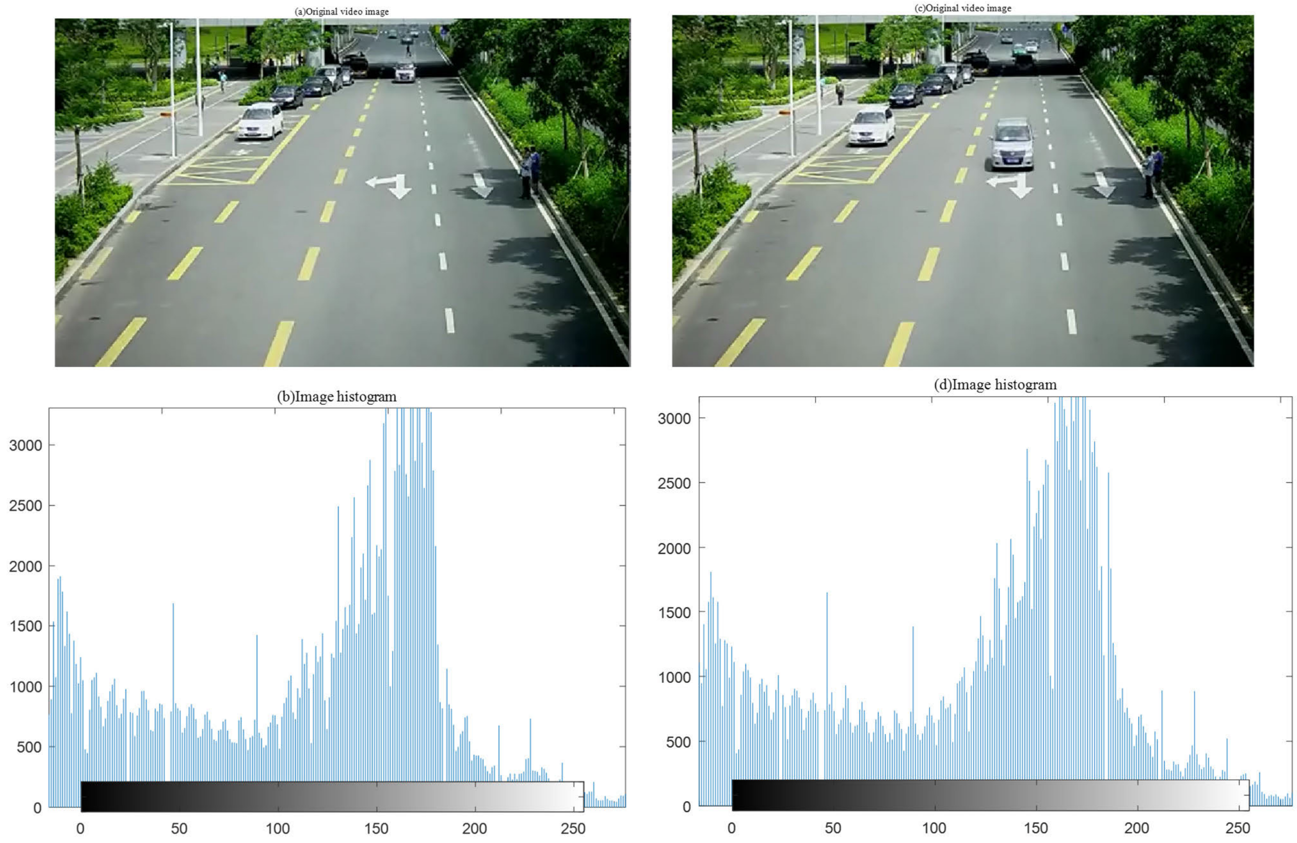

**FIGURE 3.** Key frames and background frames in surveillance video.

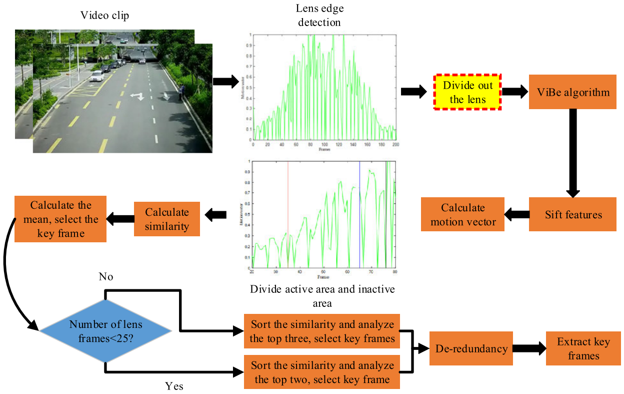
Video clip
Lens edge detection
Divide out the lens
ViBe algorithm
Sift features
Calculate motion vector
Calculate similarity
Calculate the mean, select the key frame
Divide active area and inactive area
No
Number of lens frames<25?
Yes
Sort the similarity and analyze the top three, select key frames
Sort the similarity and analyze the top two, select key frame
De-redundancy
Extract key frames

**FIGURE 4.** Algorithm structure diagram.

and formula (3), the conventional adjacent frame subtraction algorithm and the image two-dimensional entropy calculation method are combined to define the difference measurement as follows:

$$SS = \sum_{i=0}^{RGB} (gray_i \log gray_i - gray_{xy} \log gray_{xy}) \tag{4}$$

To sum up, this method can not only reflect the information contained in the difference image obtained by the frame subtraction algorithm of the surveillance video image, but also highlight the gray level information of the pixel position in the difference image and the comprehensive feature of the gray level distribution in the pixel neighborhood.

Due to the feature of video shot switching, we can judge whether there is enough obvious difference value according to the difference value between video frames obtained by formula (4), that is, whether there is shot switching.

In the practical application, the influence of background and noise needs to be taken into account, so a simple mathematical morphological operation and threshold processing is usually carried out before to remove some interference in the image, such as the background noise caused by the changes of vehicle lights, street lamps and shadows. Therefore, next, this article adopts ViBe algorithm to model the background.

