Hindawi
Advances in Multimedia
Volume 2023, Article ID 9789754, 1 page
https://doi.org/10.1155/2023/9789754

Retraction

Retracted: A Hierarchical Children's Dance Movement Pose Estimation Method Based on Sequence Multiscale Feature Fusion Representation

Advances in Multimedia

Received 15 August 2023; Accepted 15 August 2023; Published 16 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Qin, T. Huang, and G. Tang, "A Hierarchical Children's Dance Movement Pose Estimation Method Based on Sequence Multiscale Feature Fusion Representation," *Advances in Multimedia*, vol. 2022, Article ID 2445210, 10 pages, 2022.

Hindawi
Advances in Multimedia
Volume 2022, Article ID 2445210, 10 pages
https://doi.org/10.1155/2022/2445210

Research Article

A Hierarchical Children's Dance Movement Pose Estimation Method Based on Sequence Multiscale Feature Fusion Representation

Yanan Qin,[1] **Tao Huang**,[2] **and Guanzhen Tang**[3]

[1]*Preschool Education, Xi'an University, Xi'an 710065, Shaanxi, China*
[2]*School of Information Science, National University of Defense Technology, Changsha 410015, Hunan, China*
[3]*College of Art, Yuncheng University, Yuncheng 044011, Shanxi, China*

Correspondence should be addressed to Yanan Qin; qinyannan@xawl.edu.cn

Received 25 May 2022; Revised 12 June 2022; Accepted 25 June 2022; Published 24 August 2022

Academic Editor: Qiangyi Li

Aiming at the problem that traditional human motion pose estimation methods cannot accurately capture and estimate the movement changes of children dancers, a hierarchical dance pose estimation method for children based on sequence multiscale feature fusion representation is proposed. By comparing the pose feature extraction algorithm with the actual recognition effect, the recognition rates of the dancer's upper body, infiltration, and whole body have increased by 14.2, 10.6, and 12.6, respectively. The experimental results show that the proposed pose estimation algorithm achieves good pose estimation results on both the standard human pose estimation dataset and the self-built dance dataset.

1. Introduction

Children's dance is a culture, and modern culture is one of the important aspects of education. In most dance clubs in our country, many children study, so, in the lessons, the teacher only evaluates the movements of the children. The body language and facial expressions of the students when dancing convey the students' mood fluctuations [1]. The truth of a young child's dance experience is beyond comprehension. Therefore, using modern scientific data and technology, dancers' movements, body shape calculations, and dancer status can be obtained. The use of training materials in the classroom can support independent learning. In recent years, technology and culture have continued to deepen, and predicting dance movements and body language has become a research topic. Technical integration not only corrects dance in time but also increases personal discipline [2]. However, there is always a certain error in the measurement method, and its performance has a great influence on the absolute quality of the target detector. Therefore, a step-by-step approach to dance music pose assessment based on a series of multiscale features is planned (Figure 1). The method improves the ability to accurately assess dancer positions by analyzing the geometry of the human skeletal joints and creating a layered model. Position is calculated based on joint geometry.

2. Literature Review

Many calculations can now be divided into two categories: upper and lower. The first usually detects the human body by looking for the frame in the picture according to the purpose and then calculates each human body to find the frame that makes up the human body. Finally, joints are connected as a result of human imagination [3]. Some researchers have cited the Cascaded Pyramid Network (CPN) method as a way to calculate the nodes of the Pyramid and RefineNet networks. A simple foundation is an easy and effective network for many people to predict and control behavior. Many people in the region wanted to complete the pose estimation (RMPE) calculation. Ge et al. request the Hourglass network to collect properties of different scales;

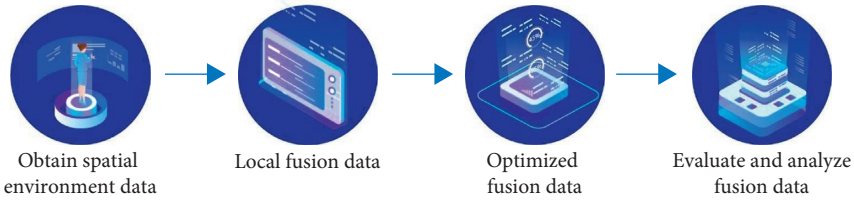

Figure 1: Sequence multiscale feature fusion.

HRNet was asked to compute human design using high-performance functions. The following process is generally divided into two parts: node detection and aggregation. It uses a single-character prediction algorithm to identify all nodes in an image and then combine nodes in a single body to achieve multiple individual character hypotheses [4].

Some studies have proposed the use of Facial Dynamics Map (FDM) to characterize the microexpression sequence. In this method, the microexpression sequence is first aligned at the pixel level, and then each expression sequence is divided into space-time cuboids according to the selected particle size, and then the main optical flow direction of each cuboid is calculated through correction and optimization. FDM generated by these main optical flow directions can show the subtle changes of microexpressions to a certain extent. This method has a large amount of calculation and high feature dimension [5]. It is considered that the vertex frame is the most obvious microexpression frame in the microexpression sequence, while the initial frame is the best reference frame with neutral expression. Therefore, the optical flow is calculated by using these two frames, and a biweighted optical flow descriptor (BI-WOOF) is proposed. The optical flow histogram features are obtained by local and global weighting according to the optical flow size and optical strain size, which are used to characterize microexpressions.

