Research Article

Real-Time Dance Posture Tracking Method Based on Lightweight Network

Zhigang Wang

School of Music and Dance, Guangzhou University, Guangzhou 510006, China

Correspondence should be addressed to Zhigang Wang; ywwangzhigang@gzhu.edu.cn

Received 21 December 2021; Revised 4 January 2022; Accepted 5 January 2022; Published 28 January 2022

Academic Editor: Xin Ning

Copyright © 2022 Zhigang Wang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Video analysis of human motion has been widely used in intelligent monitoring, sports analysis, and virtual reality as a research hotspot in computer vision. It is necessary to decompose and track the movements in the process of movement in order to improve the training quality in dance training. The traditional motion tracking decomposition method, on the other hand, is unable to calculate the visual changes of adjacent key nodes, and the contour of 3D visual motion tracking remains ambiguous. This paper applies the human posture estimation algorithm in computer vision to the detection of key points of rectangular objects and obtains the heat map of key points of rectangular objects by adding a lightweight feature extraction network and a feature pyramid layer integrating multilayer semantic information, on the basis of summarizing and analyzing related research work at home and abroad. Because of the fusion of multilayer information, the network's design not only reduces the amount of calculation and parameters but also improves the accuracy of the final detection result. The test results show that the proposed algorithm's recognition accuracy has improved.

1. Introduction

Motion recognition is a very challenging topic in the field of computer vision, and it has become a very popular research direction in recent years because of its high research value [1]. When discussing a certain action or technical key points in dance teaching, dance teachers should make corresponding actions and give corresponding explanations, which causes a lot of inconvenience [2]. The traditional basic dance training adopts the uncalibrated global visual feedback method, that is, the traditional video decomposition training. When the visual motion is decomposed, the visual changes of adjacent key nodes cannot be calculated, and there are problems such as large tracking error of 3D visual motion and unclear decomposition outline. Traditional dance pose estimation and key point detection techniques mostly rely on complex image processing techniques and postprocessing skills, and the inference speed and accuracy are low. With the rapid development of deep learning (DL) and computer vision [3,4], lightweight network has gradually become the main development direction of attitude estimation and key point detection, which has more accurate

results and more possibilities than traditional methods. At present, the research on the combination of action recognition technology and dance action is still in its infancy. At the same time, due to the high complexity of dance action and the problems of human self-shielding when performing dance, the research progress in dance video action recognition is relatively slow.

In the current field of computer vision research, motion recognition is a very difficult subject. Its goal is to recognize human motion in video data using image processing and classification recognition technology [5]. Human posture estimation is the technology for inputting pictures or videos, detecting and locating key points of human bones in pictures or videos, and outputting key points of human bones in pictures or videos. With the advancement of video-based research in the field of vision, computer graphics and vision researchers can obtain dance motion data from the contour changes of dance movements and analyze the motion changes of dancers' joints, guiding and correcting dance trainers' movements and enhancing dancers' learning effect [6]. Action recognition has become a popular research topic in recent years due to its high research value. It has attracted

a large number of scientific research institutions and scholars to conduct research in this area in recent years. Action recognition technology can be used in a variety of video scenes [7]. Traditional dynamic recognition methods have a low recognition rate, are unable to recognize joint changes in detailed dance movements, and have a poor recognition effect [8]. A real-time dance posture tracking method based on lightweight networks is proposed to address the low recognition rate of traditional dance gait contour recognition methods.

As one of the most popular research directions in the field of computer vision, motion recognition has been applied to intelligent human-computer interaction, virtual reality, motion aided analysis, video surveillance, video retrieval, and so on [9]. At present, a large number of application achievements based on action recognition have appeared in the field of virtual reality, which has brought subversive changes to some scene applications in our life [10]. Dance pose estimation is a very important and difficult field in computer vision. The successful application of motion recognition technology in other fields also provides a sufficient theoretical basis for us to apply it to dance video motion recognition [11]. In view of the current large number of music and dance video materials, professionals need to spend a lot of time analyzing these dance video materials through listening and looking, which is undoubtedly very inefficient. If the motion recognition technology is applied to the analysis of these music and dance videos, so as to obtain organically connected music and dance action clips, it can not only reduce the work intensity of dance professionals and facilitate the retrieval of music and dance video data but also make the automatic dance arrangement system more efficient and the results more colorful [12]. This paper presents a lightweight dance pose estimation network. Firstly, the network selects the appropriate lightweight feature extraction network, reasonably reduces the redundant feature processing module when processing the feature map, and uses only one initial layer and one fine-tuning layer to generate the final key point heat map. The combination of several small convolution layers is used instead of large convolution layers, and the operation of hole convolution is added to improve the receptive field of the network. It makes the dance pose estimation network lighter and ensures the accuracy of key point detection.

