Motion Balance Ability Detection Based on Video Analysis in Virtual Reality Environment

JILAN ZHOU, YANG YOU, AND YANMIN ZHAO
Department of Physical Education, China University of Petroleum (East China), Qingdao 266580, China
Corresponding author: Jilan Zhou (zhoujl@upc.edu.cn)

This work was supported by the Teaching Reform of the China University of Petroleum: Research on the Supply Side of College Physical Education-Big Data Analysis Based on Academic Performance and Physical Characteristics under Grant JY-B201888.

ABSTRACT In recent years, smart camera devices under the Virtual Reality (VR) environment have been widely popularized. These devices can be equipped with fast and effective computer vision applications, including the detection of the balance ability of moving targets. Moving target balance ability detection plays an important role in public security, traffic monitoring and other fields, and is also a basic technology for many vision applications. Therefore, the requirements for accuracy and completeness of detection are getting higher and higher. This article proposes a tracking method Motion Model and Model Updater (MMMU) based on the balance acquisition and model update and intelligent adjustment of the motion model. Improved Motion Model (IMM) is a background sample balance acquisition algorithm based on simple linear iterative clustering, completes the abstraction of background images. Different from other update strategies with a fixed number of frames, the update strategy based on image histogram contrast relies on the human selective forgetting mechanism to better avoid burst frames and process similar frames. Since the data used to detect the balance ability of moving targets is inherently unbalanced, the idea of dealing with imbalance in data mining is introduced into it, and the problem of balance ability detection of moving targets is studied from the perspectives of downsampling and oversampling. In addition, temporal and spatial oversampling of the foreground and selective downsampling of the background are performed to reduce the imbalance of the data set, and the resampled data set is used for modeling and classification. The feasibility of the MMMU algorithm is tested through experiments, and the motion balance ability of the foreground target is detected relatively completely.

INDEX TERMS Movement target balance ability detection, movement balance ability, balance acquisition, VR environment.

I. INTRODUCTION
Judging from the development of computer vision technology in recent years, Virtual Reality (VR) technology has shown unique advantages since its birth in the 1980s [1]. Its purpose is to establish a user interface that allows users to get an immersive feeling in the three-dimensional space environment. Specifically, virtual reality refers to the use of computer technology to simulate a virtual world or augmented reality environment to provide users with sensory simulations of vision, hearing, and touch [2]. Users use data gloves, helmets, glasses, headsets and other equipment to interact with the virtual world, conduct interactive activities to achieve goals such as practice, design, or training in the “real” experience.

The flexibility, creativity, and repeatability of virtual reality technology in security control, scene planning, interaction frequency and intensity settings, etc. provide new ideas for the development of many fields [3], [4]. Virtual reality technology has been initially applied in the fields of military training, psychotherapy and rehabilitation of the disabled. It has produced virtual reality exposure therapy, virtual occupational therapy and situational teaching systems, and achieved good results [5].

Moving target balance ability detection technology is the key to computer vision. The quality of moving target balance ability detection in a video sequence will have a huge impact on subsequent detection, processing and analysis [6]. The purpose of moving target balance ability detection technology is mainly to extract the target information of interest from the video sequence, and then perform subsequent higher-level
detection or analysis according to the purpose [7], [8]. It plays an important role in military, aviation, medical and other fields. General moving target balance ability detection technology can be divided into moving target balance ability detection in static background and moving target balance ability detection in dynamic background according to whether the camera is moving [9]. Most of the current algorithms are based on target detection under static background conditions and cannot be applied to dynamic backgrounds [10]. However, with the rapid development of science and technology today, portable smart mobile devices have penetrated into all aspects of people, and drones are also commonly used to shoot some scenes [11]. These have produced a lot of video information. If this information can be processed effectively, it will bring great convenience to monitoring, intelligent transportation and other work [12]. When shooting with drones or handheld devices, the camera will inevitably cause jitter, offset, zoom, etc., resulting in the shooting background in a non-stationary state [13]. Therefore, it is very important to study target detection algorithms suitable for dynamic backgrounds on mobile terminals [14].

