Key Frame Extraction Algorithm of Motion Video Based on Priori

QI ZHONG, YUAN ZHANG, JINGUO ZHANG, KAIXUAN SHI, YANG YU, AND CHANG LIU
Department of Physical Education, China University of Geosciences (Beijing), Beijing 100083, China
Corresponding author: Yuan Zhang (zhangyuansCI@163.com)

This work was supported by the China University of Geosciences (Beijing) through the Undergraduate Education Quality Improvement Plan Off Campus Practice Base Construction Project, College Students Outdoor Sports Practice Teaching Base, under Project SJJD202006.

ABSTRACT Key frame extraction technology is one of the core technologies of content-based video retrieval. For video types with complex content, various scenes, and rich actions, the performance of existing key frame extraction methods is not ideal. Based on the Visual Geometry Group (VGG), this article proposes an image saliency extraction model assisted by deep prior information, and uses a large-scale data set for training on the server to obtain a trained model, and then integrates multiple features. The saliency extraction algorithm is combined with the image saliency extraction model assisted by deep prior information, and a saliency extraction algorithm based on multi-feature fusion and deep prior information is proposed. A new method for extracting key frames of motion video is introduced in detail. Taking into account that sports videos in real applications are susceptible to interference from various factors, resulting in poor picture quality, this article constructs a new visual attention model for moving targets in sports videos, which integrates images. The combination of multiple features of the bottom-level features and the skin color confidence map of the moving target overcomes the problem that a single feature cannot fully express the moving target. Since the processing object in this article is for the moving target in the video of the sports room, the extracted moving target can provide samples for video post-processing. The experimental results show that the proposed key frame extraction algorithm can quickly grasp the pedestrian information in the motion video and provide effective processing samples for the motion target for video post-processing.

INDEX TERMS Key frame extraction, deep prior information, zero-sample learning, motion video, visual attention model.

I. INTRODUCTION

With the further development of Internet technology and multimedia technology, as well as the popularization of digital devices and large-capacity storage devices, a large amount of video information is generated every day [1], [2]. In daily life, humans mainly obtain information through the visual system. Video information is widely used in all areas of our lives. People hope that the retrieval, browsing, and storage of videos can be as efficient as text data, and they can get what they are interested in through quick browsing [3]–[5]. The video data is much more complex than text and text data, and the content is particularly rich and changeable, which brings great difficulties to subsequent processing [6]. The variety of video content and formats on the Internet makes it an urgent problem to quickly and effectively find content of interest from a large amount of video data [7]. However, due to the different subjective perceptions and values of each person, different people have different understandings of the same video. This limitation of text description is easy to produce error. People need to perceive the content of the video by directly watching the real picture of the video and find the target they are interested in [8]. The amount of video data is large and there are many redundant information. It will waste a lot of time to find interesting content by directly watching each video, and the efficiency is very low [9].

The cluster-based method takes into account that different shots may also have a strong correlation [10]. The clustering algorithm is used to aggregate the shots of the low-level structural unit into the scene of the high-level structural unit. The original video is compared at the height of the scene layer [11]–[13]. The content is summarized and described, which is closer to advanced semantic understanding and human brain’s understanding of video content [14], [15].
Therefore, it is necessary to extract multiple key frames within a shot to adapt to the dynamic video semantic content. However, because computer vision is still a very difficult research challenge, most of the existing work chooses to use some low-level visual features, such as color and motion, instead of understanding high-level semantics [16], [17]. In the video stream, the camera emphasizes its own importance by staying in a position or a short stay in a certain action of a certain person [18]–[20]. However, this method requires a large amount of calculation when analyzing motion, and the local minimum is not necessarily accurate [21], [22]. Researchers select multiple key frames based on the significant changes between frames [23]–[25]. Relevant scholars calculate the energy and standard deviation characteristics of each sub-band after contourlet conversion to form a feature vector to represent the video frame, thereby extracting key frames [26]–[28]. This method greatly reduces the amount of calculation, but only compares the contours of the video frames, which is easy to extract redundant key frames [29]. Related scholars proposed a new space-time saliency model for surveillance video. The model combines top-down and bottom-up visual attention mechanisms, and analyzes static saliency and motion saliency respectively. After the foreground is extracted, the foreground is further extracted by the multi-scale Gaussian pyramid and the features are merged as static saliency, and the statistical analysis characteristics of the motion vector field are used to calculate the motion saliency. Finally, the static saliency and the motion saliency are merged through the center-surround mechanism defined by the approximate Gaussian function [30]. Scholars use the mutual information between frames as a method to measure the degree of frame difference [31]. However, the frame changes of the shots are continuous and consistent, and the degree of difference between adjacent frames is very small. These methods are based on calculations between adjacent frames and cannot accurately determine the specific position of the key frame [32]. Most bottom-up saliency detection algorithms cannot obtain results close to the true value on images with complex content and texture. This is because bottom-up algorithms only use known low-level features, and some algorithms have the problem of image features interfering with each other during processing.

Based on the idea of deep neural network and zero-sample learning, this article proposes an image saliency model based on deep prior information. The whole model is mainly divided into feature extraction module and saliency map prediction. Among them, the feature extraction module is composed of a modified VGG16, and the saliency map prediction is implemented by the nearest neighbor classifier in the feature space. We use the original image and its true value for training, so that the network learns the relationship between the image pixel feature and the true value area feature. Specifically, the technical contributions of this article can be summarized as follows:

First: We use the original image and the saliency map generated by other bottom-up methods as the input of the network. By setting the number of iterations, a better saliency map can be obtained. Combined with the model based on deep prior information, a saliency extraction method based on multi-feature fusion and deep prior information is proposed. We design a comparative experiment on the two data sets, and draw a graph of evaluation indicators. The results show that the algorithm in this article makes full use of image features and is more effective.

