Hindawi
Mobile Information Systems
Volume 2022, Article ID 3034663, 10 pages
https://doi.org/10.1155/2022/3034663

Research Article

A Dance Somersault Pose Recognition Model Using Multifeature Fusion Algorithm

Li Kang

Shaoyang University, Shaoyang Hunan 422000, China

Correspondence should be addressed to Li Kang; 161847247@masu.edu.cn

Received 16 February 2022; Revised 3 March 2022; Accepted 7 March 2022; Published 18 March 2022

Academic Editor: Mian Ahmad Jan

With the continuous progress of society, dance art has gradually entered the lives of ordinary people, and people's appreciation level and artistic attention have gradually improved. How to improve the level of dance art has become an urgent problem to be resolved. In order to solve the problems encountered in dance tumbling pose recognition, this paper proposes a dance tumbling pose recognition model based on multifeature fusion algorithm. According to the characteristics of dance movements, an effective feature extraction method is studied. By combining the key information of bones, the fusion features of the relative position of human joints, the angle of joints, and the ratio of limb length are selected to classify the movements in the dance scene. Through the residual block automatic motion detection method, a multifeature fusion module is used to fuse multiple features, so as to improve the estimation of complex pose by the algorithm and finally complete the dance action identify. The test and analysis results on the data set show that the algorithm can identify the dance somersault pose and can effectively improve the accuracy of dance movement recognition, with strong recognition performance. It is convenient for people to recognize dance tumble posture in the video, thereby realizing the action correction function for dancers.

1. Introduction

The research of human action identify can be summarized into three levels according to the content of motion from simple to complex, namely, mobile vision, motion vision, and behavioral vision [1]. At present, the research on human motion vision in video mainly has the following two problems: (1) In the process of recognition, most of the motion frames in video are duplicated or have low correlation with action identify, which not only increases the computational complexity but also affects the accuracy of action identify [2]. (2) In the process of feature selection and understanding, the main methods include representation fusion and gesture-based features [3]. At present, a large number of scientific research institutions and researchers have done in-depth research on this topic and achieved some good results. However, the research on the combination of action identify technology and dance movements is still in its infancy [4, 5]. Due to the changeable stage background and costumes of dance movements, occlusion and self-occlusion are easy to occur in the performance process, and dance movements have high complexity and other problems, so the representation information fusion cannot accurately and completely express the information of human movements in most cases, while the static information of human bodies is often overlooked based on posture features. Therefore, the research progress in dance video action identify is relatively slow. Among them, the extraction and construction of human features are the key to human action identify, but the current methods of feature extraction and construction usually have the problem of low accuracy.

With the continuous progress of society, dance art has gradually entered the lives of ordinary people, and people's appreciation level and artistic attention have gradually improved [6]. How to improve the level of dance art has

become an urgent problem in recent years. Human pose estimation is a key technology in the field of human action identify. Its principle is to identify human pose by extracting features from images [7]. Through gesture recognition, human motion analysis and information preservation can be realized. This technology can be used in intelligent dance-assisted training. By extracting the features of dancers' images, the dancers' posture skeleton map can be obtained, so as to identify the dancer's dance movements and evaluate and correct the dancer's posture [8]. In the past, there was little research on dance movements in many researches on human action identify. The main reason is that dance is a way of expressing emotions to the public through body movements. It is complex in expression form, with a wide variety of movements, and many dance movements with their own characteristics are included in different types of dance movements [9]. Moreover, due to the high complexity of dance movements and the self-occlusion of dance movements, the research on action identify based on dance videos needs further development [10]. Based on the research of multifeature fusion algorithm, this paper constructs a dance somersault pose recognition model.

Human action identify technology is very important for intelligent video analysis, and it has strong practicability, wide application range, and high requirements for professional technology [11, 12]. However, due to the complexity and changeability of human motion posture, it has some limitations in practical application, and it is impossible to describe human motion effectively [13]. Accurate description of motion posture by effective feature information is the key factor to improve the recognition effect of human motion posture. Dance somersault skill is an important index to evaluate dancers' basic skills and comprehensive abilities, and identifying dance somersault posture is of great significance to art teaching and research [14]. Therefore, this paper constructs a dance somersault pose recognition model based on multifeature fusion. Aiming at the problem of the combination of action identify and dance video, this paper focuses on the feature extraction, representation, and action identify methods based on dance video. The extracted feature information is optimized and fused by genetic algorithm, in which a fitness function is constructed based on mean variance ratio to select feature information with good separability among multiple categories. Draw the fused feature data into a line chart, and find that the fused features can reduce redundancy and ensure the separability between classes and the stability within classes. In order to verify the effectiveness and feasibility of the multifeature fusion dance somersault pose recognition method proposed in this paper, we established a basic dance somersault pose data set. Design a comparative experiment on this data set. The experimental results show that the multifeature fusion dance somersault pose recognition method proposed in this paper has higher recognition accuracy.