Pose estimation method is used to estimate human motion in video frame sequence, and then human pose information is used as the input of motion recognition. Some scholars changed the traditional separate training and combined pose estimation and motion recognition sequentially and proposed a framework combining pose estimation and motion recognition [6]. The precision of motion recognition reaches the first-class standard, and the attitude estimation is improved. It is proposed that, compared with the apparent feature, the motion recognition based on the attitude feature is better than the motion recognition based on the underlying representation in the video data, even in the data with very serious noise. Pose based motion recognition can solve the problem of "in-class spacing" which is perplexing based on appearance feature recognition, especially the invariance of 3D skeleton pose in appearance feature and perspective. Using gesture feature based representation greatly simplifies the learning of motion recognition itself, but representational features are more general than gesture features [7]. Therefore, many researchers generally choose to combine the two features in order to achieve higher accuracy and achieve the universality of the motion recognition method, but the disadvantage is that the motion recognition based on posture will bring high computation [8].

3. The Description Algorithm Based on Children's Dance Pose Feature Is Introduced

3.1. Problem Presentation and Analysis. The algorithm improvement in this chapter is derived from the p-CNN feature descriptor, which uses the human body posture information to segment the image region, obtains the left hand, right hand, upper body, whole body, and the whole image region of the human body, and then carries out feature extraction and calculation according to each region. The segmentation of each region of the image by human posture can effectively solve the influence of occlusion on motion recognition [9]. However, the disadvantage of this method is that the upper body of the human body is regarded as the main information position of the movement, and the situation that the lower body of the human body determines the movement type is ignored. Especially in dance movement recognition research, there are many movement types determined by the obvious change of legs. The extraction of the whole image information is to make full use of the background information as the auxiliary of motion recognition, while ignoring the influence of complex background in the extraction of motion features [10]. At the same time, the background will not change much in the collected video, and the information will be ignored due to the small change of the background during motion feature extraction. The operation of the whole image will increase the calculation time instead, affecting the efficiency of motion recognition. Although the human body image region is also determined by the position of main joints generated by attitude estimation, the human body region is divided into upper body, lower body, and whole body region. The reason is that dance movements are mainly represented by human limbs, and the extracted regions will greatly reduce the influence of background information and reduce the amount of calculation [11]. Considering the difference between dance movements and human daily movements, SIFT was selected to replace RGB representation features proposed in P-CNN. One reason why SIFT feature is selected to replace RGB feature and optical flow fusion is that it has its own expansibility and is easy to combine with other forms of feature vectors. At the same time, the scale, illumination, and rotation invariance of SIFT features can make up for the shortcomings of optical flow sensitivity to illumination changes, and the recognition of different scale movements formed by children dancers with different body types is more

robust. The retention of optical flow characteristics is mainly due to the fact that the continuity of dance movements is stronger than the daily movements of human body, and optical flow can well represent the dynamic information of continuous movement of dance movements [12]. At the same time, dense optical flow and frame difference method are used to calculate the optical flow, which can correct the mismatching problem of SIFT corner points and the influence of unstable edge response points to a certain extent. Finally, the joint Angle and motion speed are calculated by using the main joint position obtained by human body posture estimation to determine the overall motion category and current motion speed.

3.2. Feature Extraction

3.2.1. Human Posture Characteristics. Human body posture features are derived from human body posture information, and there are two ways to obtain children's posture information. One is the coordinates of each joint of the human body obtained by the motion capture device when the dance video is collected. The other is to use pose estimation to obtain child joint positions and calculate joint angle information in the test set. The reason why the actors in the image are not divided into regions according to their limbs is that the dance movement has its own particularity, which is different from daily movements such as walking or running. The motion state of one arm can be inferred from the motion of the other arm. In most children dance movements, two arms and two legs are needed to jointly determine a dance movement, and there is no fixed corresponding relationship between arms and between legs [13].

3.2.2. Optical Flow Feature Extraction. The optical flow is to use the variational method to calculate the displacement vector field d_t of pixel (x, y) between two consecutive frames at moments t and $t+1$ and find the function u, v that minimizes the energy function. Optical flow calculation is used to optimize the global energy function composed of data items and smoothing items. The mathematical form is

$$E_{Global} = E_{Da\ te} + \lambda E_{Smooth}. \tag{1}$$

In the above formula, $E_{Da\ te}$ is the data item, which measures the consistency of optical flow and input image, E_{Smooth} is the smoothing term, indicating the flow field tending to smooth slip change, and E_{Global} represents optimized global energy. The specific calculation process of each item is described below.