Second: We propose a key frame extraction method based on the saliency of moving targets in sports videos. This method comprehensively considers the problem of poor picture quality of sports videos in real applications, and proposes a feature description method of moving targets that integrates multiple characteristics to achieve the goal of fully expressing the characteristics of moving targets. The specific operation is to extract three underlying features of color, texture, and shape of the moving target. The purpose of adding the skin color confidence map is to highlight the area of the moving target.

Third: The experiment proves that the multi-feature fusion image can more fully express the moving foreground and the key frame extracted from the multi-feature fusion image can represent the more prominent and clear state of the target in the video, so the key frame of the moving target can provide more effective processing samples for video post-processing. In the later period, the research object of the key frame extraction system will be further expanded from moving pedestrians to other fields. This will be worth studying and challenging work.

The rest of this article is organized as follows. Section 2 analyzes the key technologies of sports video retrieval. Section 3 proposes a motion video key frame extraction algorithm based on multi-feature fusion and depth prior information. Section 4 discusses the simulation results. Section 5 summarizes the full text.

II. KEY TECHNOLOGIES OF SPORTS VIDEO RETRIEVAL
A. VIDEO DATA CONTENT CHARACTERISTICS AND STRUCTURAL ANALYSIS
The so-called content-based video retrieval is the process of finding images that meet specific visual feature descriptions in large-scale video databases based on the scenes, shots, frames, moving objects in the video data and the color, texture, and shape features in the image data. Its research goal is to provide algorithms that can automatically understand or recognize image visual features without human involvement. Content-based video retrieval has broad application prospects. It is currently mainly used in the following areas: embedding content-based video retrieval engines into conventional database management systems to achieve multimedia data retrieval and retrieval of video libraries in special fields; content-based retrieval of multimedia data contained on HTML pages in the Web information network on the Internet. The early full-text information retrieval, the identification and management of criminal avatars, and the identification and management of fingerprints were all attempts based on
content retrieval. Now this technology will be extended to any media and wider fields. Image retrieval divides a video sequence into several shot sequences, and then finds several key frames in each shot sequence to represent the main visual content of the shot; after the video sequence is structured, the visual features (colors) of each key frame are extracted. Texture, shape and motion parameters are stored in the feature database; the system similarity matching module processes the query constructed by the user, finds the matching image in the video database, and feeds the result back to the user. The parameters are adjusted to perform a step-by-step refinement query, and finally a satisfactory query result is obtained.

Usually what we call video refers to the digital video that is easy to be processed by the computer. It stores the analog signal in a digital format of “0” or “1” after being processed by an analog-to-digital converter. Digital video has many advantages over analog video. It has strong scalability. The digital video data is easy to edit, process, store and transmit. Digitization refers to the process of capturing an analog video signal through a specific device and converting it into a digital form, sampling each pixel separately, and finally saving the numerical result. Nowadays, there are many video encoding and decoding technologies, but the huge amount of data is still a bottleneck to be solved in the video retrieval system. The video data processing methods before and after compression are also very different.

Video data is a collection of sequential images in continuous time, which integrates multimedia information such as images, text, and audio, and is a kind of comprehensive media information. Compared with text type data, there are many differences, and its particularities mainly include:

1) LARGE AMOUNT OF DATA
For example, an image with a resolution of 640*80 and a color depth of 24bit/pixel has a data volume of about 1M. If the playback speed is set to 30 frames/second, it will generate about 30M data in one second. It can be seen that the amount of data contained in video data is very large, so the research on the compression, encoding and decoding of video data is of great significance.

2) COMPLEX STRUCTURE
Text data is a collection of pure character data, which can be regarded as one-dimensional data. Each pixel in the image data is under the two-dimensional coordinates in the image, which can be understood as two-dimensional data. Video data is a sequence of images, which changes with time. Therefore, video data has dual attributes of space and time. It is three-dimensional data, and its expression and processing methods are the most complicated.

3) RICH CONTENT
The content of the video data is extremely rich. It contains not only audio information such as music and voice, but also visual information such as color and shape, as well as spatial location and movement information of the subject. Due to differences in understanding ability and knowledge levels between people, the interpretation of video data by different individuals is highly subjective. For video data, it is difficult to find a unified and objective description standard.

As shown in Figure 1, video data is an unstructured image stream sequence. As the storage object of the video database, we call the video data entered into the computer the original video stream. In order to quickly retrieve video objects, video data is stored in a hierarchical structure, from high to low, the video data is divided into video sequence, scene, shot, and frame at a time.

It is necessary to extract one or more images from a shot to describe the content of the shot. We call this kind of special images a key frame. The superior key frame extraction technology can greatly reduce the redundancy while greatly improving the retrieval efficiency.

B. FEATURE EXTRACTION AND SIMILARITY CALCULATION
Figure 2 shows a common content-based video retrieval system. To retrieve the video, we must first analyze the structure of the video stream, divide the video into shots. Next, we extract the dynamic features of the shots. At the same time, we extract the static special frames of the key frames. The class characteristics are recorded and stored in the database. Finally, when the user submits a search query instruction, the system returns the query result to the user according to the matching degree. The quality of feature extraction and retrieval algorithms will also affect the performance and efficiency of the entire retrieval system.

Video feature extraction is the basis of video clustering and retrieval, which reflects certain characteristics of the shot content. The static features of video are usually also available.
in static images, while dynamic features are unique feature attributes of video, namely motion features.

