Hindawi
Advances in Mathematical Physics
Volume 2021, Article ID 1268747, 11 pages
https://doi.org/10.1155/2021/1268747

Research Article

Extraction of Human Motion Information from Digital Video Based on 3D Poisson Equation

Yilin Wang[1] **and Baokuan Chang**[2]

[1]*Henan Kaifeng College of Science Technology and Communication, Kaifeng, Henan 475001, China*
[2]*Henan University, Kaifeng, Henan 475001, China*

Correspondence should be addressed to Baokuan Chang; 104753130484@vip.henu.edu.cn

Received 15 November 2021; Accepted 11 December 2021; Published 28 December 2021

Academic Editor: Miaochao Chen

Based on the 3D Poisson equation, this paper extracts the features of the digital video human body action sequence. By solving the Poisson equation on the silhouette sequence, the time and space features, time and space structure features, shape features, and orientation features can be obtained. First, we use the silhouette structure features in three-dimensional space-time and the orientation features of the silhouette in three-dimensional space-time to represent the local features of the silhouette sequence and use the 3D Zernike moment feature to represent the overall features of the silhouette sequence. Secondly, we combine the Bayesian classifier and AdaBoost classifier to learn and classify the features of human action sequences, conduct experiments on the Weizmann video database, and conduct multiple experiments using the method of classifying samples and selecting partial combinations for training. Then, using the recognition algorithm of motion capture, after the above process, the three-dimensional model is obtained and matched with the model in the three-dimensional model database, the sequence with the smallest distance is calculated, and the corresponding skeleton is outputted as the results of action capture. During the experiment, the human motion tracking method based on the university matching kernel (EMK) image kernel descriptor was used; that is, the scale invariant operator was used to count the characteristics of multiple training images, and finally, the high-dimensional feature space was mapped into the low-dimensional to obtain the feature space approximating the Gaussian kernel. Based on the above analysis, the main user has prior knowledge of the network environment. The experimental results show that the method in this paper can effectively extract the characteristics of human body movements and has a good classification effect for bending, one-foot jumping, vertical jumping, waving, and other movements. Due to the linear separability of the data in the kernel space, fast linear interpolation regression is performed on the features in the feature space, which significantly improves the robustness and accuracy of the estimation of the human motion pose in the image sequence.

1. Introduction

Obtaining and analyzing various parameters of human motion from video image information is a key research direction of multidisciplinary fusion. The core of its research is to detect and track the human body from a single or multiple video sequences, obtain human motion data, and reconstruct the human body for three-dimensional motion or description and understanding of human motion [1]. The research of video human motion analysis has broad application prospects in the fields of human animation, games, virtual reality and augmented reality, human-computer interaction, video surveillance, sports motion analysis, and auxiliary clinical medical diagnosis [2–5]. Due to the relative ease of acquisition and wide range of applications, many existing video human motion tracking studies are dedicated to tracking the moving human body from the image sequence acquired by a single camera. Most of these methods are based on extracting different features of the human body from image frames for matching and tracking. The features used for tracking are usually points, regions (image blocks), or contours. There is also a class of methods

that use the framework of model-based motion analysis to design different forms of three-dimensional human models for tracking human motion in monocular videos [6].

The visual analysis of human motion has broad application prospects in human-computer interaction, video conferencing, medical diagnosis, virtual reality, etc., which makes it a frontier direction that has attracted the attention of researchers in recent years [7]. The main purpose of visual analysis is to detect, identify, and track the human body from a set of image sequences containing people and to understand and describe its behavior. In general, this process can be divided into the underlying vision module level vision, data fusion module, and high-level vision module. Among them, the underlying vision module mainly includes motion analysis methods such as motion detection and target tracking; the data fusion module mainly solves the fusion processing of multicamera data: the off-layer vision module mainly includes target recognition and semantic understanding and description of motion information [8–11].

This article focuses on key technologies such as the two-dimensional acquisition of human motion posture in monocular video and three-dimensional reconstruction. The work focuses on obtaining the three-dimensional motion of the human body from the monocular video image containing the human motion and recovering the three-dimensional motion of the human body represented by the joint skeleton model. Human action behavior analysis is one of the frontier directions that have attracted more attention in the field of computer vision in recent years. The human body movements in the video can be seen as a combination of different movements of the moving torso and limbs. This paper analyzes and studies the three parts of the target detection of the moving human body, the feature extraction of the human action sequence, and the learning and classification of the human action feature according to the order of the analysis of the human action behavior. Thus, it is possible to conduct research on the human body's motion technology, human body motion laws, and motor capabilities, which is expected to help designers to make a reasonable evaluation of human factor engineering factors in the early design stage, reduce design rework and the production of physical prototypes, and shorten the design process. In addition, placing the digital human body model in a virtual production environment can well solve the practical problems of human factor engineering such as the accessibility, safety analysis, and standardization of operating actions for workers in production and assembly operations.

2. Related Work

In a large number of research experiments, the analysis and recognition of human movements mainly include three aspects: the structural analysis of the target human body, the detection and tracking of moving targets, and the analysis and recognition of human movements based on image sequences. These three aspects are mutually advancing. The structural model used is determined through the analysis of the human body structure. The human body target is detected and tracked in the video based on the human body structure model to obtain a series of action image sequences of the human body structure model and finally the sequence performs analysis and identification.

Keceli [12] surrounds each pedestrian with a rectangular frame and selects the center point of the rectangular frame as the tracking feature. In the tracking process, the current position is estimated by the position of the center point in the previous frames. If there is occlusion, as long as the speed of the center of mass can be distinguished, the tracking can still be carried out successfully. Stuart et al. [13] also proposed a system to track the human body's two-dimensional translational motion. From the restored static or changing background image, the motion of the foreground image is estimated by matching the line segments in the background image, and finally, the center of the bounding box is used to achieve tracking. The human body tracking system of Li et al. [14] uses the corner points of the motion contour as the corresponding features. These feature points use distance metrics based on position and curvature to perform forward and reverse matching between consecutive frames. This matching process implicitly assumes a certain degree of rigidity in human motion and small motion between consecutive frames, so the applicability is greatly restricted, and feature extraction itself is also difficult. Blair et al. [15] extended the region-based motion estimation framework and introduced the joint chain constraint in robot control. Using exponential mapping and integrating it into partial differential equations, the tracking problem is reduced to solving a simple linear system. Their method can track high-degree-of-freedom human joint movement from cluttered video sequences with obvious noise. The user needs to manually mark the initial frame, and then, the program automatically tracks it.

