Human Behavior Feature Representation and Recognition Based on Depth Video

Miao He[1], Guangming Song[1,*] and Zhong Wei[2]

[1]*School of Instrument Science and Engineering, Southeast University, Nanjing, 210018, China*
[2]*School of Automation, Nanjing University of Information Science and Technology, Nanjing, 210044, China*
E-mail: mikesong@seu.edu.cn
[*]*Corresponding Author*

Received 18 September 2020; Accepted 22 October 2020;
Publication 12 December 2020

Abstract

With the continuous development of computer artificial intelligence technology, various applications based on artificial intelligence emerge in an endless stream, among which video image recognition technology is the most widely used in life. This article starts from the process of image recognition, based on the composite characteristics of artificial intelligence and video images, to discuss human gesture recognition technology.This article uses the feature extraction algorithm for image composite feature extraction as a method, and conducts human body movement collection experiments, analyzes the database and The gesture recognition step. This paper mainly introduces the extraction method of image composite features and the basic requirements of gesture recognition, and through the algorithm calculation of feature extraction, the function of human gesture recognition video and image composite features is completed, and the human action collection experiment is carried out to confirm. The results of images and data show the advantages of the algorithm support used in this article. We will Dmti. MsHOG is compared with other methods in the three subsets. In terms of the accuracy of all tests, our method performs better than other methods. The results show

Journal of Web Engineering, Vol. 19_5–6, 883–902.
doi: 10.13052/jwe1540-9589.195614

that the MSHOG (Multi-scale Histogram of Oriented Gradients) descriptor can represent the unique characteristics of human behavior, reflecting the effectiveness of our proposed method. In particular, this method achieved 100% recognition accuracy in Test, with an average recognition accuracy of 94.91%, which is significantly better than existing methods.

Keywords: Artificial intelligence, composite features of video images, motion recognition, feature extraction algorithm.

1 Introduction

In the past few decades, people have done a lot of in-depth research on the methods of moving object detection, feature extraction, classification and behavior understanding in video motion analysis, and made some achievements, and put forward many effective methods. Relevant departments in the United States have carried out relevant research, aiming at the battlefield and civil scenes, in-depth research in order to achieve the automatic analysis technology of real-time monitoring [1, 2]. The common representation methods of action features include shape representation, appearance representation and motion information representation [3, 4]. The method of shape representation is to establish two-dimensional or three-dimensional parametric model of human body, and use the model parameters obtained by analysis to represent human body posture. For example, AF bobick et al. Put forward the concept of motion energy map and motion history map based on silhouette contour features. Wang Heng et al. Put forward a method of using dense set trajectory and motion boundary descriptor to represent the whole video motion features [5, 6]. Yang Xiaodong et al. proposed another method based on depth image, which models 3D skeleton nodes in depth image sequence to represent actions. Compared with the previous researchers and foreign researchers, many domestic research institutions and University researchers have also carried out the research of human motion recognition, made some achievements, and put forward many effective methods. Cai Yanglei et al. Put forward an image target classification algorithm based on multi-scale context information, and Deng Liqun put forward two new feature extraction methods based on singular value decomposition (SVD). In recent years, NSFC has successively funded research on video understanding, moving target detection, feature extraction, feature classification, etc. [7, 8]. Through these studies, there are some practical systems such as intelligent video surveillance, which make a great contribution to the field of visual

analysis of moving objects. In the case of a central frequency equal to 3 standard deviations, the bandwidth is approximately one times the frequency (the ratio of the upper frequency to the lower frequency, where the cutoff frequency is defined as the upper and lower limits when the transfer function drops to half of the maximum). It can be interpreted as, for the Gaussian, the point that goes down to half of the maximum is about plus and minus a standard deviation. Thus, the upper and lower frequencies are approximately and respectively, and the resulting bandwidth is one times the frequency. This bandwidth limitation means that you need a lot of Gabor filters if you want to get a wide spectrum coverage.

Performance Analysis of Log Gabor Function

The purpose of this paper is to explore the recognition of human gesture based on artificial intelligence combined with the composite features of video images. This paper mainly introduces the extraction method of image composite features and the basic requirements of gesture recognition, and through the algorithm calculation of feature extraction, completes the function of human gesture recognition of video and image composite features, and carries out the experiment of human motion acquisition to confirm it. Through the results of image and data, it shows the advantages and advantages of the algorithm proposed in this paper The efficiency solves the problem that the traditional method relies too much on manual extraction of action feature data [9, 10]. It provides a more accurate and effective method for the future human motion acquisition experiment. The validity and accuracy of hog algorithm for human gesture recognition. It has high theoretical value and practical significance for the future development of artificial intelligence video image and the research of virtual human body and robot recognition.