2. Related Work

Literature [15] establishes a high-level description model of human structure information based on the spatiotemporal characteristics of human posture and uses image feature recognition method to recognize human posture. Literature [16] puts forward a new method of 3D human action identify, that is, using Kinect, a depth vision sensor, to overcome the problems of light changes and poor reliability of human-computer interaction under complex background conditions, so that the robot has safe and friendly interactive perception ability. Literature [17] proposed an improved ViBe algorithm combined with Langsky function. It can effectively solve the ghost problem, reduce the interference of noise, and obtain more complete and clear human moving targets. Literature [18] extracts human behavior characteristics, establishes spatiotemporal and OR graph model by using human behavior characteristics, and then uses video sequence moving target detection and recognition method to identify human posture in spatiotemporal and OR graph model. Literature [19] proposed a method of human behavior recognition based on multifeature fusion. This method combines acceleration features and plantar pressure features and uses hierarchical support vector machine to identify six behaviors, including standing, sitting down, walking, going upstairs, going downstairs, and running. The overall recognition accuracy rate is over 92%. Literature [20] studies the multifeature fusion dance action identify method based on directional gradient histogram feature, optical flow directional histogram feature, and audio signature feature. Literature [21] proposed a multifeature fusion model of human motion and posture based on eight-star model, Hu invariant moment, Zernike moment, and wavelet moment. Literature [22] uses emotion-oriented speech recognition method to recognize human posture features. This method selects a single feature to recognize human posture, which leads to the inability to accurately distinguish the target area from the background area, and the accuracy of the recognition result is poor. Literature [23] realized the application of gesture-based action identify technology in the field of dance movements and completed the recognition of some dance movements. Literature [24] proposed a recognition method based on the fusion of time, space, and CSI signals of the motion in video. Literature [25] summarizes the methods of human behavior recognition and gait phase recognition and analyzes the advantages and disadvantages of each method. Literature [26] proposes a gait recognition method for complex road conditions based on multifeature fusion to meet the needs of gait phase recognition for complex road conditions.

Based on the previous research, this paper designs a multifeature fusion algorithm for dance somersault pose recognition. Divide each dance action video in the data set into equal parts, and at the same time, accumulate the edge features of all video images in each segment into one image, and extract the direction gradient histogram features from it. Before human body region segmentation, the depth image is denoised and smoothed to remove the noise that affects human body region segmentation and make the segmentation effect better. After the human body region in the depth image is obtained, the area threshold denoising is performed. The extracted features are reduced in the dimension and normalized to obtain the feature vectors of various features,

and then, the feature vectors are fused as the input of the classifier to realize the recognition of dance tumbling posture.

3. Methodology

3.1. Dance Recognition-Related Technology. Human action identify technology has important applications in many fields today. Through the continuous exploration of relevant researchers, human action identify technology has made great progress. From the initial characteristics of traditional manual design to modern popular deep learning methods [27]. These methods need to extract the corresponding features based on video information and then select the appropriate classifier to classify these features. Hierarchical model provides an intuitive and simple interface to integrate prior knowledge and understanding of action structure.

Most of the first dance action identify methods rely on artificial design features; that is, specific algorithms are used to extract motion features from videos [28]. Video is composed of many video frames, in which video frames contain the appearance information and background information of dance movements, and adjacent video frames have a progressive relationship in time. The video action identify process includes video capture, video processing, feature extraction, feature construction and analysis, and motion classification. The structure of human action identify system is shown in Figure 1.

The local feature value of dance takes the moving target as a set, which is composed of many image blocks. When selecting features, only the parts of interest need to be selected to extract features. The global feature of action refers to the need to describe the moving target as a whole.

At the initial stage of human action identify, some features of images or video sequences are extracted, and action identify is realized by feature matching. Later, researchers found that the bottom features combined with human posture features can improve the accuracy of action identify. In order to prevent the subsequent human posture recognition from being disturbed by the background, it is necessary to preprocess the initially collected images so that the images can be subjected to subsequent feature extraction. Video processing is mainly based on image processing such as foreground and background segmentation, and denoising after video is converted into image, and the main method of segmentation of action sequence in video is planning and design before data collection. However, feature extraction always tends to the recognition method of multifeature fusion, and the pose-based feature proposed in recent years has outstanding advantages.

There are roughly two ways to obtain human posture information. (1) Information such as joint coordinates, human skeleton, and motion trajectory can be accurately obtained by motion capture equipment. (2) Through the estimation method of human posture, the approximate joint positions and skeleton of human body in images or videos are obtained. Action identify first needs to detect moving objects from video information; the detected video sequence of human moving target is preprocessed and analyzed to obtain a clear moving posture image. Then, according to different feature extraction methods, the human motion posture features are obtained, and related processing is carried out to ensure that these features are simple and effective, so as to form the feature data set under various postures. According to the obtained multipose feature data set, the classifier is trained to form feature templates, and finally, the motion features of individuals to be identified are compared with the feature templates in the feature database to realize classification and identification.

Feature extraction refers to extracting feature information from the dance motion data set to describe the target motion in the video, which is an essential step for the research of dance action identify. It seems that the extracted features play a vital role in the accuracy of dance movement recognition results and the robustness of recognition methods. Generally speaking, dance movement recognition falls into two categories: (1) directly use the information of main joints in dance movements to calculate the similarity, and determine the category of movements by matching the similarity with known movements. (2) The human body in the image is segmented by using the obtained posture of human body, and the local images are obtained by taking the positions of the main joints as the key points, and then, some features of the local images are extracted for recognition. The dance somersault gesture recognition process in this paper is shown in Figure 2.