Motion balance ability detection refers to locating the target position in the image and accurately identifying the target category as a pedestrian [15]. Commonly used target detection feature sources mainly include spatial information and geometric information. Among them, the detection methods of motion balance ability based on spatial information mainly include background subtraction and inter-frame difference methods [16]. The application scenarios of the two are generally static cameras and fixed backgrounds. The approximate area of the moving foreground is extracted through the difference between the current frame image and the background, and it is sensitive to changes in lighting, shadows, and background shaking [17]. The inter-frame difference method is similar to background subtraction, and the moving target area is obtained through the inter-frame image difference [18]. The application of this type of method has large limitations, and its compatibility with different scenarios is weak [19]. The basic idea of template matching is based on a certain similarity measurement method, the target template is windowed and traversed in the search window, and the best matching position is determined through the highest similarity, so as to detect the target. Common similarity measurement methods include mean square error, normalized mean square error, correlation error, and normalized correlation error [20], [21]. The method based on template matching is suitable for the detection of rigid targets. Because pedestrian targets are flexible targets with diverse poses, related scholars propose to use the texture information of pedestrian target contours to form templates, and then use the overall matching method for detection [22]. However, this method needs to construct a large number of target contour templates, the construction process is cumbersome, the flexibility is poor, the occlusion is sensitive, and the detection accuracy and robustness are not high [23], [24]. In response to the above-mentioned overall template matching problems, related scholars have proposed a method that only uses the head and shoulder contours as the matching template to avoid the influence of pedestrian posture changes and improve the pedestrian detection effect to a certain extent [25]. Related scholars introduced CF (Correlation Filter) into the field of target detection, and proposed MOSSE (Minimum Output Sum of Square Error filter) [26], [27]. The online training filter template is used as the detection model, and the peak position of the response graph obtained by the image and the template is used to locate the target center. In the process of minimizing the objective function by derivation, the matrix inversion in the time domain space requires a large amount of calculation [28]. To avoid this complicated operation, the MOSSE algorithm extends the detection window in the time domain periodically [29]. The transformation is converted to frequency domain for solving, which greatly reduces the calculation amount of the objective function. The most representative optimization of the MOSSE algorithm is KCF (Kernelized Correlation Filter) [30]–[32]. It turns the optimization problem of CF into a classification problem, uses the ridge regression method to solve it, introduces a circulant matrix to increase the number of negative samples, improves the quality of classifier training, and adds Gaussian kernels to the ridge regression to convert the nonlinear problem into high linear problem in the dimensional space, makes the algorithm more general [33], [34].

In this article, aiming at the imbalance problem in the detection of the balance ability of the moving target, the imbalance algorithm based on the data level is applied to solve the imbalance problem in the detection of the balance ability of the moving target. We synthesize new foreground samples, then down-sample the background samples, and use the rebalanced sample set for modeling and classification. The frame difference method is used for down-sampling the background samples, which mainly deletes the background sample points in the foreground and background overlapping area, that is, the potential foreground area. This reduces the imbalance of the two types of samples. Specifically, the technical contributions of this article can be summarized as follows:

First: In order to obtain a balance between the appearance of the target and the description of the surrounding scene, this article proposes a combination of image segmentation and saliency area detection on a spatial scale to deal with the precise expression of target samples and the comprehensive acquisition of background samples.

Second: We align the target and historical target for image similarity processing, and propose a simple and effective strategy to judge whether the current target is reliable on the time scale, and use this as the basis for updating the classifier on the current frame.

Third: Based on the data set, experiments are carried out on the algorithm of this article and related classic detection algorithms, which show that the MMMU algorithm
does improve the detection effect of the balance ability of moving targets. The comparative test with the excellent algorithm proves the feasibility of the method in this article in improving the overall performance and increasing the integrity of foreground detection.

The rest of this article is organized as follows. Section 2 discusses the related theories of video structuring in the VR environment. Section 3 constructs a visual inspection model based on the balance acquisition of the motion model and the intelligent adjustment of the model update. In Section 4, simulation experiments and result analysis are carried out. Section 5 summarizes the full text.

II. RELATED THEORIES OF VIDEO STRUCTURING IN VIRTUAL REALITY ENVIRONMENTS

A. VIRTUAL REALITY TECHNOLOGY

Essentially, virtual reality technology is a relatively advanced computer user interface that uses computers to express the real world. It can provide users with a series of real-life perceptions such as sight, hearing, touch, and smell. It constructs a three-dimensional human-computer interaction scene through a computer, reconstructs real life, and enables users to have an immersive experience. The architecture of the virtual reality system is shown in Figure 1.

Virtual reality is an all-round three-dimensional simulation of reality or real objects. When the user is in a virtual environment, it is close to the real environment. Users can roam immersively in the virtual world through virtual reality equipment.

The virtual reality environment is a dynamic and concrete three-dimensional digital space. In a virtual environment, humans and computers are not abstract and boring interactions. Not only can information in the virtual environment be generated through sensors, but virtual objects can also be manipulated through sensors.

Virtual reality technology is not only a form of showing the real world, but also a new way for people to understand the world through immersion and dialogue. It stimulates people’s creative thinking activities and can expand people’s knowledge.

The modeling technology based on geometric data is based on computer graphics and combines the geometric data of the scene through mathematical models such as curves, lines, surfaces, polygons, and triangles, and combines the geometric contours of the virtual scene according to certain rules. At the same time, a material model and a lighting model were added to the model outline to construct a virtual scene very similar to the real scene. The modeling technology for
constructing virtual scenes based on video images is to collect video image data of the scene, then classify the video images, extract terrain coordinates and scene textures, and import the data into professional modeling software to automatically generate virtual scenes. The video data is less, and the image is easy to use under the condition of collecting. The two modeling technologies have different data applications, and their application scenarios and technologies are also very different. The modeling technology of constructing virtual scenes based on video images can be applied to small-scale modeling scenes. This article needs to establish an intelligent test scene for unmanned vehicles, because it only needs to shoot the test field on the spot, with little workload and little video data. Therefore, this technology can be used to model the test field in three dimensions. The vehicle model and small three-dimensional model in the scene can be modeled by using the modeling technology of constructing virtual scene based on geometric data. Therefore, in this article, in order to improve the authenticity and reliability of the scene, a combination of geometric data and image data is used.