### B. BACKGROUND MODELING ViBe ALGORITHM

ViBe algorithm [30] is a pixel-level background modeling algorithm proposed in 2011. Compared with familiar foreground detection algorithms such as hybrid Gaussian model [31], ViBe algorithm has more advantages in foreground detection and background model updating.

The ViBe algorithm is mainly divided into three parts: firstly, it initializes the background model of pixels in the image. For each pixel initialized N times from its eight-neighborhood, the background model of the pixel can be expressed as

$$ViBe = \{V_1, V_2, \ldots, V_N\} \tag{5}$$

where, variable V represents the pixel value in the sample model, and there are altogether N samples in the model.

Secondly, the image sequence is segmented into foreground objects. The pixel value is compared with the pixel value in the sample set. If its distance from the pixel value in the sample set is greater than a certain threshold, the pixel point is considered as the foreground pixel point; otherwise, it is the background pixel point. In Figure 5, It(x) refers to the pixel value of the current frame, threshold is set as threshold, threshold = 20, and I1, I2., it is the pixel value in the sample set.

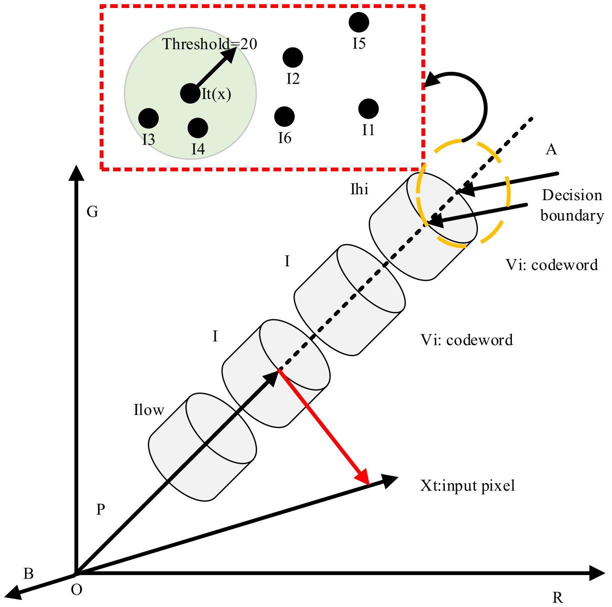

FIGURE 5. ViBe classification diagram.

Finally, the background model is updated.

SIFT features of images containing moving objects extracted by ViBe algorithm are extracted. SIFT feature is a very stable local feature. The feature vector extracted by SIFT feature point extraction algorithm has the characteristics of invariant to rotation, scale scaling, brightness change and angle of view change, affine transformation, and stability to noise. The acquisition of scale space used in SIFT feature point extraction algorithm requires Gaussian blur to achieve.

The definition of two-dimensional Gaussian fuzzy function is as follows:

$$Gaussian = e^{-(x^2+y^2)/2\text{Ȼ}^2}/2\pi \tag{6}$$

where, the variable Ȼrepresents the standard deviation of a normal distribution. Then the scale-space representation of a two-dimensional image at different scales can be obtained by the image and Gaussian convolution kernel:

$$G\_I(x, y) = G(x, y)^{*}I(x, y) \tag{7}$$

where, $(x, y)$ represents the pixel position of the image, and $G\_I(x, y)$ represents the scale space of the image.

SIFT feature point extraction algorithm detects local extremum as feature points simultaneously in the two-dimensional plane space and difference of Gaussian (DOG) scale space of the image. The DOG operator is shown as follows:

$$DOG(x, y) = (G(x, y, i) - G(x, y))^{*}I(x, y) \tag{8}$$

The generation of an image SIFT feature vector generally includes 4 steps:

(1) Carry out scale space detection to preliminarily determine the location and scale of key points.

(2) The 3d quadratic function was fitted to accurately determine the position and scale of key points, and the key points with low contrast and unstable edge response points were removed.

(3) Make use of gradient direction distribution characteristics of pixels in the neighborhood of key points to specify direction parameters for each key point. And the modulus value of gradient at $(x, y)$ is as follows:

$$\partial(x, y) = \sqrt{\begin{array}{l}(G\_I(x+1, y) - G\_I(x-1, y))^2 \\ +(G\_I(x, y+1) - G\_I(x, y-1))^2\end{array}} \tag{9}$$

$$\Delta(x, y) = j\tan 2\frac{G\_I(x, y+1) - G\_I(x, y-1)}{G\_I(x+1, y) - G\_I(x-1, y)} \tag{10}$$

At this point, the key point detection has been completed, and the scale used by L is the respective scale of each key point.

(4) Generate SIFT feature vectors.

Through the above steps, each key point has three information: position, scale, and direction. Standard $4 \times 4$ sub-regions are set for each key point. Each sub-region uses 8 intercell orientation histograms. 128 data are generated from such a key point, namely, 128-dimension SIFT feature vector is formed.

### C. ACTIVE REGION DIVISION BASED ON MOTION VECTOR

For the cameras-based video key frame extraction method, first the video sequence is divided into shots by the lens boundary detection method proposed in Section 3.1, and then the detection process of active regions are improved by the motion vector measure based on color histogram. Within a shot, areas of the video with significantly varying motion vectors are marked as active and other areas as inactive. We extract a key frame from each active region and a sufficiently long inactive region respectively.

