

Received July 20, 2020, accepted August 20, 2020, date of publication August 25, 2020, date of current version September 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019233

# Feature Extraction of Human Motion Video Based on Virtual Reality Technology

MIMI ZHOU 

School of Philosophy, Heilongjiang University, Harbin 150080, China  
IMUN Institute of Physical Education, Inner Mongolia University for Nationalities, Tongliao 028000, China  
e-mail: zhoumimi20100402@163.com

This work was supported by the Project of Inner Mongolia University for Nationalities: The Experimental Research on the Introduction of Healthy Physical Fitness Into Calisthenics Teaching under Project BS578.

**ABSTRACT** In virtual reality scenes, the premise of moving target recognition in video is accurate target segmentation and extraction of target low-level features. In order to distinguish moving targets, it is not enough to use the underlying features. Further extraction of structural features that reflect the target structure can improve the recognition and tracking of moving targets. In order to ensure the stability of the feature areas extracted from the three-channel most stable extremum region, this article proposes an improved algorithm of the three-channel most stable extremum region to improve the three-channel most stable extremum region. The algorithm can adaptively select the filters of each channel to filter the feature regions extracted from the three most stable extreme value regions. An action cycle is generally 30~50 frames, so it is faster and more advantageous to directly use the first 50 frames of video for processing. In this article, two feature representation methods of variance gait energy graph algorithm and image splitting algorithm are proposed. The variance energy graph algorithm significantly improves the recognition rate; image splitting enhances the robustness of behavior classification. This article proposes a feature representation algorithm of “distance from contour line to center line” to improve the recognition rate. By analyzing the feature extraction methods of principal component analysis, Fisher linear discriminant analysis and maximum divergence difference discriminant analysis, the main component discriminant analysis of row maximum divergence discriminant column and the two-dimensional two-dimensional maximum divergence discriminant analysis of row and column are proposed. This further enhances the ability to classify behaviors.


**INDEX TERMS** Human motion video, feature extraction, improved three-channel most stable extreme value area, virtual reality.

## I. INTRODUCTION

Virtual reality is generally an application that solves certain problems in a specific field. To solve these problems, not only need to understand the needs of the application, but also need to have a wealth of imagination, immerse people and acquire new knowledge, improve emotional and rational understanding, so as to deepen the concept and germinate new ideas [1], [2]. At present, there are many methods for detecting and tracking moving objects in video, and they have been well applied in specific applications, such as airborne visible light target tracking and infrared target tracking [3]. But these methods are not ideal in other applications. Especially in military applications, the scenes have complicated changes, such

as wild jungle backgrounds, desert backgrounds, urban block backgrounds, and sea and sky backgrounds under severe weather conditions [4]. The main reason for this difficulty is that the moving process of the moving target in the video is a nonlinear and non-Gaussian dynamic process. If the characteristic information of the moving target itself is not fully utilized, it is difficult to achieve effective and robust detection and tracking.

After years of research, the face recognition technology in video has made great progress and development [5]. With the development of video surveillance, information security, access control and other application fields, video-based face recognition has become one of the most active research directions in the field of face recognition [6]–[8]. “Video-image” face recognition refers to the use of face video as input (query) to use the still image face database for recognition or

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

verification [9]. Since most of the existing face databases are still image face databases, how to make full use of facial information in the video for better face recognition is an urgent problem to be solved at this stage [10], [11]. The traditional methods for solving such problems can be divided into two categories: one type of method tracks the face in the input video, finds the face image that meets certain rules (such as size, pose, clarity, etc.), and then uses the image based on the still image face recognition method [12]. The disadvantage of this type of method is that the rules are difficult to define, and the time and space continuous information in the face video is not used to the maximum [13]. Another method uses the spatial information in the video for face recognition [14]. By using a face recognition method based on still images for each face or several faces in the input video, various joint rules (such as majority voting or probability/distance accumulation methods) are used for final recognition [15], [16]. The disadvantage of this type of method is that the joint rules are often quite random. Relevant scholars use the time series model to describe the dynamic changes of the face, use the identity variables and motion vectors as state variables to introduce time and space information, and use the Sequential Importance Sampling (SIS) method to effectively estimate the combination of identity variables and motion vectors [17]. The probability distribution of identity variables is extracted by marginalization. Experiments prove the effectiveness of the algorithm, but the recognition rate is only 57% when the posture changes [18]. The reason why the recognition rate is low when the posture changes is because the use of time continuity is reflected in the consistent appearance of the face, and the change in lighting or posture will cause a significant difference in appearance [19]. Relevant scholars further proposed an adaptive appearance change model and adopted an adaptive motion model to more accurately deal with changes in posture, and updated the variance of the noise in the motion model and the number of particles in the sampling algorithm according to the error of the appearance model obtained by calculation [20]. Using the likelihood function based on Bayesian face recognition method to update the weight makes the whole algorithm more effective [21].

In the image sequence, because there are often translations and zooms of the moving human body, many recognition tasks must first eliminate these changes [22], [23]. Moments are an effective tool for handling translation, rotation, and zoom changes. In the early 1960s, the classic Hu moment was used in 2D image processing [24]. In recent years, moments have also been used in gait analysis and recognition [25]. Relevant scholars use Zernike velocity moments to describe motion [26]. However, because Hu moments and Zernike moments are all based on geometric moments, high-order moments are difficult to calculate and have poor stability. Therefore, only low-order moments are used for description and analysis, and it is difficult to characterize the subtle features between signals. The two-dimensional low-resolution identification is mostly based on the appearance method, which does not require high video quality, and the

algorithm is simple and easy to implement. However, changes in viewpoint, viewing angle, lighting, etc. must be considered. The model-based method assumes a priori model and matches the 2D image sequence with the model, which is often used for high-resolution image sequences [27], [28]. The researchers used the spatio-temporal slice method for human tracking [29]. First, they observe the spatiotemporal interweaving pattern generated by the trajectory of the human lower limbs, then locate the motion projection of the head in the spatiotemporal domain, then identify the trajectories of other joints, and finally use these joint trajectories to outline the contours of a pedestrian. Related scholars extract joint objects in the video sequence for recognition [30], [31]. The monotonic operator that calculates the displacement vector field initializes the human body model and extracts the outer contour lines of different body parts for tracking. To automatically locate different body parts, a monotone operator is used in the first two frames of the motion sequence, the purpose is to calculate the vector field displacement. They extract the contour lines of different parts to track the body movement, estimate the shape and position of a simple 3D block, and match the 2D contour line extracted from the image and the projection contour of the 3D block projected onto the image, thereby extracting the 3D shape and position of the body [32].

In this article, the basic gait energy map is improved, and the characterization methods of “variance energy map” and “image splitting” are proposed. The feature of “distance from contour line to center point” is improved to “distance from contour line to center line”, and the variance of the vertical distance of center point is also used as a feature vector to identify behavior. Specifically, the technical contributions of this article can be summarized as follows:

First: This article proposes an improved three-channel feature extraction algorithm for the most stable extreme value region. The algorithm uses support vector machines to train several indexes of the most stable extreme value region with good stability, and obtain filters with different precisions for each channel. A filter with adaptive precision selection is used to filter the feature regions extracted by each channel, so as to fully utilize the information of each channel of the color image while ensuring the quality of the feature regions extracted by the three channels.

Second: Two dimensionality reduction methods are proposed. One is the dimensionality reduction algorithm that extracts the maximum divergence difference identification information in the column direction and then performs the row direction principal component global feature extraction, and the other is the dimensionality reduction method for row and column two-dimensional two-dimensional maximum divergence difference identification.

The rest of this article is organized as follows. Section 2 discusses the related work of human motion video feature extraction under virtual reality technology. Section 3 presents the improved three-channel most stable extreme value region feature extraction algorithm.