2. Related Work

Literature [13] combines motion capture technology with Chinese puppet shows and proposes a technical scheme for puppet show digitization. The motion capture system was used to study the hip and trunk movement of golfers during swing, and the relationship between trunk and hip movement and swing was calculated, providing theoretical support for golf scientific training. Literature [14] used motion capture technology to create a virtual basketball training system that allows users to practice their own free throw courses. Literature [15] proposed a motion capture-based animation production method and process, as well as animation synthesis and the elimination of sliding steps in the

animation production process. Literature [16] studied the motion posture of the main joint points of the human body and proposed a method of human motion posture simulation, which can use human motion data to drive the virtual human model, based on the normal range of motion of human joints. According to the graphic structure framework method proposed in the literature [17], the appearance of human key points and spatial modeling components is critical to the model's final overall performance. To model each part of the human body, it uses the shape descriptor obtained through intensive sampling and the AdaBoost classifier trained differently. This method was very effective at the time, but it also had a lot of flaws. Its attitude estimation model was not based on image data and had a low level of robustness. A method for real-time rectangle detection in high-resolution images was proposed in the literature [18]. Using the rectangular fitting method proposed in this paper, this method proposes a new technology to improve the performance of the progressive probabilistic Hough transform, extract long line segments in the image, and detect rectangles. Although the final results show that the method has good real-time performance and robustness, the false edge problem caused by sundries in the image will still have a significant impact on the outcomes. Literature [19] uses video motion information to make the human posture in adjacent images form a loop between frames and then optimizes the loop by using the consistency of image information to eliminate the misestimated human posture samples, so as to obtain the optimal human posture. Literature [20] regards complex articulated human posture tracking as a special visual target tracking problem, realizes tracking by using time analytic Markov chain, realizes human posture optimization in a single frame by using space analytic Markov network, and realizes human posture tracking in video by joint training of two Markov models. Literature [21] proposed a global and local human posture expression model based on variable structure. The local layer focuses on visual target tracking based on motion features, and the global layer focuses on human posture optimization based on image features, so as to achieve the purpose of human posture online tracking. Literature [22] proposed to characterize human actions in video by combining motion history map and appearance information and achieved good results. Literature [23] combines global features and local features to identify human actions. The global features use the contour features based on the action energy graph, the local features use the target cells, and finally use the multiclass classification method of support vector machine to classify the feature points. Literature [24] expanded the motion history map from two-dimensional space to 3D space and proposed an action template for perspective independent human motion recognition. Inspired by gait energy map, literature [25] proposed cumulative action map to represent the temporal and spatial characteristics of action. Cumulative action map is the average value of image difference. At present, the research of lightweight neural network model still faces many problems and challenges. Firstly, the design space of lightweight network model is small, and the ability of feature representation is insufficient.

Secondly, the information capacity of lightweight network model is small, and the extracted feature information is less. Aiming at the problem of attitude estimation and key point detection based on lightweight network, this paper discusses the design and practical application of lightweight network for attitude estimation. This paper analyzes the “sparse connection” convolution operation commonly used in lightweight network model and proposes a real time dance posture tracking model based on lightweight network. In this paper’s dance estimation experiment, the key point detection experiment of rectangle is mainly carried out on the data set mixed by program synthesis data and real data. The experimental results show that the method proposed in this paper is feasible and efficient.

3. Methodology

3.1. Action Recognition Method. One of the traditional methods of dance pose estimation is the graphic structure framework. It uses a group of “parts” to represent an object. The arrangement of these “parts” is variable rather than rigid, and “parts” can be understood as the key points of the human body in the image. When the position and direction of pixels of human body shift due to motion, the graphic structure can model the joint part, thus realizing the task of attitude estimation [26]. The task of dancing posture tracking in video sequence is to estimate the specific dancing posture in each frame image by using video image information and to track the dancing posture change accurately in continuous video.