2 Proposed Method

2.1 Extraction of Image Composite Features

Extracting and selecting appropriate image features in computer vision systems, images are composed of discrete pixels. The essence of extracting the features of an image is to count the distribution rules of discrete image pixels, and classify and divide the pixels. The process of feature extraction is also the process of reducing the pixel dimensions of pictures and videos. The importance of feature extraction is to reconstruct the expression and find a

relatively better place to map the initial data object to this place. According to the video image feature extraction to make different distinctions, focusing on introducing several commonly used feature extraction methods:

1. Color-based feature extraction and application
2. Feature extraction and application based on texture
3. Shape-based feature extraction and application.

2.2 Demand of Gesture Recognition System

The main functions of the computer vision-based gesture recognition system include image acquisition, video analysis and recognition, and recognition result output. Among them:

1. Image collection: Collect the operator's gesture video image through the camera.
2. Video analysis and recognition: According to the video image information collected in real time, the image is preprocessed to analyze the operator's gestures.
3. Recognition result output: After gesture tracking and recognition processing, after correctly and steadily recognizing the operator's operation intention, the gesture signal is converted into a gesture motion control command, and transmitted through the computer.

2.3 Feature Extraction Algorithm

The choice of function is the main point of action pattern recognition. The purpose of feature extraction is to distinguish various gestures and actions represented by EMG's specific feature data as much as possible. The characteristics of the time domain have a better resolution effect, and the calculation is relatively less, which can immediately obtain such advantages, so the characteristics of the time domain are selected as the classification criteria in this article. This is the mean absolute value (MAV), zero crossing point (ZC) and the length of the waveform (WL). In view of the above defects of Gabor filter, a Log Gabor filter with gaussian distribution on the logarithmic scale is introduced, and its amplitude-frequency response is defined as follows:

Where is the center frequency of the filter. In order to ensure that the shape of the filter is constant, the value should be kept unchanged for different filters. For example, when set 0.74, the bandwidth of the filter is approximately one frequency, when set 0.55, it is approximately two times frequency, and when set 0.41, it is approximately three times frequency.

In the final analysis, image analysis is to serve people, so it is natural to use computational mechanisms consistent with the cognition of human visual system. The study on the characteristics of human visual cells shows that the human visual system is nonlinear, and the nonlinearity has logarithmic properties. It is also found that the transfer function of the filter is roughly symmetric on the logarithmic frequency scale. Field also points out that the filter whose transfer function is Gaussian on the logarithmic frequency scale can encode the image more effectively, while the transfer function of Gabor is Gaussian on the linear frequency scale, that is, the Log Gabor function can more truly reflect the frequency response of natural images

The average absolute value is shown in equation (1):

$$MAV = \frac{1}{N}\sum_{i=1}^{N}|x(i)| \tag{1}$$

The number of zero crossings is shown in (2):

$$\begin{cases} ZC = \dfrac{1}{N}\displaystyle\sum_{K=1}^{N-1} f_k \\ f_k = \begin{cases} 1, & x_k x_{k+1} < 0, |x_k - x_{k+1}| > \varepsilon \\ 0, & \text{else} \end{cases} \end{cases} \tag{2}$$

The waveform length is shown in (3):

$$WL = \sum_{i=1}^{N-1}|x(i+1) - x(i)| \tag{3}$$

During the experiment, record the sample data of the 24 dimensional features of a posture.

3 Experiments

The data set includes 10 experimenters, using RGB. D (Depth image) camera shoots 20 times. Each person completes each action 2 to 4 times, and there are a total of 567 depth-mapped data sequences. The resolution of the depth map is 340x220. Table 1 lists the 20 kinds of movements,

As listed in Table 1, they are the data of human gestures in the experiment.

Target video target tracking is a basic core technology in the field of computer vision, which is the focus of many research institutions and researchers. It is the basis of many high-level video processing.

Table 1 A subset of actions and tests used in the experiment

Action Set 1	Action Set 2	Action Set 3
Wave hand (2)	Raise arm and wave hand (1)	Throw up (6)
Beat (3)	Excerpt (4)	Front kick (14)
Boxing (5)	Draw fork (7)	Side kick (15)
Throw up (6)	Draw fork (8)	Jogging (16)
Clapping (10)	Draw fork (9)	Waving a tennis racket (17)
Bend over (13)	Waving hands (11)	Serve (18)
Serve (18)	Front kick (14)	Waving Golf (19)
Pick up and throw (20)	Side hit (12)	Pick up and throw (20)

There are many tracking algorithms for video targets. According to the different tracking principles, the tracking algorithms can be briefly divided into the tracking based on contrast analysis, the tracking based on feature matching, the tracking based on core and the tracking based on motion detection. The following is a brief introduction to several commonly used target tracking algorithms.