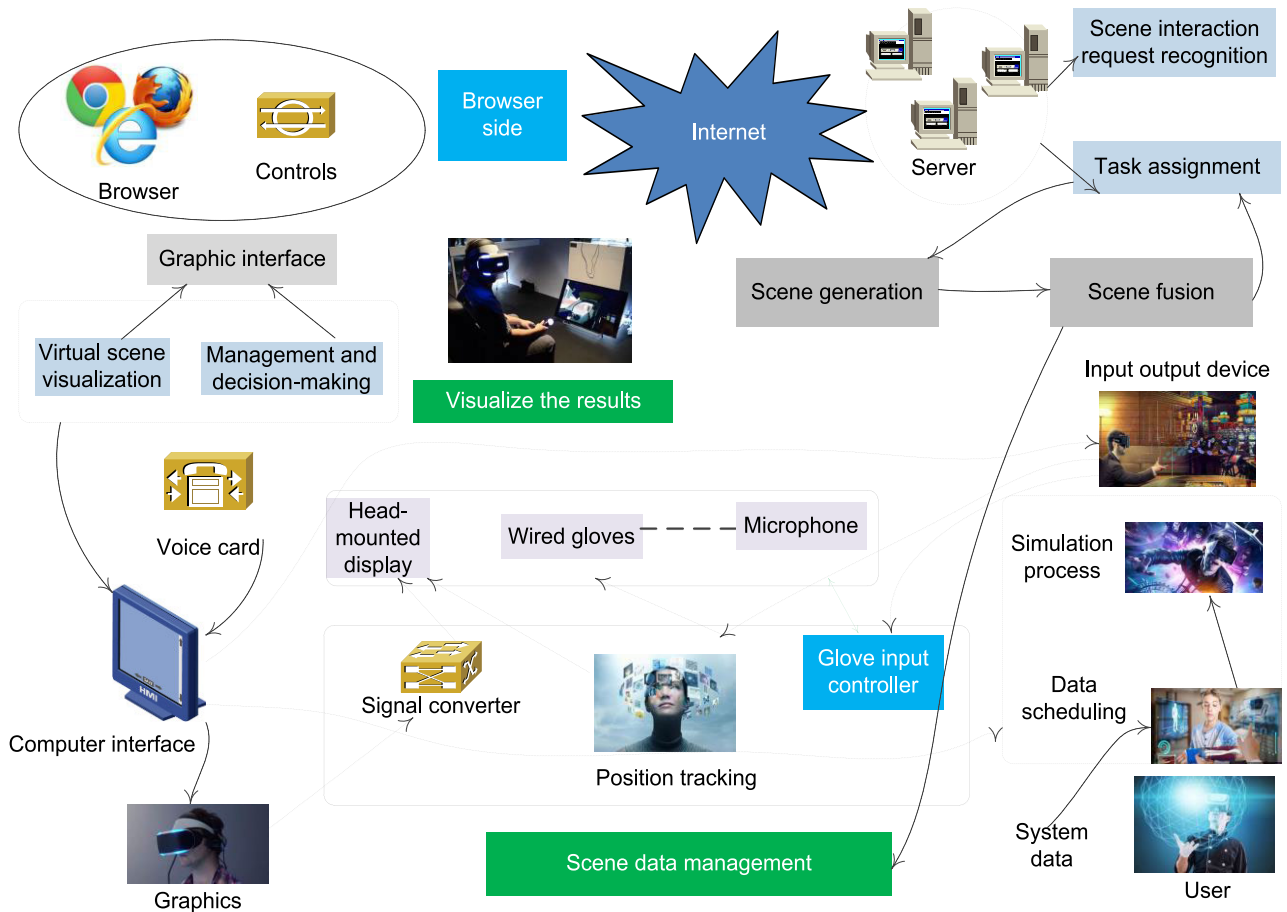


FIGURE 1. Architecture of virtual reality system.

Section 4 conducted simulation experiments and analyzed the results. Section 5 summarizes the full text.

## II. VIRTUAL REALITY TECHNOLOGY RELATED WORK OF HUMAN MOTION VIDEO FEATURE EXTRACTION

### A. VISION-BASED HUMAN-COMPUTER INTERACTION TECHNOLOGY OF VIRTUAL REALITY

The key to the human motion video interaction problem is to solve the highly realistic hand-computer interaction technology in the virtual reality system. According to the analysis of the characteristics of the virtual environment of human motion video, the virtual reality environment is constructed according to the principle of “seeing is virtual and touching is real”. The core of the human-computer interaction problem of human motion video is the non-contact accurate positioning of the fingers, which achieves the consistency of the operation of the real hand and the virtual hand in the virtual scene. The architecture of the virtual reality system is shown in Figure 1.

There are many kinds of human-machine interactive positioning technologies for virtual reality systems, such as the hand interactive positioning methods used by data gloves and 3D tracking locators, computer vision gesture recognition and positioning. However, there is still a major problem that it

cannot provide accurate and realistic finger force, palm force, arm force feedback and tactile sensation while accurately locating finger and other hand position information. For example, a data glove with force feedback can provide partial force feedback and more accurate hand positioning, but it still cannot provide a satisfactory feedback effect of coordinated finger force, palm force, and arm force. At the same time, due to the isolation of the glove material, it affects the realistic touch, and may cause collision in a small space. Computer vision gesture recognition can make the computer understand sign language, but it cannot provide force feedback and touch.

Humans mainly obtain information from the outside world through vision. We hope that computer vision can play an important role in human-computer interaction technology. The vision-based interactive technology requires only simple contactless peripherals, is inexpensive, and does not increase user burden. It conforms to human natural operating habits and makes the entire system more immersive. Therefore, based on the existing computer vision positioning technology, this subject builds a semi-virtual reality cockpit based on the actual interaction characteristics of human motion video. Among them, all the visual scenes inside and outside the cabin are provided by virtual reality, and the operator’s hand in the visual scene is also generated by a computer model,

called a virtual hand; force feedback and tactile sensation are provided by physical objects. In the end, what people see is the virtual dashboard and virtual hands in the virtual scene. The actual hands are the real hands, and the virtual hands and the real hands are unified, that is, the virtual and the real are consistent.

### **B. FEATURE EXTRACTION AND DESCRIPTION IN BEHAVIOR RECOGNITION**

The trajectory is used as the moving target carrier, and the feature extraction and description of the trajectory are used as the moving target. The behavioral feature is different from other types of behavioral feature description methods. That is, the video area mostly uses the distant view, and the trajectory extraction object is mostly the moving target centroid, so it is separate. The use of trajectories for behavior feature extraction and expression is mostly a macro representation, and often does not pay attention to local micro behavior changes. The description and representation of the trajectory can be divided into two categories, namely, quantitative feature description and qualitative feature description. Quantitative description refers to the use of accurate quantitative description and representation of the trajectory shape, mutual relationship, etc. It is a microscopic description and representation form. Qualitative description refers to the trajectory attributes and change trends. It is different from quantitative description and adopts precise quantitative relationship. It is a macro description representation form.

Among the data-driven feature extraction and description methods, deep learning is one of the most representative models. It has become one of the popular mainstream research directions of machine learning, and has even surpassed the tradition in many application fields of computer vision. The advantage of deep learning over traditional machine learning methods is that you can learn features directly from the original data without manual intervention for feature detection and description. Deep learning models can be roughly divided into two categories, namely, generation models and discriminant/supervised models. The following deep learning model will be of great significance for further research and promotion of deep learning. The reason is that the subject's knowledge of the object is through observation rather than being informed, and most people's learning is unsupervised.

The middle-level feature is the transitional form between the image presentation form with no semantic randomness in the bottom layer.

### **C. VIDEO CLUSTERING AND SELECTION OF REPRESENTATIVE FRAMES**

After the video is divided into shots, it is necessary to perform feature extraction on each shot to obtain a feature space that fully reflects the content of the shot as much as possible. This feature space will be used as the basis for video clustering and retrieval. The characteristics of video data are divided into static characteristics and dynamic characteristics.

The extraction of static features is mainly for representative frames. The representative frame is a key image frame used to describe a shot. It reflects the main content of a shot. On the one hand, the selection of the representative frame must reflect the main events in the shot. On the other hand, in order to facilitate management, the amount of data should be as small as possible, and the calculation should not be too complicated.