Define pixel (x, y) and its brightness at time t as

$$I(x, y, t). \tag{2}$$

Flow is defined as follows:

$$(u(x, y, t), v(x, y, t)). \tag{3}$$

Then the brightness remains unchanged and is expressed as

$$I(x, y, t) = I(x + u, y + v, t + 1). \tag{4}$$

The above formula is simplified by first-order Taylor expansion, and the right side of formula (4) is linearized to obtain approximately

$$I(x, y, t) = I(x, y, t) + u\frac{\partial I}{\partial x} + v\frac{\partial I}{\partial y} + 1\frac{\partial I}{\partial t}. \tag{5}$$

After simplification, the optical flow constraint equation can be obtained as follows:

$$u\frac{\partial I}{\partial x} + v\frac{\partial I}{\partial y} + 1\frac{\partial I}{\partial t} = 0. \tag{6}$$

The calculation of formulas (6) and (4) may lead to errors in each pixel, which will eventually lead to the problem of gathering errors in the whole image [14]. Therefore, all pixel errors are aggregated and the penalty function is selected to minimize the error. A basic approach is to use the $L2$ paradigm:

$$E_{Date}(u, v) = \int \Omega\Psi\left(\left|I(\vec{x} + \vec{w}) - I(\vec{x})\right|^2 + \lambda\left|\nabla I(\vec{x} + \vec{w}) - \nabla I(\vec{x})\right|^2\right)d\vec{x}. \tag{7}$$

In the above formula, Ω represents the entire image region.

$$\vec{x} \coloneqq = (x, y, t)^T, \vec{w} \coloneqq = (u, v, 1)^T, \left|I(\vec{x} + \vec{w}) - I(\vec{x})\right|^2. \tag{8}$$

The above formula represents the assumption that the grey value is constant. $\left|\nabla I(\vec{x} + \vec{w}) - \nabla I(\vec{x})\right|^2$ assumes that the gradient is constant, and $\gamma \geq 0$ represents the weight of the two hypotheses.

If a shift in an image sequence is required, the smoothing constraint can be applied only to the spatial domain (if only two frames are available) or to the spatial-temporal domain [15]. In order to optimize the discontinuity of the displacement field on the object boundary in the scene, the smoothness hypothesis is summarized by piecewise smoothing of the optical flow field. The simplest smoothing term is the flow field in favor of the first derivative (gradient), which is defined using $L2$ norm as

$$E_{Smooth}(u, v) = \int_{\Omega} \Psi\left(\left|\nabla_3 u\right|^2 + \left|\nabla_3 v\right|^2\right)d\vec{x}. \tag{9}$$

When the image has two frames,

$$\nabla: = (\partial x\ , \partial y)^T. \tag{10}$$

If the image sequence is more than two frames, the time smoothing item should be added; namely,

$$\nabla_3: = (\partial x\ , \partial y\ , \partial t)^T. \tag{11}$$

Suppose that f is a vector that connects horizontal and vertical pixels to make up each pixel stream, and the goal is to optimize E_{Global} with respect to f. The simplest gradient descent method is the extreme descent algorithm, which performs steps in the negative gradient direction $-\partial E_{Global}/\partial f$. There are two ways to solve this problem. One is to constantly adjust the step size according to the energy change. If the energy decreases, the step size increases, and if the energy increases, the step size decreases. Another way is to set the step size to a fixed value:

$$step = -\omega \frac{1}{T} \frac{\partial E_{Global}}{\partial f}. \tag{12}$$

Many nonquadratic formulas can be solved by iterative weighted least squares. That is, they constitute a sequence of quadratic optimization problems related to the data, which are constantly changing through iteration, and the weighted quadratic is iterative solution and weight reestimation [16]. The dense optical flow algorithm will be used to extract the optical flow information of each frame in the action sequence for normalization according to the following equation:

$$f_m(R) = \frac{1}{|R|} \sum_{i \in R} f_m(i). \tag{13}$$

If $f_m(R) < \alpha$ ignores the optical flow information of its pixels, where i represents each pixel, α is set to 0.3 according to the empirical value. The optical flow feature extraction algorithm is described in Algorithm 1:

After Difference of Gaussian (DoG) method is used to determine key points, each key point needs to be checked to remove points with low contrast and unstable edge response. In this paper, in addition to removing pixels with very asymmetric local curvature of DoG, pixel optical flow values extracted in the previous section are also used for filtering [17]. The optical flow threshold was selected as the empirical value 0.3, so the initial value was set as 0.3 in this paper. However, the final threshold was obtained after the experiment. If the optical flow value of the selected key point was less than the threshold, the key point would be removed. Directional assignment is to determine the direction and gradient for each key point. The 2D gradient size and direction of each pixel are defined as follows:

$$m_2 D(x, y) = \sqrt{L_x^2 + L_y^2}, \theta(x, y) = \tan^{-1}\left(\frac{L_y}{L_x}\right). \tag{14}$$

In the above formula, x, y is the coordinate of pixel in the image; L_x and L_y are obtained through finite difference approximate calculation:

$$\begin{aligned} L_x &\approx L(x+1, y, t) - L(x-1, y, t) \\ L_y &\approx L(x, y+1, t) - L(x, y-1, t). \end{aligned} \tag{15}$$

L_x, L_y, and L_t are used to calculate the gradient size and direction of 3D:

$$\begin{aligned} m_3\ D(x, y, t) &= \sqrt{L_x^2 + L_y^2 + L_t^2} \\ \theta(x, y, t) &= \tan^{-1}\left(\frac{L_y}{L_x}\right) \\ \varphi(x, y, t) &= \tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}\right). \end{aligned} \tag{16}$$

