Hindawi
Journal of Sensors
Volume 2024, Article ID 9824980, 1 page
https://doi.org/10.1155/2024/9824980

Retraction

Retracted: Visual Sensing Human Motion Detection System for Interactive Music Teaching

Journal of Sensors

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

Copyright © 2024 Journal of Sensors. 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.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1].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) Manipulated or compromised peer review

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] X. Chang and L. Peng, "Visual Sensing Human Motion Detection System for Interactive Music Teaching," *Journal of Sensors*, vol. 2021, Article ID 2311594, 10 pages, 2021.

Hindawi
Journal of Sensors
Volume 2021, Article ID 2311594, 10 pages
https://doi.org/10.1155/2021/2311594

RETRACTED

Research Article

Visual Sensing Human Motion Detection System for Interactive Music Teaching

Xunyun Chang and Liangqing Peng

Shaoyang University, Hunan Province, China

Correspondence should be addressed to Xunyun Chang; 3582@hnsyu.edu.cn

Received 17 August 2021; Accepted 1 October 2021; Published 19 November 2021

Academic Editor: Haibin Lv

Copyright © 2021 Xunyun Chang and Liangqing Peng. 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.

The purpose is to study the interactive teaching mode of human action recognition technology in music and dance teaching under computer vision. The human action detection and recognition system based on a three-dimensional (3D) convolutional neural network (CNN) is established. Then, a human action recognition model based on the dual channel is proposed on the basis of CNN, and the visual attention mechanism using the interframe differential channel is introduced into the model. Through experiments, the performance of the system in the process of human dance image recognition based on the Kungliga Tekniska Högskolan (KTH) dataset is verified. The results show that the dual-channel 3D CNN human action recognition system can achieve high accuracy in the first few rounds of training after the frame difference channel is added, the error can be reduced quickly, and the convergence can start quickly; the recognition accuracy of the system on KTH dataset is 96.6%, which is higher than that of other methods; for $3 \times 3 \times 3$ basic convolution kernel, the best performance of the classification network can be obtained by pushing forward 0.0091 seconds in the calculation. Thereby, the dual-channel 3D CNN recognition system has good human action recognition accuracy in the dance interactive teaching mode of music teaching.

1. Introduction

Music teaching mode is a professional term in teaching theory. It is a kind of teaching activity structure, which reflects the specific logic of teaching theory and aims to achieve a relatively stable teaching task [1]. In the traditional teaching mode, teachers show students and students imitate and practice. Although this teaching method has been recognized to a great extent in the practical application of education for a long time, its limitations are increasingly obvious with the change of students' knowledge and skill structure and the increase of music teaching content and subjects [2]. Too much emphasis on the rigid quantitative trend of music and dance skills training means that even if some students can perform high-level dance, they cannot effectively use dance, such as improvisation, which requires a high degree of freedom. The music teaching method under the information technology environment can effectively integrate the method into the interaction of dance teaching through the integration of human movement recognition and traditional dance teaching concept and through the design of the educational framework of multimedia interactive dance course learning. It can help students understand the connotation of dance deeply and guide students to master the learned movements. It can also effectively improve teachers' demonstration and students' imitation of traditional teaching methods and increase the interaction between teachers and students [3].

Pham et al. (2018) [4] showed that the computer vision community is currently committed to solving the problem of action recognition in real videos, which contain thousands of challenging samples. In this process, a deep convolutional neural network (CNN) plays a crucial role in various vision-based action recognition systems. Zhang et al. (2019) [5] reviewed the latest methods of human behavior recognition, including the artificial design progress of action characteristics in *R* (red), *G* (green), *B* (blue), and deep data, the current action feature representation method based on deep learning, the progress of human-computer interaction recognition method, and the outstanding research topics of current action detection methods. Chen et al. (2016) [6] proposed a human motion recognition method based on a depth motion map. Each depth frame in the depth video sequence is projected on three orthogonal Cartesian planes. In each projection view, the absolute difference between two consecutive projections is accumulated by forming a complete video sequence in depth. At present, there are some achievements on human action recognition, but there is a lack of research on dance action recognition in music teaching. The dance action recognition model established aims to give new inspiration for the traditional music teaching mode.