The calculation of color feature extraction is small, which is more conducive to image feature extraction. Color histogram features are commonly used in video retrieval to express its characteristics. The histogram describes the amount of pixels contained in each color interval in the image frame, reflects color-related statistical information, and shows the occupation of different colors in the entire image ratio. Color moments are another simple and effective way to express color features. In addition, there are color feature extraction methods such as cumulative color histograms and block primary colors.

We summarizes 6 texture expression characteristics. Roughness refers to the size of the texture interval. The larger the image resolution, the thicker the texture. The local grayscale changes of the image are measured by contrast. The linearity describes the direction of the texture. These three texture feature components play an important role in image retrieval. Effect. The LBP method has multi-scale and rotation invariance, and the computational complexity is small, and the effect is significant when used for texture classification. The wavelet-based textural analysis method enables the image to obtain multi-resolution characteristics in space and frequency.

After the image feature extraction is completed, the extracted feature vectors can be used to characterize the corresponding images. The similarity measurement in image retrieval is performed by judging whether these feature vectors are similar.

When the components of the feature vector are related or have different weights, the Mahalanobis distance can be used to calculate the similarity.

\[ d_{AB} = C^{-1}(A - B)^T(A - B) \]

In the above formula, \( C \) is the covariance matrix of the eigenvector. If the components of the feature vector are not correlated, the Mahalanobis distance can be simplified to:

\[ d_{AB} = \sum_{k=0}^{n-1} \frac{(B_k - A_k)^2}{c_i} \]

In the above formula, \( c_i \) is the variance of each component. Compared with Euclidean distance, Mahalanobis distance takes into account the correlation of each component of the feature vector, but also exaggerates the effect of variables with small changes.

C. KEY FRAME EXTRACTION TECHNOLOGY

It is used to statically describe the theme and important content of the original video stream. Corresponding services can be provided according to user needs, such as obtaining the specific part required in the video.

It is very difficult for the extraction of key frames to satisfy both validity and accuracy. At present, the principle of key frame extraction is relatively conservative. That is to give priority to the effectiveness of key frames. On the premise of ensuring that the shot can be represented as much as possible, the number of key frames is minimized and redundant frames are removed to improve the efficiency of video retrieval and browsing. Generally, we select different criteria for different video types, and different selection methods can also establish the most suitable criteria for the method according to different principles.

1) METHOD BASED ON CONTENT ANALYSIS

Due to the influence of operational changes such as target movement in the scene and camera zoom during video shooting, one key frame is not enough to fully express the content of the shot, and it is often necessary to extract a few more representative key frames. From the point of view of the key frame extraction principle, the key frame should provide a rich and comprehensive lens summary as much as possible. Therefore, the key frame extraction can be regarded as an optimization process.

A lens is a collection of image frames that are continuous in time and highly correlated in content. The least relevant image frames in this collection are selected as the key frames of the lens to maximize the maximum information.

2) EXTRACTION METHOD BASED ON COMPRESSED VIDEO STREAM

Before extracting key frames from a compressed video stream, the above methods must perform decompression operations, and then analyze and process video data. This type of method requires a large amount of calculation and is not efficient. In response to the above problems, the exploitation of key frame extraction methods in the compressed domain began. Such methods directly analyze certain types of features in compressed video data, and their computational complexity has been significantly reduced. In the MPEG
format compression standard, a group of pictures is a basic block that can perform arbitrary access operations on a video stream. Its English abbreviation is GOP. Generally, a typical GOP can only have one shot change.

Based on motion analysis, the image change brought about by camera motion can also be used as an important basis for key frame extraction. The reasons for this image change can be divided into changes in camera focus, and one type is changes in camera angle.

III. MOTION VIDEO KEY FRAME EXTRACTION ALGORITHM BASED ON MULTI-FEATURE FUSION AND DEEP PRIOR INFORMATION

A. VGG NETWORK MODEL

Considering that there are relatively few methods to combine the bottom-up and top-down models, this section considers the combination of the two models and uses the saliency map based on superpixel fusion multi-feature as the prior information map. The deep convolutional neural network is used to extract the deep uncertain features of the image. This section uses a guided learning method to further obtain more accurate saliency maps.

The Visual Geometry Group (VGG) network mainly studies the effect of the depth of the convolutional neural network and the size of the convolution kernel on the performance of the network. The final conclusion is that the 3*3 convolution kernel is used to increase the depth of the network to 16. At the 19th layer, the performance of the network can be significantly improved, and the generalization ability to other data sets is strong. The VGG series of network configurations have a total of six columns. Each column represents the composition of a network. All networks contain a varying number of hidden layers, as well as three identical fully connected layers and a softmax layer. The difference is that the depth is different. Among them, the network A has the least number of layers, and the network E has the most 19 layers in total. Due to the unstable gradient in the deep convolutional neural network, the shallow network is trained first, and then the deep network is initialized with the weight of the shallow network. This can greatly accelerate the convergence speed of the deep network.

Because VGG16 can not only achieve quite good results in image classification tasks, but also has much fewer parameters than VGG19, and the final trained model occupies a smaller memory. Therefore, when studying image-related classification tasks, you use VGG16 to modify according to actual needs. Figure 3 shows the graphical structure diagram of the VGG16 network.

B. ZERO SAMPLE LEARNING

Supervised learning is a large-scale labeled training. Specific features can be learned on the set, but with the deepening and refinement of the computer vision field, many small fields lack large-scale data sets, which is a difficult problem for supervised learning. Therefore, we introduce zero-sample learning to solve learning tasks that lack a labeled training set.

In layman’s terms, traditional supervised learning is to learn the characteristics of the same target from the training set, and then test on the test set of the same kind, that is, the learning and extraction are the same target. What the zero-sample learning needs to do is to learn the characteristics of similar targets from the training set, and then extract different types of targets with the same attributes, that is, the categories in the training set and the test set in zero-sample learning are disjoint, which is the same as supervised learning. Zero-sample learning is actually simulating human reasoning thinking and inferring more advanced results based on known low-level attributes.