The frame averaging method is to take the average value of the pixel values of all frames at a certain position from the shot, and then take the frame with the pixel value closest to the average value at that point in the shot as the representative frame. The histogram averaging rule is to average the statistical histograms of all frames in the shot, and then select the frame closest to the average histogram as the representative frame. For shots with a lot of motion, if one or two representative frames cannot be fully described, multiple representative frames can be selected according to the significant changes between frames. The schematic diagram of human motion video tracking structure is shown in Figure 2.

### **D. EXTRACTION AND EXPRESSION OF IMAGE FEATURES**

For a particular image feature, there are usually many different expression methods. Due to the vast differences in people's subjective understanding, there is no so-called optimal expression for a certain feature. In fact, different expressions of image features characterize some of the features from various angles.

#### **1) TEXTURE FEATURE EXTRACTION AND EXPRESSION**

The method of expressing texture features by the co-occurrence matrix has studied the spatial dependence of gray levels in image texture from a mathematical point of view. It first builds a symbiotic matrix based on the directivity and distance between pixels, and then extracts meaningful statistics from the matrix as texture features. The Gabor filtering method can minimize the uncertainty of space and frequency, and can also detect edges and lines in different directions and angles in the image.

To calculate the color histogram, the color space needs to be divided into several small color intervals, each of which becomes a bin of the histogram. This process is called color quantization. The color histogram can be obtained by calculating the number of pixels where the color falls in each cell. The most common method of color quantization is to evenly divide the components (dimensions) of the color space. In contrast, the clustering algorithm will take into account the distribution of image color features in the entire space, so as to avoid the situation where the number of pixels in some bins is very sparse and make quantization more effective. In addition, if the image is in RGB format and the histogram is in HSV space, we can build a lookup table from the quantized RGB space to the quantized HSV space in advance to speed up the calculation of the histogram.



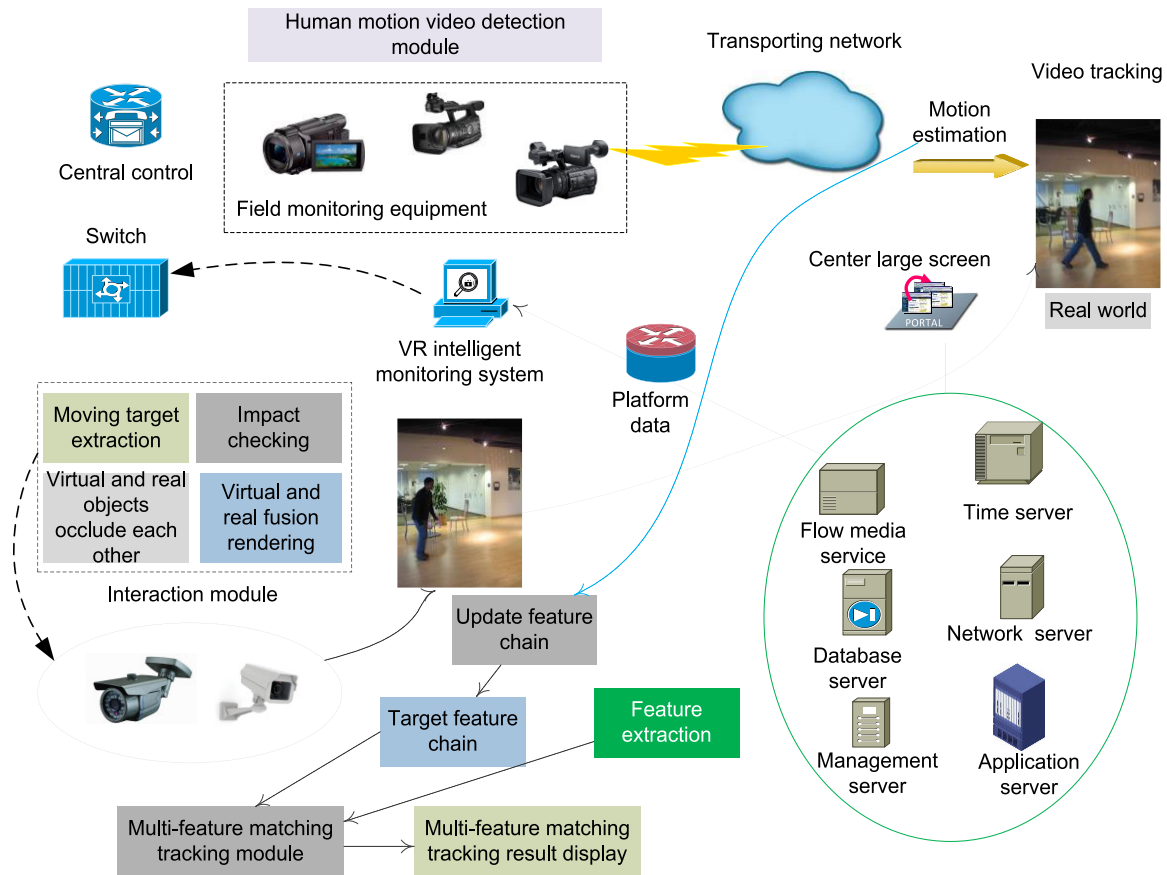


FIGURE 2. Schematic diagram of human body motion video tracking.

## 2) SPATIAL RELATIONSHIP FEATURE EXTRACTION AND EXPRESSION

In order to avoid the difficulty of accurate image automatic segmentation, and at the same time provide some basic information about the spatial relationship of the image area, a compromise method is to divide the image into several sub-blocks in advance, and then extract the various features of each sub-block separately. In retrieval, we first calculate the similarity between the corresponding sub-blocks of the image according to the features, and then calculate the total similarity by weighting. A similar method is the quadtree method, which regards the entire graph as the structure of the quadtree, and each branch has a histogram to describe the color characteristics. This method can support the retrieval of the spatial relationship of objects. Although this kind of method is very simple in concept, this kind of ordinary regular partitioning cannot accurately give the information of local colors, and the calculation and storage are relatively expensive.

### E. DYNAMIC CHARACTERISTICS OF VIDEO IN VIRTUAL REALITY

Most of the videos are made by connecting many shots, some videos switch frequently and the duration of shots is short,

such as TV news programs and feature films. There are few video switches, and the duration of each shot is longer, such as the broadcast of sports programs, and the surveillance videos used for bank security and traffic supervision have almost no switch of lenses. For these videos, people mainly care about the objects in the lens exercise. The coordinate system in the virtual reality system is shown in Figure 3.

The motion within the lens includes local motion caused by object motion and global motion caused by camera motion. The movement of the camera will often have a global impact on the video image. For example, moving the lens horizontally will cause all pixels to move horizontally. Lengthening the focal length will cause the pixels to diverge from the center to the surroundings, and shortening the focal length will cause the pixels to move from the surroundings. When only the object is moving, most of the background pixels remain unchanged, but only the moving object and the part that is occluded change.

During the video shooting process, the camera can move in different ways to achieve a specific shooting effect. Camera movements include panning, rotating, moving, pushing and pulling. The movement of objects varies with the actual situation, but it is also an important aspect of video retrieval, especially for surveillance video. For example, the user may need to retrieve a video clip of an object being moved

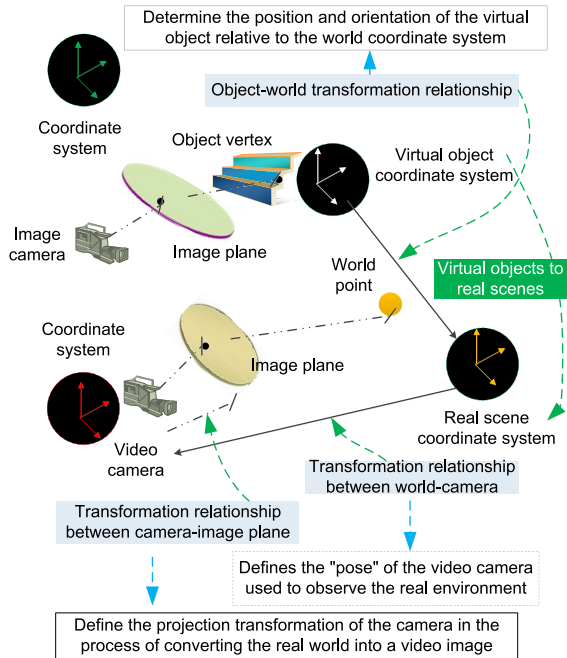


FIGURE 3. Coordinate system in virtual reality system.

or a video clip of a car starting. According to the object motion characteristics, some object motion events can be defined.