The method of establishing a model and verifying the results in the dataset is adopted. The innovation is that the new computer vision sensing technology is introduced into the dance action teaching in music teaching, which is adopted to identify the dance action. It is verified that the model recognition accuracy is high. It breaks the traditional teaching mode of teacher demonstration and student practice, increases the interaction between teachers and students, and provides new ideas for future music teaching.

2. Research Method

2.1. The Human Action Recognition Method. Human motion recognition mainly refers to the analysis and recognition of the species and behavior patterns of the monitored organisms [7]. Human vision system can recognize and analyze human motion in one millisecond. Human motion recognition based on computer vision needs to analyze, sort, and calculate the basic information of human behavior in video by using computer vision technology, so as to extract high-level semantics, which can represent the characteristics of human motion position, work, and behavior. Their behavior can be described and inferred, so that the device has the ability to understand human behavior in the video [8]. Figure 1 displays the human body recognition process.

Previous human motion recognition mainly focused on visible image sequences. Visible images are easy to see, and the interference of lighting conditions and complex background will affect the posture of the human body in color, texture, and light output in the actual scene; so, motion modeling has great challenges. Human motion has rich visual, temporal, and spatial information, but the lack of scene depth information makes it difficult for computers to fully simulate human motion [9]. The existence of visual sensors makes the computer simulation of human motion segments become diverse, which can well express the spatiotemporal performance. The task of recognizing human actions is no longer so difficult with the accumulation of information technology and computer animation.

Figure 1 reveals that in the video for computer vision recognition, human motion is a dynamic process, and different dance movements have different rules. Luminous flux, motion path, spatiotemporal interest points, and other methods can be used to express human motion. After the features that can represent human dance movements are obtained, it is essential to determine the type of human dance movements according to the extracted features; that is, the recognition and classification of dance movements can be realized. The methods used in typical classification motion recognition include template matching, state space, and deep learning [10].

2.2. Scheme of the Human Action Recognition System. An action recognition system is designed based on the above research, which includes an action capture module, action segmentation module, action training module, and action recognition module. Figure 2 displays the overall architecture of the system. After the action capture module, it enters the action segmentation and judges whether the object starts to execute the action by changing the motion state of the human body. If it is, the action is segmented and transferred to the action recognition module. In the action recognition module, the dual-channel 3D (three-dimensional) CNN human action recognition method is adopted to calculate the dance action sequence data through the code and matrix transformation obtained from the training module, and then the action recognition is carried out. Finally, the result of the action is set and returned to the user interface to complete the whole process.

2.3. System Development Environment. The realization of an action recognition system needs a development environment, including hardware environment and software environment. The main hardware environment is ThinkPad laptop, Windows 10 operating system, CPU Inter(R) Core(TM)i5-4200 processor, 2.6GHz main frequency, 8G memory; A Kinect2_0 for Windows. Software environment is Microsoft Visual Studio 2010 development platform, C# programming language, using WPF development framework Kinect for Windows SDK v2.0, which is released by Microsoft for Kinect 2.0 driver and development kit; MatlabR2016b and some *C* language code.

2.4. Motion Capture and Segmentation Module. Kinect SDK 2.0 [11] is employed to obtain skeleton data. A skeleton stream needs to be initialized to get skeleton stream data; then, the skeleton frame ready event [12] or all frames ready event is recorded, and the skeleton frame is obtained by event or polling mode; besides, the Skeleton Copy Data To method is adopted to copy the data from the skeleton frame of the skeleton object to the object array.

In practical application, the device captures real-time motion and transmits it to the software application in the form of the data stream, and there is no start and end action time; so, it cannot be adopted by the action recognition frame. Moreover, the action segmentation step should be added in the working process of the recognition system. According to the characteristics of human actions, the Kinect data stream provided by actions is divided into action segments, each of which contains complete action information, so as to ensure the accuracy of recognition.