There are two models for zero-sample learning: Direct Attribute Prediction (DAP) and Indirect attribute prediction (IAP). The DAP model can be regarded as a three-layer model. In the original input layer of the first layer, there are K known classes y and L unknown classes z. After the input layer, M classifiers are trained, and the output of the second layer contains the attribute a_m, an m-dimensional feature space. For each attribute a_m, you use the training data to learn the classifier p(a_m | x), and use the following formula to calculate the classification probability during testing:

\[ p(z | x) = \frac{P(z)}{P(a_1^m)} \prod_{m=0}^{M-1} p(a_m^m | x) \] (3)

In the formula, p(z) represents the prior probability of the unknown class. You classify data x into the class with the largest posterior probability:

\[ l(x) = \arg \max_{m=0}^{M-1} p(a_m^m | x) \] (4)

Different from the DAP model, the IAP model uses the known class y as an intermediate layer to indirectly learn the mapping between unknown features and attributes. The method is to first use supervised learning to learn a probability classifier p(y_k | x), and then used to predict the attributes of unknown data, as shown in the following formula:

\[ p(a_m | x) = \sum_{k=0}^{K-1} p(a_m | y_k)p(y_k | x) \] (5)

Compared with the two models, the IAP model has more layers and fewer practical applications. Many existing researches are conducted on the basis of the DAP model. Therefore, in this article, aiming at the key problem that there are fewer data sets in the saliency extraction field, the idea of the DAP model is also applied to the saliency extraction model assisted by prior information.

C. SALIENCY EXTRACTION MODEL ASSISTED BY DEEP PRIOR INFORMATION

1) SALIENCY EXTRACTION NETWORK MODEL

Deep convolutional neural networks have achieved good results in various tasks in the image field, especially the VGG
Q. Zhong et al.: Key Frame Extraction Algorithm of Motion Video Based on Priori

model. This model automatically extracts image features with high versatility. Applying this model to image saliency extraction, the extracted image features are more complex and more advanced than manually designed features. Therefore, based on the VGG16 model, this section designs a new network to extract the deep features of the image.

This article adds a 1\( \times 1 \) convolution kernel after each convolution block of VGG16 to reduce network parameters, improve the speed of network training and testing, and remove the last pooling layer to retain more high-dimensional features; at the same time, we make full use of the image features of different resolutions extracted by each convolution block. This article refers to the network model of extracting multi-resolution feature information based on the 4\( \times 5 \) grid structure, and deconvolution expands the feature maps of the five convolution blocks. Then we use two convolutional layers for feature fusion; use the fused feature vector in the feature space to iteratively train a nearest neighbor classifier, and the output of the classifier is the binary significant label distribution of the input image. The input and output dimensions of this model are the same.

The overall structure of the saliency extraction model assisted by deep prior information designed in this article is shown in Figure 4. The true value image uses the same feature extraction module as the original image. When the network is trained, the input is the original image and its true value image, and when the network is tested, the input of the network is the original image and its prior information map. The prior information map is incomplete by other existing bottom-up methods. The saliency map corresponds to the DAP model. The input includes some known data and labels and some unknown data.

The size of the feature map output by the last convolution module is only 1/16 of the original image, which is inconsistent with the size of the input image. Therefore, this article only retains the first three pooling layers, so that the smallest feature map size becomes 1/8 of the original image, and then the output of each convolution module is up-sampled using the deconvolution operation, and the feature map is output. The size is adjusted to be consistent with the input image, and then two convolutional layers containing 512 neurons are used to convert the feature map into a feature vector.

This network can be regarded as a feature mapping function with a parameter of \( \theta \). For each pixel \( x_{mn} \) on the image \( X_m \), its feature vector can be obtained:

\[
\phi_{mn} = \phi(\theta, x_{mn}) \tag{6}
\]

Similarly, for the true value image area \( C_{mk} \), the same neural network is used to extract features and the same subsequent convolutional layer as the pixels to obtain its...


\[ \lambda_{mk} = \varphi(C_{mk}; \delta) \tag{7} \]

In the formula, \( \delta \) represents the parameters of the network, and \( C_{m1} \) and \( C_{m2} \) represent the foreground and background areas of the image \( X_m \), respectively.

It belongs to this area, and the degree of membership is obtained by the softmax layer at the end of the network, which is expressed as follows:

\[ p(C_{mk} | x_{mn}) = \frac{-\exp(d\phi_{mn})}{\sum_j -\exp(d\phi_{mj})} \tag{8} \]

In the formula, \( j \) represents the total number of image areas; \( k \) represents the corresponding area is the background or foreground.

2) FEATURE EXTRACTION MODULE

The feature extraction module is the core of the saliency extraction model assisted by deep prior information. The main operations included in this part include 1*1 convolution and deconvolution, which are specifically described as follows:

(1) 1*1 convolution kernel: 1*1 convolution kernel is widely used in deep networks. Because its size is only 1*1, it does not consider the influence of neighboring pixels in actual operation. It can be used to adjust the number of feature maps output by the convolution module, and perform nonlinearity on pixels at corresponding positions on different feature channels. Combination can not only increase the non-linear expression ability of the network, but also achieve the purpose of dimension increase or dimension reduction while significantly reducing the parameters.