Georgiou et al. [16] uses a 2D ribbon model to represent the human body, and a ribbon area represents a certain part of the body. In the model, 5 U-shaped ribbon areas are used to construct the human body, which represent the human head and limbs. For the input gray-scale image sequence, we first segment the foreground moving target. On this basis, each area is detected and described through the area abstraction process, and the detected ribbon area is tracked frame by frame. Then, the belt-shaped human body model is used to match these tracked areas, and the appropriate areas are marked as the human head, arms, and legs. Finally, a two-dimensional human skeleton sequence can be obtained as an output. The Pfinder (person finder) system regards the human body as being composed of small areas corresponding to the head, torso, limbs, etc., and uses Gaussian distribution to establish a statistical model of these small areas and backgrounds. A mapping table is used to indicate the attribution of the zombies, and the location of the small area is determined by the attribution of pixels in the image frame. They used such small area features to perform indoor single-person motion tracking [17]. The rule is a real-time visual monitoring system that tracks multiple people in an outdoor environment and monitors their behavior. Under the monocular gray-scale video or infrared video, multiple people and their heads, hands, feet, and torso are located through regional shape analysis and tracking to establish

the appearance model of each person to realize the tracking of multiple people [18–20].

3. Construction of Digital Video Human Motion Information Extraction Model Based on Three-Dimensional Poisson Equation

3.1. Distribution of the Solution Set of the Three-Dimensional Poisson Equation. By calculating the second-order partial derivative of U in the three-dimensional Poisson equation, the local orientation information of each part can be extracted, thereby dividing the human body into parts with different orientations. The Hessian matrix constructed for each pixel can roughly estimate the second-order spatio-temporal characteristics at each pixel. Figure 1 is the distribution of the solution set of the three-dimensional Poisson equation.

The eigenvector of the Hessian matrix represents the local principal direction at the pixel point, and the eigenvalue of the Hessian matrix represents the local curvature of the point in the direction of the corresponding eigenvector.

$$\left[\frac{\partial(\rho u)}{\partial x}+\frac{\partial(\rho v)}{\partial x}+\frac{\partial(\rho w)}{\partial x}\right]+\frac{\partial \rho}{\partial t}=0. \tag{1}$$

Compared with the RGB color space, the HSV color space is closer to the visual model of the human eye and can more directly reflect the brightness information of the color, so the use of the HSV color space can better represent the difference between the shadow part and the moving target. For the background image, we first initialize the HSV color space model and then compare each pixel of the foreground area detected with the corresponding background pixel for each frame of image.

$$\frac{\partial q/\partial x+\partial q/\partial x+\partial q/\partial x}{f_x}=d(\rho x q), \tag{2}$$

$$\frac{\partial q/\partial x+\partial q/\partial x+\partial q/\partial x}{f_y}=d(\rho y q), \tag{3}$$

$$\frac{\partial q/\partial x+\partial q/\partial x+\partial q/\partial x}{f_z}=d(\rho z q). \tag{4}$$

For background modeling, a single Gaussian model or a mixed Gaussian model can be used. The mixed Gaussian model generally uses 3 to 5 single Gaussian models to describe the characteristics of each pixel in the image. Each single Gaussian model has its own different weights and priorities. These Gaussian models are higher in priority to the lowest order.

$$tf(w_i,D_i)=N_{w_1,D_j}\otimes\lim_{K\longrightarrow\infty}\sum_{n=1}^{K}N_{w_n,D_j}\otimes N, \tag{5}$$

$$idf(w_i)=\ln\frac{|D|}{\left|\left\{D_i:D_j\in w_j\right\}-1\right|}. \tag{6}$$

For the background model update of the mixed Gaussian model, in addition to the variance and mean of each single Gaussian distribution, it is necessary to update their weights and priorities. Compared with the complex calculation of the Gaussian mixture model, the calculation of the single Gaussian model is relatively simple, and the single Gaussian model is suitable for occasions with a single constant background.

$$\nabla^2\alpha-\left(1-\frac{1}{a^2}\right)\times\frac{\partial^2\alpha}{\partial t^2}+\nabla^2\frac{\rho}{\varepsilon}=0, \tag{7}$$

$$y(n)=Ax(n)-tx(n-1). \tag{8}$$

One of the main characteristics of the multigrid method is that the error correction process on the coarse grid layer can be carried out recursively. We iterate on the grid and then attenuate the subhigh frequency components.

$$\mathrm{TVDI}=\frac{\mathrm{LST(NVDI)}-\mathrm{LST(min)}}{\mathrm{LST(max)}-\mathrm{LST(NDVI)}}, \tag{9}$$

$$\mathrm{LST(NVDI)}=A+B*\mathrm{NVDI}. \tag{10}$$

When the two-dimensional image coordinates are known, this reverse correspondence will not have a unique point, and the solution space is a straight line. It can also be seen from the pinhole model of perspective projection that a straight line is obtained by connecting the optical center of the camera and the imaging point. This ambiguity when modeling from 2D to 3D is caused by the morbidity of the problem itself, so our difficulty lies in how to use some prior knowledge to eliminate this ambiguity. By dividing the human motion sequence into several subsequences, the interference solution is deleted.

3.2. Digital Video Algorithm Flow. When dealing with digital video, if we use*c*, two cameras observe the point*P*at the same time and can determine the point*P*on the*c*-camera image, and the point*P*on the*c*-camera image is the imaging point of the same object point*P*in their respective images (also called the corresponding point); then, we can know that the spatial point*P*is located on01*P*and is also located on 02*P*. They include the system's state transition probability matrix and system cost function.

$$\begin{bmatrix}u(k)\\u(h)\end{bmatrix}=\begin{bmatrix}f(k,h)+w(k)\\\mathrm{LST}(k)\end{bmatrix}, \tag{11}$$

$$\frac{\partial^2\alpha}{\partial x^2}+\frac{\partial^2\alpha}{\partial y^2}+\frac{\partial^2\alpha}{\partial z^2}=f(x). \tag{12}$$

Therefore, point P is the intersection of the two rays. If the geometric positions of the two cameras are known and the cameras are linear, then the position of the object in space can be calculated using the principle of triangulation. The above-mentioned differences between the extracted descriptor and the detection feature for detection are mainly due to the difference in the predictor used in the postprocessing and the different discriminant criteria, which lead to the original feature in the detection problem. The problem of