Therefore, the goal of the video frame sequence segmentation is to segment the human posture sequence of small movements, so that the dance movement sequence is as short as possible but contains enough movement information to identify the movement type. Feature-level fusion refers to the comprehensive analysis and processing of multisource feature vector information extracted from different feature extraction methods through feature fusion algorithm to form a new fused feature vector group.

Besides manual segmentation, video sequence segmentation technology mainly includes segmentation of dance action video based on sliding window, boundary, and modeling method. The video frame sequence segmentation method based on sliding window and boundary belongs to local segmentation, that is, using the local characteristics of video frames to determine the similarity of adjacent frames or frame sequences and to determine whether they belong to the same kind of actions. The movement of human body in video is equal to the movement of pixels, so the optical flow can be regarded as the displacement change of pixels between video frames. The optical flow method is to calculate the displacement vector of the point of interest between the current frame and the previous frame to obtain the information change of the moving target, so as to obtain the time characteristics of the moving target. The data flow is assumed to be a sequential or temporal structure, which consists of continuous data blocks, and the data points in each block originate from the same underlying distribution. The task of segmentation is carried out unsupervised; that is, there is no label or segmentation boundary given a priori.

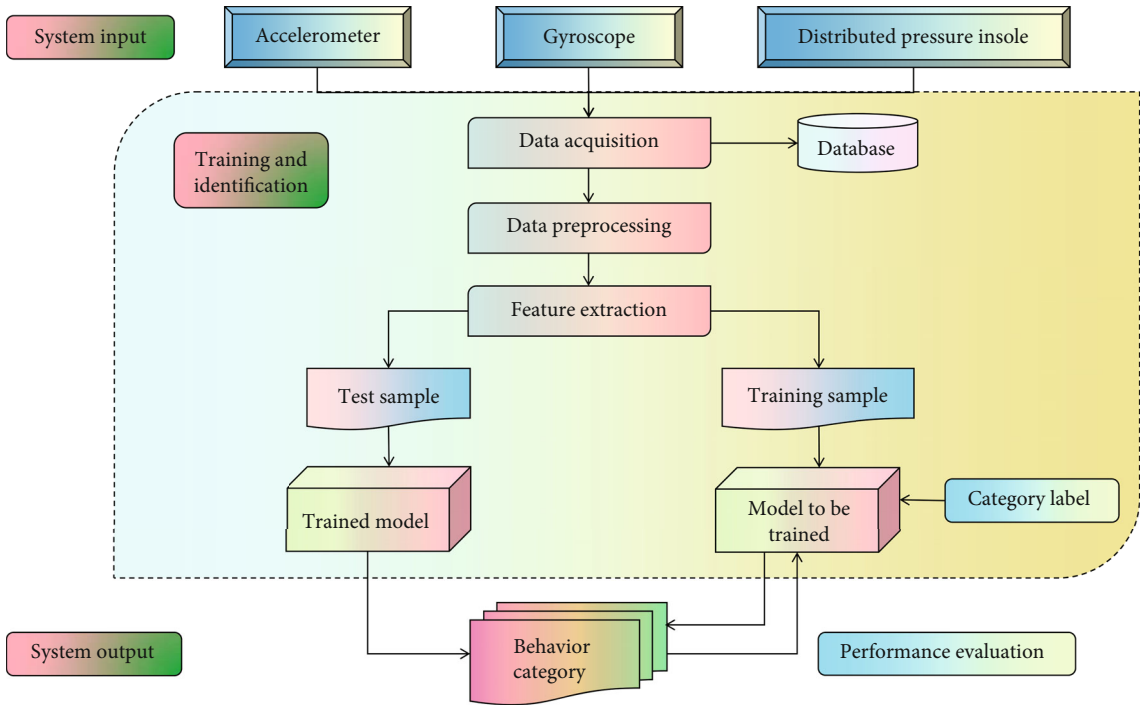

Figure 1: Structure of human action identify system.