(2) Deconvolution: Deconvolution and convolution are corresponding. In fact, it is the back propagation of convolution. The forward propagation and back propagation of the two are just interchangeable. The deconvolution is from top to bottom. Convolution is bottom-up. It should be noted that the use of deconvolution does not get the same input image as the convolution operation, but can only make the output feature map size larger, that is, only the output of the same size as the input can be obtained, which can help us saliency in this article achieve the end-to-end purpose in the extraction model. The main implementation process of deconvolution is shown in Figure 5. First, the input image is filled, then the convolution kernel is used for convolution, and finally the crop is performed to obtain an output twice the size of the original image.

The model designed in this article uses a 1*1 convolution kernel to perform dimensionality reduction after each convolution block to obtain the same number of feature maps as the previous convolution module.

3) NETWORK TRAINING AND TESTING

In the actual training process, the newly added convolutional layer is randomly initialized, the output of each convolution block is expanded and convolved, and then the feature vector is merged. We train the entire network on the current larger data set for saliency extraction. During training, the original image and ground-truth image of the data set are divided into training set and validation set. A total of 32 epochs are trained. The training process is shown in Figure 6.

During training, the convolution block extracts the feature vector of the image area and the image pixel, and by constructing the nearest classifier in the feature space, it is determined which area of the image the pixel belongs to. The whole process is equivalent to using a large number of images and truth values to enable the network to learn the relationship between the pixels and regions of the salient targets in common images, that is, through learning with few samples, the network has a certain reasoning ability like...
a person, and can be based on a small amount. The prior information infers the salient area.

In the network training stage, the input is the original image and the corresponding true value image, and the pixel-level true value image in the data set is used to extract regional features. However, when the trained model is used for testing, the input is not the true value but the incomplete saliency map and the corresponding original image generated by other classic methods, so that the feature vector of the background or salient area extracted from the incomplete saliency map is inaccurate, so the result of the nearest neighbor classifier needs to be used to correct the initial saliency map.

4) IMAGE SALIENCY EXTRACTION BASED ON MULTI-FEATURE FUSION AND DEEP PRIOR INFORMATION

The image saliency extraction algorithm based on multi-feature fusion and deep prior information mainly includes two major steps: multi-feature fusion and deep prior information correction. These two steps are performed separately. First, the color feature saliency map and the texture feature saliency map of the image are merged, and then the pre-trained saliency extraction model is used to improve the effect. The whole process combines two classic saliency extraction ideas, bottom-up and top-down. Of course, these two steps can also be independently saliency extraction.

The multi-feature fusion part is completely bottom-up. The depth prior information correction part uses the pre-trained model in this article, takes the original image and the fusion saliency map output from the multi-feature fusion part as input, and iteratively improves the effect of the input saliency map, this part is top-down. Therefore, the image saliency extraction algorithm based on multi-feature fusion and depth prior information is a combination of bottom-up and top-down algorithms.

IV. SIMULATION RESULTS AND ANALYSIS
A. MOTION VIDEO DATA

In order to analyze the specific situation of wrong matching, Figure 7 shows an example of the wrong matching moving target in case video 1. The movement state of the whole body area of the moving target remains stable during the movement, only a brief slight turning of the head occurs. Some viewers ignore the subtle changes in the orientation of the face area, and other viewers notice the face orientation. However, it is impossible to make a correct judgment on the frame with the greatest degree of change in orientation, so 8 out of 20 observers chose frame 2734 as the candidate key frame. The motion range is relatively large. Five people chose the 2830 frame of the moving target, because the moving target was close to the camera at frame 2830 and the side face area was large; the remaining 7 viewers chose the others. According to the above-mentioned discrimination conditions, candidate key frame 1 is a valid artificial key frame. According to the key frame extraction algorithm in this article, the algorithm in this article is sensitive to changes in the moving face area, so it can obtain the frame with a large proportion of the moving target face area in the picture as the key frame. According to the idea of this article, the 2742th frame is obtained as the key frame, the frame cannot be matched with the valid candidate key frame (frame 2734), so the match fails. The match rate of the experimental video in this article is shown in the last column of Table 1.

B. KEY FRAME EXTRACTION

In order to provide effective pedestrian processing samples for surveillance video post-processing, this article proposes a new key frame extraction method suitable for surveillance video based on the existing key frame extraction methods. The main feature of this method is to extract key frames based on a new visual attention model. The visual attention model integrates the underlying features of the image and the skin color confidence map of the moving target, and uses the dynamic fusion method to fuse multiple features to generate a comprehensive feature image. In this article, the comprehensive feature image is regarded as a significant feature of pedestrian moving targets. Finally, according to the saliency of the pedestrian moving target, the frame that can better represent the pedestrian moving target is extracted as the key frame.

Figure 8 and Figure 9 are the application examples of the key frame extraction algorithm in this article. Figure 8 depicts a moving target taken from the case video and its key
TABLE 1. Sports video data used in this experiment.

| Sports video number | Number of frames | Size     | Frame rate | Number of key frames | Video features | Match rate |
|---------------------|------------------|----------|------------|----------------------|----------------|------------|
| 1                   | 4000             | 180*120  | 25f/s      | 4                    | Outdoor        | 0.99       |
| 2                   | 4200             | 170*110  | 15f/s      | 6                    | Outdoor        | 0.87       |
| 3                   | 5000             | 290*120  | 27f/s      | 9                    | Outdoor        | 0.94       |
| 4                   | 4700             | 260*120  | 29f/s      | 12                   | Outdoor        | 0.76       |
| 5                   | 3300             | 320*120  | 34f/s      | 4                    | Outdoor        | 0.81       |
| 6                   | 3900             | 280*120  | 20f/s      | 5                    | Outdoor        | 0.92       |
| 7                   | 2800             | 190*80   | 22f/s      | 9                    | Outdoor        | 0.87       |
| 8                   | 2100             | 130*100  | 37f/s      | 11                   | Outdoor        | 0.74       |
| 9                   | 4300             | 190*150  | 36f/s      | 5                    | Outdoor        | 0.98       |
| 10                  | 3980             | 180*140  | 22f/s      | 3                    | Outdoor        | 0.97       |