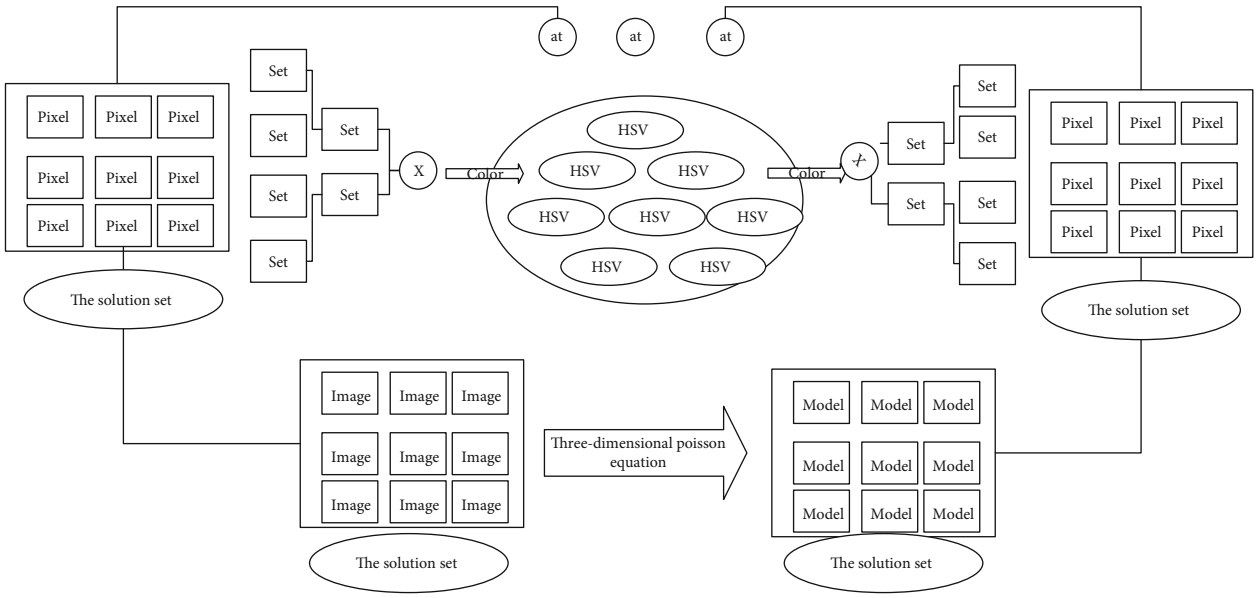

Figure 1: Distribution of the solution set of the three-dimensional Poisson equation.

motion tracking is different. Figure 2 is the flow structure of the digital video algorithm.

After multiple cycles of processing, the grid becomes thicker layer by layer until the thickest layer. At this time, various frequency components have been attenuated, and then, we start from the thickest grid and then return to the finer grids at all levels. And finally, we obtain the solution of the required equation on the finest grid. Therefore, the multigrid method requires an iterative solution on a series of grids of different sizes.

$$\begin{cases} \dfrac{\alpha(x) - \alpha(x-1)}{\sum a(t) \times w(x)} = 1, \\ \dfrac{\beta(x) - \beta(x-1)}{\sum b(t) \times w(x)} = 1. \end{cases} \tag{13}$$

Each iteration of the coarse grid provides a more accurate error correction result for the fine grid of the next layer. For the multigrid V-cycle algorithm, in each cycle, only one iteration (presmoothing) is performed before the error residual is limited to the next layer of the coarse grid, and the error correction result is interpolated and returned to the next.

$$g(x,k) = \lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{i=1}^{n} (1 - \alpha(x))(1 - \beta(x)) \times x(t), \tag{14}$$

$$f(x+1) - f(x) = \partial \frac{y(x,n)}{\|x(n)\|^2} * x(n). \tag{15}$$

There is only one iteration before the fine mesh layer (postsmoothing). The two-dimensional image coordinates of the joint points and the known depth coordinates of the three-dimensional skeleton model of the skeleton human body in the image sequence are manually obtained. Then, using the a priori data of the perspective projection relationship and the length of the human skeleton, the coordinates of each three-dimensional feature point on the human model are solved, and the three-dimensional human motion skeleton sequence under the perspective projection is established. This method is not aimed at specific human motion modes (simple motion modes such as walking and jumping) and can analyze large-scale motions of various parts of the human body in a complex and changing background and has the advantages of wide sources of motion information.

3.3. Extraction of Human Motion Information. Once the corresponding relationship of human body motion information is determined, the position of the object points represented by these image points in space can be easily calculated. But for a given matching primitive in one picture, more than one possible primitive can often be found to match with it in another picture, which leads to the problem of ambiguity in matching or false target of matching. For a multicamera tracking system, it is necessary to determine which camera (several) or which image (several) to use at each moment.

The movements of the human body can simply describe the combination of different movements of the torso and limbs of an adult's body. Since the human body's torso can mostly be regarded as approximately static when the human body is moving, and the limbs have a greater range of motion relative to the human body's torso, the main reason for the diversification of human movements is the diversification of the movements of the human limbs. Then, the features extracted from the movements of human limbs can approximate the features of human body movements. Figure 3 shows the distribution of human motion information extraction factors.

The purpose of motion tracking is to establish corresponding feature matching between consecutive image frames in a video sequence and obtain continuous