A partitioning method based on human motion state is proposed, that is, to judge whether the human body is in motion state and then to divide the motion sequence in motion state. The main principle of this method is to count the sum of combined displacements in unit time. If the total

RETRACTED

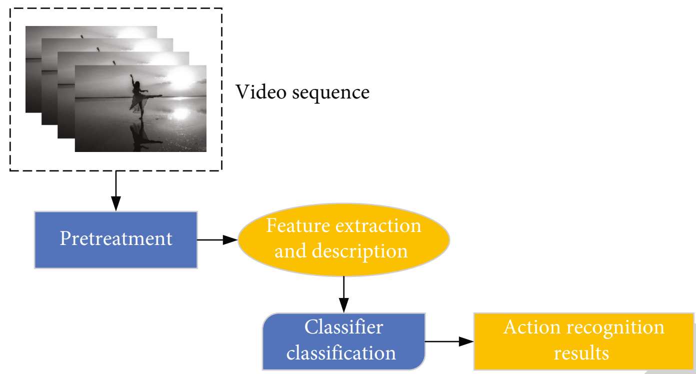


Figure 1: Human action recognition process.

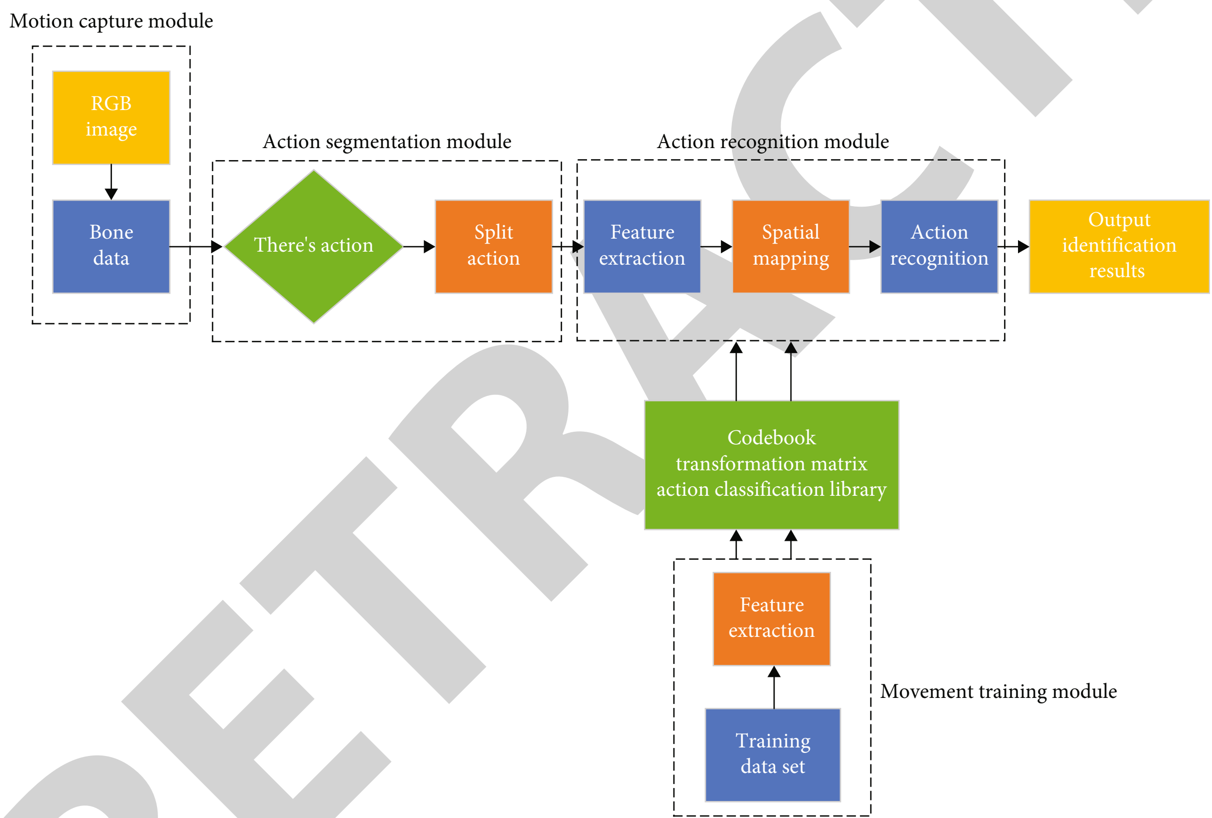


Figure 2: Overall architecture of human action recognition system.

displacement exceeds a predetermined threshold, the current object is judged to execute the action.