According to the extraction algorithm in this article, each target pedestrian on the front or a certain degree of side will be selected as a key frame corresponding to a frame during the movement. Figure 8 is a histogram of the saliency of the moving target. The histogram represents the saliency of all states of the moving target in the video. This article uses the normalization operation for the calculation of the saliency, making the saliency value range from 0 to 1. It can be seen from the figure that as the target moves, the saliency of the
moving target presents a tortuous trend, because the target gradually approaches the camera during the movement, its proportion of the area in the video screen is increasing. When the moving target is in the 3200th frame, the corresponding saliency of the moving target is the greatest, and as the moving target continues to walk, the moving target in the moving video screen gradually changes from a front face to a side face and the body area gradually leaves the camera to capture the range of the screen to be reached. Therefore, the saliency value of the target after 3200 frames starts to decrease. In the 3200th frame, the moving target is close to the camera and the physical features of the moving target are in a clearer state, so the corresponding frame is selected as the key frame. This coincides with the saliency histogram in the first row and is consistent with common sense when contracting. From the saliency value, it can be seen that the saliency of the face area is correlated with the clarity of the body area. Facial regions corresponding to images with higher body saliency often have a higher degree of clarity, and vice versa. This proves that the algorithm in this article is effective in obtaining pedestrian information. At the same time, the moving target extracted from the key frame has a relatively clear state, which can provide effective processing samples of the moving target for video post-processing.

The motion state of the moving target in Figure 9 pauses at the beginning of the motion, and turns the head in the middle of the motion, so the saliency value of the saliency curve will change up and down in the midway. It can be seen from the motion saliency curve that the motion target is most significant when it moves to 3500 frames. At the same time, at 3500 frames, the face and body area of the moving target are in a relatively clear state, which is consistent with the saliency curve in Figure 9.

Figure 10 and Figure 11 are two moving targets in the same video. The moving conditions of the moving targets are moving from far and near to the camera and finally disappear from the lens. It can be seen from the saliency curve that the overall trend is still that the frames close to the lens and not disappearing from the lens are key frames. Because the two people’s movement patterns are not exactly the same during...
the exercise, the saliency curve of the two people rises as they get closer to the camera while showing different details of changes.

C. COMPARATIVE ANALYSIS WITH OTHER KEY FRAME EXTRACTION ALGORITHMS

First, we extract the key frames of the case video 2. According to the algorithm in this article, when the research object is set as the person in the video, the 578th frame is extracted as the key frame image, which means that the saliency of the moving object is the largest in the whole movement process. According to the optical flow method, the key frame of case video 2 is extracted, and the extracted key frame is the 1184th frame. Comparing and analyzing the extraction results of the two algorithms, the algorithm in this article can achieve a higher compression rate, and a clearer description of the shape of the moving target can be achieved with one frame of video. It finds the distance between each frame and the cluster center to select the closest to each cluster center. The frame with the greatest similarity between cluster centers is used as the key frame. The video background color changes weakly, resulting in the key frames not being extracted until frame 1324, and the motion trajectory of the moving target and the morphological characteristics of the moving target cannot be highlighted, and the degree of redundancy is large. The comparison of the results of the three key frame extraction methods of case video 2 is shown in Figure 12.

Analyzing the two PR curves in Figure 13 shows that the proposed model shows a theoretically consistent upward trend. The FLICSP curves in both figures are higher than the FLICS curves. This shows that after the effect of the FLICS algorithm has been improved, the accuracy and recall rate have been improved. At the same time, the PR curve after the improvement of the Principle Component Analysis (PCA)
algorithm is higher than before the improvement, while the improvement of the PR curve after the improvement of the Histogram Contrast (HC) and Saliency Filters (SF) algorithms is relatively small, which indicates that the improvement effect is not enough. Therefore, analyzing the initial saliency maps generated by the HC and SF algorithms, it is found that the saliency maps obtained by these two algorithms on some images with small saliency regions are relatively messy. This phenomenon shows that the saliency extraction model based on the depth prior information is more dependent. By comparison, it is found that the PR curve of the Multiscale Deep Features (MDF) algorithm is the highest, which also shows that the features extracted entirely by the neural network are more adequate.

As shown in Figure 14, the proposed model improves the F value more obviously, but it still has a certain gap with the MDF algorithm. Based on the analysis of the above three indicators, the saliency extraction (FLICSP) algorithm based on multi-feature fusion and deep prior information can obtain better extraction results than the multi-feature fusion (FLICS) algorithm, and the saliency extraction model can effectively improve the extraction results of HC, SF and PCA algorithms.

The algorithm complexity analysis in this article is mainly the analysis of processing time. Considering the motion video from the whole framework, the FLICSP algorithm takes the least time. The comparison of algorithm running time is shown in Figure 15.

V. CONCLUSION

The original image and the true value are used as input for training, and the relationship between the image true value area feature and the pixel feature is learned through the network. The original image and the initial saliency map generated by other methods are used as input for testing, and the zero-sample learning method is used to iteratively improve the extraction effect. A new key frame extraction method suitable for motion video is proposed. We construct a new visual attention model for the moving target in the sports video. The visual attention model integrates the underlying features of the image and the skin color confidence map of the moving target. The integration of multiple features overcomes the inability of a single feature to fully express the moving target. Since the object processed in this article is for the moving target in the moving video, the extracted moving target can provide target samples for video post-processing operations such as face super-resolution reconstruction. In practical applications, it is based on the compressed domain, so the key frame extraction of the video data stored in the compressed domain requires further research in the future.