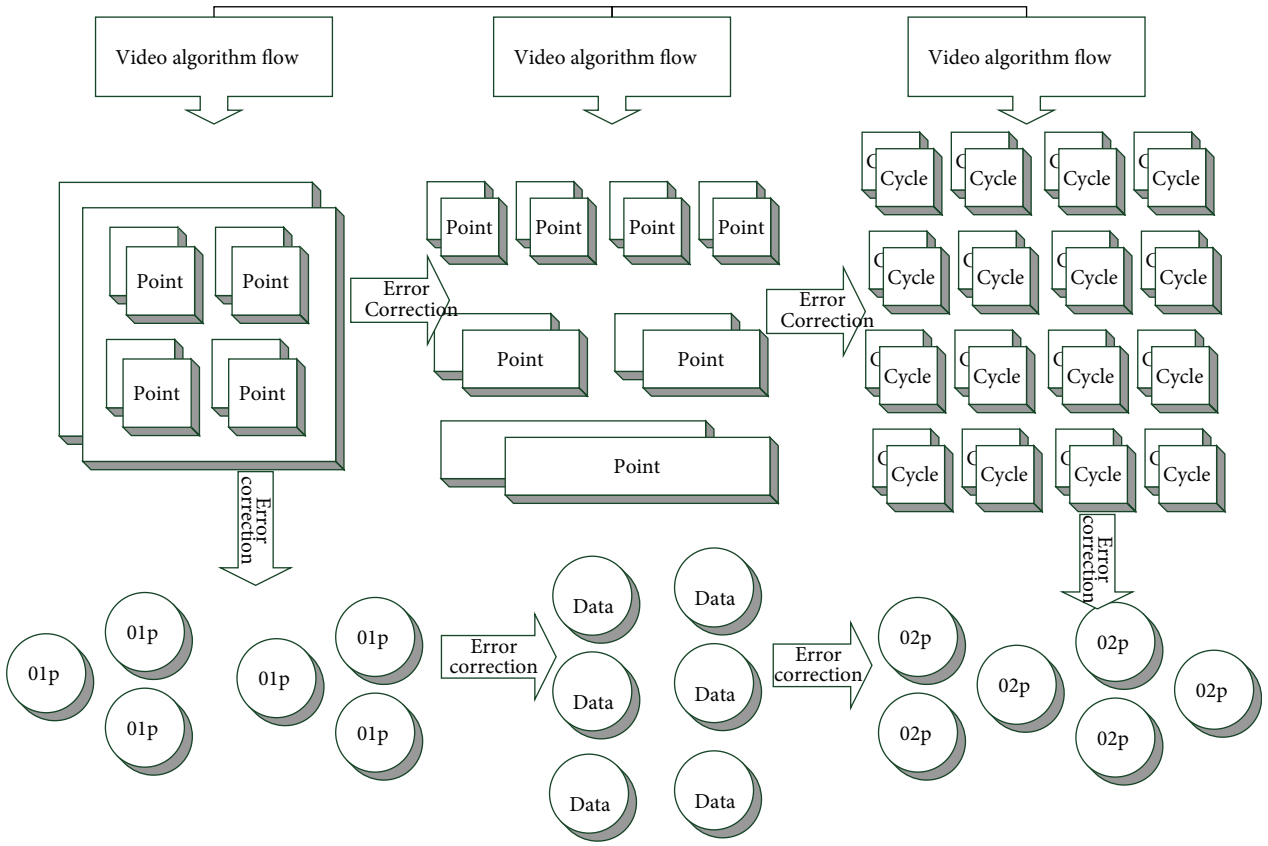

Figure 2: Digital video algorithm flow structure.

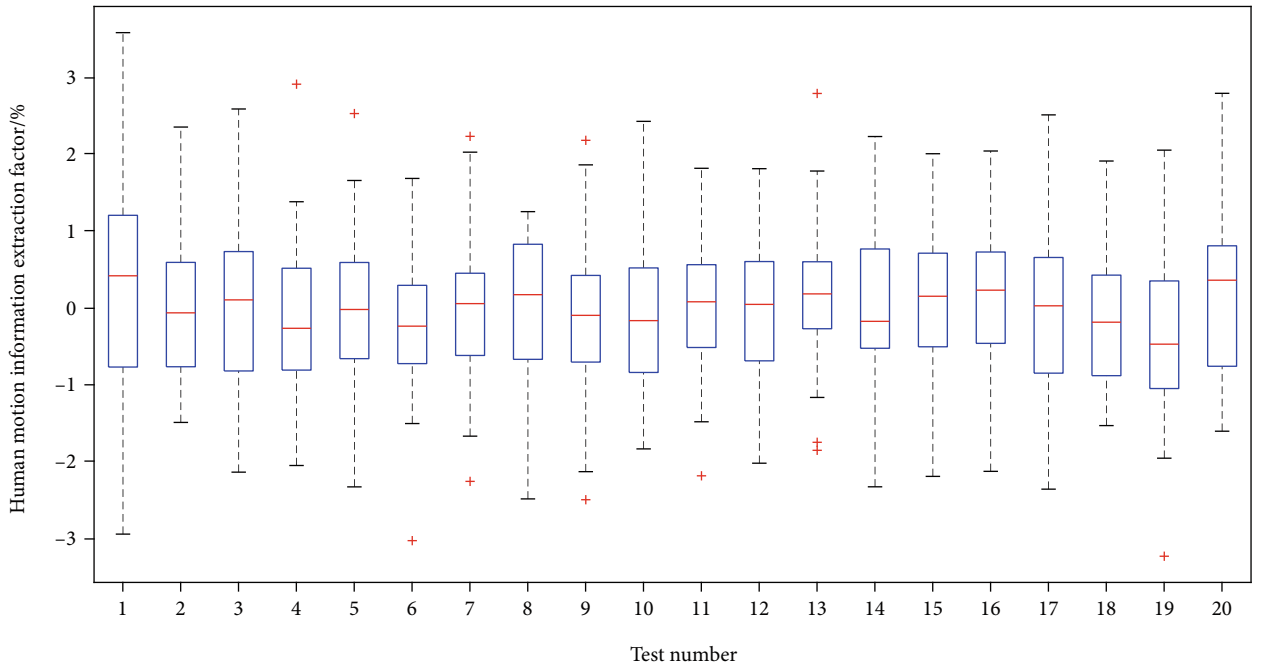

Figure 3: Human motion information extraction factor distribution.

information about position, speed, shape, texture, or color. These tracked and matched features can be points, lines, image blocks, contours, and 3D feature elements. Human motion tracking is to establish the correspondence between these features of the human body in the image sequence. Human motion tracking is to establish the corresponding

relationship of the human body features in the image sequence. Tracking is the basis of human motion acquisition and behavior understanding. What we want to establish is a classification rule; for any sample with unknown category, this rule can be applied to its feature vector x to determine the category the sample belongs to.

It is the continuous and comprehensive trend estimation of the specific real value in the real number space. In this problem, the corresponding label is a continuous space, such as human movement. The machine learning method of human body motion tracking is to estimate the value of the three-dimensional posture Y corresponding to the image space x. Due to the influence of the image noise of the video and the complexity of the human movement in the video, the tracking of the human movement is extremely difficult. At present, there is no ideal and universal tracking method for human body movement. Because of this, for complex human movements, manual calibration of joint points is still not a guide.

3.4. Model Weight Factor Distribution. According to experience, the Sobel operator is better than the Prewitt operator in the accuracy of edge detection. The Sobel edge detection operator is based on the first-order derivative, and the effect of this operator in smoothing noise is very good. This is because the operator adds the operation of local averaging to the image, which can greatly reduce the influence of noise. The effect of image processing shows that this operator has relatively accurate positioning of the edge of the image, but it is susceptible to noise. It is very suitable for image segmentation with obvious edge information and relatively little noise. The Robert operator is used to detect the local differences of edges, and the smoothness of the edge information obtained after the image is processed is poor.

And the response is wide, so when using the Robert operator for edge detection, it is necessary to refine the image to improve the accuracy of edge positioning. The template of the Sobel operator consists of two sets of 3×3 matrices, which represent the detection of vertical and horizontal edges, respectively. By convolving the two templates with the image, the approximate value of the difference in the two directions can be obtained. Figure 4 shows the fan-shaped distribution of model weighting factors.

Image data has a discontinuous characteristic, which is reflected by the edges of the image. For depth images, at the edges of the image, the depth values corresponding to pixels have changed. The Sobel operator finds the point adjacent to the pixel and then weights its gray value to calculate the gray value of a point. The edge detection is performed by setting the threshold of the edge point. Compared with the general algorithm that seeks the average first and then performs the difference, the Sobel operator has a better suppression effect on independent noise points.

It can provide more accurate edge direction information, and the edge of the image has greater brightness, making it easier to identify. The characteristics of edge information include amplitude and direction. Along the trend of the edge curve, the pixels change smoothly, while when the trend of the vertical edge curve is vertical, the pixels change drastically, and this drastic change may be a slope shape or a step shape. In actual processing, the edge detection operator is often used to detect the presence or absence of an edge and its direction.