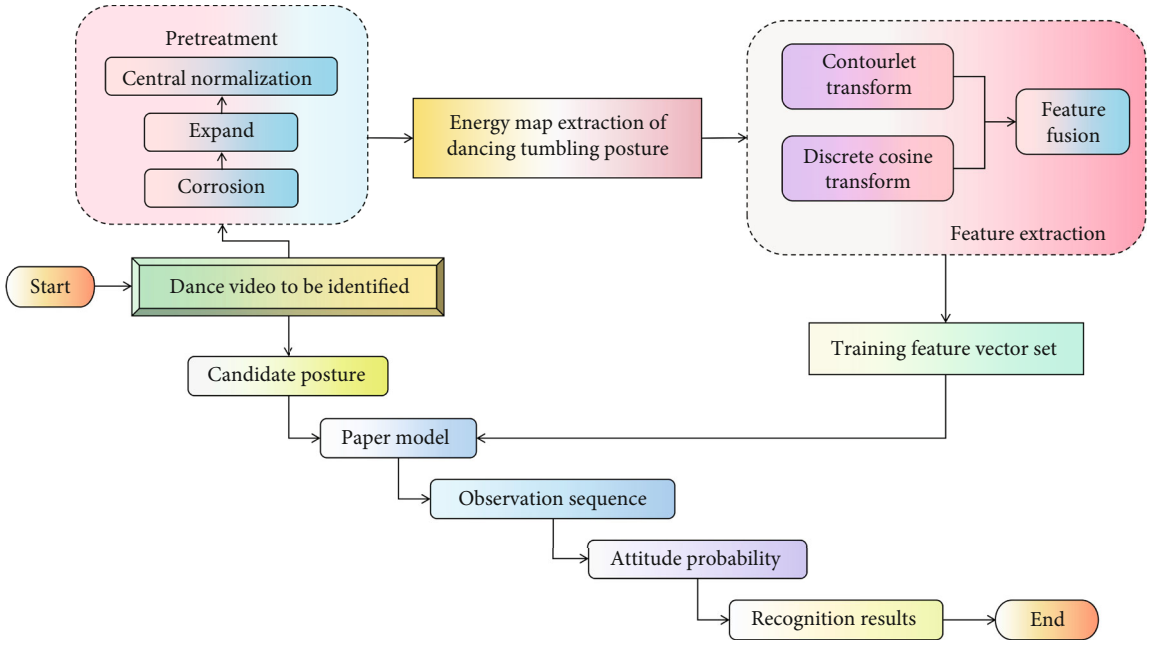

Figure 2: Dance somersault pose recognition process.

3.2. Realization of Dance Somersault Pose Recognition Model Based on Multifeature Fusion. There are many kinds of video states, different time periods, and different frame rates. However, most videos contain many kinds of actions, including both the same actions and different actions. The acquisition of effective actions in video will greatly reduce the amount of computation in the recognition process. Calculate that number of pixel included in each image of the video to be identified, obtaining the frame numb with the minimum number of target pixels, setting three adjacent image frames with the minimum number as a complete posture period, and obtaining the frame number of posture images in the period. The dance video data set used in this paper is directly recorded, with high resolution. At the same time, it also contains a lot of noise during the conversion, which will affect the extraction of edge features in dance

video images. Therefore, we need to preprocess the dance data set first.

Assuming that a dance tossing posture cycle has n frames of images, after preprocessing by the center normalization method, the t-th frame of dance tossing posture images can be obtained as $B_i(x, y)$, and the dance tossing posture energy map corresponding to the t-th frame of dance tossing images can be obtained. The formula is as follows:

$$G(x, y) = \frac{1}{n}\sum_{t=1}^{n} B_t(x, y). \tag{1}$$

In the formula, $G(x, y)$ is the grayscale image, and the grayscale value of each pixel in the image is the energy in the dance tossing posture period of the point, that is, the frequency of the pixel appearing here during the dance tossing process. Let $\{f(x, y), x, y = 0, 1, 2, \cdots, N-1\}$ represent a dance image of size $N \times N$, and the calculation formula of the two-dimensional discrete cosine transform of the image is as follows:

$$F(u, v) = a(u)a(v) \cdot \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x, y) \cos\frac{(2x+1)u\pi}{2N} \cos\frac{(2y+1)v\pi}{2N}. \tag{2}$$

In the formula, $f(x, y)$ represents the gray value of the pixel coordinates (x, y) in the dance tossing posture energy map. u represents the transformation rate of the pixel point (x, y) in the horizontal direction. v represents the vertical transformation rate of the pixel point (x, y), that is, the vertical spatial frequency. $F(u, v)$ represents the frequency coefficient after discrete cosine transform. Among them, $F(0, 0)$ represents the DC part of the image frequency domain feature. The gesture recognition network is used for iterative prediction, and the calculation method of the prediction is as follows:

$$S^I = \rho^I\left(F, S^{I-1}, L^{I-1}\right), \quad \forall_t \geq 2, \tag{3}$$

$$L' = \Phi^t\left(F, S^{t-1}, L^{t-1}\right), \quad \forall_t \geq 2. \tag{4}$$

In the formula, ρ and Φ represent the convolution operations corresponding to the S and L branches, respectively.

The detailed information such as the speed and shape of the human body's dance somersault posture can be reflected by the dance somersault posture energy map, and the extracted dance somersault posture energy map can filter out random noise, which has high robustness. Background subtraction is a method for detecting foreground objects in the video. Background subtraction is usually to establish a background model first and learn the parameters of the model. After learning the parameters, the image of the current frame is compared with the background model. Any area with large differences is considered as a foreground object.

In order to obtain richer semantic features, this paper uses residual module to enhance feature extraction of each layer. Discrete cosine transform is used to extract the frequency domain features of the energy map of dance somersault pose, and the frequency domain information is used to divide the high- and low-frequency components in dance somersault pose. The spatial contour features in the energy map of dance somersault posture are extracted by Contourlet transform.

$L2$ norm is used to optimize the loss function of Euclidean distance, as shown in

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^{N} \|F(X_i; \theta) - F_i\|_2^2. \tag{5}$$

In the formula, θ represents the parameters of the dancer action recognition network to be optimized. N is the total number of dancer pictures participating in the learning and training. X_i represents the currently learned dancer image sample i. F_i represents the true value of the heat map of the ith dancer image. $F(X_i; \theta)$ represents the heat map of the dancer skeleton keypoints regressed by the model.

Local feature description methods such as local binary method cannot describe the complete dance somersault posture. However, the discrete cosine transform method can distinguish the high-frequency and low-frequency components in the dance somersault posture and distinguish the low-frequency parts such as the head and shoulders with small movements from the high-frequency limb swinging movements in the dance somersault posture, thus effectively extracting the frequency domain features of the dance somersault posture. Specifically, the kernel function can be used to map the classification function into the expression (6). That is, the kernel function $K(x, x_i)$ is used to replace the inner product operation in the case of linear classification, and the output decision function shown in formula (7) is obtained.