At present, there are two kinds of motion recognition methods: single-layer method and hierarchical method. Single-layer-based methods usually regard actions as characteristic categories of videos and use classifiers to identify actions in videos. The image sequence in the video is regarded as being generated by a specific action category. Therefore, the single-layer motion recognition method mainly involves how to represent videos and match them. The hierarchical method mainly identifies high-level actions by identifying simple actions or low-level atomic actions in the video. High-level complex actions can be decomposed into a sequence of subactions, and subactions can continue to be decomposed as high-level actions until they are decomposed into atomic actions. Figure 1 is a classification diagram of action recognition methods.

The spatiotemporal method, which primarily captures the temporal relationship of observations, is not the same as the single-layer sequential model method. Human movements are thus incorporated into observation sequences. An observation is usually linked to local or global features extracted from a frame or a set of frames. The methods based on sample sequence and the methods based on state model are the two main types of sequence model methods. The motion capture system uses a computer to process data, places trackers on key parts of moving objects, records the movement process of the objects with the motion capture system, and then uses the computer to obtain 3D spatial data. Mechanical motion capture, acoustic motion capture, electromagnetic motion capture, and optical motion capture

FIGURE 1: Classification diagram of action recognition method.

are the four types of motion capture [27]. All videos are made up of a time series of two-dimensional images in the spatiotemporal motion recognition method. A video can be thought of as a three-dimensional space-time body that contains all of the information needed for human motion recognition and matching. Most spatiotemporal volume-based methods currently use the entire 3D volume as a feature or template to match unknown actions in the video in order to classify it. Noise and meaningless background information, on the other hand, will affect the above methods.

Using human body capture marker points, multidirectional double camera space plane imaging system and computer analysis and calculation system to track and decompose the reverse 3D visual motion, introducing inverse kinematics to calculate key action lattice and vector parameters of each key action lattice, and using multidirectional double camera space plane imaging system to realize the 3D tracking and decomposition of dance visual motion [28]. Optical motion capture system gives performers plenty of space to perform freely, which is not limited by space and mechanical equipment and can capture high-speed movements or objects. The flow of 3D motion data capture is shown in Figure 2.

The sample sequence method represents an action by using a template sequence or a group of sample sequences, and there are no restrictions on how to extract observations. The comparison of a new input video with a template or action sample sequence is the focus of the sample sequence-based method. The dynamic regularization algorithm was widely used in the research of action recognition based on sample series, in which the similarity between input actions and action templates was measured by the correlation coefficient of reduced-dimension actions and was widely used in the research of action recognition based on sample series, in which the similarity between input actions and action templates was measured by the correlation coefficient of reduced-dimension actions. In contrast to the related methods described above, the hierarchical method based on description can explicitly model the spatiotemporal structure of human actions. As a result, without being limited to sequence actions, this method can identify sequences and simultaneous actions [29]. Human actions are represented as the appearance of some subactions in description-based methods, and these appearances must adhere to specific temporal, spatial, and local relationships. In description-based hierarchical

FIGURE 2: 3D motion data capture process.

action recognition methods, context-free grammar is commonly used to represent actions.

The extraction of video information usually depends on two steps. Firstly, relevant visual features in the video are extracted. Then, the extracted features are studied to generate corresponding description labels. In this technology, the most important thing is to extract features effectively, and DL algorithm is one of the most efficient methods to extract video features at present. However, traditionally, the extraction method based on this method pays more attention to the spatial domain of video, that is, the extraction of pixel information in video frames, but ignores the change of action state of video actions in time domain.

Most of the existing dance pose estimation networks pay attention to the accuracy of human body key point detection, while ignoring the calculation amount of the network. The lightweight network design makes the network parameters smaller and can be used on other lightweight devices. In this paper, motion information is used to strengthen the correlation between video frames, abundant image spatial information is used to optimize the dance posture pertinently, and a lightweight network is constructed to strengthen the recognition of the most flexible arm, so as to realize the high-precision tracking of the dance posture.