4. Application and Analysis of Digital Video Human Motion Information Extraction Model Based on Three-Dimensional Poisson Equation

4.1. Data Analysis of Three-Dimensional Poisson Equation. When processing the three-dimensional Poisson equation data, the experiment quantified the angle tracking error of the joint points of the human body in the motion tracking problem and used more than 2000 images generated by the POSER software for training and testing. Among them, 1927 silhouettes are used for training data, and the other 418 images are used for test data. At the same time, the corresponding PoserBVH 3D motion data is generated as the corresponding data label. It is composed of 15 joint points and 3 degrees of freedom for each joint, plus 1 global direction variable to show the relative direction of movement, a total of 55 data dimensions. For the test data, the human posture results estimated by the motion tracking are stored in an XML page file and placed on the server of the laboratory for processing.

Then, the Bellman equation shown can be used to obtain the optimal solution of the problem P1 through the classical value iteration algorithm or the strategy iteration algorithm. The error between the estimated value of each frame and the actual motion posture of the motion capture database can be obtained. At the same time, artificial parameter adjustment (tuning) is avoided to make the result tend to be optimal. Unknown true values and a unified quantification platform are also the best database suitable for human motion analysis. Figure 5 shows the data node distribution of the three-dimensional Poisson equation.

It can be seen from the results that for some frames in the video database, a single Gaussian model is used to model the background, and then, an adaptive background subtraction method is used to extract the foreground area, and then, the shadow area is eliminated according to the characteristics of the smaller chromaticity change of the shadow area. The method can extract the moving human body very well. The background part of the video in the Weizmann video database is mainly geometrical shapes such as walls, roads, and a small part of the window. The background is relatively simple, so the single Gaussian model is used to model the background of the video in the database, and then, using the background subtraction method to extract the target is feasible in theory, and the actual effect is also possible.

The main reason is now analyzed as follows: the more the decomposition layer, the more the characterization ability of the characteristics of the low-frequency approximation coefficients of the previous layer is lost in the extraction of the descriptor which is far greater than the

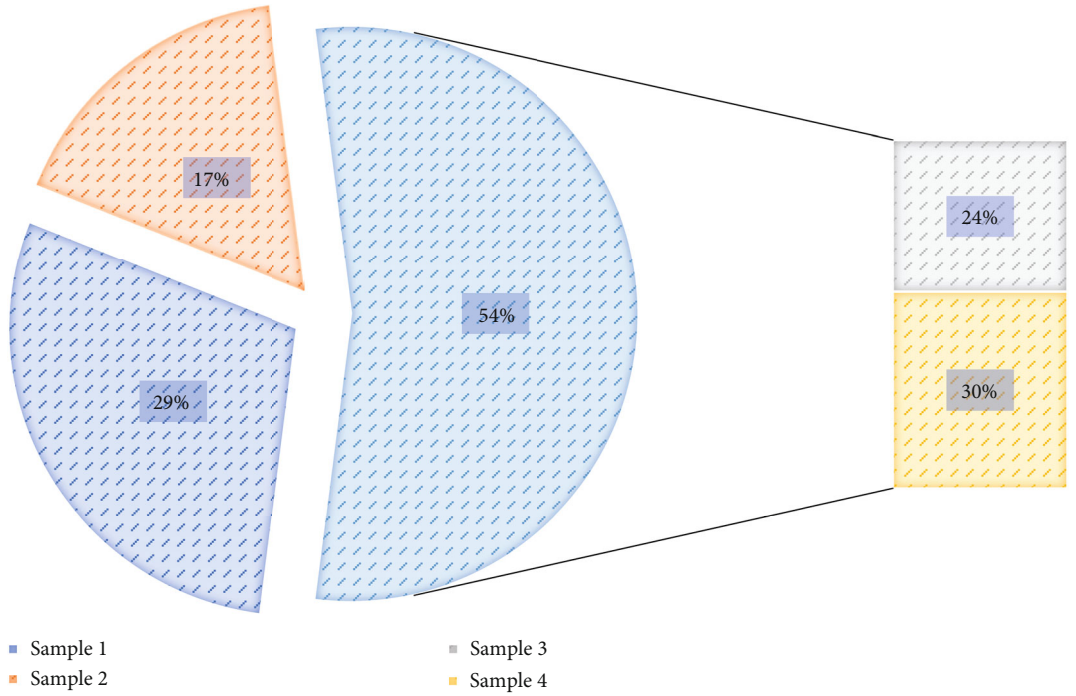

Figure 4: Fan-shaped distribution of model weight factors.

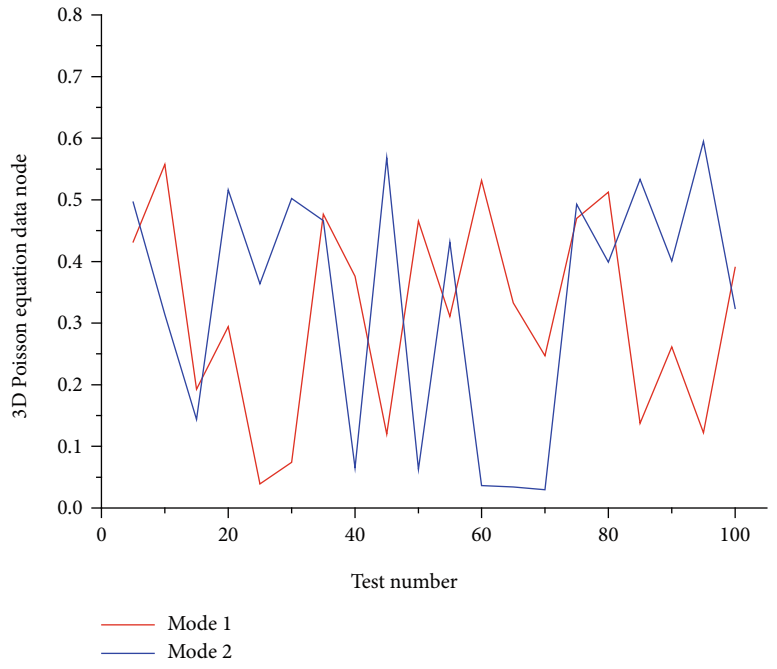

Figure 5: Data node distribution of 3D Poisson equation.

characteristic characterization ability of the obtained high-frequency detail coefficients. Using only one layer of two-dimensional wavelet transform is also conducive to maintaining a uniform quantization interval for selecting the scale range of the feature during the regression mapping process of the tracked predictor, and avoiding the instability caused by the excessively large variation interval of the kernel parameters.