Because $\sqrt{L_x^2 + L_y^2}$ is positive, $\varphi \in (-(\pi/2)\Delta(\pi/2))$ is always there and each angle is represented by a unique (θ, φ) pair, so the gradient direction of each pixel in 3D is represented by two values. There are many ways to construct a weighted histogram for the 3D neighborhood of a point of interest. In this paper, the meridian-collateral-parallel method, which is simpler and quicker to find the extremum of the directional histogram, is used, followed by quadratic interpolation to find the true extremum. Bins need to be regularized by solid angle ω when using the meridian parallel method [18, 19]. Values are added to each bin by normalizing the bin region, also known as solid angles. If solid angles are not normalized, the orientation histogram will get the wrong weighting. Solid angle is calculated as follows:

$$\begin{aligned} \omega &= \int_{\varphi}^{\varphi+\Delta\varphi} \int_{\theta}^{\theta+\Delta\theta} \sin\ \theta d_\theta d_\varphi \\ &= \Delta\varphi \int_{\theta}^{\theta+\Delta\theta} \sin\ \theta d_\theta \\ &= \Delta\varphi [-\cos\ \theta]_{\theta}^{\theta+\Delta\theta} \\ &= \Delta\varphi (\cos\ \theta - \cos(\theta + \Delta\theta)). \end{aligned} \tag{17}$$

Add normalized values to the histogram as follows.

(x, y, t) represents the coordinates of points of interest, (x', y', t') represents the coordinates of pixels added to the orientation histogram, and the peak value of the histogram is the main direction. The main peak is stored because it can be used to rotate key neighborhoods, creating rotation-invariant features. It is expressed as follows:

$$\begin{aligned} &\text{hist}\left(i_\theta, i_\varphi\right) + \\ &= \frac{1}{\omega} m_{3\ D}(x', y', t') \\ &e^{\frac{-\left((x - x')^2 + -(y - y')^2 + -(t - t')^2\right)}{2\sigma^2}}. \end{aligned} \tag{18}$$

3.3. Experimental Results and Analysis. The databases used in this paper include the Northeast Yangko dance video collected by motion capture equipment and the popular JHMDB and MPII Cooking databases, respectively. JHMDB is a subset of HMDB database, which contains 21 human actions, such as combing hair, mountain climbing, golf,

Input: body parts (up, down and body) of key image sequence, $\lambda = 1$, $\omega \approx 1.95$, $a = 0.3$, down-sample ratio = 0.5
Output: displacement vectors $f(u(x, y, f) \vee (x, y, t, 1)^T$
(1) Fort = 1,2,. . .,$m - 1$ (all the images)
(2) pyramid of images with a scaling factor 0.5 till width = 40:
(3) For $i = 0:17$
(4) grey value constancy assumption equation (4.4)get data term equation (4.5):
(5) Smooth term equation (4.6) to avoid the problem of aperture:
(6) Global function equation (4.1);
(7) The second derivative of energy function bounded above T;
(8) Calculate step k equation (4.7):
(9) be solved with SOR iterations, $u^* += u^* + du+ = v^* + dv^*$:
(10) $w^{k+1} = w^k + dw^k$
(11) Normalization as equation (4.8);
(12) if $fm < (R) < a$ ignore;
(13) End For
(14) Coarse vector to color:
(15) End For
(16) PCA to reduce dimension
(17) Until minimize global energy

Algorithm 1: Optical flow feature extraction algorithm.

running, and sitting. According to the duration of the action, each action basically ranges from 36 to 55 pieces, with a total of 928 pieces. Each segment is 15–40 frames and 320×240 in size, and each frame is marked with human posture. MPII Cooking Activities consists of 64 finely textured actions and an additional background class that takes place in a kitchen with a static background. The database contains 5609 action fragments with frame size of 1624×1224. Some actions are very similar, such as dicing, slicing and cutting, and washing hands and things. Northeast Yangko is a database of dance moves collected by motion capture equipment and used as a training set. There are 15 combinations of dance moves in total. Each set contains about five to ten dance moves. There are two types of video: one is the video of dance movement combination, and the other is to collect the data of the combined movement according to the movement type, and each video clip is about 10 seconds. The test set used in this paper comes from the video on the Internet, and there are eight types of movements from low to high, such as breaking round and breaking step, back and kicking step, moving head, pointing and standing step, cross pull step, jumping and squatting cross step [20].

In this paper, we first describe the process of creating dance experience by defining the upper, lower, and whole body of the human body. Body movement is based on the structure of the human body. Following the procedure, follow the third chapter and finally obtain a partial procedure. Each qualification was performed by reducing and normalizing the PCA size, and the important data of the show were combined as a sequence explanation. Finally, a special comment is used as an SVM input to inform the dance. During the experiment, it was clear that the color spectrum represented in the JHMDB and MPII datasets was recognized, but the difficulty behind the pixels in recognizing the dance movement affected the output and reduced the perception of performance. Using 3D-SIFT to represent static data improves the experience of working with the above data [21]. At the same time, recognizing the benefits of a combination of 3D-SIFT and optical performance is better than RGB and optical flow. Using P-CNN interpretation on the JHMDB and MPII datasets can significantly improve validation values, so combining representations and representation-based criteria can improve recognition performance. At the same time, the identification of limb regions divided into children also leads to different perceptions. When using a part of the human body, the recognition is higher than the other, and the recognition is the highest of all the images. This is due to the fact that JHMDB and MPII are configured as support functions for the base data in the dataset. In the following MPII test data, the lower part of the human body is not included in the comparative test because the lower part of the human body is closed with a table. A comparison of gratitude from p-CNN comments can be found in Table 1.