### III. IMPROVED THREE-CHANNEL MOST STABLE EXTREME VALUE REGION FEATURE EXTRACTION ALGORITHM

#### A. IMPROVED THREE-CHANNEL FEATURE EXTRACTION OF THE MOST STABLE EXTREME VALUE REGION

To solve the problem that the three-channel most stable extreme value region can increase the number of extracted feature regions when performing feature extraction, but cannot guarantee the quality of these extracted features, this article proposes an improved three-channel most stable extreme value region method. The fundamental strategy of this method is to filter the feature regions extracted from the three-channel most stable extreme value regions among the three channels.

The training of the three-channel filter is a process of machine learning using support vector machines. We use some feature areas in the image related to the image to be processed for consistency comparison, and obtain the consistency area that meets the overlap error standard. Then, the threshold of overlap error can be set according to the need to divide these feature areas into two categories of good quality and bad quality. Then we use the stability parameters of the feature area (average edge strength, squareness, aspect ratio and circularity of the smallest circumscribed rectangle) as training features, and perform the classifier (filter) training. After the filter is trained, the most stable extreme value region extracted in the three spaces of the image to be processed

can be predicted using the filter, so as to filter out the feature regions with better quality.

The stability of the feature areas extracted by the three channels fluctuates greatly with the change of the image type. Therefore, we propose a strategy for adaptive selection of filter accuracy for each channel. Adaptively using different precision filters in each channel is the most stable. The extreme value region  $s$  is filtered, so that we can obtain more stable extreme value regions under the premise of ensuring the quality of the characteristic region.

The improved three-channel method of the most stable extreme value region can be divided into three steps:

First, you train various precision filters through training images related to the image to be processed;

Second, you perform the extraction of the most stable extreme value region of the three channels;

Third, the adaptive selection filter in each channel filters the most stable extreme value region extracted by each channel, and finally merges the information of the three channels.

#### B. MODELS OF MACHINE LEARNING PROBLEMS

Machine learning is the process by which any computer program obtains experience when completing a task. This experience continuously improves its performance. Feature extraction and classification of human motion video based on machine learning methods are shown in Figure 4.

The condition of traditional statistical pattern recognition problem research is that the number of data sets is large enough. However, in practical applications, the number of data sets is usually limited, and the statistical learning law under the small sample data represented by SVM can be very good solution to such problems. In this article, the filters in the three channels of the three-channel most stable extreme value region are trained by the SVM method.

The specific solution is to construct an optimization problem under constraints, specifically a restricted quadratic programming problem, solve the problem, and obtain a classifier.

#### C. FEATURES FOR LEARNING

A stable and most stable extreme value region has better repeatability and accuracy in matching. In order to predict the repeatability and accuracy of the most stable extreme value region, in the SVM training stage, this article introduces average edge strength, rectangularity, aspect ratio and circularity of the smallest circumscribed rectangle (minimum circumscribed rectangle) to train each space filter.

Here we use the Sobel operator to calculate the intensity of each point on the edge of the most stable extreme region. The Sobel operator uses two different convolution kernels in the  $x$  and  $y$  directions.

If the convolution pixel values of a pixel obtained by using the convolution kernels in the  $x$  and  $y$  directions are  $x$  and  $y$ , respectively, the calculation formula of the gradient value  $t$  of the pixel is:

$$t = \sqrt{x^2 + y^2} \tag{1}$$

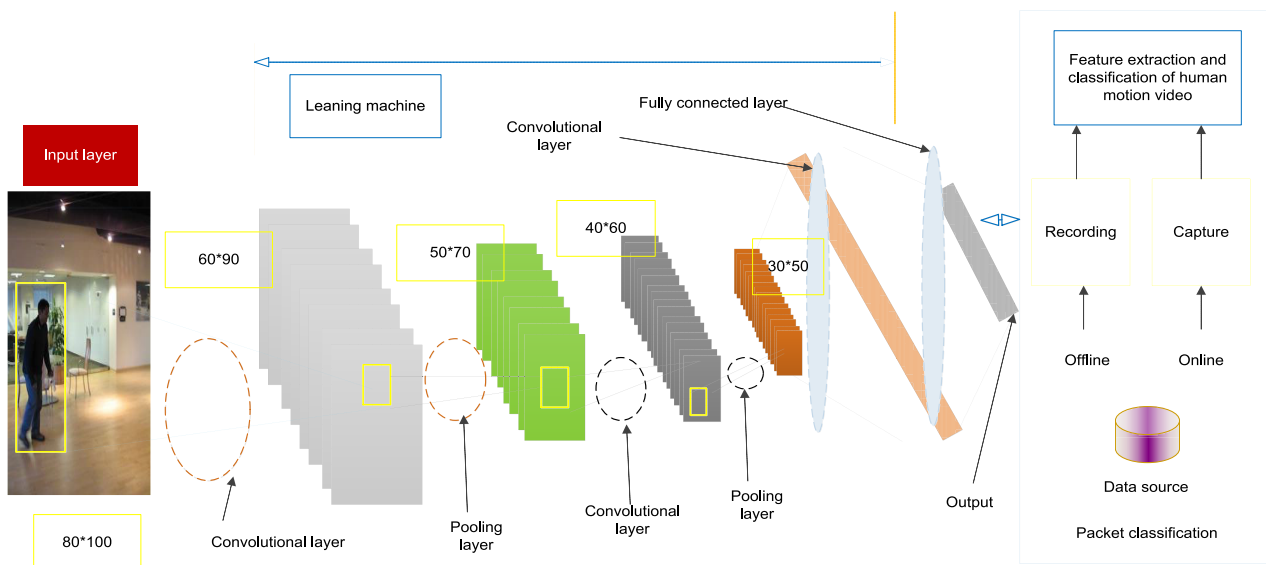


FIGURE 4. Feature extraction and classification of human motion video based on machine learning methods.

Then the formula for calculating the average edge strength is:

$$\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i \quad (2)$$

where  $n$  is the number of points on the edge of the most stable extreme region and  $t_i$  is the gradient value of a point on the edge of the most stable extreme region.

A parameter that reflects the rectangularity of an object is the rectangular fitting factor:

$$R = S_0/S_R \quad (3)$$

Among them,  $S_0$  is the area of the object, and  $S_R$  is the area of its smallest circumscribed rectangle (smallest circumscribed rectangle).  $R$  reflects the fullness of an object to its smallest circumscribed rectangle.

#### D. FILTER TRAINING

The three-channel filter training process in this article is a process of using SVM for machine learning. By learning feature regions with better stability, a filter (classifier) with distinguishing stability of feature regions is formed.

Under ideal conditions, the two images that have undergone some changes, following the principle of affine changes, the shapes of the two regions with the same content should coincide after the image is changed and automatically fitted. However, due to the different algorithm principles of the various detectors, the accuracy of the description of the same area in the two figures is different. In this way, when the two areas overlap after automatic fitting, the coincidence of the two areas will also be different. The more accurate the two regions described by the detector for the region coincide, the higher the degree of coincidence. Therefore, this article uses

overlap error to perform consistency detection to judge the performance of the extracted feature regions. When the overlap error between the two regions is small enough, the two regions are consistent.

Since the size of the measurement area is arbitrary, we can get different overlap errors by scaling the measurement area size of the two feature areas that will be corresponding to the overlap error judgment. Relatively large measurement areas have a higher degree of overlap. It can be understood that, assuming that the salient area is an ellipse, the measurement area is also an ellipse and its center is in the salient area. Because the size of the measurement area is arbitrary, it can be scaled at any scale. For a significant area, changing the scale of the measurement area can determine a cone. The two salient regions that are related in the two images can identify two such cones.