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b, \tag{6}$$

$$y = \operatorname{sgn}\left(\sum_{i=1}^{n} a_i Y_i K(x_i, x) + b\right) 0 \leq a_i < < C, \tag{7}$$

where, a_i is the Lagrangian coefficient corresponding to each training sample, C is the penalty parameter, b is the bias, and $K(x_i, x)$ is the kernel function. The function of cross entropy is to compare the difference between the real label of the action video and the predicted action label through Softmax. If the difference between the two is very close, it can prove that the model optimization effect is good. The cross entropy can be expressed by

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) \log\left(q(x_i)\right). \tag{8}$$

In the formula, p represents the actual action of the sample, n represents the number of categories contained in the data set, and q represents the predicted action output by Softmax. Use stochastic gradient descent to fine-tune the built two-stream network model, and the loss function is shown in

$$\text{Loss Function} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_{ij}) \log\left(q(x_{ij})\right). \quad (9)$$

In this paper, $m = 15$ is chosen to represent the size of a small batch.

When studying the method of action identify based on dance video, we should fully consider the differences between dance video and previous action identify video data sets, and at the same time, we should extract relevant features from the dance data sets to represent dance movements according to the characteristics of dance itself. The discrete cosine transform method can be used to accurately distinguish various frequencies of human dance somersault posture, and the calculation is simple, and the frequency domain features of dance somersault posture can be easily extracted. Features that can describe the shape of dance movements can be extracted to represent the dance movements in the data set. Considering that the dance movements in the dance video are composed of a group of frame sequence images, and according to the characteristics of the dance itself, the dance movements are continuous, and the shapes of the movements between adjacent frames have little difference.

For a dance action video, we can divide the dance action video into equal parts. In the image sequence of each equally divided video segment, the edge features of the target in each frame image can be extracted and accumulated into a pair of images, and then, the directional gradient histogram features can be extracted from the images. When all equal segments of the video have completed the operation of accumulating edge features, a group of directional gradient histogram feature vectors can be formed to represent the local appearance and shape features of the dance movement. The low-frequency component in the dance somersault posture energy map is in the upper left corner of the frequency amplitude spectrum, which indicates the region with larger pixel values, that is, the region with slower transformation, which is the main part of the dance somersault posture energy map; the high-frequency component in the energy map of dance somersault pose is located in the lower right corner of the frequency amplitude spectrum, which indicates the area with small pixel value in the map, and this area reflects the details and edge parts of the energy map of dance somersault pose.

4. Result Analysis and Discussion

In this experiment, the dance somersaults recorded by the motion capture device are taken as the data set. The action video contains 786 frames. At the beginning of the dance action, the actor has a long period of preparation action, and the process of the actor returning to the initial state from the last posture after the action is over leads to a gentle change of the joint distance at the end of the above picture and then a sharp change again. The algorithm in this paper is compared with the residual network four-channel algorithm and HOC algorithm, and the comparison result of recognition quantity is obtained, as shown in Figure 3.

The following results can be obtained from the figure: the number of recognition obtained by this algorithm is the closest to the number of actual dance somersaults. This verifies the effectiveness of this algorithm. The algorithm is used for training and recognition. The accuracy of the algorithm on the training set and the test set is shown in Figure 4.

It can be seen from the figure that the average recognition accuracy of this research method on the test set is higher, and the overall recognition accuracy is higher. Because there are many wrong segmentation positions in the initial segmentation position, the difference constraint between frame sequences and the difference constraint between adjacent minima and maxima are added. Compared with other algorithms, this algorithm needs two metrics, namely, precision and recall. The precision rate is the accuracy of the video frame sequence segmentation position, and the recall rate is the percentage of all the correct action sequences found in the manually determined action sequences. After feature extraction, the feature vector has a high dimension, so it is necessary to select the feature vector. In this paper, the method of direct feature selection is used to directly screen the feature vectors used in each layer according to the classification effect of each feature in the actual test. In order to verify the superiority of this algorithm, residual network four-channel algorithm and HOC algorithm are used for identification test on the test set, and the results are shown in Figure 5.

As can be seen from Figure 5, compared with the comparison algorithm, the accuracy rate of this algorithm is higher, reaching more than 94%, which shows that this algorithm has ideal recognition effect on dance movements and higher accuracy rate.

In the research of dance action recognition, only the key positions in the action sequence cannot accurately represent a dance action, and one action sequence contains many simple actions, and there are many unrelated actions between two simple actions, so it is difficult to screen the transitional actions between simple actions. Therefore, in this paper, the whole motion sequence obtained by segmentation is selected to represent a dance motion, which is the basis of the following feature extraction image sequence. In order to make a reasonable evaluation for each feature, the recognition results of a single feature on two dance data sets and the recognition results of feature fusion method using multicore learning are given and compared, and the influence of the three features on the experimental results is also analyzed. In order to further test the recognition performance of this model, this model is compared with hierarchical model and dynamic path model. The recognition errors of the three models are shown in Figure 6.