The displacement between the current block and the best matched block is calculated as the motion vector of the current block. If there is a matching block displacement of (p0, q0), so that $Best_b$ gets the minimum value, then the matching block is the $Best_b$ match for the current block.

$$Best_b(p, q) = \frac{1}{H^{*}W}\sum_{h=1}^{H}\sum_{w=1}^{W}\begin{array}{l}|F_{before}(h, w) \\ -F_{before+1}(h+i, w+j)|\end{array} \tag{11}$$

where, the variable (P, Q) is the displacement between the current block and the matching block, namely the motion vector $Best_b$ of the current block. The variables W and H are the width and height of the block, respectively. The variable $F_{before}$ represents the current frame, and $F_{before+1}$ Fbefore

+1 represents the next frame. The variable $F_{before+1}(h, w)$ represents the pixel value of the pixel point at the coordinate $(h, w)$. The sum of the motion vectors of all blocks of the current frame represents the motion vectors of the current frame, which reflects the variation degree of the content of the current frame.

$$Best_f = \sum_{i=1}^{U} Best_b(i) \tag{12}$$

$$Best(i) = Best_f(i)/\max(Best_f) \tag{13}$$

where, the variable U is the total number of blocks divided by the current frame. The variable $Best_f$ is the sum of the motion vector $Best_b$ of all blocks in the current frame. The variable $Best(i)$ is the normalized measure of the current frame motion vector and the final result obtained from the original video.

To detect the boundary of a wave segment, the absolute value of the motion vector difference between two successive frames in a shot is calculated to assess the degree of the wave.

$$CZ = |Best_g - Best_{g+1}| \tag{14}$$

In order to ensure the integrity of motion and reduce redundancy, this article locates the boundary of active region by matching the pattern of weak continuous high CZ value. Weak continuous mode means that the CZ value of continuous frames allowed in an active region is less than a given threshold, but the number of consecutive frames L in this situation must be less than 12.

In the case of a short pause in motion, a complete motion process will be divided into several small motion regions if it is strictly divided according to the threshold, resulting in redundancy. Therefore, the weak continuous mode is used to combine the small interval regions in these complete motions. Since the test video is played at 24 frames per second, half a second is the appropriate length to attract the eye's attention. Therefore, according to this principle, the repeated test results show that L is the most appropriate when the value is 12.

### D. KEY FRAME EXTRACTION ALGORITHM

For each video shot, if there is at least one active region inside the shot, all active regions and inactive regions with a length greater than 24 frames are extracted, and the number of extracted frames is one frame for each region. If there is no active region inside the lens, all the areas of the lens are considered as inactive regions, and a key frame extraction is performed for an inactive region of the lens.

Assume that the active interval is the interval from frame P to Q, and the relative entropy of all frames in this interval and their next frame $SS\left(F_{before}, F_{before+1}\right)$ has been calculated, where the value of before is P to Q-1. Taking the active interval as an example, the selection and calculation of key frames are as follows:

$$F_{keyframe} = \arg\min_{before} \Big|SS\left(F_{before}, F_{before+1}\right) - \sum SS\left(F_{befor}e, F_{before+1}\right)/(Q-P)\Big| \tag{15}$$

**TABLE 1.** Video information.

| Video number | The total number of frames | The lens number | The length |
|---|---|---|---|
| football | 7561 | 55 | 303 s |
| basketball | 5213 | 18 | 204 |
| gymnastics | 3766 | 21 | 152 |
| physical education | 2788 | 32 | 110 |

## V. EXPERIMENTAL VERIFICATION

### A. DATA SET SELECTION AND EVALUATION CRITERIA

In order to ensure the accuracy and objectivity of the video data, this article selected a variety of videos with different characteristics and made repeated tests to evaluate the motion vector-based video key frame extraction algorithm proposed in this article.

In this article, 20 videos of different types of sports such as football, basketball, gymnastics, and physical education were tested. Due to space limitation, four video clips are selected as research cases. The experimental video information is shown in Table 1.

In the test results of the data set, the number of wrong selected frames, the number of missed selected frames, the number of redundant frames and the number of selected frames are given. The evaluation criterion selects F1- scores, precision ratio, and recall ratio, and the formula is as follows:

$$\Pr ecision = \frac{TN}{TN + FP} \tag{16}$$

$$\text{Rec}all = \frac{TN}{TN + FN} \tag{17}$$

$$F1 - scores = \frac{2\text{Rec}all^{*} \Pr ecision}{\text{Rec}all + \Pr ecision} \tag{18}$$

For the evaluation of the display effect in the virtual reality environment, 20 volunteers were invited to conduct the test. The devices used in the experiment include a set of high-performance desktop and virtual reality headset. In order to prevent the unexpected interference in the experiment from affecting the accuracy of the experimental results, the key frame extraction results and the objective evaluation indexes in this experiment are the mean results after removing the maximum value after several experiments. The display mode experiment in the virtual reality environment also averaged the data of each volunteer for the accuracy of the experiment, and the final display mode result was the comprehensive average result of 20 people.

### B. THE EFFECT ANALYSIS OF DIFFERENCE ENTROPY AND IMAGE TWO - DIMENSIONAL ENTROPY IS INTRODUCED

In order to verify the effectiveness of the algorithm proposed in this article, the simulation experiment of key frame detection for a surveillance video is carried out by using the neighboring frame subtraction algorithm based on the difference entropy proposed above.

**TABLE 2.** Statistical characteristics of difference image after binarization.

| Image number | 1 | 2 | 3 |
|---|---|---|---|
| Image size (number of pixels) | 300*600<br>(180000) | 300*600<br>(180000) | 300*600<br>(180000) |
| Number of background pixels | 170761<br>(0. 981) | 161234<br>(0. 961) | 160012<br>(0. 957) |
| Number of foreground pixels | 101<br>(0. 0009) | 387<br>(0. 287) | 467<br>(0. 412) |
| Difference image entropy | 0. 0299 | 0. 2176 | 0. . 2811 |

In order to quantify these minor changes in the difference image, the previous two-dimensional entropy calculation formula of the image was used to calculate the entropy of the difference image. The calculated results and some statistical characteristics of the binarization of the difference image by the adjacent frame subtraction algorithm were shown in Table 2.