For comparison, after the warranty connection, the upper, lower, and all parts of the body were selected to exclude RGB functions, SIFT functions, flow features, and SIFT and optical flow features. By recognizing the value obtained by various methods, it can be seen that the order is close to being received after the upper body or lower body is used for special mining, but the body acceptance is slightly lower [22]. The combination of human body parts has the highest cognitive value, while the combination of human body parts has the lowest cognitive value. This is because when choosing a movement level, the number of dances held at the top or bottom is close to the number of dances at the bottom. When all parts of the body are used, the two zones interact when the energy is only in the upper part of the body or in the lower part of the body, and the dance is often called the body and upper body. Separate training can improve

Table 1: Comparison of recognition effects of p-CNN feature description.

The body part	JHMDB			MPII Cooking Activities		
	RGB	Flow	RGB + flow	RGB	Flow	RGB + flow
The upper body	53.8	59.9	66.1	32.3	47.6	51.9
The lower body	51.2	60.6	61	—		
The whole body	54.3	54.7	68.1	28.8	56.2	56.5
Whole image	59.4	68.1	74.4	43.6	57.4	60.8

Table 2: Recognition effects of different features extracted from different human body regions.

The body part	Northeast Yangko			
	RGB	SIFT	Flow	SIFT + flow
The upper body	11	23.3	32.9	47.1
The lower body	11.3	24.8	35.5	47.9
The upper body + the lower body	24.3	29.3	44.6	57.9
The whole body	23.4	31.4	401.7	55.3

validation accuracy. As shown in Table 2, various effects from different regions of the human body are recognized.

The algorithm determines each part of the human body in the video through posture recognition, extracts color features and optical flow features from each part and the overall image, and uses CNN to form feature descriptors for motion recognition. Secondly, the application of the two features in the algorithm is analyzed, and it is understood that the color information as a static feature plays a little role in dance movement recognition, and a 3D-SIFT feature is proposed to replace the color features. Then, the 3D-SIFT feature, optical flow feature, and the way to obtain human pose information are introduced, respectively, and the feature vector, which is the feature descriptor of a certain movement type pose sequence, is formed by PCA and normalization processing of several features. Finally, by analyzing the experimental results and comparing with the basic methods, the advantages of the improved algorithm in this paper are summarized [23].

Similar Movement Experiments. The experiment took the Northeast Yangko dance movements recorded by motion capture equipment as the dataset and selected the eleventh group of small wrong steps as the sample data. It can be seen from the following table that the first segmentation position is in frame 104, between the sequences of preparatory action frames, so there will be no loss of action information. In the segmentation results of other movement sequences, there is an obvious movement folding process between the little false step and the leg lift, so it is mistaken as a single simple movement from frame 104 to frame 219. However, under normal circumstances, the little false step and the back leg lift belong to the same dance movement sequence. Taking the forward movement of small wrong steps as an example, the manual segmentation position is compared with the segmentation position determined by the algorithm in this paper, as shown in Table 3.

Table 3: Comparison between manual segmentation and automatic segmentation.

Motion segments	The starting frame	End frame	Automatic segmentation result	Motion description
1	0	120	104	Anticipation
2	120	290	219	Lift your leg after a misstep
	281	350	358	Unrelated action
3	341	420	499	Lift your leg after a misstep
	411	490		Unrelated action
4	481	554	622	Lift your leg after a misstep
	545	608		Unrelated action
5	608	677	709	Lift your leg after a misstep
	678	740	725	Unrelated action

Table 4: Comparison of segmentation locations with added constraints.

Motion segments	The starting frame	End frame	Automatic segmentation result	Motion description
1	0	120	104	Anticipation
2	121	290	219	Lift your leg after a misstep
	291	350	358	Unrelated action
3	351	420	499	Lift your leg after a misstep
	421	490		Unrelated action
4	491	554	622	Lift your leg after a misstep
	555	608		Unrelated action
5	609	677	725	Lift your leg after a misstep
	678	740		Unrelated action

Table 5: Effect comparison of segmentation algorithms.

Algorithm	Correctly split quantity	All partition quantity	Precision (%)	Recall ratio (%)
PCA	96	110	80.60	87.30
Clustering	85	110	87.20	86.40
Algorithm in this article	1033	110	93.80	91.80

Table 6: Action sequence segmentation.