As shown in Figure 5, the three-channel most stable extreme value region is improved. First, the images used for training are randomly paired to form an image pair, and the consistency check of the feature region is performed to find the most stable extreme value region that meets the consistency requirements. Then we set a threshold of overlap error according to our needs (the accuracy of the filter), classify the most stable extreme value region on the match. We divide the most stable extreme value region into one category. Finally, we take the average edge strength, rectangularity, aspect ratio and circularity of the smallest circumscribed rectangle of the most stable extreme value region as the training labels, and carry out LIBSVM training to obtain model\_file as a filter.

Changing the overlap error threshold can train filters with different accuracy. The flow of the filter shows that lowering the overlap error threshold can result in a filter with higher accuracy. This article uses 20%, 30%, 40%, and 50% overlap error thresholds for filter training.

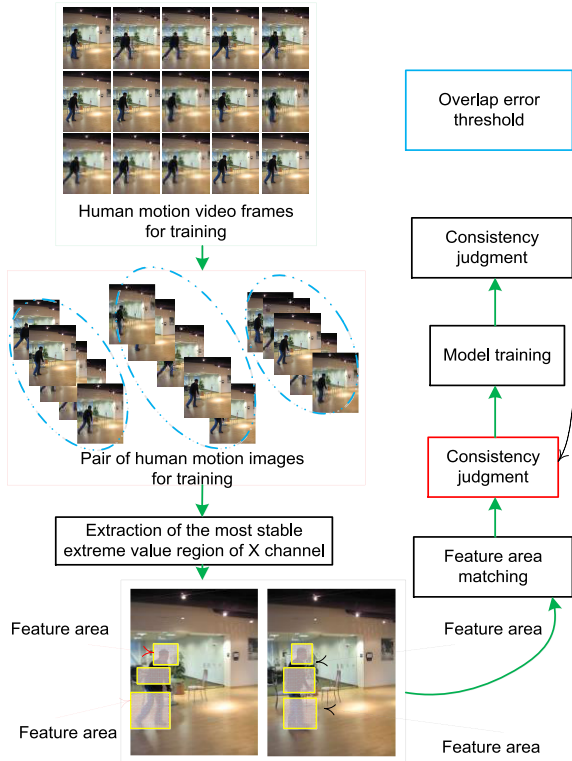


FIGURE 5. Filter training flowchart.

**E. ADAPTIVE SELECTION OF FILTERS**

Due to the characteristics of each spatial information, the stability of the most stable extreme value regions extracted from different channels is not the same. Generally speaking, the most stable extreme value region of the I channel is more stable than the most stable extreme value region extracted from the other two channels. This is also one of the reasons why the traditional most stable extreme value region is only extracted in the I channel. Therefore, in order to ensure the stability of the most stable extreme value region extracted by each channel, we will adaptively select each channel filter to filter the three-channel feature region.

The filter can filter out the unstable most stable extreme value region extracted from each channel, but because the stability of the most stable extreme value region extracted from each channel is different, in order to ensure the most stable extreme value extracted from each channel, the value area can meet our stability requirements. We adopt an adaptive method to adaptively select different precision filters in the three channels of H, S, and I to filter the most stable extreme value area. In this article, the repeatability of each algorithm is used as the evaluation index of the overall stability of the extracted feature area. When calculating the repeatability of an image pair, we only consider the feature regions detected from the regions representing the same scene in the two images.

Before the filter selection, we used the training group image to perform 20%, 30%, 40%, and 50% precision filters

on the group according to the method of filter training. Then, we will use the evaluation group images from the same sample channel to select the filter. First, the image sequence is randomly combined in pairs to form an image pair. Then we divide these image pairs into two paths for processing. After extracting the most stable extreme value regions all the way, the region consistency judgment is performed to calculate the number and repeatability of the most stable extreme value regions with consistency. The other way is to extract the most stable extreme value region on the H channel, filter through different overlapping error accuracy filters, and then perform consistency judgment to obtain the most stable extreme value region that can be judged to be consistent under each overlapping error accuracy. The processing results of the two channels are compared, and the filter whose corresponding repeatability value is greater than or equal to the corresponding value of the first channel, which is selected from the second channel. Among these selected filters with different overlapping error accuracy training, the corresponding number of consistent most stable extreme value regions is the largest, that is, we are looking for the H channel filter. If no suitable filter is found in the H channel according to this method, then we give up the most stable extreme value region of this channel.

**IV. SIMULATION EXPERIMENT AND ANALYSIS**

**A. EFFECT OF VIDEO EXTRACTION METHOD ON RECOGNITION RATE**

The purpose of this experiment is to verify the effect of video extraction on recognition rate. We use two methods for video extraction, one is to first judge the cycle, and then select all the videos in a cycle as the original data. Another method is to directly select the first 50 frames of the video sequence for processing without calculating the period. The benefits of doing so can save the trouble of period judgment. Taking the shooting speed of 25 frames/second as an example, generally, a behavior period is roughly between 30-50 frames, so selecting 50 frames can meet the requirements of one period without generating redundant information. Figure 6 shows the human motion recognition of 10 frames in the video.

Because it is only to check the influence of the extraction method on the recognition rate, the basic gait energy map is selected as the feature. The recognition rate of principal component analysis and its derived algorithms is shown in Figure 7. The two-dimensional Fisher linear discrimination algorithm is shown in Figure 8. Figure 8 shows a graph of the recognition rate as the feature vector dimension increases.

It can be seen from Figure 7 and Figure 8 that the recognition rate of directly extracting 50 frames of video is significantly better than that of extracting only one cycle. The reason is that the 50-frame video provides more and richer information, and also saves the tedious calculation of the cycle. Therefore, the following experiments all directly extract 50 frames of video as the initial input, and no longer consider the problem of periodic extraction.



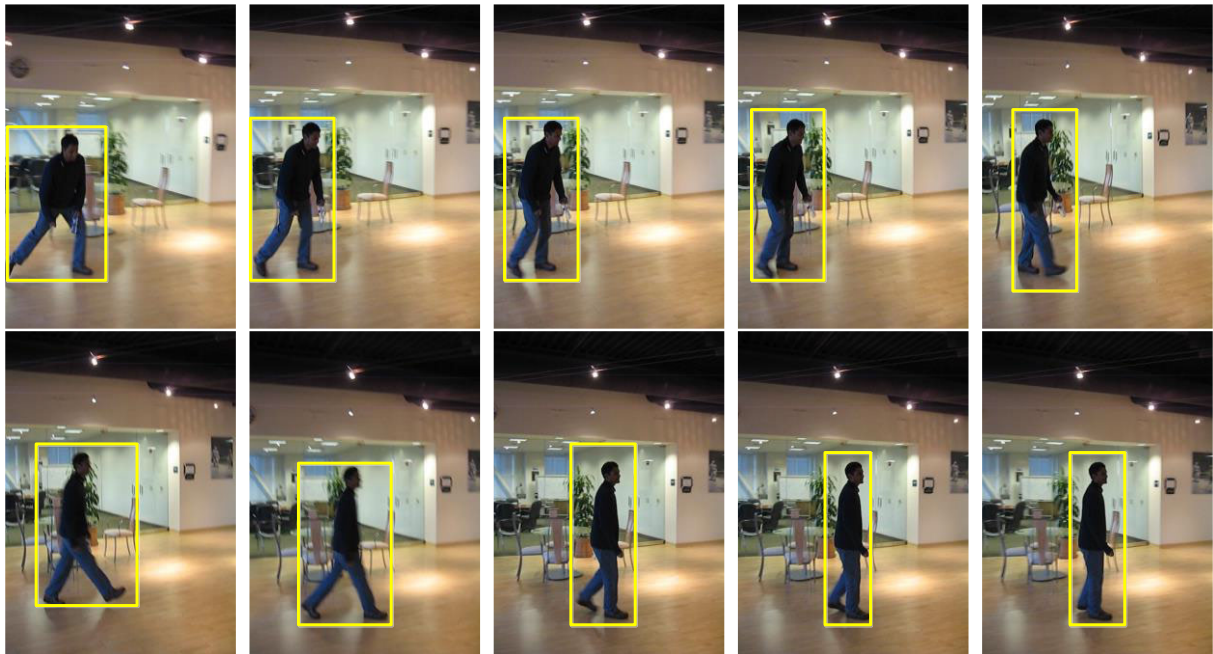


FIGURE 6. 10 frames of human motion recognition in the video.