4.2. Human Motion Information Extraction Model Simulation. In the learning classification experiment, the 93 samples are divided into 9 parts according to different people, and each part is all the actions performed by one person. Then, for these 9 sample sets, we select 1, 2, 4, and 8 sample sets as the training sample sets to train the AdaBoost classifier and use the remaining sample sets are used as the final test sample sets. For the experiment of selecting

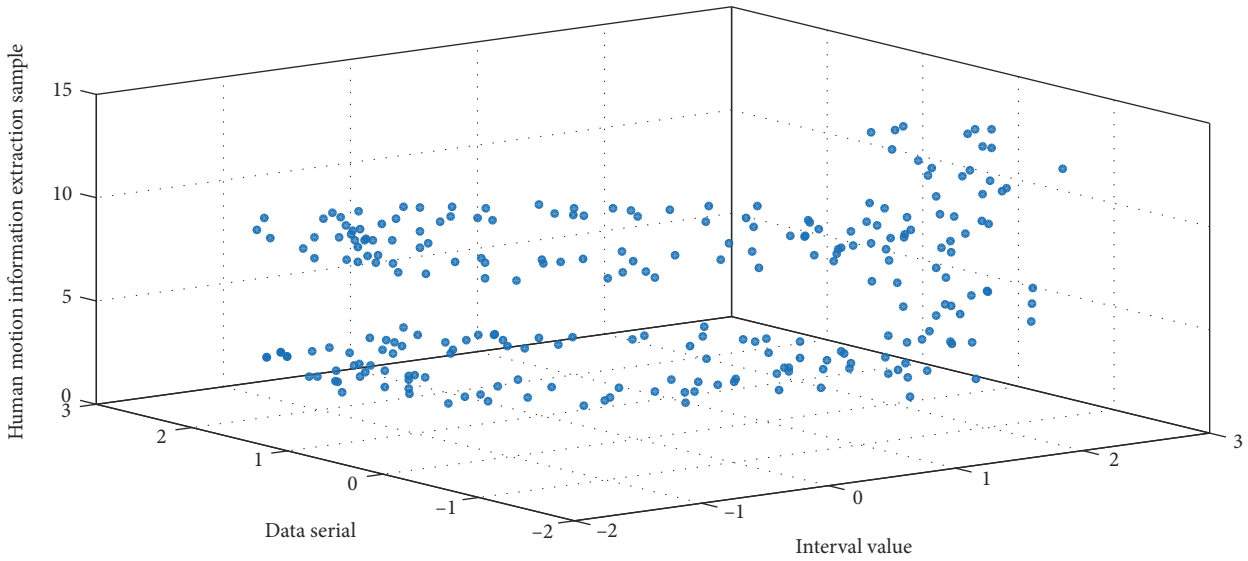

Figure 6: Sample distribution of human body motion information extraction.

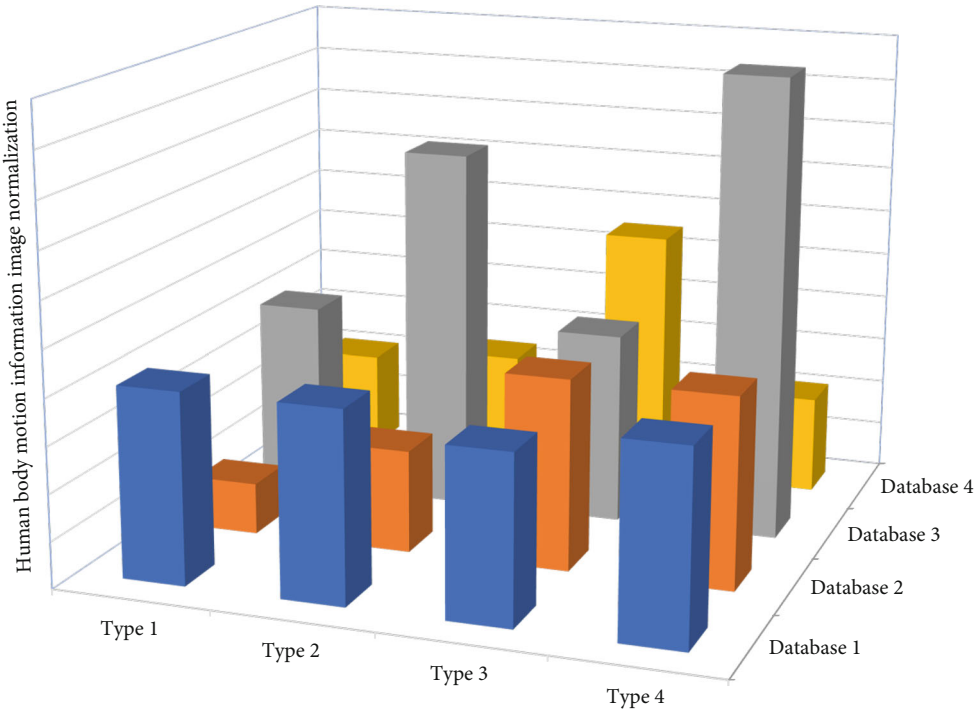

Figure 7: Image normalization processing of human motion information.

1 and 8 samples as the training sample set, separate experiments were carried out on 9 different classification situations. Each sample is used as a training sample to perform an experiment; that is, for an experiment that takes 1 sample as a training sample, for 9 then, the correct rate of 9 experiments is averaged as the classification correct rate when 1 sample is selected as a training sample.

However, in the real network environment, it is difficult to obtain as the main user of the agent. Similarly, for the case where 8 samples are selected as training samples, 9 experiments are also carried out to calculate the classification accuracy rate. For the case of selecting 2 or 4 samples as training samples, because there are many combinations, only some combinations are selected for experiment, covering most of the samples. Figure 6 shows the sample distribution of human body motion information extraction.

It can be observed that when a person's center of gravity is relatively low, the shadow of the human body is relatively obvious, and the change in gray information is relatively large compared with when there is no shadow. Therefore, when extracting foreground information, the shadow parts cannot be eliminated by setting the threshold, and these

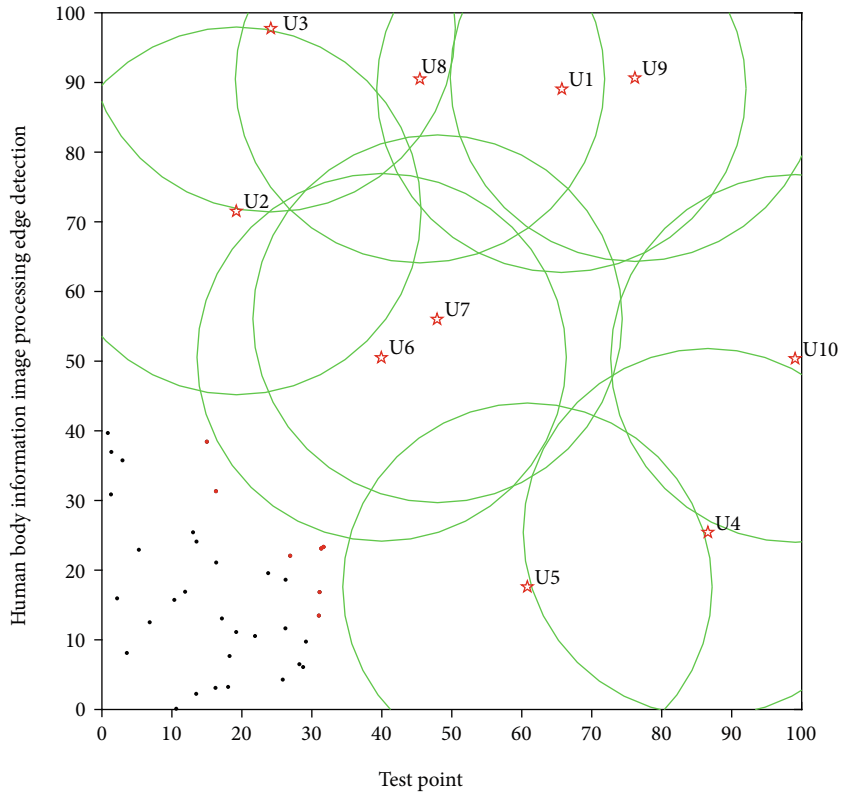

Figure 8: Human body information image processing edge detection.