Kalman filtering tracking algorithm: The basic idea of Kalman filtering algorithm is essentially a recursive algorithm for state prediction of linear dynamic system with noise. The process of target tracking is a process of constant prediction and correction. The conventional Kalman filtering algorithm has many limitations in the application process. It requires that both the system state model and the observation model are linear and conform to the Gaussian model, and the noise must conform to the gaussian distribution in order to obtain better results in the tracking process.

Particle filter tracking algorithm: Particle filter tracking algorithm is a Bayesian filter algorithm based on Monte Carlo simulation. It can solve some tracking that the Kalman filter algorithm cannot achieve, and it can mainly solve the nonlinear and non-gaussian tracking. Particle filter algorithm mainly includes three steps of particle sampling, particle weighting and particle resam. Its core idea is to express the distribution of random state particles extracted from the posterior probability. In short, particle filter refers to the process of finding a group of random samples propagating in the state space to approximate the probability density function and replacing the integral operation with the sample mean to obtain the minimum variance distribution of the state.

Target tracking algorithm based on motion detection: The basic idea of this algorithm is to track the target by detecting the different areas of the target and the background motion in the sequential image. Optical flow method is the representative of this kind of algorithm. It studies the change of image gray level in time and the relationship between object structure and motion in scene to achieve the purpose of tracking the target.

Target tracking algorithm based on kernel: The basic idea of kernel tracking algorithm is to use direct continuous estimation for similarity probability density function or posterior probability density function to track the target. Its main representative algorithm is mean-Shift tracking algorithm. Compared with other algorithms, it has obvious advantages and can get better tracking effect. This paper mainly studies and improves the algorithm, and only gives a brief introduction here.

4 Discussion

4.1 Analysis of Gesture Recognition Steps

(1) Action recognition method

Analyzing human movement is one of the important topics in the field of computer vision. Current behavior analysis and recognition methods are divided into template-based methods and state space-based methods. The template-based method compares the analyzed pose modeling with the existing model, and analyzes the classification of the detection model according to the two differences. Template matching or dynamic programming and other suitable methods fall into this category. It can be seen from the figure that the Log Gabor filter has no DC response and has a Gauss-shaped distribution on the logarithmic scale, which ensures its local analytical capability similar to that of the Gabor filter, and at the same time makes up for the above shortcomings of the Gabor filter.

In image processing, the DC component corresponds to the average gray level of the image. In the representation and recognition system of human behavior characteristics, due to the influence of image acquisition equipment and environmental light changes and other factors, the average gray level of images is constantly changing and cannot represent the essential information of the representation and recognition of human behavior characteristics, so it cannot be regarded as the characteristics of the representation and recognition of human behavior characteristics.Template matching means that, according to a certain time rule, the action sequence of the test posture is compared with

the posture form of the basic library installed, and the posture is analyzed according to the matching degree. The difference between the dynamic scheme method and the template-like method is that there is no time rule involved in the comparison. The template and sample are compared and checked at any time in the library, and the best fit between the two styles is analyzed, and the analysis and style of the results are achieved. The difficulty of the dynamic programming method is that it is not easy to get an excellent sample pose mode. Then, set a few samples in each posture, and compare the check posture with the samples one by one to find the best method. Therefore, as the size of the training model becomes larger, the analysis difficulty of the dynamic planning method will also increase, and this method is not robust and is easily affected by noise.

(2) Principle of parallax

The binocular stereo vision system is a computer that simulates the human eye to perceive the world, and obtains the depth information of the scene from two two-dimensional images through the principle of parallax. The difference in the projection position of the spatial point in the three-dimensional scene on the left and right cameras is called parallax. The parallax value of the spatial point in the image can restore the depth information of the point in the scene.