Action sequence	*A*0	*A*1	*A*2	*A*3	*A*4	*A*5
*A*0	320.7389	1903.5753	1610.405	2586	452	1197.076
*A*1	562.2397	0	467.2325	363.552	784.8848	233.901
*A*2	3509.3268	534.9004	0	108.0585	50.661	233.3639
*A*3	2773.5539	278.5316	70.551	0	138.311	233.3246
*A*4	4638.3941	898.7726	79.6687	2,193,55	0	423.8951
*A*5	1878.3056	313.7521	108.225	165.9044	1,707,066	0

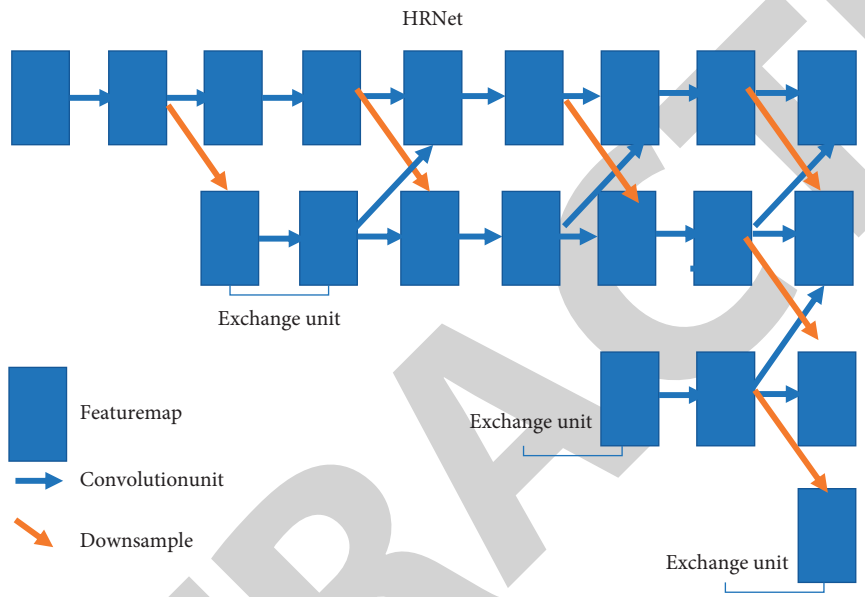

Figure 2: HRNet backbone network.

Because there are many nonfunctional positions in the initial segmentation position, the order difference limit and the difference limit of the adjacent minima and maxima increase. Comparison of segmentation positions after additional constraints is shown in Table 4.

Video frame segmentation algorithms typically include the PCA algorithm and the cluster algorithm. Comparing the segmentation of this type of algorithm and other algorithms requires two dimensions: accuracy and inverse. The actual value is the location of the temporary actual segmentation of the video frame, and the return value is the percentage of total success obtained by cutting made regularly. The results of the comparative segmentation of the application algorithm and other algorithms are shown in Table 5.

At the end of segmentation, similarity comparison is made between the segmented action sequences to determine the similarity of each action sequence in the same video, so as to determine the main action sequence that can represent the video. DTW algorithm is used for comparison in this article. The smaller the cumulative regular distance is, the more similar the action sequences are [24]. If six action sequences *A*0, *A*1, *A*2, *A*3, *A*4, and *A*5 are separated from each other in the small wrong step, the distance between two action sequences can be obtained through experiments, as shown in Table 6.

Among the six action sequences, *A*0 is a preparatory action, which does not contain obvious action information. In addition, it can be seen from the table that the regular distance between *A*0 and the other action sequences is large, while the regular distance between the other action sequences is relatively small, so it can be confirmed that similar actions are generated in the other action sequences. By selecting the first 10 minimum distances (5*2), it can be judged that similar actions occur in *A*2, *A*3, *A*4, and *A*5, among which *A*2, *A*3, and *A*4 are the most similar. Finally, the longest action sequence *A*3 (358–488 frames) is selected to represent the video information of small wrong steps moving forward. The above 5*2 selection is based on the fact that all action sequences are similar (except for preparatory actions), so it is necessary to ensure that the distance between all action sequences is obtained.

4. A Hierarchical Dance Pose Estimation Method Based on Geometric Relation of Joint

4.1. Multiscale Feature Fusion Representation. As the attitude estimation task is pixelwise level node estimation problem, it needs to use low-level and high-level features to locate different scale nodes. The high-level features are conducive to the location of large-scale nodes, while the low-level

features are very important for the location of small-scale nodes. In order to improve the robustness of pose estimation to the scale change, a sequence multiscale feature fusion model was proposed.

4.2. Sequence Multiscale Feature Fusion. Taking HRNet network as the backbone network, as shown in Figure 2, it is composed of four parallel multiresolution subnets; each network using ResNet module design principle is made up of four residual unit.

In a special representation, low-density functions have richer semantic data when the data source is coarse, while high-resolution lower-level functions have richer semantic data. Text spacing is true even if semantic data is weak. Therefore, this paper proposes multicomponent function integration (SMF) processes to create high-resolution and low-resolution combinations and increase the efficiency of depicting network features. The multifunction integration process is used for the 4 outputs of the final integration unit on the HRNet network. The solutions vary depending on convolution, interpolation, and deconvolution.