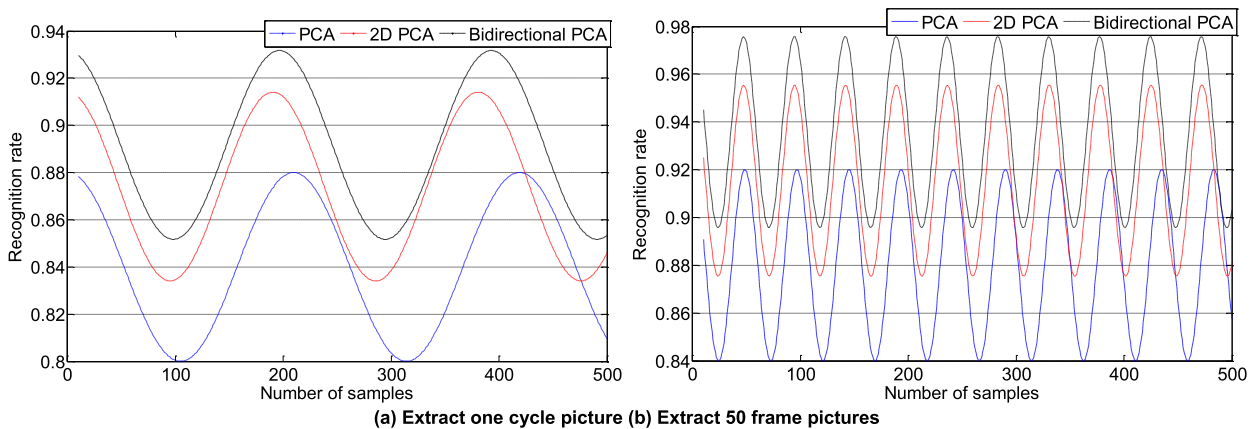


FIGURE 7. The effect of video extraction on the recognition rate in principal component analysis.

**B. RECOGNITION RATE OF VARIANCE ENERGY MAP IN VARIOUS DIMENSIONALITY REDUCTION METHODS**

1) Comparison of recognition rate between variance energy map and basic gait energy map

We still take principal component analysis and Fisher linear identification as examples, using the basic energy map and the variance energy map as input features to examine the impact of the difference energy map on the recognition rate. Figure 9 and Figure 10 are the recognition rates of the two energy maps when using principal component analysis and its derivative algorithms to examine the sample changes. It can be seen that when the number of training samples increases, the recognition rate changes, but the recognition rate of the variance energy map is generally higher than the basic gait

energy map. The advantage of variance is that it reflects the distribution information of the target. Each target is different, its variance is different, so the variance contains a lot of identification information. The basic energy map contains more energy shared by various behaviors, and the energy is larger, which masks the identification information with smaller energy. Therefore, the variance energy map eliminates the shared energy of various behaviors, and reflects the identification information.

Figure 11 shows the recognition rate of two energy maps when using 2D Fisher linear discriminant analysis algorithm to examine the change of feature vector number. It can be seen from this that when the feature vectors choose different dimensions, the recognition rate of the variance

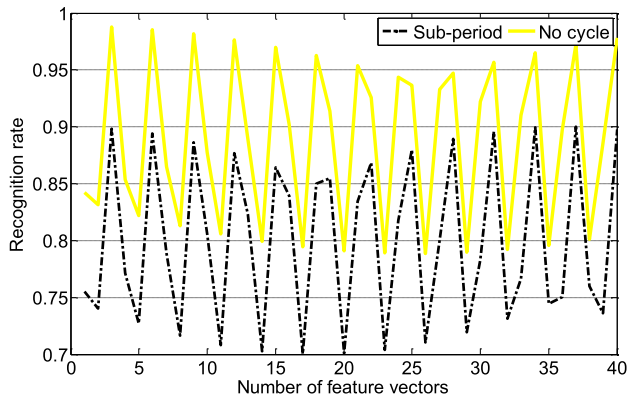


FIGURE 8. The effect of video extraction on the recognition rate under 2D Fisher linear discrimination analysis.

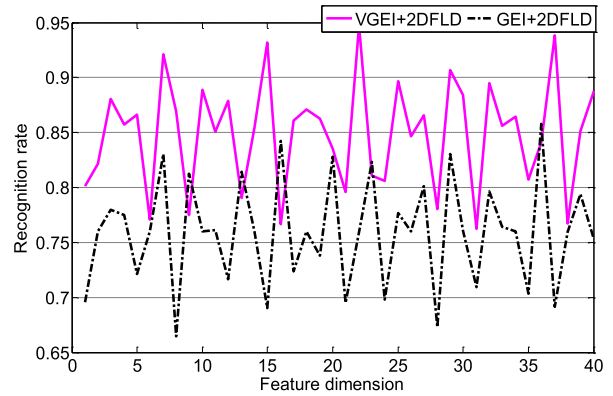


FIGURE 11. Recognition rate of two energy maps under 2D Fisher linear discriminant analysis.

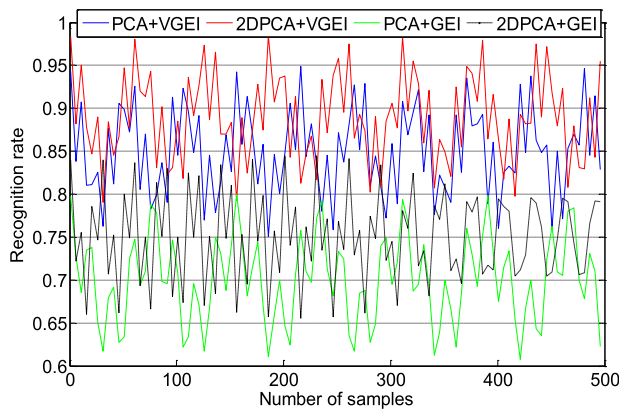


FIGURE 9. Comparison of recognition rate between variance energy map and basic energy map.

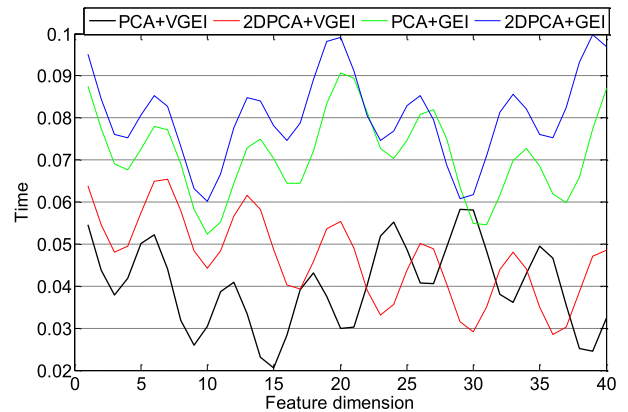


FIGURE 12. Comparison of time used by different algorithms.

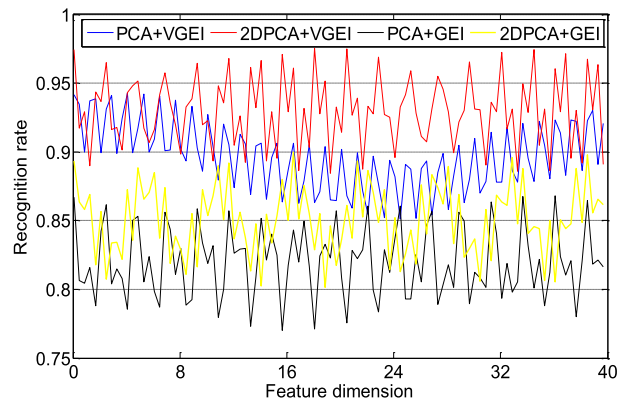


FIGURE 10. Comparison of recognition rates of different algorithms.