(3) Camera imaging model

Camera imaging principles are divided into two types: linear cameras and non-linear cameras. Now, the commonly used pinhole camera samples belong to the principle of linear cameras. The lens is not distorted, the imaging path is linear, and there is no lens distortion. The non-linear camera sample analyzes the undesirable conditions such as chaos and centrifugal mirrors. The purpose of camera imaging is to convert points in space into pixels in a two-dimensional image through a series of processes such as mapping transformation. In the process of imaging, it is actually the world coordinate system, camera coordinate system, imaging coordinate system, and image coordinate system. Between two coordinate systems. Among them, the world coordinate system is used to determine the position of the calibration object, and as the reference system of the binocular vision system, the relative relationship between the two cameras is given, and at the same time as the three-dimensional coordinate system that holds the reconstructed object; the purpose of the camera coordinate system is to explore The mathematical principle that the camera becomes a graph. Taking the optical center of

the camera as the origin, the camera observes the coordinate data of the object in an independent direction. The imaging coordinate system is the data coordinates of the image plane. Set the location of the object in the special photo of the coordinate system set in the data table. The Log Gabor function has significant advantages over the Gabor function. Firstly, its function always has no DC component, so the image processing is not affected by the brightness condition. Secondly, its transfer function has an elongated tail at high frequencies, which can make up for the shortcoming of the ordinary Gabor function that low frequency means excessive and high frequency means insufficient. Field's research on the statistics of natural images shows that the amplitude spectrum of natural images decays approximately at the point, and it is necessary to use a filter with similar spectrum to encode images with such spectral features. So the Log Gabor function with the long tail should be able to encode natural images more efficiently than the normal Gabor function. Thirdly, it can cover a larger frequency range, and the Log Gabor function can save about half of the computation when the parameters are selected properly.

In addition, it is more in line with human visual characteristics than Gabor function. The frequency bandwidth of the human eye equivalent space filter seems to be related to many factors, including motion, brightness and spatial frequency of the excitation source. It is difficult to draw a general conclusion from these data, but a more unified view is that the bandwidth of the spatial filter ranges from about 0.5–3 times frequency, and 1.5 times frequency is considered as the "typical" bandwidth, and the filter's transfer function is roughly symmetric on the logarithmic frequency scale. The performance analysis of the Log Gabor function shows that the bandwidth range corresponding to the minimum filter space size is exactly 1–3 times frequency, which is very consistent with the human eye equivalent model. And its spatial width at 3 octave is roughly the same as that of Gabor function at 1 octave, indicating that it can use compact spatial filter to capture wider spectral information.

(4) Camera calibration

The purpose of mark calibration is to obtain the internal and external parameter matrix of the camera, which is an indispensable step in binocular ranging. After the camera parameters are obtained through calibration, the image taken by the camera can be corrected, and the relationship between the two-dimensional image coordinates and the three-dimensional world coordinates can be obtained.

(5) Binocular correction

To get the visual difference constructed by the target points in the east and west side views, start to align and coordinate the points with the two pixels corresponding to the east and west views. However, the matching of matching points in the quadratic space takes a very long time. In order to reduce the scope of the matching search, the extension limit can be used to convert the comparison of the corresponding points from the second element to the first element to escape all-round inspection. The function of binocular correction corresponds to the two sample photos after removing distortion. Therefore, the extension lines of the two sample pictures are correctly on the same horizontal straight line, and the points of one sample photo are different from the points of the other sample photos. Corresponding points must have the same code, only need to perform a meta search with lines to find the corresponding points.

(6) Process steps for gesture recognition

CBIR (Content based image retrieval) video image system usually means that the video picture database contains a lot of video or picture sample data. When the user enters a video or picture, the system can immediately and effectively search for the one that is most similar to the video or picture entered in the video picture database. The search method uses the similarity of visual features between videos to create similarity feature vectors of photos in a video picture library, and counts and finds similarities between pictures and database video pictures, thereby realizing the search for videos and images. The general functions of CBIR video image system are shape, color, stripes and three-dimensional dynamic spatial relationship.

The recognition of human gestures based on the combined features of artificial intelligence and video images generally has the following branches: video picture collection; video picture preprocessing (human gesture detection, vest detection module, and human area modeling module); motion tracking Module; gesture recognition module (static gesture recognition module, dynamic gesture recognition module and hand recognition module). The specific module diagram is shown in Figure 1 below.

As shown in Figure 1 above, it is a flowchart of the gesture recognition module. Through the flowchart, it can be learned that the research on human gesture recognition based on the composite features of artificial intelligence and video pictures is to collect photos or images through one or more cameras, and use video pictures to solve Methods to explore and distinguish the meaning of gestures. It is characterized by non-contact gesture