2.5. Human Action Recognition Module Based on Dual-Channel 3D CNN. The action recognition unit is the central unit of the whole system, which is mainly responsible for human motion feature extraction and action recognition, and then returns the final results to the user interface.

2.6. CNN. CNN is similar to the biological neural network. Compared with the general deep neural network, its complexity and weight are smaller. Using CNN can realize end-to-end training, so that the image can be directly used as network input, and the network can extract images from features, including color, texture, and shape [13]. Feature layer extraction can form a bottom-up abstract feature. Moreover, the complexity of processing in the algorithm feature extraction, and the traditional data extraction and reconstruction feature are avoided. It has good durability and running efficiency.

Figure 3 displays the overall structure of CNN. The general structure is a classifier and feature extraction. The extracted features include multiple convolutional layers and overlapping subsampling layers. Classifiers generally use a complete connected neural network, and the image data does not need much processing as the direct input of the network. The output is obtained through several steps of extracting features and connecting to the classifier.

RETRACTED

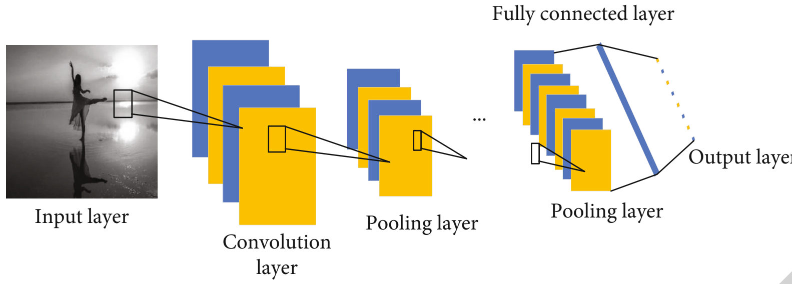


Figure 3: CNN structure diagram.

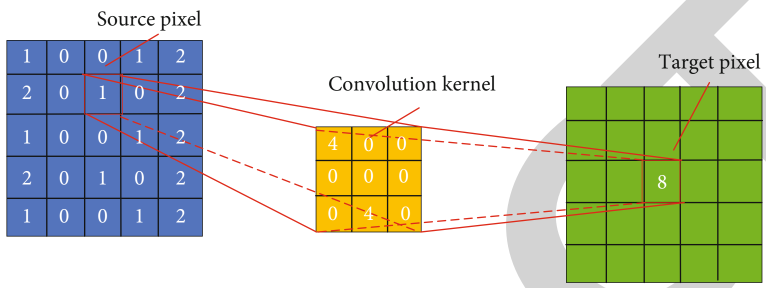


Figure 4: Two-dimensional convolution process.

Convolutional layer is a crucial layer to construct CNN which is composed of a series of learning convolution kernels. In CNN, with the spatial correlation between local layers, the neurons between adjacent layers will not be fully connected, but only connected to the upper neurons in similar regions, that is, local connection. As shown in Figure 4, the basic convolution is to slice the whole image through a very small sensory area in a certain step. Then, the offset of the convolution process shown in equation (1) is adopted to calculate the dot product pixel, and then the result graph is activated through the activation function.

The result of convolution process is calculated by equation (1).

$$x_{mn} = f\left(\sum_{i=0}^{I-1}\sum_{j=0}^{J-1} y_{m+j,n+i} v_{ij} + a\right)(0 \le m \le M, 0 \le n \le N). \quad (1)$$

In (1), y is the image of the input matrix, v is the deviation of the kernel weight matrix $I \times J$, a is the deviation, x is the size of the output $M \times N$, and $f(^*)$ is the activation function.