3.2. Design of Real-Time Dance Posture Tracking Method.

The initial feature map generated by the backbone network will be upsampled by linear interpolation to obtain a feature map equivalent to 1/4 of the size of the original map. Because of the small range of human key points, if the resolution of the feature map is small, it may be difficult to detect whether pixels correspond to human key points in the feature map. The second stage is the initial layer for generating the initial two-dimensional heat map. In this layer, the initial heat map corresponding to the key points of the human body is

obtained after a series of convolution operations by inputting the up-sampled characteristic map. The initial heat map has 16 channels, which, respectively, correspond to the 16 key points of the human body. Due to the low accuracy of the initial heat map, the fine-tuning layer will adjust the initial heat map in the third stage, so that the final heat map contains more information. The fine-tuning layer will connect the upsampled feature map with the heat map generated by the initial layer as the input, instead of just using the output of the initial layer as the input, and finally output the heat map of human key points predicted by the network.

It is necessary to express the human body contour in order to make better use of the image and motion features of the human body region in the video. Because the human body's outline is a complex and hinged figure shape, expressing the posture of the human body with a single model is impossible. In this chapter, the image contour corresponding to the corresponding limbs can be calculated using the joint points of human posture by analyzing and training the principal components of the contour, allowing the image features to be calculated more accurately. The spatial limit range of the 3D motion tracking system should be selected first, and then the extreme point of the 3D motion tracking system should be located in the process of motion tracking in basic dance training. The lattice data and vector parameters of each key action lattice are obtained in the extremum lattice, and the feature vectors of the key lattice are calculated, followed by the 3D tracking decomposition of the dance visual action using the multidirectional dual-camera space plane imaging system.

Let $I(x, y, \tau)$ be the variable Gaussian function of the visual action space, representing the dance action path and displacement, where τ is the spatial scale coordinate and (x, y) is the spatial plane coordinate point. The spatial extreme points of the detection of the 3D motion tracking system are as follows:

$$D(x, y, \tau) = \frac{\sqrt{I(x, y, \tau) - (x, y) \times O(x, y)}}{L(x, y, \tau)}, \quad (1)$$

where $O(x, y)$ is the initial image coordinates and $L(x, y, \tau)$ is generated by image convolution.

The relationship between the contour C_i of each part of the human body i (i represents one of the six human parts, $i \in [1, 2, \dots, 6]$) and the position of the joint point k_i of the human body and the pretrained human body shape model is

$$\begin{bmatrix} C_i \\ k_i \end{bmatrix} = B_i z_i + m_i. \quad (2)$$

In the formula, k_i represents the position of the joint point in the human body part i . The matrix B_i and the vector m_i represent the parameters of the human body shape model, which are obtained by training the 3D contour of the 3D model based on the principal component analysis method. The vector m_i represents the mean value of the contour of the part i and the positions of the joint points. The matrix B_i contains the main eigenvalues of the eigenvectors of the training data. z_i is the deformation coefficient, which

is calculated by inheriting the relative position deformation between the joints of the upper body of the human body during decoupling. The contours of the body parts after decoupling can be calculated using the changed joint point positions and contour model parameters, namely,

$$C_i = f(k_i | B_i, m_i, z_i). \quad (3)$$

The image features, motion features, and other information contained in a specific dance posture can be extracted from the image using the decoupled human posture contour. It can more effectively compare the features of the same human body part in adjacent frames in visual target tracking and obtain consistency information in the video, making dance posture tracking more targeted. To reduce the amount of data needed for judgment processing, the chosen image is tested twice, each time using the image block's space-time condition information to improve test accuracy. After that, using the image block marking matrix, the dance motion target is calculated as follows:

$$L_{IB}(m, n) = \begin{cases} 2, & \text{SAD}(m, n) > t_{\text{SAD}}, \\ 1, & \text{IBSCI}(m, n) > t_{\text{IBSCI}}, \\ 0, & \text{else,} \end{cases} \quad (4)$$

where L_{IB} represents the image block labeling matrix. m, n is the coordinate position of the dancer in the image. t_{SAD} represents the image block differential classification threshold. IBSCI represents the spatiotemporal condition information of the image block. t_{IBSCI} is its classification threshold. 1 means foreground; 0 means background; 2 means select background. When the selected image block of the video dance action is the foreground candidate area, the second detection is performed using the spatiotemporal condition information of the image block. Use the Camshift algorithm to calculate the gait contour target of the dance action by using the video image foreground after the inspection:

$$Z_k(x, y) = \begin{cases} 255, & |F_k(x, y) - B_k(x, y)| \geq \text{Th}, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

In the formula, $F_k(x, y)$ and $B_k(x, y)$ are the pixel gray values between the original video and the foreground candidate area, respectively.