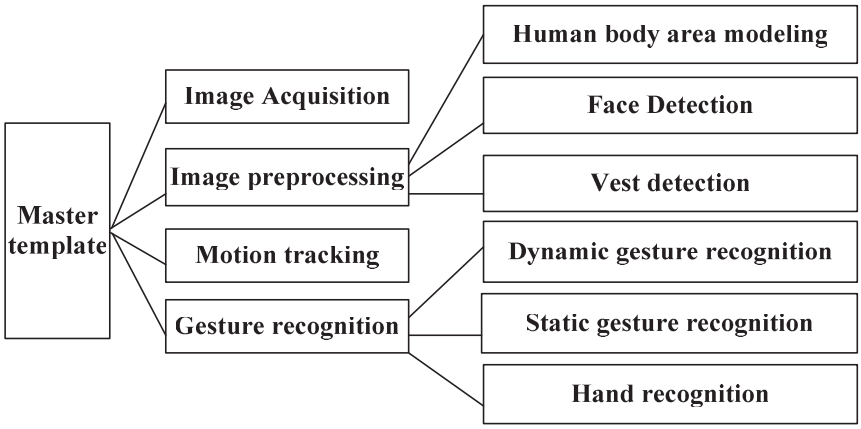

Figure 1 Flow chart of gesture recognition module.

discrimination that makes the interaction between humans and computers more natural and harmonious without any additional pressure on human gestures. Initially, the data of the image and picture are obtained through one or more camera equipment. If there is an arm posture, the posture is separated from the image information, the frequency response of the real (even symmetric) and imaginary (odd symmetric) parts of the Log Gabor filter. Near the center frequency, the imaginary part response has a positive part and a negative part respectively, thus ensuring that the imaginary part response is a better edge detector. The real part response has a large positive part and two small negative parts on both sides, thus ensuring that the real part response is a better BLBO target detector. It can also be seen that Log Gabor wavelet should be able to detect human behavior feature representation and texture feature well, and it also has some applications in image edge detection and recognition. the model of the trained posture sample is selected for analysis and discussion of human gestures, and the feature values of the human gestures are extracted during the discussion, and these are recognized using artificial intelligence grammar Characteristic data and data are necessary. As a result, a description of human gestures is generated.

4.2 Analysis of Human Body Movement Experiment

For each subset, there are three different tests: Test One, Test Two, and Cross Test. Because each action is performed 2 or 3 times by the same performer,

in test 1, we use the first action of the same action performed by the same person as training data, and the remaining one or two Actions are used as test data; in test two, we use the first two actions of the same action performed by the same person as training data, and the rest as test data; in cross-testing, we use the first action of the same action , 3, 5, 7, and 9 performers 'actions are used as training data, and 2, 4, 6, 8, and 10 performers' actions are used as test data. Cross-testing is challenging because different action performers have their own styles, so there is a big difference between training data and test data.

For B (Bins) and L (HOG scale) in HOG, we can see in the figure and graph. With the growth of B, the recognition accuracy also increases, because with the increase of B, DMTI. MsHOG (Depth Motion Trace Image and Multi-scale Histogram of Oriented Gradient) becomes more and more sparse, this is L2.The core idea of regularized collaborative representation (L2. CRC).When B exceeds 18, L2. The sparse contribution of CRC becomes smaller, and the recognition accuracy rate reaches more than 94.5%. It keeps stable with the increase of B, but the calculation cost increases with the increase of B, so B = 18 has higher recognition accuracy and has more Low calculation cost. Considering the accuracy of recognition, L = 3 is better than L = I2, especially in AS2 of the cross-person test, 92.9% is much higher than 78.76% (L = 1) and 84.07% (L = 2), 86.73% (L = 4). Therefore, we finally use B = 18 and L = 3 for the experiment.

The experiment results of different experimenters as test sets are compared with L2.CRC (Collaborative representation classifier) classifier and SVM (Support Vector Machine) classifier. When the action data of the first-ranked experimenter is used as the test set, the recognition effect is not ideal. This is because different experimenters have different behavior habits when making the same gestures, resulting in large intra-class differences. When the motion data of the 9th and 10th experimenters were used as the test set, both classifiers achieved nearly 100% recognition effect. We can simply think of the comparison of the gestures and movements of the two experimenters. Table 2 shows the performance comparison on the MSR Gesture3D database. MSR Gesture3D is a deep hand gesture data set collected by the depth camera. The design of filter Banks involves the relationship between the filter bandwidth and the scaling factor of the center frequency between adjacent filters. The aim is to obtain a spectrum with reasonable width and uniform coverage with fewer filters. As mentioned earlier, the maximum bandwidth available from the Gabor filter is about one frequency. In order to obtain a uniform spectral coverage, the ratio between the center frequencies of

Table 2 Performance comparison on the MSR Gesture3D database

Method	Accuracy (%)
Action Graph on Occupancy	80.50%
Action Graph on Silhouette	87.70%
Random Occupancy Pattern	88.50%
Depth Motion Maps	89.20%
HON4D	92.45%
SVM	94.17%
L2-CRC	94.70%

adjacent filters should not be greater than 1.5. In this way, 8 filters are needed to construct a filter bank spanning 4 frequency.