shadow parts are also judged as the foreground. Since the chromaticity change of the shadow part is small, the shadow part can be separated from the human body part in the foreground based on this point. For some frames where the shadows have been removed after the foreground extraction, after the shadow elimination process, the extracted human body part and the background are different in color except for the color difference, and the human body part is still judged as the foreground. Therefore, after the shadow elimination process, the human body has not changed much. Therefore, the shadow removal method can effectively remove lighter shadows from the foreground. Figure 7 shows the normalization of human motion information images.

There will be a certain error in the depth coordinate of the known point obtained by measuring the depth value. In order to verify that this error will not affect the reconstructed 3D posture, we adjust the distance between the camera and the measured object during image acquisition, and shoot three images at depth values of 1280 mm, 1300 mm, and 1320 mm. But in the process of calculating the three-dimensional coordinates, the depth value is still 1300 mm for calculation, and the three groups of reconstructed line segment lengths and their angles are obtained, as shown in the text. When the depth coordinates of the point with a known depth value take three different values, compare the length of the reconstructed line segment and the value of the angle L and n between the line segments.

Therefore, we can think that the reconstruction result is not sensitive to the error of the depth coordinate value. In image coding, T is used to control the compression ratio. On the other hand, the selection of the quantized value affects the number of coefficients greater than T during the one-dimensional wavelet transform in a certain direction when searching for the optimal direction of the geometric flow. Therefore, choosing T to be too large or too small is not conducive to finding the best geometric flow direction.

4.3. Example Application and Analysis. By selecting the block grid size, the feature can be reduced to the lowest 144 dimensions (corresponding 4×4 size, 9-direction grid). Of course, the selection of the number of grids will affect the actual extraction effect of the descriptor. We further verified that the 5×6 size feature extraction is suitable for its later operation, because the extracted human head is large and it is easy to get a complete contour detection. The basic feature extraction process is as follows: by extracting the foreground and cutting the background of the original image, then binarizing the image to find the corresponding silhouette as the preprocessed image, and then further processing the image to eliminate the redundant shadow noise part and extract the final, which is encoded by HOG descriptors.

These are perfect a prior knowledge about renewable energy, computing tasks, channel status, etc. The edge information usually marks the end and the beginning of the area, and the area and the edge represent the basic image features, and many other features of the image can be obtained by deriving the basic features. The early processing content in the field of image processing included edge detection. Figure 8 shows the edge detection of human body information image processing.

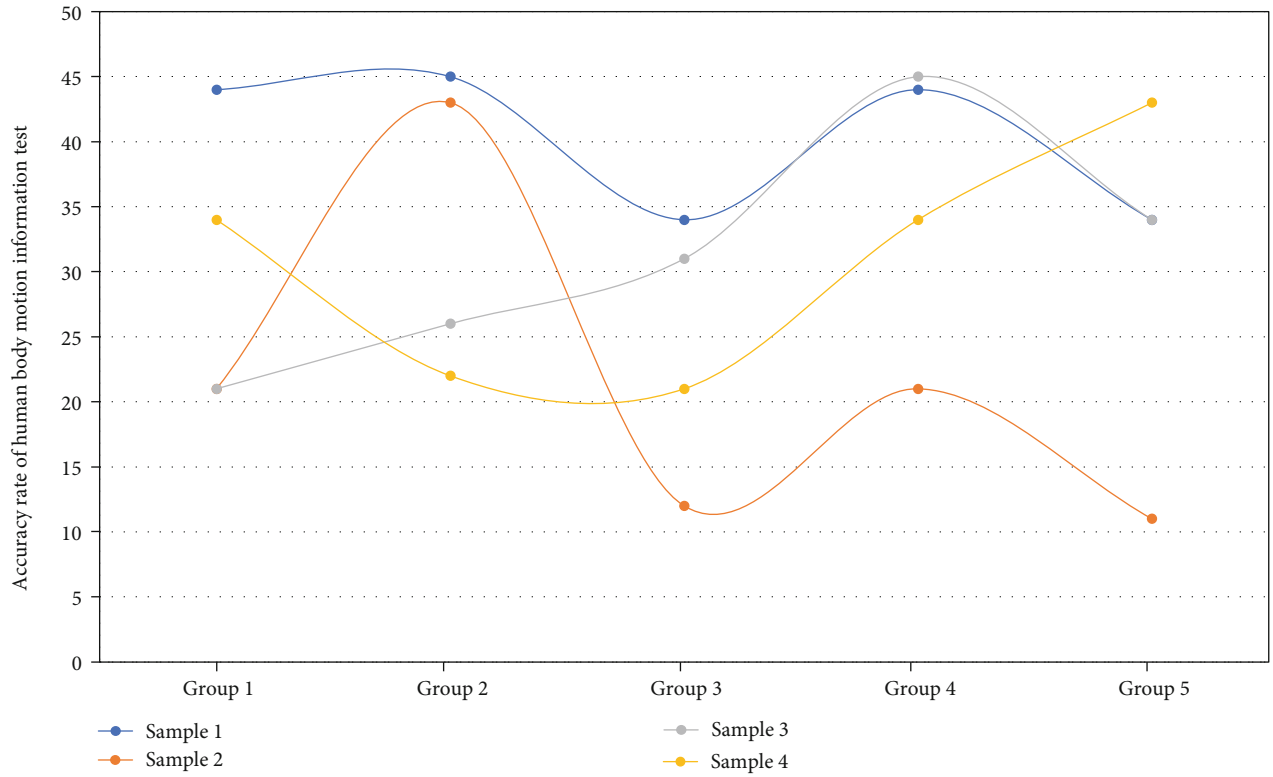

Figure 9: Comparison of the accuracy of human body motion information test.

It can be seen that in the case of a small number of training samples, as the number of iterations of the AdaBoost classifier increases, when the number of iterations is relatively small, such as the first 5, 10, and 15 times, the classification result improves significantly. It shows that when the AdaBoost classifier first iterated, due to the small number of training samples (1 sample for each action), the difference between each action could not be effectively obtained. After several iterations, the differences between the actions began to show up, which made the test accuracy rate increase significantly. As the number of iterations continues to increase, for a small number of training samples, the difference between the actions is basically determined, so the improvement of the classification result is not obvious at the beginning, and the curve gradually becomes flat. For the jump in the middle of 40 iterations, the preliminary analysis is caused by the misjudgment of some similar actions in the case of a small number of samples. Figure 9 is a comparison of the accuracy of human body motion information testing.