From the experimental results in the figure, it can be seen that the misidentification rate of the dance tumbling

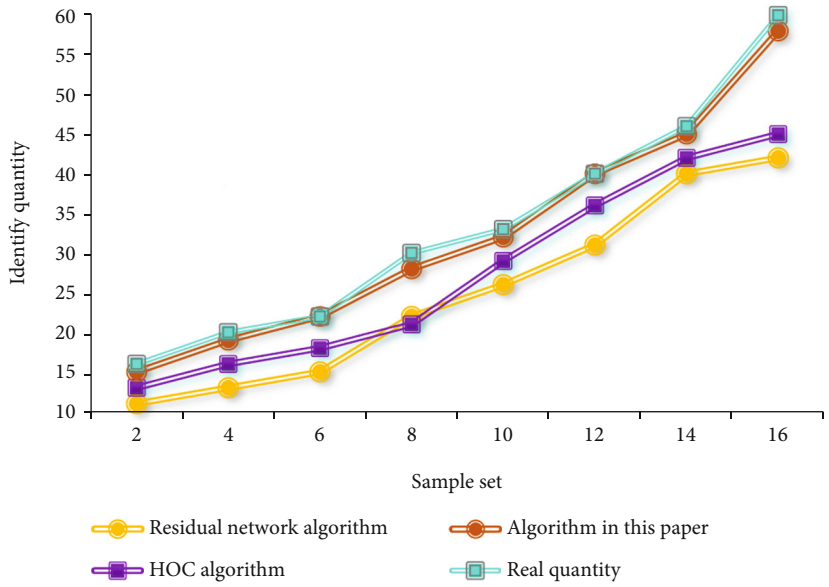

Figure 3: Comparison of recognition quantity of different algorithms.

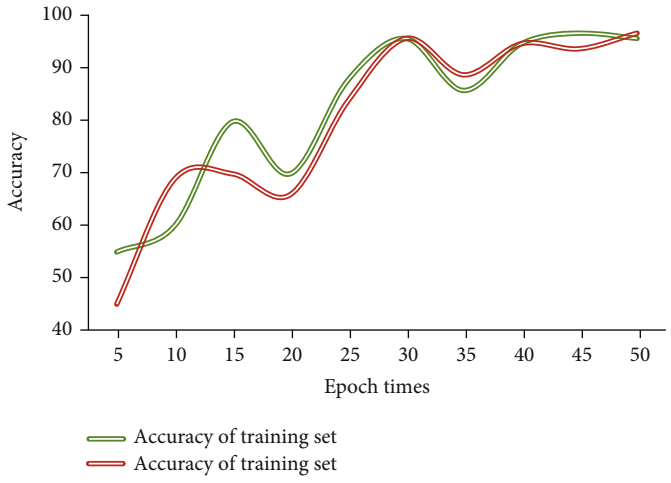

Figure 4: Accuracy of training set and test set.

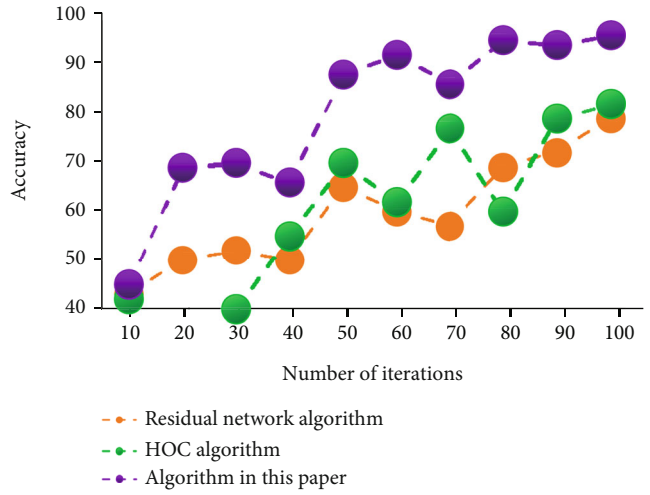

Figure 5: Comparison of accuracy of different algorithms.

pose in the sample video identified by this model is all below 1%, which is significantly lower than the misidentification rate of the dance tumbling pose identified by the hierarchical model and the dynamic path model. This result verifies the accuracy of the dance tumbling pose identified by this model in the video.

According to the posture of the human body, the upper body, lower body, and whole body image areas and main video action sequences are determined, and then, 3D-SIFT features and optical flow features are extracted, respectively. Each feature is reduced in dimension and normalized by PCA, respectively, and the main information of the obtained features is combined to form a descriptor of an action sequence. Classify the first classifier, and input the classification results to the classifiers of the next level, respectively, and so on until finally, divide each object to be classified into a separate category. This method makes full use of the apriority of human behavior. During each classification, the categories within subclasses are similar, while the categories between subclasses are obviously different, thus ensuring the effectiveness of classification.

In order to further verify the recognition efficiency of this algorithm, this algorithm is compared with other algorithms, and the running efficiency of each algorithm is obtained. The comparison is shown in Figure 7.