In Table 2 we can see that the DMTI-HOG (Depth Motion Trace Image and Histogram of Oriented Gradient) descriptor is based on L2. The CRC classifier achieves an average recognition accuracy of 94.70%, which is better than existing methods. The confusion matrix of the test results of the HOG descriptor on the MSRGesture3D dataset, we can see that the recognition rate on almost all gestures exceeds 90%, especially in the three types of gestures such as Z, Where, and Pig. Recognition accuracy, only very few gestures (such as Store, Finish, J) recognition rate is slightly less than 90%. We can see that our method performs well on almost all gestures. Figure 2 below shows the average accuracy under cross test.

$$\frac{\partial G}{\partial x} = kx \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(-\frac{y^2}{2\sigma^2}\right) = h_1(x)h_2(y) \tag{4}$$

$$\frac{\partial G}{\partial y} = ky \exp\left(-\frac{y^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2}\right) = h_1(y)h_2(x) \tag{5}$$

We can see in Figure 2 that for B (Bins) and L (HOG scale) in HOG. With the growth of B, the recognition accuracy also increases, because with the increase of B, HOG becomes more and more sparse, which is the core idea of regularized cooperative representation (L2. CRC). When B exceeds 18, L2. The sparse contribution of CRC becomes smaller, and the recognition accuracy rate reaches more than 94.5%. It keeps stable with the increase of B, but the calculation cost increases with the increase of B, so B = I 8 has higher recognition accuracy and also has Lower calculation cost. Considering the accuracy of recognition, L = 3 is better than L = I, 2, 4, especially in AS2 of the cross-over test, 92.9% is much higher than 78.76% (L = 1), 84.07 %

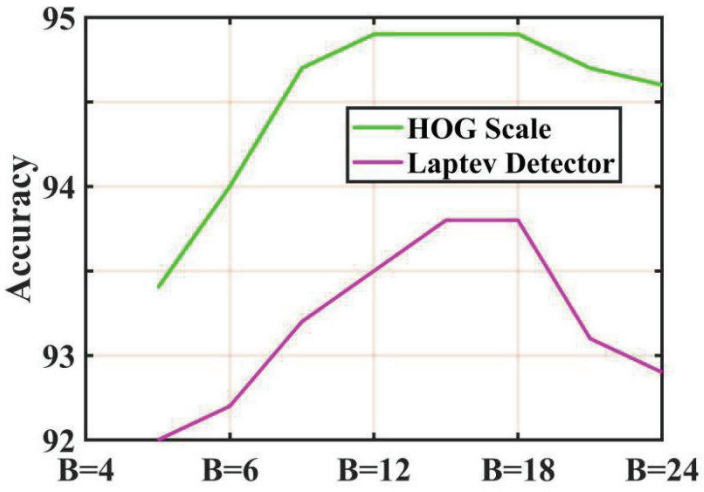

Figure 2 HOG descriptor cross-test average accuracy under different B (HOG).

(L = 2), 86.73% (L = 4). For the performance of different dimensions, see Figure 4.3 below. The large bandwidth of the Log Gabor function provides greater flexibility in the design of the filter bank, and the spikes in its shape provide precise positioning of the spatial domain. For example, in order to obtain a 4-octave filter bank, instead of eight 1-octave filters, only four 2-octave filters with an adjacent center frequency ratio of 2.6 can be used (a 3-octave filter can be used and still have flat spectrum coverage). Using fewer filters means less computation. However, it should be noted that increasing the bandwidth of the filter will reduce the resolution of frequency amplitude and phase information. This is because the amplitude and phase information will be averaged over a wider range when using a high-bandwidth filter. In practice, this increases the quantization of amplitude and phase information in the frequency domain, but as the quantization program increases, the increase in the average program will lead to points where the phase change is particularly strong and cannot be detected. Thus, there is a trade-off between the frequency resolution of amplitude and phase information and the amount of computation (number of filters) required.

Figure 3 shows the L2 based on different test sets. The recognition accuracy of MsHOG with different sizes of CRC. Interestingly, in different dimensions, the overall recognition rate of various test sets is stable. For each dimension, our method performed well in all tests. In particular, the performance of AS2 cross-testing is significantly better than existing methods.