Key frame extraction is only one step in the realization of content-based video retrieval. To build the entire content-based video retrieval system still needs a lot of work. This article only does a small part of the research. The proposed key frame extraction algorithm still has many areas to be improved. In video retrieval, due to the huge amount of video data and rich content, it is very difficult to retrieve the video you need quickly, efficiently and accurately. More research is needed to truly realize video retrieval. In addition, in the key frame extraction part of the motion video, due to the limited judging effect of the skin color model for the positioning of the moving target face area, the processing of this part is still in a semi-automatic state, and the precise positioning of the target face cannot be achieved. Therefore, the next step needs to continue to optimize the face positioning of the moving target in the sports video.

REFERENCES

[1] R. Hu, R. Min, and Y. Pi, “A video-SAR imaging technique for aspect-dependent scattering in wide angle,” IEEE Sensors J., vol. 17, no. 12, pp. 3677–3688, Jun. 2017.
[2] J. Lu, V. E. Liong, and J. Zhou, “Cost-sensitive local binary feature learning for facial age estimation,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5356–5368, Dec. 2015.
[3] 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.
[4] C.-A. Deledalle, L. Denis, S. Tabb, and F. Tupin, “MuLoG, or how to apply Gaussian denoisers to multi-channel SAR speckle reduction?” IEEE Trans. Image Process., vol. 26, no. 9, pp. 4389–4403, Sep. 2017.
[5] J.-L. Lai and Y. Yi, “Key frame extraction based on visual attention model,” J. Vis. Commun. Image Represent., vol. 23, no. 1, pp. 114–125, Jan. 2012.
[6] M. Fei, W. Jiang, and W. Mao, “A novel compact yet rich key frame creation method for compressed video summarization,” Multimed. Tools Appl., vol. 77, no. 10, pp. 11957–11977, May 2018.
[7] J. T. Case, M. T. Ghasr, and R. Zoughi, “Optimizing two-dimensional uniform spatial sampling for microwave SAR-based NDE imaging systems,” IEEE Trans. Instrum. Meas., vol. 60, no. 12, pp. 3806–3815, Dec. 2011.
[8] X. Wang, R. Wang, Y. Deng, P. Wang, N. Li, W. Yu, and W. Wang, “Precise calibration of channel imbalance for very high resolution SAR with stepped frequency,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 8, pp. 4252–4261, Aug. 2017.
[9] J. Lu, V. E. Liong, and J. Zhou, “Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 8, pp. 1979–1993, Aug. 2018.
[10] Y. Zhang, R. Tao, and Y. Yang, “Motion-state-adaptive video summarization via spatiotemporal analysis,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 6, pp. 1340–1352, Jun. 2017.
[11] P. Sidirooulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, “Temporal video segmentation to scenes using high-level audiovisual features,” IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 8, pp. 1163–1177, Aug. 2011.
[12] G. Guan, Z. Wang, S. Lu, J. D. Deng, and D. F. Deng, “Keypoint-based keyframe selection,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 4, pp. 729–734, Apr. 2013.
[13] J. Lu, G. Wang, and J. Zhou, “Simultaneous feature and dictionary learning for image set based face recognition,” IEEE Trans. Image Process., vol. 26, no. 8, pp. 4042–4054, Aug. 2017.
[14] D. A. Yocky, R. D. West, R. M. Riley, and T. M. Calloway, “Monitoring surface phenomena created by an underground chemical explosion using fully polarimetric VideoSAR,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 5, pp. 2481–2493, May 2019.
[15] H. Jacob, F. L. C. Pádua, A. Lacerda, and A. C. M. Pereira, “A video summarization approach based on the emulation of bottom-up mechanisms of visual attention,” J. Intell. Inf. Syst., vol. 49, no. 2, pp. 193–211, Oct. 2017.
[16] Y. Zhang and D. Zhu, “Height retrieval in postprocessing-based VideoSAR image sequence using shadow information,” IEEE Sensors J., vol. 18, no. 19, pp. 8108–8116, Oct. 2018.
[17] H. Wang, Z. Chen, and S. Zheng, “Preliminary research of low-RCS moving target detection based on ka-band video SAR,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 6, pp. 811–815, Jun. 2017.
[18] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, “Learning compact binary face descriptor for face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 10, pp. 2041–2056, Oct. 2015.
Q. Zhong et al.: Key Frame Extraction Algorithm of Motion Video Based on Priori

[19] L. Zhang, L. Sun, W. Wang, and Y. Tian, “KaaS: A standard framework proposal on video skimming,” IEEE Internet Comput., vol. 20, no. 4, pp. 54–59, Jul. 2016.

[20] J. Zhang, Z. Suo, Z. Li, and Q. Zhang, “DEM generation using circular SAR data based on low-rank and sparse matrix decomposition,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 724–728, May 2018.

[21] Y. Cong, J. Liu, G. Sun, Q. You, Y. Li, and J. Luo, “Adaptive greedy dictionary selection for Web media summarization,” IEEE Trans. Image Process., vol. 26, no. 1, pp. 185–195, Jan. 2017.

[22] X. Zhang, C. Zha, H. Wu, Z. Liu, and Y. Xu, “An imbalance compensation framework for background subtraction,” IEEE Trans. Multimedia, vol. 19, no. 11, pp. 2425–2438, Nov. 2017.

[23] X. Song and W. Yang, “Processing video-SAR data with the fast back-projection method,” IEEE Trans. Aerosp. Electron. Syst., vol. 52, no. 6, pp. 2838–2848, Dec. 2016.