After modeling the shape of the experimental scene model, the rendering is only a single model without specific details. To increase the details of the object, it is necessary to add texture information to optimize the display details. Texture is a very important element in computer graphics. Texture can not only express the color and pattern of the surface of the object, but also express the position and depth of the object in the scene, especially for some similar shapes and structures, but the actual objects are different. In large cases, the difference in texture is the best way to distinguish.

B. VIDEO STRUCTURE
Since the video data itself is not structural, it is necessary to combine multiple technical means to extract the key information of the video hierarchically. Video structuring technology refers to converting the key content of video data into hierarchical structured information through specific algorithms, such as key frame extraction, video segmentation, target detection, image description, image segmentation and other technical means. It is based on video advanced semantics classifies and stores key information, so that users can quickly retrieve the video content they need.

In the research of video structuring technology, there is often no uniform standard for the form and content of structured information. Establishing a video structured information model and providing a way to describe video information will become a key research issue.

1) THE STRUCTURE OF THE VIDEO
The structure of the video can be simply divided from large to small: video, scene, shot and frame. Video is a data format in which a series of continuous images are stored and recorded in the form of electrical signals. Continuous images often record one or more specific stories in a continuous period of time. When continuous images are played sequentially at a faster frame rate, people will see continuous images instead of static images, which are called videos.

A scene refers to a unit composed of a series of similar shots, but the shots in the same scene are not necessarily continuous in time. This is determined by the similarity of the video shooting method and the video content.

A lens refers to a continuous picture in a video for a period of time, and the pictures have strong similarities. According to these characteristics, the lens can be detected by the lens segmentation algorithm.

Frame is the most basic component of video, and its essence is an image, and it is also the main content of video structure research. In video structuring research, the most informative video frames can be extracted through key frame extraction technology, and then the main content in these video frames can be converted into high-level semantic information for structured information storage.

2) STRUCTURED INFORMATION OF THE VIDEO
Video data contains rich information. In video structuring technology, the information extracted from video data can be divided into low-level feature information, key video image information, and high-level semantic information. The underlying feature information generally refers to the extraction of global, local and structural features of the image. Global features are the basic features of the image such as color, texture, and shape; local features extract the feature point set of the video image, calculate the expression of feature points, and use it for feature matching; structural features reflect the geometry and space-time domain between image features relationship. The overall architecture of the video transmission system is shown in Figure 2.

Key video image information refers to the extraction of key frames based on some underlying features and target information of the image. By fusing different underlying feature information to express the information difference between frames or the information richness of video frames, the most representative video frames are then screened out.

High-level semantic information refers to the semantic summary and description based on the goals and content contained in the video. It uses deep learning technology to train a targeted model based on a proper amount of picture sets, extracts target semantics, scene semantics, image semantics, etc., and comprehensively extracts semantic information, extracts text sentences and summarizes events reflected in the video.

Further, the video structuring technology in this article uses five key technologies to achieve different functions, namely: key frame extraction, target detection, action recognition, scene recognition, and image description. The key frame extraction technology can extract the key information in the video and save storage space. The target detection can identify the type, number and location of the target appearing in the key frame, the motion detection can recognize the target action, and the image description can transform abstract
image data. It is an easy-to-understand text description for easy storage and retrieval.

### C. BASIC THEORY OF VIDEO STRUCTURED

With the rapid development of current science and technology, video structuring technology has also applied many new technologies to improve information extraction efficiency and information richness. This article uses the combination of traditional image processing and new deep learning technology as a tool for mining video data information, which can effectively extract the key information of the video. This section will mainly introduce the basic theories related to video structuring.

1) **COLOR CHARACTERISTICS**

Color is an important information in the human visual system. In the video structuring, the color features in the video are widely concerned. Therefore, to obtain color features reasonably and accurately, it is necessary to select an appropriate color space and color feature expression.

Color moments, as an image feature expression method, can simply and effectively express color features through the moments in mathematical theory. In the image feature expression, the first three moments can express the features well. Their calculation methods are as shown in the following formulas. In the formula, \( p_{i,j} \) represent the i-th component of the j-th pixel of the image, and \( n \) represents the pixel total. In practical applications, the information expression ability of this method is not strong enough, so in the field of image processing and other fields, this method is generally only used as a part of multi-feature fusion or as a preliminary screening method of images.

\[
u_i = \frac{1}{n} \sum_{j=0}^{n-1} p_{i,j}
\]

(1)
which is not suitable for situations with many changes and is only modeled by a single-distributed probability model.

The former to replace the old ones. Background modeling can be divided to changes by itself, and continuously generate new models that can change at any time. Therefore, the modeled back-ground needs to face many complex environments and environments that can not change at any time. It first obtains the background ground pixel model needs to be updated at any time, can adapt to a complex and changeable environment. Commonly used background modeling methods include single Gaussian and mixed Gaussian models.

3) BACKGROUND SUBTRACTION

The background subtraction method is also a kind of frame difference method, which is generally used to detect the position of the moving target. It first obtains the background model through background modeling technology, and then subtracts the frame image from the background model by pixel. If the difference is higher than the threshold, the pixel is judged as a foreground pixel; otherwise, the pixel is classified as a background pixel.