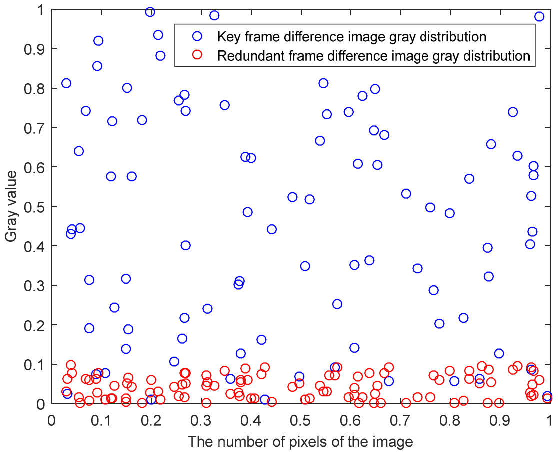

**FIGURE 6.** Detection results of conventional adjacent subtraction algorithm.

In the statistical properties shown in Table 1, if the threshold of the conventional adjacent frame subtraction algorithm is not properly selected, a large number of missing or false detection phenomena will occur. The former will lead to the loss of some key video frames containing important changes in the surveillance video, while the latter will lead to the retention of a large number of redundant video frames, causing unnecessary system burden and bringing difficulties to subsequent key frame retrieval, thus losing the significance of key frame detection in the surveillance video.

In order to clearly explain the problems of conventional frame subtraction algorithm in the process of key frame detection, and thus prove the significance of introducing image two-dimensional entropy calculation into conventional frame subtraction algorithm, a group of comparison experiments are carried out by using conventional frame subtraction algorithm and frame subtraction algorithm based on difference entropy respectively. The experimental objects are 100 key frame difference images and 100 redundant frame difference images. The experimental results are shown in Figure 6. From the experimental results shown in Figure 6, it is not difficult to find that after binarization of the redundant frame difference