The feature points and marking parameters for multi-key points and key movement parts in dance movements are set using the improved inverse kinematics method. The principle of visual movement tracking decomposition is used to compute the static replacement data of dance movements. Dynamic capture is carried out after the image segmentation marks are clear. If the image is not clearly segmented and marked, the coefficients are corrected and recalculated until the image segmentation marks of dance movements are clear. When a specific action is thoroughly analyzed, the system automatically tracks and decomposes it, and 3D action tracking and decomposition imaging can be carried out using the projection system or the teaching display without the need for the teacher to take any action.

The presentation process can be started at any time, allowing for indefinite tracking and decomposition of the movements in the basic dance training process.

Let $L'(a, b)$ be the Euclidean distance between the key mark points a and b of the dance movement. L is the characteristic distance threshold of the motion amplitude. $L'_{\min}$ is the closest Euclidean distance between the action feature point and the next action feature point. $\omega(W)$ is the length of each feature point, and $L''_{\min}$ is the subadjacent distance of the action amplitude. According to the following formula, the extreme points of the 3D motion tracking system are located:

$$\mu = \frac{L'_{\min} \times L''_{\min}}{L'(a, b)} \times \omega(W). \quad (6)$$

Let $Q(x, y)$ be the cumulative value of the gradient direction of the action feature point. $H(x, y)$ is the cluster center where the action feature point is located. $W(x, y)$ is the histogram of the gradient of acquiring action feature points. B_i is the normalized value of the feature vector length. $\varepsilon(\lambda)$ is the direction vector of the feature point. According to the following formula, to obtain the lattice vector parameters of each key action,

$$\eta(Q) = \frac{W(x, y) \times Q(x, y)}{H(x, y)} \times \varepsilon(\lambda) \times B_i. \quad (7)$$

Let $H(a, b, c)$ be the coordinates of the spatial motion feature point. W is the range of motion. A and B are the external and internal camera parameters. Calculate the feature vector of the key lattice according to

$$\partial''(x, y, z) = \frac{WK \times H(a, b, c)}{A \cdot B} \times \tau(o). \quad (8)$$

To complete visual motion tracking, the aforementioned spatial motion tracking decomposition principle is applied to basic dance training. The four corners of a rectangular object in the graph are detected in real time, and the required rectangular object is obtained by connecting the corners, allowing key point detection of rectangular objects on lightweight equipment to be realized. A suitable lightweight backbone network is chosen as the initial feature extraction layer, and a new module is added behind the backbone network to predict the two-dimensional heat map, similar to dance pose estimation. Despite the fact that rectangular objects have distinct key points, there are still differences in size, shape, and location on the map. To address this issue, we plan to add a feature pyramid to the heat map prediction module for multiscale fusion, which will take into account both low-level and high-level semantic information to provide more precise key point location.

4. Results Analysis and Discussion

At present, the research on the combination of motion recognition technology and dance has just started, and there are few available dance data sets. In order to verify the feasibility of the dance movement recognition algorithm in this paper, we use cross validation to evaluate the algorithm

in the experiment. Cross validation is a statistical method of cutting data samples into subsets. Its idea is to divide the original data set into training set and test set. Usually, the training set is used to train the classifier. After the training is completed, the test set is used to test the trained model, and the performance of the classifier, that is, the feasibility of the algorithm, is evaluated. Cross validation methods are mainly divided into K-fold cross validation and leave-one-behind cross validation.

In this experiment, we share two videos to verify the effectiveness of the algorithm. In the first video, the moving human body pasted a mark at the center of each joint, which made it easier for us to get the two-dimensional coordinate position of the joint more accurately. When the dancer is in positive phase, it means the whole process of the left foot (right foot) touching the ground, landing, and leaving the ground. On the contrary, when in the negative phase, it indicates the whole process of the right foot (left foot) touching the ground, landing on the ground, and leaving the ground. In actual tracking, by constantly changing the displacement model of the dance, we can track the information such as the step length and movement speed of the dance movement in the whole tracking space in real time, use the gait cycle model and displacement model for tracking, and finally get the footprint tracking as shown in Figure 3.