The background subtraction method relies heavily on background modeling. In actual use, background modeling needs to face many complex environments and environments that can change at any time. Therefore, the modeled background pixel model needs to be updated at any time, can adapt to changes by itself, and continuously generate new models to replace the old ones. Background modeling can be divided into two types: single-mode and multi-mode. The former is only modeled by a single-distributed probability model, which is not suitable for situations with many changes and complex backgrounds. The latter is to jointly build a model through a multi-distribution probability model, which can adapt to a complex and changeable environment. Commonly used background modeling methods include single Gaussian and mixed Gaussian models.

4) CONVOLUTIONAL NEURAL NETWORK

The main feature of convolutional neural network is to use convolution operation to calculate the neural network. This calculation method is very suitable for processing images, so it has important application value in image processing.

Convolutional neural network is a hierarchical model, and its structure is shown in Figure 3. The input of the network is generally images, etc. or information after feature extraction. It uses a series of operations such as convolution, pooling, and nonlinear mapping to automatically represent the characteristics of the data. Among them, different types of operations are represented by layers of modules in the network model, such as convolutional layers. In the last layer of the network, the loss function is used to calculate the error between the current output and the actual situation, and then the backpropagation algorithm is used to reversely update the parameters of each layer of the network, so that the calculation result can be more in line with our expectations during the next operation. The process will be repeated many times during training, and finally the model we need can be obtained when the loss function converges.

III. VISUAL INSPECTION BASED ON MOVEMENT MODEL BALANCE ACQUISITION AND MODEL UPDATE INTELLIGENT ADJUSTMENT

A. ANALYSIS OF THE DETECTION FRAMEWORK BASED ON THE PRINCIPLE OF VISUAL ABSTRACTION AND SELECTIVE FORGETTING MECHANISM

The motion model samples from the original input image and predicts the possible candidate positions of the target to determine the target search range. The accurate target sample and the balanced background sample in the current frame can not only improve the accuracy of the classifier, but also correct the classifier to a certain extent (correct the classifier from the wrong training). Therefore, the balanced selection of samples is very important for the improvement of the motion model.

This article designs a strategy based on the combination of image segmentation theory and salient region detection to obtain target and background samples. In more detail, the former image segmentation is used to obtain scene samples in a balanced manner; the latter salient area detection more accurately describes the appearance of the target. Different from the general motion model that generally uses random selection within the range or region sliding selection of samples in the general motion model, the image segmentation processing of the background not only reduces the blindness caused by random selection, but also considers different “similar regions in the background”. Similarly,
different from the method of random selection of target samples within a rectangular frame in the conventional motion model, this article adopts significant area detection, and comprehensively considers multiple attributes of the target to construct a more comprehensive, accurate and expressive target sample set.

Model update is like the brain’s analysis and identification process of the target in human eye detection, and it occupies a very important position in the detection process. Most detection-based detection methods use simple frame-by-frame or stepwise update method to train the classifier. This approach will not only reduce the detection efficiency, and when the detection target is wrong (for example, target drift), the classifier will also be “contaminated” (the classifier is trained with the wrong target sample). In this article, the image similarity principle is used to judge the target and surrounding area images according to certain strategies, so as to achieve the goal of not updating extremely similar and extremely dissimilar target scenes. The former is used to reduce the update frequency, and the latter is used to avoid the classifier.

Therefore, this article proposes a detection method Motion Model and Model Updater (MMMU) to obtain Improved Motion Model (IMM) and model update intelligent adjustment based on the balance of motion model. The current frame is in the motion model module. On the one hand, the Simple Linear Iterative Cluster (SLIC) method is used to perform super pixel segmentation to obtain atomic area blocks with higher and the same properties, and on this basis, balanced scene samples are obtained; on the other hand, Histogram is used -based Contrast method (HC) detects the salient area of the target to obtain a more comprehensive and accurate expression of the target appearance. Then the two conduct collaborative learning to form more reasonable target and background training samples.

B. ACQUISITION OF BACKGROUND SAMPLE BALANCE BASED ON SIMPLE LINEAR ITERATIVE CLUSTERING

The image segmentation technology in the field of image processing can accomplish the background division very well. Image segmentation refers to dividing an image into different sub-regions according to the similarity of different image characteristics such as brightness, color and texture. These sub-regions do not overlap each other and have certain visual significance. A small number of super pixels are used to replace a large number of other pixels in the area to represent the characteristics of the picture, and the distribution of pixels in each sub-area conforms to a predetermined rule. And this technical idea is basically consistent with the requirements for background division.

Specifically, the classic superpixel segmentation algorithm is used in this article. Simple linear iterative clustering SLIC performs simple superpixel segmentation on the current image frame, and allocates the number of background samples collected according to the number and size of the segmented atomic regions. Random collection of background samples is performed in the divided atomic regions.

The basic idea of the SLIC algorithm is very simple: First, the image is converted from the RGB color space to the Lab color space, and each pixel in the image is expressed as:

\[ S = [a, b, x, y, l]^T \]  

Among them, l, a, b represent the color value of the pixel in the Lab color space; x, y represent the coordinates. The algorithm first generates K seed points, and then in the space around the seed points, according to the distance metric D constructed by the following formula, it searches for the nearest other pixels and classifies them into one category. After that, the vector of each category of pixels in the K category is averaged, K seed points are obtained again, and the local clustering is performed again, and iterated several times until convergence.

\[ D = \sqrt{m^2 \left( \frac{Kd_k}{n} \right)^2 + d_c^2} \]  

In order to avoid selecting edge points and noise points when selecting the cluster center point in the iterative process, SLIC will move the cluster center to the minimum area of the gradient G(x, y) in each iteration.
Although SLIC can also cluster the target into several types of superpixel blocks to a certain extent, this representation method also decomposes the complete target into several parts, which is not convenient for the acquisition of the overall target sample, so it still needs to accurately extract the target sample.