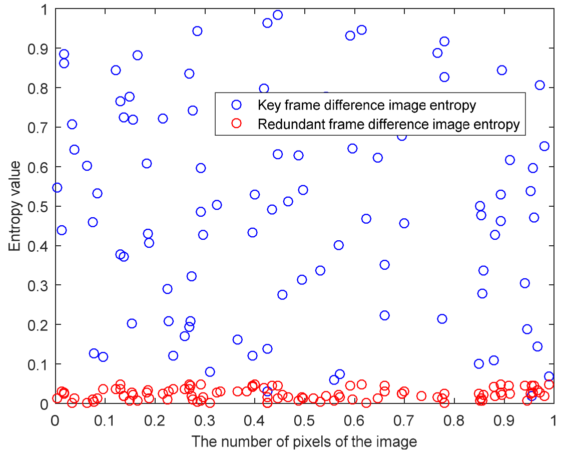

**FIGURE 7.** Detection results of adjacent frame subtraction algorithm based on difference entropy.

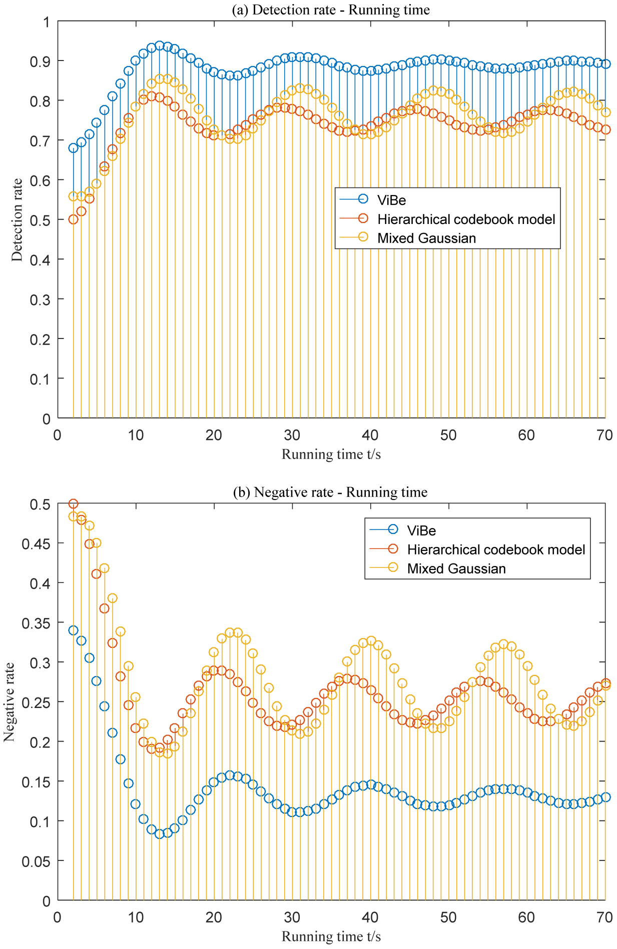

**FIGURE 8.** Detection rate and error matching rate of the three algorithms for the robustness of irregular motion.

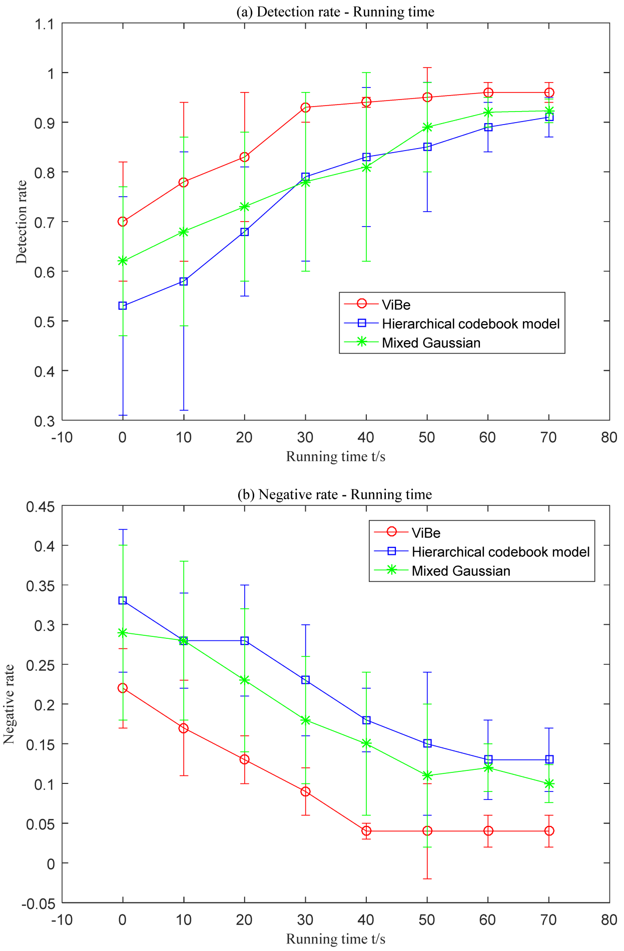

**FIGURE 9.** For the anti-interference performance of lens jitter, the detection rate and error matching rate of three algorithms are compared and analyzed in the experiment.

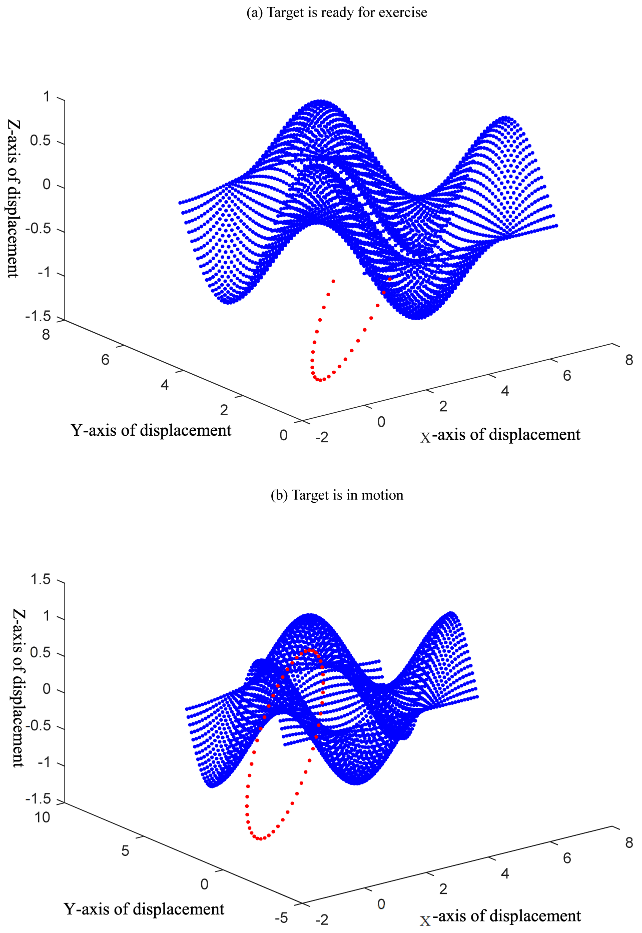

**FIGURE 10.** Multi-target tracking results for high-speed motion.

image and the key frame difference image, there is an obvious aliasing phenomenon after the calculation of the number of foreground pixels.

Figure 7 shows the experimental results of the adjacent frame subtraction algorithm based on the difference entropy. Results after binarization of redundant frame difference images and key frame difference images, after two-dimensional entropy calculation of images, the aliasing between them is not obvious, and the classification distance between the two types of image frames is greatly improved. By comparing the experimental results of the two methods, it can be seen clearly that the adjacent frame subtraction algorithm based on difference entropy has better robustness against background interference and can maintain better detection efficiency and accuracy.

### *C. THE RESULT ANALYSIS OF ViBe ALGORITHM IS INTRODUCED*

This experiment compares the experimental performance of background modeling method based on hierarchical codebook model, mixed Gaussian background modeling method and ViBe background modeling method in dealing with complex background, and describes the detection rate and error matching rate in the form of curve for comparative analysis.

Figure 8 compares and analyzes the detection rate and error matching rate of the three algorithms for the robustness of irregular motion. As can be seen from the experimental results in Figure 8, the three methods have strong detection performance for this scenario. Among them, the detection rate of the algorithm in this article is 90% on average, and the error matching rate is controlled at about 0. 1%.

For the anti-interference performance of lens jitter, the detection rate and error matching rate of the three algorithms are compared and analyzed in the experiment. The results are shown in Figure 9.

As can be seen from the experimental results in Figure 9, the change of background object position caused by camera jitter is effectively inhibited in the ViBe model method. In contrast, the ViBe model method achieved the highest detection rate, which was close to 100% at the end of the experiment. At the same time, it can also suppress the error matching rate. In comparison, the mixed Gaussian model and layered codebook model have poor inhibition ability in restraining camera jitter. The camera lens jitter can be regarded as the irregular distribution of pixel features to