4.3. Hierarchical Attitude Estimation Based on Geometric Relations of Nodes. Firstly, according to the structure of the human body, the joints can be divided into two types: the first type is the body joints with small deformation, such as shoulder, hip, and neck joints; and the second type is the deformation of obvious limb joints, such as wrist, elbow, knee, and ankle hinge joints. Then, according to the two types of joints, a hierarchical pose estimation model was designed, and all human joints were aggregated into five parts as shown in Figure 3: neck, left shoulder, right shoulder, left hip, and right hip, so as to predict the joints based on the geometric relationship of human joints [25].

The hierarchical network designed in this paper consists of three stages. In the first stage of the network, heat map prediction of all nodes of human body was carried out for the designed SMF model, and corresponding coordinate positions were calculated. Then, the heat map of joints obtained in the first stage was used as the input of the network in the second stage. In view of the small deformation of human trunk joints and the large deformation of limb joints, the SMF model was used to predict the trunk joints (KTrunk) with relatively stable deformation from all human joints obtained in the first stage. The joints of the human body are divided into five parts with trunk joints, also known as five categories (neck, left shoulder, right shoulder, left hip, and right hip). Then, all the nodes in the first stage of the network and the five types of torso joints predicted in the second stage were used as inputs to construct the third stage network. At the same time, considering the geometric correlation of human body structure, all the joints of human body are divided into five types of trunk joints by intraclass correlation, so as to realize the connection between limb joints and trunk joints.

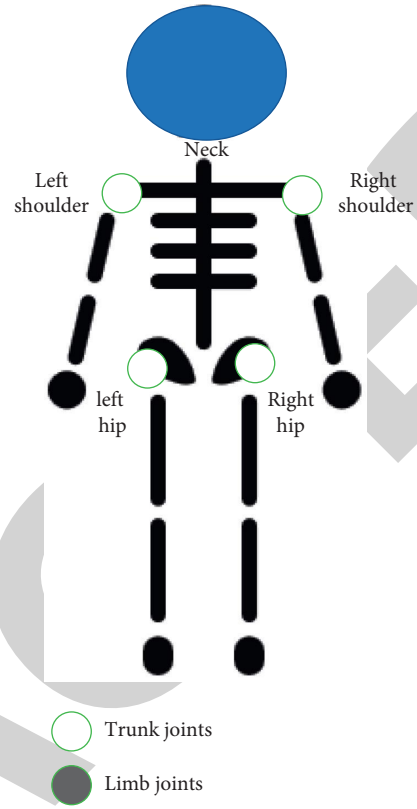

Figure 3: Geometric relations of human body joints.

Table 7: Analysis of ablation results.

Model	HRNet-W32	SMF	HPE	AP
A	√			74.46
B		√	√	75.21
C	√	√		75.76
D	√	√	√	76.29

4.4. Analysis of Experimental Results. As shown in Table 7, when the two important components of the proposed model are used simultaneously, as shown in Model D in Table 7, the proposed algorithm obtains the best mAP value and improves the positioning accuracy of mAP key nodes by 1.83% (76.29~74.46) on the basis of hrNET-W32 model in thermal mAP regression. In the ablation experiment, only SMF model and HPE model were added, respectively, on the basis of HRNET-W32 model, as shown in Model 8B and Model C. The mAP improved by 0.75% (75.21~74.46) and 1.3% (75.76~74.46), respectively. As can be seen from Table 7, the SMF model in Model B improves the ability of multiscale feature representation by orderly fusion of multiresolution and multiscale features of HRNet network, which is conducive to node estimation. In Model C, the HPE model was designed to estimate the joints according to the geometric relations of joints and to flexibly deal with different types of trunk joints and limb joints. By solving the optimal matching problem of the connection set of all candidate joints within the class, the optimal matching of the

connection between the trunk joint and the limb joint is obtained. The accuracy of attitude estimation can be improved by inference of occluded nodes.

5. Conclusion

Hierarchical dance poses approach the way you plan to integrate a wide range of functions in sequence. The goals of children dance, adaptation, and major design changes create many interconnected designs. In order to improve the robustness of the evaluation algorithm for large-scale transformation of the dancer skeleton, this time we focused on the approximate model of the stepped pose based on the large deformation of the dance pose, the main obstruction, and the geometry dance period. This is to improve the results of the dance pose assessment. The results of the experiment show that the evaluation algorithm has achieved the performance of a standard human data prediction set and a self-developed single- and multiplayer dance dataset. It is then possible to provide dance instruction for use in performances and activities, to correct the dance in real time, and to help train and educate the dancers. This is very important for China's cultural heritage.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by Xi'an University.

References

[1] G. Huang, X. Chen, J. Chen, W. Lin, and Z. Wang, "Multi-person Pose Estimation under Complex Environment Based on Progressive Rotation Correction and Multi-Scale Feature Fusion," *IEEE Access*, vol. 8, no. 99, p. 1, 2020.

[2] J. Zhang, M. Zhang, L. Shi, W. Yan, and B. Pan, "A multi-scale approach for remote sensing scene classification based on feature maps selection and region representation," *Remote Sensing*, vol. 11, no. 21, pp. 2504–2506, 2019.

[3] C. Guo, J. Zhou, W. Du, and X. Zhang, "Multi-scale collaborative network for human pose estimation," *International Journal of Humanoid Robotics*, vol. 16, no. 04, pp. 1850002–1850039, 2019.