C. ACQUISITION OF TARGET SAMPLE BALANCE BASED ON HISTOGRAM CONTRAST

The saliency research of computer vision originates from the human eye’s attention mechanism. Without prior knowledge, the image is extracted by analyzing the global contrast and spatial correlation of various attributes such as color, texture, edge, intensity, and gradient of the image. Specifically, HC is used to perform an image pixel saliency detection algorithm based on histogram contrast on the image.

The saliency value of a pixel in this article is defined as:

\[
R(S_i) = \sum_{S_j \in S} D(S_i, S_j)
\]

(7)

Among them, S is the image, \(S_i\) is any of the pixels, and \(D(S_i, S_j)\) is the distance measurement between the pixels \(S_i\) and \(S_j\) in the Lab color space.

Figure 4 shows a comparison diagram of using HC to detect the salient area of a frame of the video sequence in the test set. The target benchmark of the video sequence is the head of the man in the figure. If the conventional method of obtaining target samples is used, the head area in the target frame is randomly sampled. Since the target frame itself has a small range, it will cause a large number of repetitive target samples, which will result in a waste of computing resources. Therefore, this conventional sampling method is incomplete and inaccurate. In addition, this kind of sampling based on the target frame only considers the distance measurement of the target, and does not consider other measures such as color, texture, and intensity. Therefore, this sampling method does not meet the requirements of multiple angles. Figure 4(b) shows the saliency map after saliency area detection. It can be seen that the man’s body, arms and other body parts are displayed well together with the target head.

At this time, sampling the man’s target is not necessary. Confined to the target frame and extended to the whole body part with strong relevance, the selected target sample meets comprehensiveness and accuracy at the same time. The outline of the entire human body is well displayed, and the target sample can be selected based on this outline, thereby reducing the selection of other background samples. Moreover, the saliency detection comprehensively considers the various metrics of the image (color, texture, intensity, etc.), so the multi-angle is also reflected.

Although the saliency detection method based on the histogram contrast can well complete the effective selection of target samples, it also has shortcomings. For example, in Figure 4, in addition to the human body itself, some of the surrounding backgrounds also show high saliency. Therefore, it is necessary to fuse distance measurement and superpixel segmentation to reduce such non-target low-correlation but highly significant background regions.

D. IMAGE SIMILARITY MEASUREMENT BASED ON DISCRETE COSINE TRANSFORM COMPRESSION

How to quickly and effectively determine whether the current frame belongs to a stable frame or a frame that changes drastically during this period? The simplest and most effective method is to use image similarity to measure the correlation between pictures, that is, to compare the similarity between the current frame image and other frames in the period. A big difference indicates that the frame changes drastically, and a small difference means that the current frame is very stable. Specifically, this article uses Discrete Cosine Transform (DCT) to compress images in order to quickly and effectively calculate the similarity between images.

The picture is essentially a set of two-dimensional signals, which contain different frequency components, in which the low-frequency part describes the overall information of the picture; the high-frequency part describes specific details such as edges. According to the above analysis, finding the similarity between pictures only needs to keep the low-frequency part of the picture, and discarding...
the high-frequency part can also reduce the amount of calculation. As a classic image compression algorithm, due to the high energy concentration of the cosine function, the discrete cosine transform DCT algorithm can provide a higher compression rate under the same image quality, so it has a great advantage in energy concentration.

Because most images have a lot of redundancy and large spatial correlation, when DCT transforms the image from the pixel domain to the frequency domain, most of the frequency components are 0 or tend to 0. As shown in Figure 5, the correlation of the image processed by the DCT algorithm is greatly reduced. Figure 5(b) is the DCT coefficient matrix after transformation. The frequency is getting higher and higher from the upper left corner to the lower right corner. It can be seen that the image energy is mainly concentrated in the upper left corner area. At this time, only the data of this part of the area needs to be obtained.

In order to better compare the similarity of pictures after DCT transformation, we extract the corresponding similarity “fingerprint” (hash value) for each $S_{t,DCT}$.

$$H_t(k) = \begin{cases} 1 & m(S_{t,DCT}) < S_{i,DCT}(i,j) \\ 0 & Otherwise \end{cases}$$  \hspace{1cm} (8)

It can be seen that for the sub-matrix $S_t$ after DCT compression of the image, the $H_t$ obtained after DCT processing cannot fully display the real low-frequency part of the picture. $H_t$ simply represents the ratio of the picture to the average low-frequency part. Therefore, as long as the overall structure of the picture does not change significantly, the corresponding $h_t$ value remains basically unchanged. This not only enhances the robustness of the similarity comparison, but also avoids the impact on image processing (such as color histogram equalization adjustment, etc.).