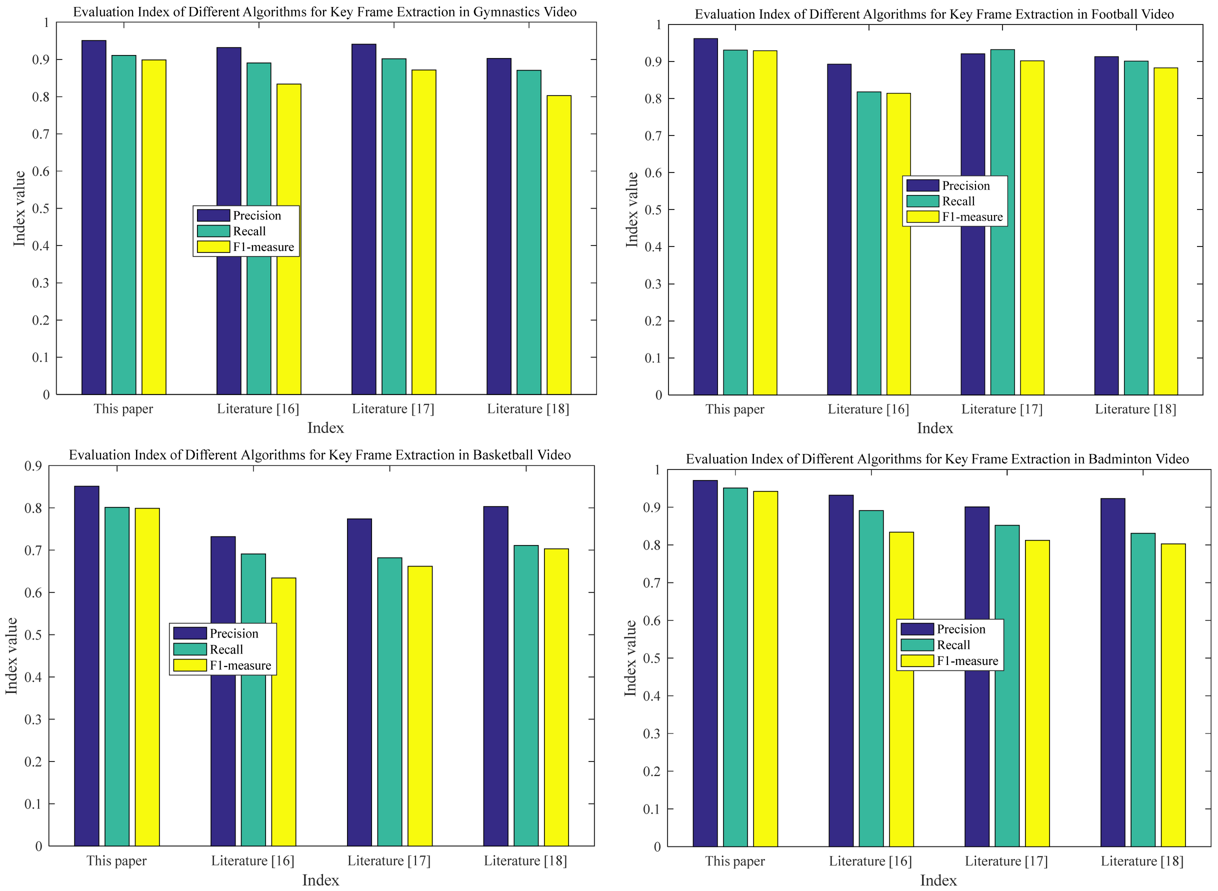

**FIGURE 11.** Comparison of evaluation indexes of key frame monitoring results under different video scenes.

a certain extent, and the ViBe model has higher detection rate and lower error matching rate. In order to avoid false detection caused by screen jitter, the jitter detection module should be added to the improved algorithm.

### D. KEY FRAME EXTRACTION RESULTS

Figure 10 shows the 3d event stream output by the video key frame. The blue points are the events generated by the motion changes at the pixels of the array, and the red points show the motion trajectory of the target in the video key frame, which can clearly see the motion trajectory of a single target.

In order to objectively and comprehensively evaluate the monitoring effect of the algorithm in this article, the algorithm in this article and the algorithms in literature [16], [17] and [18] are statistically analyzed. The specific results of evaluation indexes are shown in Figure 11.

Figure 11 shows the comparison results of evaluation indexes of moving target detection algorithms under different video scenes. It can be seen from the detection results that the comprehensive evaluation index F1-scores in literature [17] is the lowest in most video scenes. This is because literature [17] only detects the target contour, which generates a lot of missing detection inside the moving target. In reference [16], the algorithm did not have an appropriate background update strategy to update the changed background information, resulting in a large area of false detection, resulting in a low accuracy, and thus affecting the F1-scores, the comprehensive evaluation index. Compared with the other three algorithms, the algorithm proposed in this article achieves the best effect in most scenes. Great progress has been made in dynamic background and target intermittent motion scenes through adaptive judgment threshold and ViBe background model update rate, and the comprehensive evaluation index F1-scores is also higher than the other three classical algorithms.

### E. DISPLAY OPTIMIZATION RESULTS UNDER VIRTUAL REALITY ENVIRONMENT

In the experiment, for universality, we used the Stroop task information of the psychological test task to display instead of the video key frame, so as to prevent the inaccurate display mode result caused by a single type of image

#### 1) CONTRAST BETWEEN BRIGHT SCENE AND DIM SCENE

The test results showed that different levels of light caused the following effect: the task appeared in the same position, and the response time was shorter in the bright scene. This indicates that the degree of light affects the subjects' emotional state to a certain extent, thus affecting the subjects' task performance efficiency, as shown in Figure 12, the reaction time of the subjects when answering the Stroop task question becomes longer.

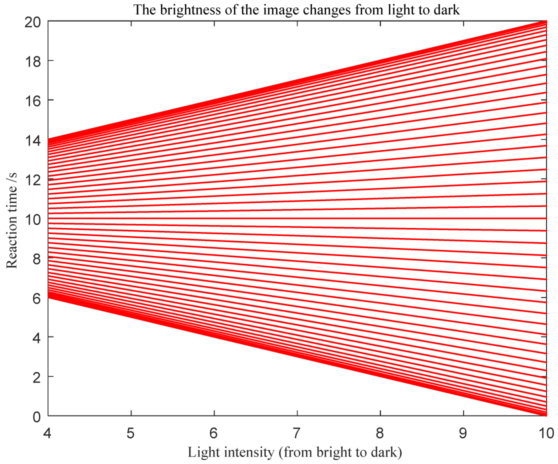

**FIGURE 12. Relationship between light intensity and reaction time in different scenes.**

At the same time, we found that the effect was more significant when the exposure to light was a single factor. Specifically, compared with the indoor scene familiar to the subjects, the subjects in the unfamiliar outdoor scene had a lower reaction time prolongation caused by the same degree of light change. These findings indicate that the change in the level of light will obviously lead to the decline of task performance efficiency of the subjects. When there are other influencing factors (such as environmental familiarity), the effect of the change in the level of light will be reduced, but when the level of light is extremely low, it will become the main factor affecting the response time.