High-frequency marker points are suitable for small movements and inconvenient observation of skillful positions, such as finger and muscle exertion positions, because they have a high transmission frequency and are sensitive to position accuracy information. Low-frequency transmission has good anti-interference and transmission stability, making it ideal for transmitting information between major joints like the shoulder and knee. The intermediate frequency marker has good overall performance and is primarily used for body contour marking. It is suitable for the transmission and analysis of body contour information in motion. This paper uses the intermediate frequency marker as the main marker to ensure the accuracy of the data. A comparative experiment was used to determine the environmental variables, control condition variables, and characteristic data. For two experimental analyses, use the same dance movement and the same basic training course and compare the action tracking decomposition method developed in this paper to the traditional action tracking decomposition method. Judging by the decomposition success rate, the experimental results are shown in Figure 4.

According to the experimental data in Figure 4, with the increase of time, the decomposition success rate of traditional motion tracking decomposition method gradually decreases. However, the motion tracking decomposition method designed in this paper is not affected by time and complexity of motion because it uses human body to capture marker points and corresponding computing equipment.

Choose the same dancer, with the increasing difficulty coefficient of dancing. Under the same experimental environment, the action tracking decomposition method designed in this paper is compared with the traditional action tracking decomposition method. The contour

FIGURE 3: The final footprint tracking results.

FIGURE 4: Methods of the experimental results were compared.

extraction time is selected for judgment, and the experimental results are shown in Figure 5.

According to the experimental data in Figure 5, the decomposition success rate of the motion tracking decomposition method proposed in this paper increases with the complexity of the dance, the increase in contour extraction time is not noticeable, and the decomposition success rate tends to be stable. With increasing dance complexity, the contour extraction time increases linearly, and the decomposition success rate decreases linearly, according to the traditional motion tracking decomposition method. It demonstrates that the motion tracking decomposition method proposed in this paper can clearly track and decompose dance movements. On two dance data sets, we test the algorithm's recognition effect and all single features. The similarity of geometric shapes between objects can be

FIGURE 5: Methods of the experimental results were compared.

attributed to the similarity of dancers and standard movements in this paper. A feature plane is defined by three feature points, and the skeleton of human posture is primarily made up of seven feature planes. The angle relationship between joint points is used to compare dancer's action to standard action. The spine is the main axis of the human body, the spine is the z-axis of the spatial rectangular coordinate system, and the x-axis and Y-axis of the horizontal plane are the ground plane of the motion capture equipment, according to ergonomics. The standard analysis of human motion can be simplified as the comparison of the similarity of edge vectors in the same plane and the similarity of normal vectors between planes. Figure 6 shows the features on the DanceDB dataset and the experimental comparison results of this method.

The audio features are not affected by the target and background when performing dance movement classification, and the optical flow direction histogram feature characterizes the movement information of the dance movement, and the interference from the background and the target is less than the direction gradient histogram feature. Finally, the experimental results also show that the experimental results of this data set based on the feature fusion method are better than the results of a single feature. Figure 7 shows the comparison of the experimental results of this method and the benchmark method on DanceDB.

In the benchmark method based on trajectory feature fusion, the trajectory feature is not robust to the mixing of target and background, resulting in low recognition rate. The fusion method in this paper can reduce the influence of background mixing and other factors and ensure the recognition accuracy of dance movements to a certain extent.

Using the video time information, the tracking process calculates the motion information of the target area of each part of the human body, and the motion information is used to transfer the posture contour of each part of the human body from frame t to frame $t + 1$ as the candidate sample of human body posture in frame $t + 1$. Given that

FIGURE 6: The characteristics of the DanceDB dataset and the experimental comparison of the methods in this article.

FIGURE 7: Experimental comparison between the method in this paper and the benchmark method on DanceDB.

the method based on spatial information from a single frame image can better detect human parts with fixed shapes, such as the trunk and head, but the detection effect of the arm is poor, it is necessary to concentrate on resolving the arm detection problem. The guidelines are based on a 20-step dance gait contour data set that has been prepared. A successful identification is defined as the successful identification of landing time, landing time, footprint area, landing time, average pressure value of single frame, flight time, and maximum pressure value of single step, and single frame. Record the number of 20 groups of parameters that can be identified by the three methods to obtain the final recognition rate of the three identification methods, as shown in Figure 8.

According to the experimental results of recognition rate in Figure 8, the final average recognition rate of traditional recognition methods is lower with the increase of recognition data sets. With the increasing number of data sets, the

FIGURE 8: Experimental results of recognition rate of three methods.

dynamic recognition method of gait contour of dance movements in this paper has higher recognition rate and is more suitable for gait contour recognition of dance movements.