For the AdaBoost classifier, a weak classifier is generated for each iteration, and the final strong classifier is formed by fusing together the weak classifiers generated in all iterations according to the error obtained in each iteration. And each iteration is based on the results of the previous iterations, redistributing the samples to generate a new classifier. And each iteration will focus on the samples that were misclassified in the previous iteration, so as the number of iterations of the AdaBoost classifier increases, the result of each iteration will gradually approach the actual situation; that is, the iteration accuracy will become more and more accurately high. In the end, the accuracy of the classification test of the AdaBoost strong classifier will also increase as the number of iterations increases.

Analyzing the main reason, it may be that the original image itself has little difference in the image space, which causes the textures to be too similar during feature extraction, and the differences cannot be distinguished well. But overall, the error level has been controlled at an average of about 50 mm in this sequence, which has reached the current good quantitative index. In fact, there is almost no large posture difference between adjacent image frames in human vision.

5. Conclusion

Based on the existing research, this paper has conducted an in-depth analysis of the motion capture algorithm and technology. Aiming at the shortcomings of the current hot spot motion capture method, a motion capture algorithm combining depth map and 3D model is proposed. First, we use Kinect to collect the depth image, remove the background from the depth image, recover the 3D model from it, and then build the 3D model database. Secondly, the statistical characteristics of the Bandelet in the optimized Bandelet transform are used as the characteristics of the image to return to learning and predict the 3D human movement posture in the monocular video image. Experiments have verified the optimal parameters and additional statistical features of Bandelet transform when it is used in the extraction of human image features, and then various regression methods are used for parameter learning, and the prediction performance and human tracking tests are carried out. In

view of the single target, single lens, and simple background characteristics of the Weizmann database, the background subtraction method of the single Gaussian model was selected to extract the human body motion area, and a relatively complete human motion silhouette was obtained, and the shadow area was used to eliminate the shadow by the feature of small chromaticity change. The simulation experiment shows that the use of the feature representation based on the efficient matching kernel obtains excellent results. The equations are learned to obtain the mapping relationship between image features and three-dimensional poses, which effectively reduces the time complexity of motion estimation.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Qian, J. Zhou, Y. Mao, and Y. Yuan, "Recognizing human actions from silhouettes described with weighted distance metric and kinematics," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 21889–21910, 2017.

[2] H. Pan, J. Li, H. Wang, and K. Zhang, "Biomechanical analysis of shooting performance for basketball players based on computer vision," *Journal of Physics: Conference Series*, vol. 2024, no. 1, article 012016, 2021.

[3] X. Fan, B. Zhou, and H. H. Wang, "Urban landscape ecological design and stereo vision based on 3D mesh simplification algorithm and artificial intelligence," *Neural Processing Letters*, vol. 53, no. 4, pp. 2421–2437, 2021.

[4] J. Zhang, Z. Ren, W. Hu et al., "Voxelated three-dimensional miniature magnetic soft machines via multimaterial heterogeneous assembly," *Science robotics*, vol. 6, no. 53, p. 14, 2021.

[5] M. Deng, C. Wang, and T. Zheng, "Individual identification using a gait dynamics graph," *Pattern Recognition*, vol. 83, pp. 287–298, 2018.

[6] S. Jeon, A. K. Hoshiar, K. Kim et al., "A magnetically controlled soft microrobot steering a guidewire in a three-dimensional phantom vascular network," *Soft Robotics*, vol. 6, no. 1, pp. 54–68, 2019.

[7] M. Bao, M. Cong, S. Grabli, and R. Fedkiw, "High-quality face capture using anatomical muscles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10802–10811, Long Beach, CA, USA, 2019.

[8] Y. Zhang, L. Chen, F. Tan, S. Wang, and B. Yin, "An improved $$\ell _1$$ median model for extracting 3D human body curve-skeleton," *Multimedia Tools and Applications*, vol. 80, no. 24, pp. 33547–33571, 2021.

[9] B. H. Kim, K. Li, J. T. Kim et al., "Three-dimensional electronic microfliers inspired by wind-dispersed seeds," *Nature*, vol. 597, no. 7877, pp. 503–510, 2021.

[10] L. C. C. Bergamasco and F. L. S. Nunes, "Intelligent retrieval and classification in three-dimensional biomedical images – a systematic mapping," *Computer Science Review*, vol. 31, pp. 19–38, 2019.

[11] L. A. Huet, J. W. Rudnicki, and M. J. Z. Hartmann, "Tactile sensing with whiskers of various shapes: determining the three-dimensional location of object contact based on mechanical signals at the whisker base," *Soft Robotics*, vol. 4, no. 2, pp. 88–102, 2017.

[12] A. S. Keceli, "Viewpoint projection based deep feature learning for single and dyadic action recognition," *Expert Systems with Applications*, vol. 104, pp. 235–243, 2018.

[13] T. Stuart, K. A. Kasper, I. C. Iwerunmor et al., "Biosymbiotic, personalized, and digitally manufactured wireless devices for indefinite collection of high-fidelity biosignals," *Science Advances*, vol. 7, no. 41, p. eabj3269, 2021.

[14] J. Li, Q. Wang, S. Li, Q. Zhou, and Q. Zhong, "Face replacement and image animation system in cultural experience," *Journal of Physics: Conference Series*, vol. 2010, no. 1, article 012144, 2021.

[15] S. Blair, M. Garcia, T. Davis et al., "Hexachromatic bioinspired camera for image-guided cancer surgery," *Science Translational Medicine*, vol. 13, no. 592, 2021.

[16] T. Georgiou, Y. Liu, W. Chen, and M. Lew, "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 3, pp. 135–170, 2020.

[17] J. Laviada, A. Arboleya-Arboleya, Y. Álvarez, B. González-Valdés, and F. Las-Heras, "Multiview three-dimensional reconstruction by millimetre-wave portable camera," *Scientific Reports*, vol. 7, no. 1, pp. 10-11, 2017.

[18] Y. Li, R. Xia, Q. Huang, W. Xie, and X. Li, "Survey of spatio-temporal interest point detection algorithms in video," *IEEE Access*, vol. 5, pp. 10323–10331, 2017.

[19] X. Hu, X. Zeng, J. Liu, T. Peng, R. He, and C. Chen, "Research and implementation of 3D reconstruction algorithm for multi-angle monocular garment image," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 1068–1072, Dalian, China, 2020.

[20] C. Chen, C. Zhang, T. Wang et al., "Monitoring of assembly process using deep learning technology," *Sensors*, vol. 20, no. 15, p. 4208, 2020.