After obtaining the corresponding hash fingerprints, the Hamming distance is used to measure the “fingerprints” of different pictures. Specifically, $h_t$ represents the processed “fingerprint” value of the $t$-th frame video sequence, and $p_t$ represents the Hamming distance between the current frame and the previous two frames, which is calculated as follows:

$$p_t = \left| h_{t-1} + h_{t-2} \right| / 2 - h_t$$  \hspace{1cm} (9)

It should be noted that when $t < 3$, $p_t$ is meaningless. The fact is that the number of frames of all video sequences processed by the inspection institute is much greater than 3 frames, and the probability of occurrence of only 3 frames in the actual application process is very high. Based on this, the Hamming distance $p_t$ and $p_{t-1}$ used to represent the current target appearance picture and scene picture can be defined respectively. Then, we set a threshold $f_c$, which is 0 or 1, and control the update of the classifier according to the following rules:

$$f_c(k) = \begin{cases} 0 & \text{pause} \\ 1 & \text{start} \end{cases}$$  \hspace{1cm} (10)

E. MODEL UPDATE STRATEGY BASED ON SELECTIVE FORGETTING MECHANISM

Figure 6 is based on the actual situation of the detection research, using the following simple but effective selection strategy: it compares the “fingerprint” information of the current frame and the image frame in the most recent period of time to determine whether there is a huge difference or similarity situation, the former means that the current frame has undergone a brief mutation, and the latter means that the current frame has basically not changed during the recent period. In both cases, the classifier update is not required. It is a sudden change that generally only occurs temporarily. At this time, the update will “contaminate” the classifier to some extent, which will affect the subsequent stable detection. The reason for not updating when the similarity is very strong is that the frame is very stable in the recent period of time. Most of the information this brings is redundant information, and rigid updates will only increase the amount of calculation and reduce the detection speed.

Taking into account the negative impact of the improved motion model module that increases the amount of calculation and reduces the detection speed, it is required that the improved model update module meets its basic requirements and must maximize the model update speed. The main innovations of the improved model update IMM module are as follows:
First, we use the discrete cosine transform DCT to compress the picture and collect the corresponding “fingerprint” information, so as to keep as much of the overall information of the picture frame as possible while simplifying the amount of calculation;

Second, we use the “fingerprint” information of the picture for similarity comparison, and adopt the method of not updating the classifier for extremely stable frames, so as to obtain an increase in detection speed without affecting the robustness of the classifier;

Third, we use the “fingerprint” information of the picture to compare the similarity and avoid blindly updating the classifier for frames with huge differences, so as to obtain a balance between model adaptability and drift occurrence.

Therefore, thanks to the simplicity and effectiveness of the discrete cosine transform DCT in comparing image similarity and the update strategy for extremely similar frames, the Improved Motion Updater (IMU) not only offsets the negative impact of the rate reduction brought by the IMM module, but also improves the running speed of the entire detection framework. Thanks to the IMU’s evasion strategy for the hugely different mutation frames, the classifier is immune to “pollution” to a certain extent.

IV. SIMULATION EXPERIMENT AND RESULT ANALYSIS

A. EXPERIMENTAL PARAMETER SETTINGS

We select True Positive Rate (TPR) and True Negative Rate (TNR) as the measurement index. However, the true negative rate TNR is not sensitive to the false detection situation, so this article selects Precision Rate (PR) and F-measure (FM) and the true rate TPR for performance measurement.

PR can reflect information similar to TNR, but it is more sensitive to false detection. The larger the value of PR, the higher the TNR and the lower the false detection rate of the algorithm. Only when the algorithm under test has both a larger TPR and a higher PR can a higher FM be achieved.

In the skiing sequence and the rowing sequence, the incomplete algorithm for deleting the down-sampling module is tested to determine the value of the parameter N. For the ski sequence, we change N from 0 to 60, as shown in Figure 7, which is its corresponding performance curve. It can be found that TPR changes as N increases. This means that as the sampling rate increases, the correct classification rate of foreground pixels will change dynamically. It can also be seen from the graph that this improvement in TPR is accompanied by a corresponding change in PR.

![FIGURE 6. Update strategy flowchart.](image)

![FIGURE 7. The performance of the MMMU algorithm with different N values in the ski sequence.](image)

![FIGURE 8. The performance of the MMMU algorithm with different N values in the rowing sequence.](image)

The same experiment was performed on the rowing sequence, and the performance curve is shown in Figure 8. It is not difficult to find that when N > 42, all indicators no longer change significantly, which means that over-fitting has occurred, and the newly generated samples cannot further improve the representativeness of a few samples.
Based on a fixed value of N, we use different \( \theta \) on the two sequences to test the complete algorithm. In the experiment, the value is increased from 1 to 50, and the corresponding performance curve is calculated. Figure 9 and Figure 10 are the experimental results on the skiing data set and boating data set, respectively. Both figures show that the value of TPR continuously changes with the increase of \( \theta \), which is the effect of selective downsampling. PR also has several stages of decline. The reason for the slow decline of PR in the first stage is that the current foreground is not aligned with the potential foreground area. Once the real background area is included in the potential foreground area due to misalignment, deleting the background samples in this area will obviously lead to a false detection rate. The reason for the sharp decrease of PR in the second stage is the large accumulation of real background pixels in the foreground model. If the false detection rate caused by the reduction of PR exceeds the tolerance of statistical modeling, more and more false detections will appear in the subsequent video sequence, resulting in a sharp reduction in PR. Therefore, the \( \theta \) smaller than the cut-off value must be selected for selective down-sampling.