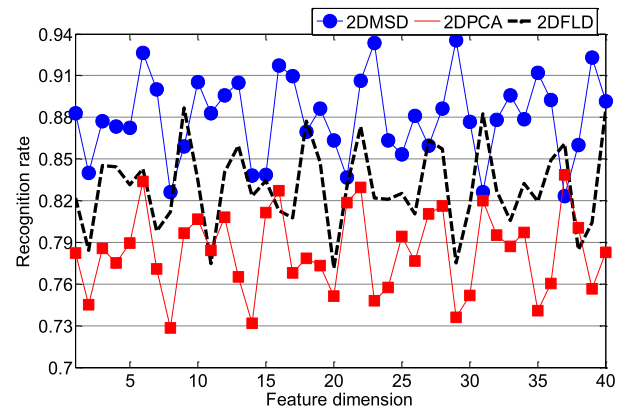


FIGURE 13. Recognition rate of different algorithms with the change of dimension.

energy map is generally higher than that of the basic energy map.

As can be seen from Figure 12, the variance energy map generally takes less time to train, because the data is more concise.

2) Using the variance energy map as input, examine the recognition rates of 2D principal components, 2D Fisher

linear discriminant analysis and 2D maximum divergence discriminant identification

It can be seen from Figure 13 that the 2D maximum divergence difference discrimination algorithm has advantages in terms of recognition rate and robustness, while the 2D Fisher linear discriminant analysis is less robust. The average recognition rates of 2D principal component, 2D Fisher

linear discriminant analysis and 2D maximum divergence difference discrimination are 0.77, 0.84 and 0.89 respectively. Their highest recognition rates are 0.84, 0.885 and 0.94, respectively.

3) Investigate 2D2 principal component maximum divergence difference identification, row maximum divergence difference identification column principal component identification analysis, row and column two-dimensional maximum divergence difference identification analysis, and 2D2 principal component identification rate

In this article, we consider the bidirectional feature extraction of images. Here, the 2D2 principal component maximum divergence difference discrimination is to perform the column principal component transformation before the row maximum divergence difference discrimination transformation; the row maximum divergence difference discrimination column principal component discrimination analysis is to perform the image column maximum divergence difference discrimination; row-column two-dimensional maximum divergence difference discrimination analysis is to perform a maximum divergence difference discrimination transformation on both rows and columns; 2D2 principal component is to perform principal component transformation on both rows and columns. Figure 14 shows the relationship between feature vector dimension and recognition rate of various algorithms. The abscissa dimension in the figure refers to the feature vector dimension after row-column transformation is  $d$  dimension, that is, the image feature information is  $d \times d$  dimension. It can be seen from the figure that the two-dimensional two-dimensional maximum divergence difference discriminant analysis algorithm has the highest recognition rate, followed by the principal component discriminant analysis of the maximum divergence discriminator column. Comparing the principal component discriminant analysis of the row maximum divergence difference discrimination column with the 2D2 principal component maximum divergence difference discrimination, it can be seen that the former has a better recognition rate than the latter, because the maximum divergence difference discrimination compression

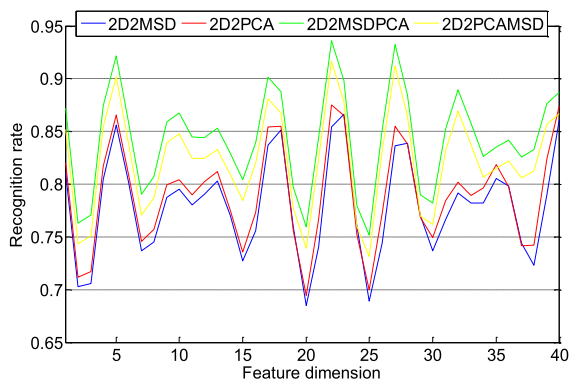


FIGURE 14. The recognition rate of several algorithms with the change of dimension.

is first, which is beneficial to the discrimination information extraction.

C. RECOGNITION RATE OF IMAGE SPLITTING IN VARIOUS DIMENSIONALITY REDUCTION METHODS

Figure 15 shows the experimental results of dimensionality reduction and feature extraction using 2D principal components and 2D2 principal components. It can be seen that the recognition rate after splitting is much higher than that before splitting, and it can be seen that splitting helps to improve the recognition rate.

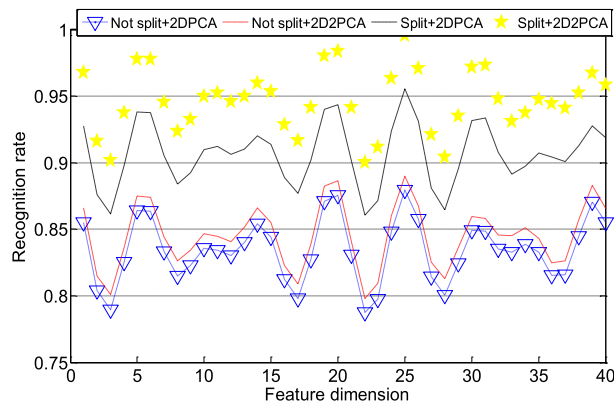


FIGURE 15. Comparison of recognition rate between image split and non-split.

D. RECOGNITION RATE OF BEHAVIOR BASED ON CONTOUR FEATURES IN VARIOUS DIMENSIONALITY REDUCTION METHODS

The purpose of this experiment is to investigate the recognition rate of the feature extraction of “Distance of Silhouette to Central Line (DSCL)” and “Distance of Silhouette to Central Point (DSCP)”. As can be seen from Figure 16, the recognition rate of DSCL is significantly better than that of DSCP.

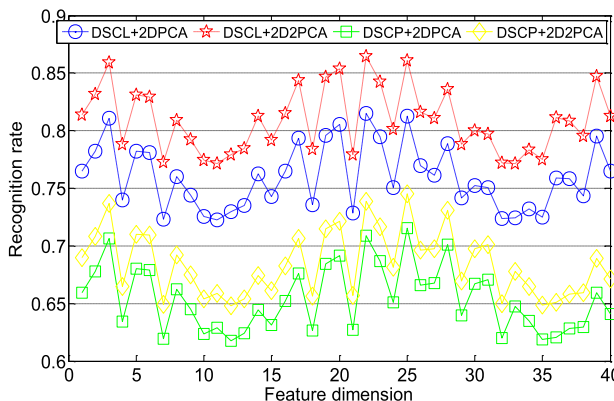


FIGURE 16. Comparison of DSCL and DSCP recognition rates.

We examine the “distance from contour to centerline” and the recognition rate of the gait energy map. As can be seen from Figure 17, although the feature extraction method of

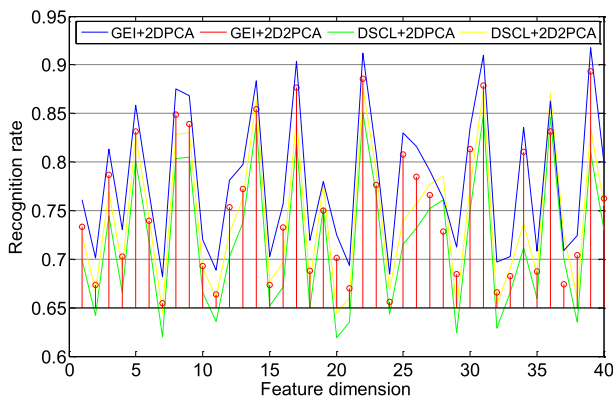


FIGURE 17. Comparison of recognition rates between DSCL and GEI.

DSCL is better than DSCP, compared with the basic gait energy map, the recognition rate is basically the same, which is obviously not as good as the enhanced energy map. The main reason is that line features provide less information than area features.