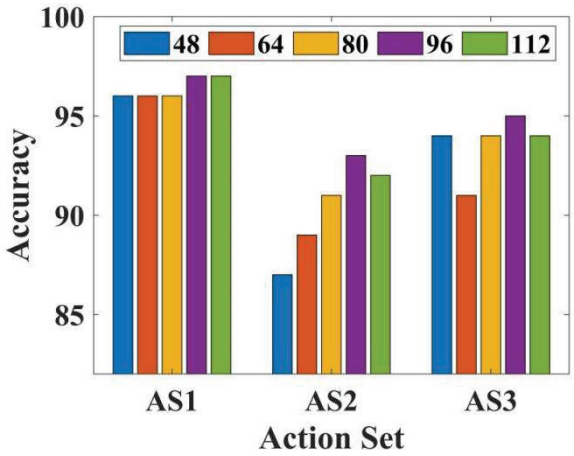

Figure 3 Comparison of recognition performance in different dimensions during cross-testing.

This may be because the actions in AS2 have many similar actions (such as drawing a fork (7), drawing a hook (8), and drawing a circle (9)). We calculate HOG through local motion and global motion, which makes the action easier to distinguish,,So our recognition rate is higher than other methods. However, the recognition rate of interdisciplinary tests is still lower than that of test one (97.8%) and test two (100%), because the actions performed by different people are quite different, and the number of experimenters tested in cross-subject tests is limited.

$$P_x[i,j] = (I[i,j+1] - I[i,j] + I[i+1,j+1] - I[i+1,j])/2 \quad (6)$$

$$P_y[i,j] = (I[i,j] - I[i+1,j] + I[i,j+1] - I[i+1,j+1])/2 \quad (7)$$

We will DMTI. MsHOG is compared with other methods on three subsets. In terms of the accuracy of all tests, our method performs better than other methods. The results show that the MsHOG (Multi-scale Histogram of Oriented Gradient) descriptor can represent the unique characteristics of human behavior, reflecting the effectiveness of our proposed method. In particular, our method achieved 100% recognition accuracy in Test 2, and the average recognition accuracy reached 94.91%, which is significantly better than existing methods. The construction of target feature model is the key of mean-Shift algorithm. Selecting good target features to construct target model can make target tracking have better results in practical application. The main target features available for selection are color features, texture features, edge

features, etc. The traditional mean-Shift algorithm selects the color feature as the basis of tracking the target. It USES the color information of the target to analyze, track and locate it.

Color features: Color features are represented by color histograms in the RGB color space. The three color features of R,G and B can be expressed quantitatively with 256 intervals respectively. The reason for using RGB colors is mainly due to the versatility of the algorithm. At present, most of the existing video test sequences and data of video capture CARDS exist in the form of RGB format. It takes a lot of computing time to convert THE RGB format into other color space representation forms. The direct adoption of RGB color space can improve the efficiency of tracking and has good real-time performance.

Edge features: The description of edge features can use Roberts operator, which is an operator that USES local difference operator to find the target edge. It only needs to calculate the current pixel field, so it has the advantage of simple calculation. Similarly, Roberts operator USES histogram to represent edge features and quantifies 256 intervals

5 Conclusions

With the advancement of technology, "intelligence" has gradually become an indispensable part of people's lives, artificial intelligence algorithms that are gradually popularized in all walks of life. Human gesture recognition based on artificial intelligence image composite features is an important research field in the field of computer 3D audiovisual and pattern recognition. In the large-scale video picture database and advanced original video picture data, starting from the discussion of artificial intelligence-based video picture synthesis feature method, it fully explains the content and timeliness of video photos and the evaluation results with certain effects.

Performance analysis, according to the former in the face of the Log Gabor function under the condition of the guarantee good results as far as possible in order to make the calculation simple, less as far as possible the amount of calculation, based on the one-dimensional Log Gabor functions, select different scales, different parameters (such as center frequency, filter bandwidth) test comparison, to determine a set of better parameter combination.

Aiming at the problem of human body positioning in gestures, especially the difficulty of squatting, it is studied two human body positioning methods based on human body positioning based on face recognition and solid

color clothes. The two positioning methods have different emphasis. Face recognition is to ensure that the system recognizes the body as a human, to prevent false detection, and also to obtain the basic spatial position of the human body. After experiment, each of the two methods of division has its own focus, which can be used to determine the area of the human body.

Acknowledgements

This work is supported by National Natural Science Foundation under grant (No. 61973076).

References

[1] Y. Lü, Zhao J, Cao F. Image Denoising Algorithm Based on Composite Convolutional Neural Network. Moshi Shibie Yu Rengong Zhineng/pattern Recognition & Artificial Intelligence, 2017, 30(2):97–105.