It can be seen that the algorithm in this paper is efficient and takes the least time. Compared with the comparative algorithm, the running efficiency of this research algorithm is the best. Therefore, this algorithm has higher efficiency and better performance. The experimental time shows that the running speed of this algorithm on the graphics card is 0.85 frames/s, and it can identify multiperson dance

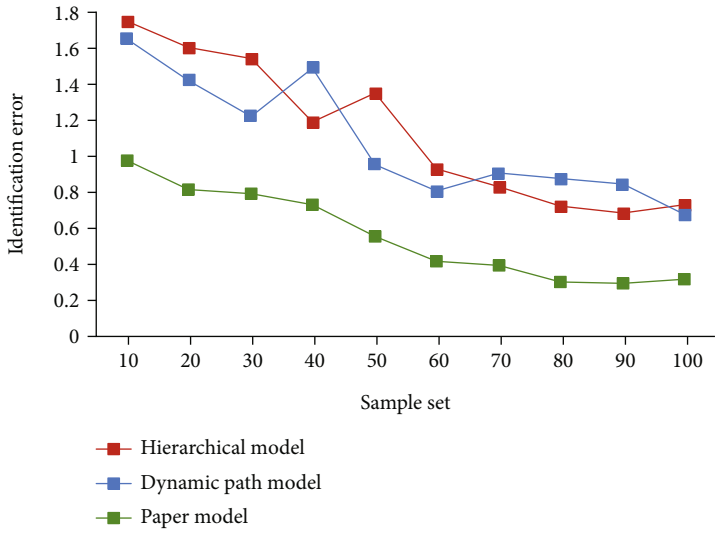

Figure 6: Identification errors of three models.

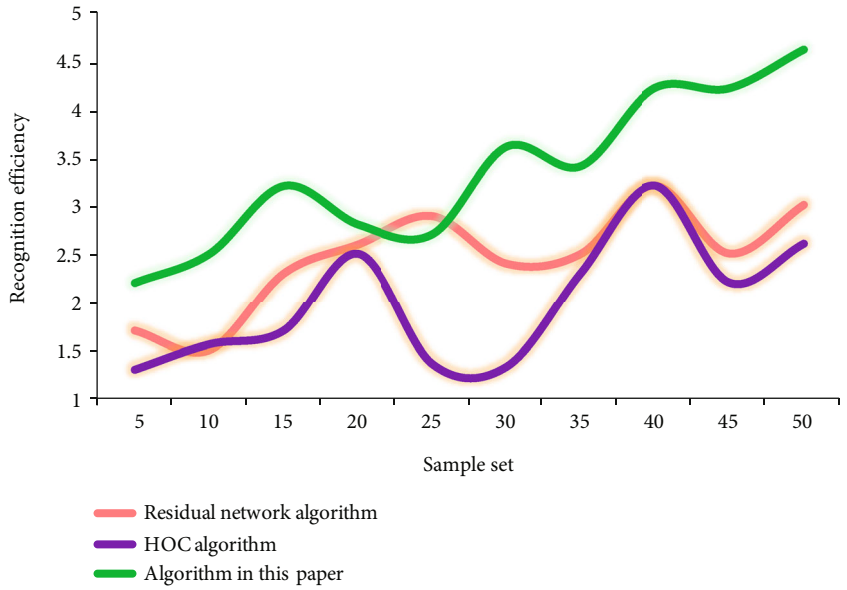

Figure 7: Comparison of running efficiency of different algorithms.

movements in a single picture. In this paper, through experiments on data sets, it is concluded that the model constructed in this paper can better extract the dance contour images with space-frequency characteristics and effectively identify the dance somersault pose in dance videos, and the recognition accuracy of the dance somersault pose in shadow dance video images is high.

5. Conclusions

Dance somersault skill is an important index to evaluate dancers' basic skills and comprehensive abilities. Identifying dance somersault posture is of great significance to art teaching and research. This paper presents a dance somersault pose recognition model based on multifeature fusion algorithm. This paper mainly studies the selection and representation of features and multifeature fusion methods in the research of dance somersault pose recognition. The discrete cosine transform is used to extract the frequency domain features of the energy map of dance somersault posture, and Contourlet transform is used to extract the spatial contour features of the energy map of dance somersault posture, which effectively improves the accuracy of identifying dance somersault posture by using hidden Markov model. A sensor data acquisition system suitable for this study is built, and the method of dance somersault gesture recognition is studied from the perspective of multisensor fusion. In this paper, the recognition model of dance somersault posture based on multifeature fusion algorithm is verified, and the detection results are analyzed. Experiments on data sets show that

the model constructed in this paper can better extract the dance contour images with space-frequency characteristics and effectively identify the dance tumbling posture in dance videos, and the recognition accuracy of the dance tumbling posture in 100 shadow dance video images is higher than 96.3%. Although this paper has made some progress in the research of dance somersault gesture recognition, there are still some areas to be improved due to my limited research time and level. For example, this paper has not studied the feature fusion in depth, so it is not sure which feature combination method is more effective when global-local, static-dynamic, and other methods are used to fuse features. In the future work, the recognition model of dance somersault posture needs further research and improvement.

Data Availability

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

I declare that I have no conflict of interest for the publication of this paper.

References

[1] W. Ding, K. Liu, H. Chen, and F. Tang, "Human action recognition using similarity degree between postures and spectral learning," *IET Computer Vision*, vol. 12, no. 1, pp. 110–117, 2018.

[2] J. Li, M. Xia, L. Chen, and L. Wang, "Human interaction recognition fusing multiple features of depth sequences," *IET Computer Vision*, vol. 11, no. 7, pp. 560–566, 2017.