[24] Y. Duan, J. Lu, J. Feng, and J. Zhou, “Context-aware local binary feature learning for face recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1139–1153, May 2018.

[25] J. Lu, J. Hu, and J. Zhou, “Deep metric learning for visual understanding: An overview of recent advances,” IEEE Signal Process. Mag., vol. 34, no. 6, pp. 76–84, Nov. 2017.

[26] J. Wu, S.-H. Zhong, J. Jiang, and Y. Yang, “A novel clustering method for static video summarization,” Multimedia Tools Appl., vol. 76, no. 7, pp. 9625–9641, Apr. 2017.

[27] X. Yang, Y. R. Zheng, M. T. Ghass, and K. M. Donnell, “Microwave imaging from sparse measurements for near-field synthetic aperture radar,” IEEE Trans. Instrum. Meas., vol. 66, no. 10, pp. 2680–2692, Oct. 2017.

[28] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, “Detecting moving objects, ghosts, and shadows in video streams,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[29] C.-R. Huang, P.-C.-J. Chung, D.-K. Yang, H.-C. Chen, and G.-J. Huang, “Maximum a posteriori probability estimation for online surveillance video synopsis,” IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 8, pp. 1417–1429, Aug. 2014.

[30] M. A. Poturwar and M. Atique, “Performance evaluation of fade, wipe and dissolve in detection of shots in video,” Perform. Eval., vol. 2, no. 2, pp. 9–276, Mar. 2015.

[31] N. Ejaz, I. Mehmood, and S. W. Baik, “Feature aggregation based visual attention model for video summarization,” Comput. Electr. Eng., vol. 40, no. 3, pp. 993–1005, Apr. 2014.

[32] X. Mao and D. Zhu, “Two-dimensional autofocus for spotlight SAR polar format imagery,” IEEE Trans. Comput. Imag., vol. 2, no. 4, pp. 524–539, Dec. 2016.

JINGUO ZHANG was born in Beijing, China, in 1965. He graduated from the Department of Physical Education, Beijing Normal University, in 1987. He has been working with the Department of Physical Education, China University of Geosciences (Beijing), since July 1987. He was promoted to an Associate Professor, in 2006, mainly engaged in physical education.

KAIXUAN SHI was born in Shandong, China, in 1987. She received the Ph.D. degree from Beijing Normal University, in 2017. After her Ph.D. degree, she has worked as a Teacher with the China University of Geosciences. Her research interest includes physical activity promoting good health and its neuroplasticity mechanisms.

YANG YU was born in Shandong, China, in 1992. She received the master’s degree from Beijing Sport University, in 2017. After her master’s degree, she has been working as a Teacher with the China University of Geosciences, since 2017, where she was a National Athletic Master. She has published three articles as the first author. Her research interests include aquatic sports, water training, and water therapy. She was awarded the National First-Class Referee’s Certificate in swimming and the Professional Qualification Certificate of swimming lifeguard.

QI ZHONG was born in Ganzhou, Jiangxi, China, in 1997. He received the bachelor’s degree in software engineering from the China University of Geosciences (Beijing), in 2018, where he is currently pursuing the master’s degree in physical education. His undergraduate project is the design and implementation of gym management system. His research interests include algorithm design and database design.

YUAN ZHANG was born in Lanzhou, Gansu, China, in 1978. He received the bachelor’s degree from Beijing Sport University, in 2000, and the master’s degree in physical education from the Beijing University of Physical Education, in 2012. Since 2000, he has been engaged in physical education with the China University of Geosciences (Beijing). In 2014, he was employed as an Associate Professor and a master’s Supervisor. He has been engaged in physical education research for a long time, published four academic articles, and participated in a project funded by the China Social Sciences.

CHANG LIU was born in 1987. She received the master’s degree in strength and conditioning from Beijing Sport University, in 2018. After her retirement in 2019, she taught at the Sports Department, China University of Geosciences (Beijing). She was a national level athlete, national level volleyball referee, and national junior volleyball coach. She was a former Bayi team volleyball player and got battalion-level cadre. She has achieved excellent results in beach volleyball competitions at domestic and abroad and also won many personal third-class merits and awards in the army.

* * *

YUAN ZHANG was born in Beijing, China, in 1965. He graduated from the Department of Physical Education, Beijing Normal University, in 1987. He has been working with the Department of Physical Education, China University of Geosciences (Beijing), since July 1987. He was promoted to an Associate Professor, in 2006, mainly engaged in physical education.

QI ZHONG was born in Ganzhou, Jiangxi, China, in 1997. He received the bachelor’s degree in software engineering from the China University of Geosciences (Beijing), in 2018, where he is currently pursuing the master’s degree in physical education. His undergraduate project is the design and implementation of gym management system. His research interests include algorithm design and database design.

YUAN ZHANG was born in Lanzhou, Gansu, China, in 1978. He received the bachelor’s degree from Beijing Sport University, in 2000, and the master’s degree in physical education from the Beijing University of Physical Education, in 2012. Since 2000, he has been engaged in physical education with the China University of Geosciences (Beijing). In 2014, he was employed as an Associate Professor and a master’s Supervisor. He has been engaged in physical education research for a long time, published four academic articles, and participated in a project funded by the China Social Sciences.

CHANG LIU was born in 1987. She received the master’s degree in strength and conditioning from Beijing Sport University, in 2018. After her retirement in 2019, she taught at the Sports Department, China University of Geosciences (Beijing). She was a national level athlete, national level volleyball referee, and national junior volleyball coach. She was a former Bayi team volleyball player and got battalion-level cadre. She has achieved excellent results in beach volleyball competitions at domestic and abroad and also won many personal third-class merits and awards in the army.

* * *