We select the four sequences in the skiing sequence and the rowing sequence to calculate FM, and calculate the relationship between it. As shown in Figure 11, the changes in FM’s imbalance are in line with our expectations, so the method in this article and the new imbalance measurement are both effective.

**B. COMPARATIVE EXPERIMENT ANALYSIS**

There have been many excellent moving target balance ability detection algorithms that have published their segmentation results on the database. This article only selects some of the algorithms for comparison, and does not compare the anonymous, unpublished, and post-processing methods. At the same time, we remove those methods that contain other functional modules, such as shadow removal, ghost removal, etc. In addition, only one or two representative methods are selected for comparison in the homology algorithm.

As shown in Figure 12, it is a comparison of the segmentation results of the ViBe and MMMU algorithms on the ski sequence. Figure 12(a) shows a typical five frames in a ski sequence. There are multiple non-rigid objects in the sequence shown, and there is weak color similarity between the background and the clothes of the person. The segmentation results of ViBe and this article are shown in Figure 12(b) and 12(c), respectively. ViBe is a non-parametric method whose sample count replaces KDE, and a random update strategy is used to accelerate and update the model. There are some adjustable parameters in ViBe, and it is necessary to adjust them for different sequences. ViBe has the advantages of fast, robust, and freely adjustable parameters. Experiments show that the method in this article has a double improvement over ViBe in terms of segmentation results:

- First, the method in this article has less segmentation noise than ViBe, which is due to the global optimization step in the method in this article, while ViBe adopts a pixel-level classification strategy. Secondly, because we have resampled the samples, the segmentation results in this article are more complete than ViBe, especially in the camouflage area, there are no holes in the ViBe segmentation results.

Observing the results of the above experiments, we noticed that there are two problems with the MMMU algorithm.
First of all, the method in this article accentuates the shaded area under certain circumstances. On the one hand, since the camouflage area and some forms of shadows (especially soft shadows) show properties similar to those of the foreground, the deletion of these sample points during selective downsampling results in the aggravation of the shadows. On the other hand, the MMMU algorithm does not consider the problem of shadow removal. Secondly, it can be seen that the detected foreground boundary is not smooth enough, and there are multiple irregular protrusions on the foreground boundary, which indicates that the foreground model in these areas is not accurate. In fact, for such non-rigid objects, it is difficult to ensure the absolute accuracy of the point detection, which causes the difference between the synthesized sample and the current foreground in some areas, and eventually leads to irregular protrusions on the border.

The KDE algorithm and MMMU algorithm were tested on the rowing sequence. The segmentation result is shown in Figure 13, where Figure 13(a) is a typical five frames in the rowing sequence. It can be seen that there are two rigid moving targets in the sequence. Figure 13(b) and Figure 13(c) are the segmentation results of KDE and this method, respectively. KDE is the first moving target balance ability detection algorithm that uses non-parametric modeling. It is the basis of a series of non-parametric modeling algorithms. Compared with KDE, the advantage of this method is that it is more robust to noise and the result of segmentation is more complete. This is due to the global optimization steps and resampling strategy we have taken. At the same time, the irregular protrusions on the boundary in Figure 13(c) are reduced because the corner detection of rigid targets is more reliable than that of non-rigid targets.

In order to further analyze the effectiveness of the algorithm, qualitative analysis through experiments is also needed. First, we calculate the TPR, PR and FM of various algorithms for four typical sequences, as shown in Figure 14. Second, we calculate the average TPR, PR and FM of the above algorithm for all sequences in the second and fourth subsets of the database, as shown in Figure 15.

In Figure 14 and Figure 15, there is a large difference between the TPR value and the PR value. It can be seen that the difference between TPR and PR is reduced by the method in this article, and in some cases the two are almost the same. This shows that the MMMU algorithm is indeed effective for class imbalance in background subtraction. However, for some sequences, the detection results after adopting the unbalanced learning strategy are still not good enough. For example, the average TPR on the ski sequence and the fourth subset are about 0.89 and 0.79, respectively. This unreliable segmentation is caused by many related factors, including dynamic background, shadows, class imbalance, etc., which means that it is impossible to repair all these influencing factors through a single algorithm, and it is necessary to design corresponding special problems.

It is not difficult to find from Figure 14 that the method in this article has the best TPR and FM on the sequence skiing and boating. Since camouflage and hidden class imbalance problems occupy a considerable part of all the challenges of the first three sequences, the detection performance can be improved by the method in this article. Figure 15 shows that the MMMU algorithm obtains higher average TPR and FM values in a series of different types of scenarios. This shows that the overall performance of the MMMU algorithm is better, and it also proves the feasibility and effectiveness of the algorithm.
Finally, it is necessary to further discuss the time complexity and space complexity of the MMMU algorithm. Since the method in this article is based on KDE’s two-class model classification, and the KDE-based method needs to calculate the membership probabilities of multiple samples, the computational complexity is high. On a PC with an Intel Core i7-4770 3.5GHz CPU and 4GB RAM, the processing time for a video with a resolution of 760 × 540 is about 7FPS. Using better hardware equipment and further optimization of the algorithm, especially the optimization of KDE, can improve the running speed of the algorithm. Regarding the space complexity, because the samples used for modeling need to be stored, some capacity is consumed for storage. Although the foreground samples are expanded...
by spatiotemporal oversampling, this is not enough to have a big impact on current memory devices.