[4] Y. Ge, Z. Yang, Z. Huang, and F. Ye, "A multi-level feature fusion method based on pooling and similarity for hrrs image retrieval," *Remote Sensing Letters*, vol. 12, no. 11, pp. 1090–1099, 2021.

[5] R. Wang, Z. Cao, X. Wanga, Z. Liu, and X. Zhu, "Human pose estimation with deeply learned multi-scale compositional models," *IEEE Access*, vol. 6, no. 99, p. 1, 2019.

[6] G. Wu, W. Chen, H. Cheng, W. Zuo, and J. You, "Multi-object Grasping Detection with Hierarchical Feature Fusion," *IEEE Access*, vol. 7, no. 99, p. 1, 2019.

[7] X. Wang, Y. Su, C. Luo, F. Nian, and L. Teng, "Color image encryption algorithm based on hyperchaotic system and improved quantum revolving gate," *Multimedia Tools and Applications*, vol. 81, no. 10, pp. 13845–13865, 2022.

[8] Z. Zhang, Q. Yang, and Y. Zi, "Multi-scale and multi-pooling sparse filtering: a simple and effective representation learning method for intelligent fault diagnosis," *Neurocomputing*, vol. 451, no. 3, pp. 138–151, 2021.

[9] Jianming, Lv, Jiajie, Zhong, L. Jintao, and Y. Zhenguo, "Ace: ant colony based multi-level network embedding for hierarchical graph representation learning," *IEEE Access*, vol. 6, no. 99, p. 1, 2019.

[10] Z. Dong and B. Lin, "Bmf-cnn: an object detection method based on multi-scale feature fusion in vhr remote sensing images," *Remote Sensing Letters*, vol. 11, no. 3, pp. 215–224, 2020.

[11] Y. Li and G. Baciu, "SG-GAN: adversarial self-attention gcn for point cloud topological parts generation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 99, p. 1, 2021.

[12] N. Yuvaraj, K. Srihari, G. Dhiman et al., "Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–12.

[13] M. Raj, P. Manimegalai, P. Ajay, and J. Amose, "Lipid data acquisition for devices treatment of coronary diseases health stuff on the Internet of medical things," *Journal of Physics: Conference Series*, vol. 2323, Article ID 012038, 2021.

[14] X. Liu, C. Ma, and C. Yang, "Power station flue gas desulfurization system based on automatic online monitoring platform," *Journal of Digital Information Management*, vol. 13, no. 6, pp. 480–488, 2015.

[15] R. Huang, *Framework for a smart adult education environment*, vol. 13, no. 4, pp. 637–641, 2015.

[16] H. Xie, Y. Wang, Z. Gao, B. Ganthia, and C. Truong, "Research on frequency parameter detection of frequency shifted track circuit based on nonlinear algorithm," *Nonlinear Engineering*, vol. 10, no. 1, pp. 592–599, 2021.

[17] X. Xu, Z. Chen, and F. Yin, "Monocular depth estimation with multi-scale feature fusion," *IEEE Signal Processing Letters*, vol. 28, no. 99, p. 1, 2021.

[18] Y. Li, Z. He, S. Wang, Z. Wang, and W. Huang, "Multideep feature fusion algorithm for clothing style recognition," *Wireless Communications and Mobile Computing*, vol. 2021, no. 4, pp. 1–14, 2021.

[19] H. Li, K. Ma, H. Yong, and L. Zhang, "Fast multi-scale structural patch decomposition for multi-exposure image fusion," *IEEE Transactions on Image Processing : A Publication of the IEEE Signal Processing Society*, vol. 29, no. 99, p. 1, 2020.

[20] H. Zhang, Z. Hu, and R. Hao, "Joint information fusion and multi-scale network model for pedestrian detection," *The Visual Computer*, vol. 36, no. 3, pp. 1–10, 2020.

[21] Y. Ji, X. Jiang, and L. Wan, "Hierarchical least squares parameter estimation algorithm for two-input hammerstein finite impulse response systems," *Journal of the Franklin Institute*, vol. 357, no. 8, pp. 5019–5032, 2020.

[22] J. Zhang, W. Ren, S. Zhang et al., "Hierarchical density-aware dehazing network," *IEEE Transactions on Cybernetics*, vol. 51, no. 99, pp. 1–13, 2021.

[23] Q. Gao and P. Liang, "Airline baggage appearance transportability detection based on a novel dataset and sequential

hierarchical sampling cnn model," *IEEE Access*, vol. 9, no. 99, p. 1, 2021.

[24] D. Jarchi, J. Kaler, and S. Sanei, "Lameness detection in cows using hierarchical deep learning and synchrosqueezed wavelet transform," *IEEE Sensors Journal*, vol. 21, no. 99, p. 1, 2021.

[25] D. Feng, Z. A. Xiao, L. A. Xian, C. Xsz, D. Aa, and D. Th, "Hierarchical extended least squares estimation approaches for a multi-input multi-output stochastic system with colored noise from observation data," *Journal of the Franklin Institute*, vol. 357, no. 15, pp. 11094–11110, 2020.