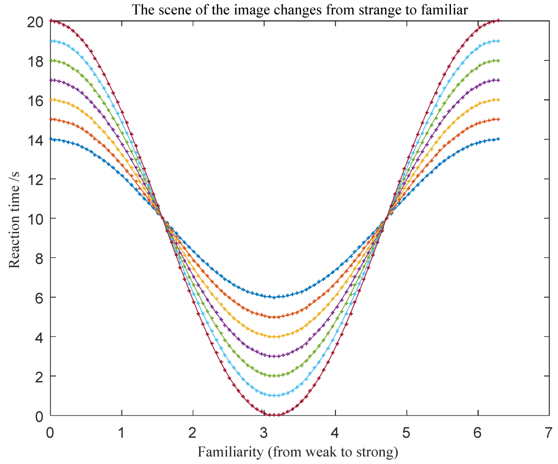

**FIGURE 13. Relationship between environmental familiarity under different lighting conditions and reaction time.**

### 2) COMPARISON BETWEEN FAMILIAR SCENES AND UNFAMILIAR SCENES

In a more unfamiliar environment, the unfamiliar environment led to more time for the subjects to complete the task and longer reaction time. When the subjects switched from the familiar environment to the unfamiliar environment, the reaction time was longer, as shown in Figure 13.

This indicates that when the subject changes from familiar environment to unfamiliar environment as a single influencing factor, the familiarity of the environment has a certain influence on the subject's task performance. Compared with the familiar environment, the unfamiliar environment leads to the decline of the subject's task performance efficiency, which is manifested as the extension of the response time. At the same time, when subjects switched from the dim and familiar indoor classroom scene to the dim and unfamiliar outdoor scene, the reaction time was slightly improved, but the change degree was not obvious. Therefore, it can be known that in the case of extremely low light degree, environmental familiarity had a low impact on task performance efficiency.

## VI. CONCLUSION

Video key frame method is a method to extract some video frames that can reflect the important video content to form a group of video frame sequences according to the time sequence. Its purpose is to condense and summarize the original video sequence by extracting key frames. Firstly, the entropy of adjacent frame difference and the two-dimensional entropy of image are introduced, and the combination of the two is used as the measurement of the difference between video frames. Secondly, outliers are detected by statistical tools to obtain the lens boundary, thus realizing the adaptive lens detection of video content. Then, ViBe algorithm is used to detect the foreground object in the video sequence and extract the scale-invariant feature transformation features of the foreground moving object. Finally, the motion vector is introduced, and the similarity of video frames is calculated according to the defined formula, and the key frames are extracted in these regions. In addition, this article also studies the display mode of key frame extraction results in virtual reality environment. The key frame display mode of video in virtual reality is optimized mainly by changing the display mode of information and changing the scene and testing the result of user task execution. Experimental results show that the proposed monitoring algorithm improves the video results with rich motion information significantly.

## REFERENCES

[1] L. Zhou, J. J. P. C. Rodrigues, H. Wang, M. Martini, and V. C. M. Leung, "5G multimedia communications: Theory, technology, and application," *IEEE MultimediaMag.*, vol. 26, no. 1, pp. 8–9, Jan. 2019.

[2] J. Wang, Y. Guo, Y. Jia, Y. Zhang, and M. Li, "Modeling and application of the underground emergency hedging system based on Internet of Things technology," *IEEE Access*, vol. 7, pp. 63321–63335, 2019.

[3] E. Bourchtein, J. M. Langberg, C. N. Cusick, R. P. Breaux, Z. R. Smith, and S. P. Becker, "Featured article: Technology use and sleep in adolescents with and without attention-deficit/hyperactivity disorder," *J. Pediatric Psychol.*, vol. 44, no. 5, pp. 517–526, Jun. 2019.

[4] K. N. T. Crowther and A. Wallace, "Delivering video-streamed library orientation on the Web: Technology for the educational setting," *College Res. Libraries News*, vol. 62, no. 3, pp. 280–286, Mar. 2001.

[5] B. Jia, L. Hao, C. Zhang, H. Zhao, and M. Khan, ''An IoT service aggregation method based on dynamic planning for QoE restraints,'' *Mobile Netw. Appl.*, vol. 24, no. 1, pp. 25–33, Feb. 2019.

[6] F. B. Goularte, S. M. Nassar, R. Fileto, and H. Saggion, ''A text summarization method based on fuzzy rules and applicable to automated assessment,'' *Expert Syst. Appl.*, vol. 115, pp. 264–275, Jan. 2019.

[7] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and I. Kompatsiaris, ''FIVR: Fine-grained incident video retrieval,'' *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2638–2652, Oct. 2019.

[8] C. Zhang, Y. Lin, L. Zhu, A. Liu, Z. Zhang, and F. Huang, ''CNN-VWII: An efficient approach for large-scale video retrieval by image queries,'' *Pattern Recognit. Lett.*, vol. 123, pp. 82–88, May 2019.

[9] B. Cui, Y. Zhang, L. Yan, J. Wei, and Q. Huang, ''A SAR change detection method based on the consistency of single-pixel difference and neighbourhood difference,'' *Remote Sens. Lett.*, vol. 10, no. 5, pp. 488–495, May 2019.

[10] D. Griffiths and J. Boehm, ''Improving public data for building segmentation from convolutional neural networks (CNNs) for fused airborne lidar and image data using active contours,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 70–83, Aug. 2019.