V. CONCLUSION

From the perspective of the global space of the current frame, this article uses image detection and segmentation methods to solve the problem of obtaining the balance between the target appearance sample and the current background sample. In particular, these two methods have their own different missions. Without prior knowledge, the former comprehensively considers the color, texture, and boundary of the image to construct a more robust sample weight expression. By using the prior knowledge of the target in the video sequence, the individual pixels in the current frame are aggregated into meaningful atomic regions. We use the complementarity of these two methods to build a more reasonable confidence map. For the model update problem, the current frame is similarly analyzed on a continuous time scale, and the current frame needs to be updated with samples to update the classifier according to the corresponding update strategy. This approach indirectly improves the effective use of samples. It not only reduces the update frequency of the model but also prevents the classifier from being “contaminated” to a certain extent. The solution at the data level is the resampling method. On the one hand, the positive samples are over-sampled to generate new samples. The over-sampling time is designed according to certain rules, which can ensure that the generated samples are more evenly distributed in the sampling interval, thereby avoiding the over-fitting phenomenon caused by sample accumulation. On the other hand, the negative samples are selectively down-sampled, and the difference method is used for the background frame and the synthesized foreground frame to reduce the number of background samples in the foreground and background overlapping regions (potential foreground regions). We improve the imbalance of the data through the above two steps, obtain a balanced data set, and use it for classification. Finally, through comparative experiments, the feasibility of the algorithm is verified and the detection accuracy is improved.

REFERENCES

[1] J. Kolic, K. O’Brien, K. Bowles, R. Iles, and C. M. Williams, “Understanding the impact of age, gender, height and body mass index on children’s balance,” Acta Paediatrica, vol. 109, no. 1, pp. 175–182, Jan. 2020.

[2] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, “Semantic model vectors for complex video event recognition,” IEEE Trans. Multimedia, vol. 14, no. 1, pp. 88–101, Feb. 2012.

[3] C.-G. Lim, “Effects of trunk control robot training on balance and gait abilities in persons with chronic stroke,” Phys. Therapy Rehabil. Sci., vol. 9, no. 2, pp. 105–112, Jun. 2020.

[4] L. Kong, X. Li, G. Cui, W. Yi, and Y. Yang, “Coherent integration algorithm for a maneuvering target with high-order range migration,” IEEE Trans. Signal Process., vol. 63, no. 17, pp. 4474–4486, Sep. 2015.

[5] S.-B. Peng, J. Xu, X.-G. Xie, F. Liu, T. Long, J. Yang, and Y.-N. Peng, “Multi-aircraft formation identification for narrowband coherent radar in a long coherent integration time,” IEEE Trans. Aerosp. Electron. Syst., vol. 51, no. 3, pp. 2121–2137, Jul. 2015.

[6] J. Zheng, H. Liu, G. Liao, T. Su, Z. Liu, and Q. H. Liu, “ISAR imaging of targets with complex motions based on a noise-resistant parameter estimation algorithm without nonuniform axis,” IEEE Sensors J., vol. 16, no. 8, pp. 2509–2518, Apr. 2016.

[7] J. Costa Pereira, E. Covio, B. A. Doyle, N. Rasiwasia, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, “On the role of correlation and abstraction in cross-modal multimedia retrieval,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 3, pp. 521–535, Mar. 2014.

[8] C. Rochefort, E. Legace, C. Boulay, G. Macartney, K. Goulet, R. Zemek, and H. Sveistrup, “Balance markers and saccadic eye-movement measures in adolescents with postconcussion syndrome,” J. Athletic Training, vol. 55, no. 5, pp. 475–481, May 2020.

[9] X. Li, G. Cui, W. Yi, and L. Kong, “Coherent integration for maneuvering target detection based on radon-Lv’s distribution,” IEEE Signal Process. Lett., vol. 22, no. 9, pp. 1467–1471, Sep. 2015.

[10] P. Huang, G. Liao, Z. Yang, and J. Ma, “An approach for refocusing of ground fast-moving target and high-order motion parameter estimation using radon-high-order time-chirp rate transform,” Digit. Signal Process., vol. 48, pp. 333–348, Jan. 2016.

[11] J. Zheng, T. Su, W. Zha, X. He, and Q. Hao Liu, “Radar high-speed target detection based on the scaled inverse Fourier transform,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8, no. 3, pp. 1108–1119, Mar. 2015.

[12] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013.

[13] N. M. Sard, S. R. Smith, J. J. Homola, J. Kanefsky, G. Bravener, J. V. Adams, C. M. Holbrook, P. J. Hrodey, K. Tallon, and K. T. Scribner, “RAPTURE (RAD capture) panel facilitates analyses characterizing sea lamprey reproductive ecology and movement dynamics,” Ecol. Evol., vol. 10, no. 3, pp. 1469–1488, Jan. 2020.

[14] J. Tian, W. Cui, and S. Wu, “A novel method for parameter estimation of space moving targets,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 2, pp. 389–393, Feb. 2014.