[2] Nguyen H T, Nguyen L T, Dreglea A I. Robust approach to detection of bubbles based on images analysis. International Journal of Artificial Intelligence, 2018, 16(1):167–177.

[3] CACM Staff. Artificial Intelligence. Communications of the ACM, 2017, 60(2):10–11.

[4] Y. Ming, G. Wang, X. Hong. Spatial-temporal texture features for 3D human activity recognition using laser-based RGB-D videos. ksii transactions on internet & information systems, 2017, 11(3):1595–1613.

[5] Hassabis D, Kumaran D, Summerfield C, et al. Neuroscience-Inspired Artificial Intelligence. Neuron, 2017, 95(2):245–258.

[6] Ma J, Yu J, Hao G, et al. Assessment of triglyceride and cholesterol in overweight people based on multiple linear regression and artificial intelligence model. Lipids in Health and Disease, 2017, 16(1):1–7.

[7] Xie T, Qin P, Yan J. Research on Artificial Intelligence Frontier Recognition Based on LDA. Open Access Library Journal, 2018, 05(12):1–13.

[8] Han M. Application of Artificial Intelligence Detection System Based on Multi-sensor Data Fusion. International Journal of Online Engineering (iJOE), 2018, 14(6):31.

[9] Li M, f financial auditing teaching mode based on artificial intelligence expert system. Zhang H, Chen B, et al. Prediction of pKa Values for Neutral and Basic Drugs based on Hybrid Artificial Intelligence Methods. Scientific Reports, 2018, 8(1):3991.

[10] Xinman Z. Construction oBoletin Tecnico/Technical Bulletin, 2017, 55(17):743–747.
[11] Wang Y, Duan H. Classification of Hyperspectral Images by SVM Using a Composite Kernel by Employing Spectral, Spatial and Hierarchical Structure Information. Remote Sensing, 2018, 10(3): 441.
[12] Li H, Luo W, Qiu X, et al. Identification of Various Image Operations Using Residual-Based Features. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(1):31–45.
[13] Latonov V V, Tikhomirov V V. Line-of-Sight Guidance Control Using Video Images. Moscow university mechanics bulletin, 2018, 73(1): 11–17.
[14] Latonov V V. Programmed Strategies to Test the Quality of Line-of-Sight Guidance Control Using Video Images. Moscow University Mechanics Bulletin, 2018, 73(6):135–140.
[15] Zhi X, Yan J, Hang Y, et al. Realization of CUDA-based real-time registration and target localization for high-resolution video images. Journal of Real Time Image Processing, 2019, 16(4):1025–1036.
[16] Gurov I P , Volkov M V , Margaryants N B , et al. Method of bringing locally varying images into coincidence in video capillaroscopy. Journal of Optical Technology c/c of Opticheskii Zhurnal, 2019, 86(12):774.
[17] Liu L , Liu G , Chu X M , et al. Ship Detection and Tracking in Nighttime Video Images Based on the Method of LSDT. Journal of Physics Conference Series, 2019, 1187(4):042074.
[18] Itakura K, Hosoi F. Estimation of tree structure parameters from video frames with removal of blurred images using machine learning. Journal of Agricultural Meteorology, 2018, 74(4):154–161.
[19] Chen H, Ye S, Nedzvedz O V, et al. Application of Integral Optical Flow for Determining Crowd Movement from Video Images Obtained Using Video Surveillance Systems. Journal of Applied Spectroscopy, 2018, 85(1):126–133.
[20] Zhang L X. Research on similarity measurement of video motion images based on improved genetic algorithm in paper industry. Paper Asia, 2019, 2(1):135–138.

Biographies

Miao He received the bacherlor's degree from Northeast Petroleum University, China, in 2018. Now she studies in the School of Instrumen Science and Engineering at Southeast University. Her research interests include human-robot interaction and computer vision.

Guangming Song received the Ph.D. degree in control science and engineering from the University of Science and Technology of China, Hefei, China, in 2004. From 2004 to 2006, He was a Research Fellow with the Robotic Sensor and Control Laboratory, Southeast University, China. Since 2006, he has been with the School of Instrument Science and Engineering, Southeast University, China. He is currently a Professor with the School of Instrument Science and Engineering, Southeast University, China. His current research interests include distributed robots, aerial manipulators, and bio-inspired legged robots. He won the second prize of the National Technology Invention Award of China in 2017.

Zhong Wei received the Ph.D. degree in instrumentation science and technology from the Southeast University, Nanjing, China, in 2019. He is currently a lecturer with the School of Automation, Nanjing University of Information Science and Technology, China. His current research interests include leg-wheel robots and bio-inspired legged robots.