In this paper, a lightweight learning network is built and trained to extract additional candidate samples of human arm parts. The research shows that when the dance movements are too complicated and there are similar movements and self-occlusion, the trajectory features of the benchmark method based on trajectory feature fusion cannot accurately represent them. The fusion method proposed in this paper can mitigate the above influence to some extent, resulting in a higher recognition rate and greater recognition accuracy for dance movements. As a result, it also verifies the algorithm's effectiveness.

5. Conclusions

With the rapid development of information technology, people can use images or videos as information carriers. When studying the real-time tracking information of dance movements and postures, it can help to identify subtle contour changes in dance movements. The traditional real-time tracking method of dance posture based on the recognition rate is not high, and it is unable to identify the small changes in dance movements. However, the dance dynamic recognition method based on this lightweight network can keep the recognition accuracy and better recognition effect with the increasing of recognition parameters, which is of developmental significance for real-time tracking of dance posture. Based on the dance pose estimation in computer vision, this paper aims to design a pose estimation network that can run on lightweight equipment to detect key points of human body, make use of the principle of tracking and

decomposition of dance visual movements to deal with reasonable marking points of dance movements, and adopt perfect and accurate marking points. The marking points of each part are displayed on the computer operating system through calculation. Inverse kinematics is introduced to calculate the key action lattice and vector parameters of each key action lattice. The 3D tracking decomposition of dance movements is realized by using the multidirectional double camera space plane imaging system. In order to ensure the rationality of the design, it is verified by experimental data. The experimental results show that the method designed in this paper has the advantages of high tracking accuracy, fast decomposition rate, accurate and stable decomposition, clear outline, and so on. In the follow-up research, we can continue to optimize the network structure and improve the performance of the algorithm.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author does not have any possible conflicts of interest.

References

- [1] S. Gu, Z. Ma, M. Xie, and Z. Chen, "Online learning of mixture experts for real-time tracking," *IET Computer Vision*, vol. 10, no. 6, pp. 585–592, 2016.
- [2] D. Groat, H. J. Kwon, M. A. Grando, C. B. Cook, and B. Thompson, "Comparing real-time self-tracking and device-recorded exercise data in subjects with type 1 diabetes," *Applied Clinical Informatics*, vol. 9, no. 4, pp. 919–926, 2018.
- [3] W. Cai, Z. Wei, Y. Song, M. Li, and X. Yang, "Residual-capsule networks with threshold convolution for segmentation of wheat plantation rows in UAV images," *Multimedia Tools and Applications*, vol. 80, no. 21, pp. 32131–32147, 2021.
- [4] M. Zhao, A. Jha, Q. Liu et al., "Faster Mean-shift: GPU-accelerated clustering for cosine embedding-based cell segmentation and tracking," *Medical Image Analysis*, vol. 71, Article ID 102048, 2021.
- [5] C. Long, Y. Cai, W. Hai, and X. Chen, "Multilevel framework to handle object occlusions for real-time tracking," *IET Image Processing*, vol. 10, no. 11, pp. 885–892, 2017.
- [6] S. Kim, J. Nam, and B. Ko, "Fast depth estimation in a single image using lightweight efficient neural network," *Sensors*, vol. 19, no. 20, Article ID 4434, 2019.
- [7] G. Li, L. Yu, and S. Fei, "A deep-learning real-time visual SLAM system based on multi-task feature extraction network and self-supervised feature points," *Measurement*, vol. 168, no. 4–5, Article ID 108403, 2021.
- [8] Y. Wang, H. Tian, W. Cai et al., "Tracking hematopoietic precursor division ex vivo in real time," *Stem Cell Research & Therapy*, vol. 9, no. 1, Article ID 16, 2018.
- [9] L. Shi, B. Li, C. Kim, P. Kellnhofer, and W. Matusik, "Towards real-time photorealistic 3D holography with deep neural networks," *Nature*, vol. 591, no. 7849, pp. 234–239, 2021.
- [10] S. Tomer and G. Pini, "Tracking a non-cooperative target using real-time stereovision-based control: an experimental study," *Sensors*, vol. 17, no. 4, Article ID 735, 2017.