[3] D. K. Vishwakarma, "A two-fold transformation model for human action recognition using decisive pose," *Cognitive Systems Research*, vol. 61, no. 6, pp. 1–13, 2020.

[4] H. Kim, S. Park, H. Park, and J. Paik, "Enhanced action recognition using multiple stream deep learning with optical flow and weighted Sum," *Sensors*, vol. 20, no. 14, p. 3894, 2020.

[5] L. P. Kirsch, D. Nadine, D. K. Sumanapala, and E. S. Cross, "Dance training shapes action perception and its neural implementation within the young and older adult brain," *Neural Plasticity*, vol. 2018, Article ID 5459106, 20 pages, 2018.

[6] D. K. Sumanapala, J. Walbrin, L. P. Kirsch, and E. S. Cross, "Neurodevelopmental perspectives on dance learning: Insights from early adolescence and young adulthood," *Progress in Brain Research*, vol. 237, 2018.

[7] T. Han, H. Yao, C. Xu, X. Sun, Y. Zhang, and J. J. Corso, "Dancelets mining for video recommendation based on dance styles," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 712–724, 2017.

[8] M. Qadri and R. G. Cook, "Pigeons and humans use action and pose information to categorize complex human behaviors," *Vision Research*, vol. 131, pp. 16–25, 2017.

[9] A. Faridee, S. R. Ramamurthy, and N. Roy, "HappyFeet: challenges in building an automated dance recognition and assessment tool," *Mobile Computing and Communications Review*, vol. 22, no. 3, pp. 10–16, 2018.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, 2016.

[11] Y. Liu, R. Ma, H. Li, C. Wang, and Y. Tao, "RGB-D human action recognition of deep feature enhancement and fusion using two-stream ConvNet," *Journal of Sensors*, vol. 2021, Article ID 8864870, 10 pages, 2021.

[12] D. Wang, J. Yang, Y. Zhou, and Z. Zhou, "Human action recognition based on deep network and feature fusion," *Univerzitet u Nišu*, vol. 34, no. 15, pp. 4967–4974, 2020.

[13] R. Kavi, V. Kulathumani, F. Rohit, and V. Kecojevic, "Multiview fusion for activity recognition using deep neural networks," *Journal of Electronic Imaging*, vol. 25, no. 4, p. 043010, 2016.

[14] J. Xie, W. Xin, R. Liu et al., "Global co-occurrence feature and local spatial feature learning for skeleton-based action recognition," *Entropy*, vol. 22, no. 10, p. 1135, 2020.

[15] H. Wei, R. Jafari, and N. Kehtarnavaz, "Fusion of video and inertial sensing for deep learning-based human action recognition," *Sensors*, vol. 19, no. 17, p. 3680, 2019.

[16] C. I. Patel, S. Garg, T. Zaveri, A. Banerjee, and R. Patel, "Human action recognition using fusion of features for unconstrained video sequences," *Computers and Electrical Engineering*, vol. 70, pp. 284–301, 2018.

[17] P. Wang, "Research on sports training action recognition based on deep learning," *Scientific Programming*, vol. 2021, Article ID 3396878, 8 pages, 2021.

[18] H. Ou and J. Sun, "Spatiotemporal information deep fusion network with frame attention mechanism for video action recognition," *Journal of Electronic Imaging*, vol. 28, no. 2, p. 1, 2019.

[19] L. Huang, Y. Huang, W. Ouyang, and L. Wang, "Part-aligned pose-guided recurrent network for action recognition," *Pattern Recognition*, vol. 92, pp. 165–176, 2019.

[20] D. Yong, F. Yun, and W. Liang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3010–3022, 2016.

[21] S. Yu, Y. Cheng, L. Xie, and S. Z. Li, "Fully convolutional networks for action recognition," *IET Computer Vision*, vol. 11, no. 8, pp. 744–749, 2017.

[22] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," *Pattern Recognition Letters*, vol. 118, no. 2, pp. 14–22, 2018.

[23] C. Pei, F. Jiang, and M. Li, "Fusing appearance and motion information for action recognition on depth sequences," *Journal of Intelligent Fuzzy Systems*, vol. 40, no. 3, pp. 4287–4299, 2021.

[24] X. Li and H. Zheng, "Target detection algorithm for dance moving images based on sensor and motion capture data," *Microprocessors and Microsystems*, vol. 81, no. 2, p. 103743, 2020.

[25] D. Zhang, L. He, Z. Tu, S. Zhang, F. Han, and B. Yang, "Learning motion representation for real-time spatio-temporal action localization," *Pattern Recognition*, vol. 103, no. 1, p. 107312, 2020.

[26] H. Xu, Q. Tian, Z. Wang, and J. Wu, "A joint evaluation of different dimensionality reduction techniques, fusion and learning methods for action recognition," *Neurocomputing*, vol. 214, no. 3, pp. 329–339, 2016.

[27] I. Ajili, M. Mallem, and J. Y. Didier, "Human motions and emotions recognition inspired by LMA qualities," *The Visual Computer*, vol. 35, no. 10, pp. 1411–1426, 2019.

[28] S. Ma, "Music rhythm detection algorithm based on multipath search and cluster analysis," *Complexity*, vol. 2021, Article ID 5627626, 9 pages, 2021.