[11] N. Kaur, S. Goyal, A. Rani, and V. Singh, ''An improved local binary pattern based edge detection algorithm for noisy images,'' *J. Intell. Fuzzy Syst.*, vol. 36, no. 3, pp. 2043–2054, Mar. 2019.

[12] Q. Luo, B. Ge, and Q. Tian, ''A fast adaptive crack detection algorithm based on a double-edge extraction operator of FSM,'' *Construct. Building Mater.*, vol. 204, pp. 244–254, Apr. 2019.

[13] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, ''A novel violent video detection scheme based on modified 3D convolutional neural networks,'' *IEEE Access*, vol. 7, pp. 39172–39179, 2019.

[14] C. Kim, D. Song, C.-S. Kim, and S.-K. Park, ''Object tracking under large motion: Combining coarse-to-fine search with superpixels,'' *Inf. Sci.*, vol. 480, pp. 194–210, Apr. 2019.

[15] J. Li and Z. Liu, ''High-resolution dynamic inversion imaging with motion-aberrations-free using optical flow learning networks,'' *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Dec. 2019.

[16] S. S. Aote and A. Potnurwar, ''An automatic video annotation framework based on two level keyframe extraction mechanism,'' *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 14465–14484, Jun. 2019.

[17] E. Shabaninia, A. R. Naghsh-Nilchi, and S. Kasaei, ''Extended histogram: Probabilistic modelling of video content temporal evolutions,'' *Multidimensional Syst. Signal Process.*, vol. 30, no. 1, pp. 175–193, Jan. 2019.

[18] S. Yu and J. C. Príncipe, ''Understanding autoencoders with information theoretic concepts,'' *Neural Netw.*, vol. 117, pp. 104–123, Sep. 2019.

[19] G. Riva, B. K. Wiederhold, and F. Mantovani, ''Neuroscience of virtual reality: From virtual exposure to embodied medicine,'' *Cyberpsychology, Behav., Social Netw.*, vol. 22, no. 1, pp. 82–96, Jan. 2019.

[20] R. Yung and C. Khoo-Lattimore, ''New realities: A systematic literature review on virtual reality and augmented reality in tourism research,'' *Current Issues Tourism*, vol. 22, no. 17, pp. 2056–2081, Oct. 2019.

[21] M. El Beheiry, S. Doutreligne, C. Caporal, C. Ostertag, M. Dahan, and J. B. Masson, ''Virtual reality: Beyond visualization,'' *J. Mol. Biol.*, vol. 431, no. 7, pp. 1315–1321, 2019.

[22] D. Dessì, G. Fenu, M. Marras, and D. R. Recupero, ''Bridging learning analytics and cognitive computing for big data classification in micro-learning video collections,'' *Comput. Hum. Behav.*, vol. 92, pp. 468–477, Mar. 2019.

[23] Z. Kastrati, A. S. Imran, and A. Kurti, ''Integrating word embeddings and document topics with deep learning in a video classification framework,'' *Pattern Recognit. Lett.*, vol. 128, pp. 85–92, Dec. 2019.

[24] K. Kumar, ''EVS-DK: Event video skimming using deep keyframe,'' *J. Vis. Commun. Image Represent.*, vol. 58, pp. 345–352, Jan. 2019.

[25] C. Ma, D. Liu, X. Peng, L. Li, and F. Wu, ''Traffic surveillance video coding with libraries of vehicles and background,'' *J. Vis. Commun. Image Represent.*, vol. 60, pp. 426–440, Apr. 2019.

[26] Y. Kong, J. Huang, S. Huang, Z. Wei, and S. Wang, ''Learning spatiotemporal representations for human fall detection in surveillance video,'' *J. Vis. Commun. Image Represent.*, vol. 59, pp. 215–230, Feb. 2019.

[27] W. Li, Y. Chen, W. Sun, M. Brown, X. Zhang, S. Wang, and L. Miao, ''A gingivitis identification method based on contrast-limited adaptive histogram equalization, gray-level co-occurrence matrix, and extreme learning machine,'' *Int. J. Imag. Syst. Technol.*, vol. 29, no. 1, pp. 77–82, Mar. 2019.

[28] D. Oliva, S. Hinojosa, V. Osuna-Enciso, E. Cuevas, M. Pérez-Cisneros, and G. Sanchez-Ante, ''Image segmentation by minimum cross entropy using evolutionary methods,'' *Soft Comput.*, vol. 23, no. 2, pp. 431–450, Jan. 2019.

[29] S. Borjigin and P. K. Sahoo, ''Color image segmentation based on multi-level Tsallis–Havrda–Charvát entropy and 2D histogram using PSO algorithms,'' *Pattern Recognit.*, vol. 92, pp. 107–118, Aug. 2019.

[30] Q. Zhang, W. Lu, C. Huang, W. Lian, and X. Yang, ''An adaptive vibe algorithm based on dispersion coefficient and spatial consistency factor,'' *Autom. Control Comput. Sci.*, vol. 54, no. 1, pp. 80–88, Jan. 2020.

[31] J. Ma, X. Jiang, J. Jiang, and Y. Gao, ''Feature-guided Gaussian mixture model for image matching,'' *Pattern Recognit.*, vol. 92, pp. 231–245, Aug. 2019.

**ZHE WANG** was born in Shandong, China, in 1979. He received the bachelor's degree from Qufu Normal University, Shandong, in 2002, and the master's degree from Shandong Sport University, in 2009. Since 2002, he has been working with Qufu Normal University. He has published six papers, which has been indexed by CNKI. His research interest includes school physical education.

**YAN ZHU** was born in Heilongjiang, China, in 1976. She received the master's degree from Jilin University, in 2006, and the Ph.D. degree from the Dongbei University of Finance and Economics, in 2009. Since 2009, she has been working with Qufu Normal University. She has published a total of seven papers, two of which has been indexed by CSSCI. Her research interest includes sports economics.