## V. CONCLUSION

Starting from solving the problem that the simple three-channel most stable extreme value region cannot guarantee the quality of the characteristic region, an improved three-channel most stable extreme value region algorithm is proposed. The algorithm uses the feature regions with good stability as training samples to train the classifier (filter). Through the filtering of each channel filter, we can anticipate the stability of the feature regions extracted from the image to be processed, thereby eliminating the stability. Since the stability of the feature regions extracted by different channels in different changes is different, we cannot use one standard to filter in three channels. In this article, the purpose of obtaining the most feature regions under the premise of ensuring regional stability is to use adaptive methods to select the filter of each channel, and then filter out the unstable feature regions in each channel. By adopting this strategy, the three-channel most stable extreme value region can be improved, while maintaining the quantity advantage of the three-channel most stable extreme value region, and ensuring the quality of the characteristic region. In terms of characterization, based on the original gait energy map, a variance energy map and image splitting algorithm are proposed. Variance energy maps have significantly improved recognition rates compared to the original gait energy maps; image splitting has better robustness on shadows and foot occlusion. It also proposes to use “distance from contour line to center line” as information feature for behavior classification, and the recognition rate of this algorithm is much higher than “distance from contour line to center point”. Several algorithms for linear manifold learning are studied, and a dimensionality reduction method is proposed for discriminating the principal component discrimination analysis of the row maximum divergence

difference and the two-dimensional two-dimensional maximum divergence difference discrimination analysis of the row and column. They not only overcome the small sample problem of linear discriminant analysis, but also enhance the algorithm’s behavior classification ability and robustness, and the highest recognition rate is close to 100%.

Through in-depth research, it is found that there are many commonalities between existing motion estimation methods. For example, the premises of the block matching method and the optical flow method are very similar; some energy-based methods are equivalent to the area matching technology; and the phase method uses the phase gradient for the calculation of the normal velocity. These phenomena are not accidental, they imply that the existing various motion vector estimation methods can basically be unified in a framework, and the research of a unified framework is another development trend in this field.

## REFERENCES

- [1] C. Jiang, H. Yin, F. Yang, and X. Jiang, “Application of 3-D sensor tracking imaging in detailed feature extraction of motion damage action,” *J. Med. Imag. Health Informat.*, vol. 10, no. 4, pp. 842–846, Apr. 2020.
- [2] A. Vouliodimos, I. Rallis, and N. Doulamis, “Physics-based keyframe selection for human motion summarization,” *Multimedia Tools Appl.*, vol. 79, nos. 5–6, pp. 3243–3259, Dec. 2018.
- [3] Z. Liu, C. Zhang, and Y. Tian, “3D-based deep convolutional neural network for action recognition with depth sequences,” *Image Vis. Comput.*, vol. 55, pp. 93–100, Nov. 2016.
- [4] F. Han, B. Reily, W. Hoff, and H. Zhang, “Space-time representation of people based on 3D skeletal data: A review,” *Comput. Vis. Image Understand.*, vol. 158, pp. 85–105, May 2017.
- [5] V. Bloom, V. Argyriou, and D. Makris, “Hierarchical transfer learning for online recognition of compound actions,” *Comput. Vis. Image Understand.*, vol. 144, pp. 62–72, Mar. 2016.
- [6] K. P. Sanal Kumar and R. Bhavani, “Human activity recognition in egocentric video using HOG, GiST and color features,” *Multimedia Tools Appl.*, vol. 79, nos. 5–6, pp. 3543–3559, May 2018.
- [7] T.-F. Su, C.-K. Chiang, and S.-H. Lai, “A multiattribute sparse coding approach for action recognition from a single unknown viewpoint,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 8, pp. 1476–1489, Aug. 2016.
- [8] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, “A new hybrid deep learning model for human action recognition,” *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 4, pp. 447–453, May 2020.
- [9] X. Ji, J. Cheng, W. Feng, and D. Tao, “Skeleton embedded motion body partition for human action recognition using depth sequences,” *Signal Process.*, vol. 143, pp. 56–68, Feb. 2018.
- [10] G. S. Taylor and J. S. Barnett, “Evaluation of wearable simulation interface for military training,” *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 55, no. 3, pp. 672–690, Jun. 2013.
- [11] Z. Merchant, E. T. Goetz, L. Cifuentes, W. Keeney-Kennicutt, and T. J. Davis, “Effectiveness of virtual reality-based instruction on students’ learning outcomes in K-12 and higher education: A meta-analysis,” *Comput. Educ.*, vol. 70, pp. 29–40, Jan. 2014.
- [12] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, “Deep learning models for real-time human activity recognition with smartphones,” *Mobile Netw. Appl.*, vol. 25, no. 2, pp. 743–755, Dec. 2019.
- [13] J. Imran and B. Raman, “Deep motion templates and extreme learning machine for sign language recognition,” *Vis. Comput.*, vol. 36, no. 6, pp. 1233–1246, Jul. 2019.
- [14] A. Dilawari, M. U. G. Khan, Z. ur Rehman, K. M. Awan, I. Mehmood, and S. Rho, “Toward generating human-centered video annotations,” *Circuits, Syst., Signal Process.*, vol. 39, no. 2, pp. 857–883, May 2019.
- [15] M. Sharif, M. A. Khan, F. Zahid, J. H. Shah, and T. Akram, “Human action recognition: A framework of statistical weighted segmentation and rank correlation-based selection,” *Pattern Anal. Appl.*, vol. 23, no. 1, pp. 281–294, Feb. 2019.



- [16] A. Zare, H. Abrishami Moghaddam, and A. Sharifi, "Video spatiotemporal mapping for human action recognition by convolutional neural network," *Pattern Anal. Appl.*, vol. 23, no. 1, pp. 265–279, Feb. 2019.
- [17] H. Oh and S. Lee, "Visual presence: Viewing geometry visual information of UHD S3D entertainment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3358–3371, Jul. 2016.
- [18] M. Abdellaoui and A. Douik, "Human action recognition in video sequences using deep belief networks," *Traitement du Signal*, vol. 37, no. 1, pp. 37–44, Jan. 2020.
- [19] W. Ding, K. Liu, F. Cheng, and J. Zhang, "STFC: Spatio-temporal feature chain for skeleton-based human action recognition," *J. Vis. Commun. Image Represent.*, vol. 26, pp. 329–337, Jan. 2015.
- [20] X. Chen, J.-N. Hwang, D. Meng, K.-H. Lee, R. L. de Queiroz, and F.-M. Yeh, "A quality-of-content-based joint source and channel coding for human detections in a mobile surveillance cloud," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 1, pp. 19–31, Jan. 2017.
- [21] L. Wang, Y. Qiao, and X. Tang, "Latent hierarchical model of temporal structure for complex activity classification," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 810–822, Feb. 2014.
- [22] K. Xu, X. Jiang, and T. Sun, "Two-stream dictionary learning architecture for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 567–576, Mar. 2017.
- [23] S. Arif, T. Ul-Hassan, F. Hussain, J. Wang, and Z. Fei, "Video representation by dense trajectories motion map applied to human activity recognition," *Int. J. Comput. Appl.*, vol. 42, no. 5, pp. 474–484, Jun. 2018.
- [24] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *Int. J. Comput. Vis.*, vol. 119, no. 3, pp. 219–238, Sep. 2016.
- [25] J. Wang, W. Wang, and W. Gao, "Multiscale deep alternative neural network for large-scale video classification," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2578–2592, Oct. 2018.
- [26] H. Rahmani, D. Q. Huynh, A. Mahmood, and A. Mian, "Discriminative human action classification using locality-constrained linear coding," *Pattern Recognit. Lett.*, vol. 72, pp. 62–71, Mar. 2016.
- [27] F. Zhou and F. De la Torre, "Spatio-temporal matching for human pose estimation in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1492–1504, Aug. 2016.
- [28] L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: A genetic programming approach," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 158–170, Jan. 2016.
- [29] Y. Xian, X. Rong, X. Yang, and Y. Tian, "Evaluation of low-level features for real-world surveillance event detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 624–634, Mar. 2017.
- [30] M. Saremi and F. Yaghmaee, "Efficient encoding of video descriptor distribution for action recognition," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 6025–6043, Dec. 2019.
- [31] B. H. Lohithashva, V. N. M. Aradhya, and D. S. Guru, "Violent video event detection based on integrated LBP and GLCM texture features," *Revue d'Intelligence Artificielle*, vol. 34, no. 2, pp. 179–187, Jan. 2020.
- [32] J. Wang, W. Wang, R. Wang, and W. Gao, "CSPPS: An adaptive pooling method for image classification," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1000–1010, Jun. 2016.



**MIMI ZHOU** was born in Tongliao, in April 1985. She is currently a Postdoctoral Fellow with Heilongjiang University and a Lecturer with the Inner Mongolia University for Nationalities. Her research interest includes philosophy of physical education.

...