- [11] M. Chen, D. Parsons, X. Gu, and W. lu, "Real-time breathing phase identification using treatment beams, EPID, and deep network," *Medical Physics*, vol. 45, no. 6, pp. 235–236, 2018.
- [12] K. P. Tsoukatos and A. Giannoulis, "Lightweight power control for energy-utility optimization in wireless networks," *Ad Hoc Networks*, vol. 63, no. 8, pp. 91–103, 2017.
- [13] R. Fnu, V. Kulathumani, R. Kavi, I. Elwarfalli, V. Kecojevic, and A. Nimbarte, "Real-time drowsiness detection using wearable, lightweight brain sensing headbands," *IET Intelligent Transport Systems*, vol. 11, no. 5, pp. 255–263, 2017.
- [14] H. Luo, Y. Gao, Y. Wu, C. Liao, X. Yang, and K.-T. Cheng, "Real-time dense monocular SLAM with online adapted depth prediction network," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 470–483, 2019.
- [15] X. Fu, F. Zhu, F. Qi, and M. Wang, "Real-time depth camera tracking with geometrically stable weight algorithm," *Optical Engineering*, vol. 56, no. 3, Article ID 33104, 2017.
- [16] H. Liu, F. Cao, C. Wen, and Q. Zhang, "Lightweight multi-scale residual networks with attention for image super-resolution," *Knowledge-Based Systems*, vol. 203, no. 4, Article ID 106103, 2020.
- [17] C. Chwala, F. Keis, and H. Kunstmann, "Real-time data acquisition of commercial microwave link networks for hydrometeorological applications," *Atmospheric Measurement Techniques*, vol. 9, no. 3, pp. 991–999, 2016.
- [18] S. I. Jung, N. K. Lee, K. W. Kang, K. Kim, and D. Y. Lee, "The effect of smartphone usage time on posture and respiratory function," *Journal of Physical Therapy Science*, vol. 28, no. 1, pp. 186–189, 2016.
- [19] R. Huang and M. Sun, "Network algorithm real-time depth image 3D human recognition for augmented reality," *Journal of Real-Time Image Processing*, vol. 18, no. 2, pp. 307–319, 2021.
- [20] F. Zhang, A. Cully, and Y. Demiris, "Probabilistic real-time user posture tracking for personalized robot-assisted dressing," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 1–16, 2019.
- [21] H. Xu and Y. Zhu, "Real-time object tracking based on improved fully-convolutional siamese network," *Computers & Electrical Engineering*, vol. 86, no. 4, Article ID 106755, 2020.
- [22] Y. Xu, J. Wang, Y. Li, Z. Miao, and Y. Zhang, "One-step backtracking for occlusion detection in real-time visual tracking," *Electronics Letters*, vol. 53, no. 5, pp. 318–320, 2017.
- [23] A. Maher, H. Taha, and B. Zhang, "Realtime multi-aircraft tracking in aerial scene with deep orientation network," *Journal of Real-Time Image Processing*, vol. 15, no. 3, pp. 495–507, 2018.
- [24] A. Shankar, G. P. Greve, B. Wu, J. K. Thompson, and M. Holland, "Continuous real-time tracking of a quantum phase below the standard quantum limit," *Physical Review Letters*, vol. 122, no. 23, Article ID 233602, 2019.
- [25] A. Mb, A. Yn, B. St, J. Kilby, and G. Mawston, "Influencing lumbar posture through real-time biofeedback and its effects on the kinematics and kinetics of a repetitive lifting task," *Gait & Posture*, vol. 73, pp. 93–100, 2019.
- [26] W. Yang, W. Wang, X. Zhang, S. Sun, and Q. Liao, "Lightweight feature fusion network for single image super-resolution," *IEEE Signal Processing Letters*, vol. 26, no. 4, pp. 538–542, 2019.
- [27] Y. Kuai, G. Wen, and D. Li, "Multi-task hierarchical feature learning for real-time visual tracking," *IEEE Sensors Journal*, vol. 19, no. 5, pp. 1961–1968, 2019.
- [28] Y. Meng, Z. Deng, K. Zhao, Y. Xu, and H. Liu, "Hierarchical correlation siamese network for real-time object tracking," *Applied Intelligence*, vol. 51, no. 9, pp. 1–10, 2020.
- [29] W. Gan, P. K. Wong, G. Yu, R. Zhao, and C. M. Vong, "Lightweight network for real-time adaptive stereo depth estimation," *Neurocomputing*, vol. 441, no. 9, pp. 118–127, 2021.