The other crucial concept of CNN is top-down sampling [14]. The data of the original image will not lose through convolution. If the features obtained in the convolution process are classified directly, massive computation will be generated. Besides, the large-scale image contains rich detailed information, which is easy to fit when it is input into the training network. The characteristic can be restored gradually by sampling. The usual method is to add a pooling layer combined with the convolutional layer, including the internal mean, maximum pooling, and pooling pyramid [15]. The size of the convolutional layer whose original data representation characteristics gradually decrease is obtained through the aggregation process. After pooling, dimension

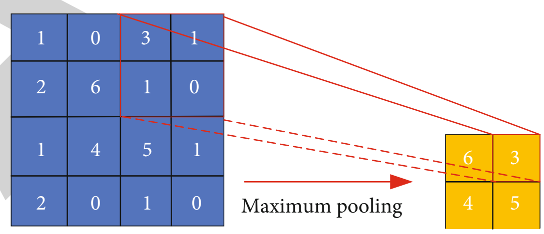


Figure 5: maximum pooling operation.

features and computing workload are reduced, and classification becomes easier. Furthermore, it makes the pan and zoom features fixed. Figure 5 below presents the maximum pooling process:

The sampling result of each window is the maximum average value of the selected sampling window, and the feature map after sampling is 1/4 times of the original size. The process is as follows:

$$\begin{aligned} x_{mn} &= \max_{0 \le i \le R_1, 0 \le j \le R_2}\left(y_{m \times R_1 + i, n \times R_2 + j}\right), \\ x_{mn} &= \frac{1}{R_1 R_2}\sum_{j=0}^{R_2-1}\sum_{i=0}^{R_1-1} y_{m \times R_1 + i, n \times R_2 + j}. \end{aligned} \quad (2)$$

y is the two-dimensional input matrix; x is the output after sampling; m and n are the target pixel positions; R_1 and R_2 represent the sample length in the horizontal and vertical directions, respectively, which are employed to specify the sample window size. Most of the phase samples adopt the nonoverlapping window strategy, that is, each window is connected without overlapping. Downsampling can reduce the size of data quickly, reduce the amount of network computation, and make the network translation and scaling uncertain.

RETRACTED

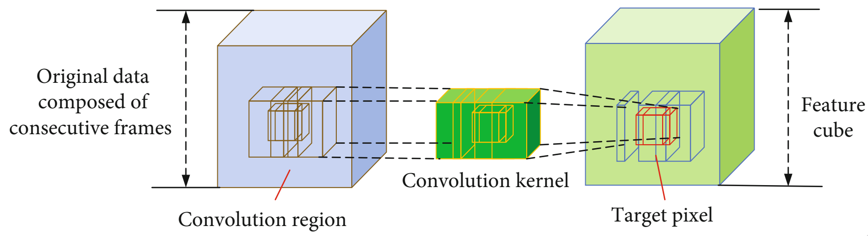


Figure 6: 3D convolution process.

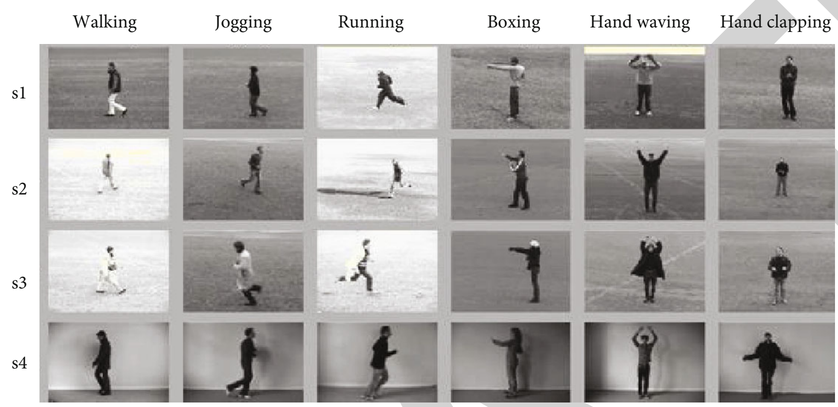
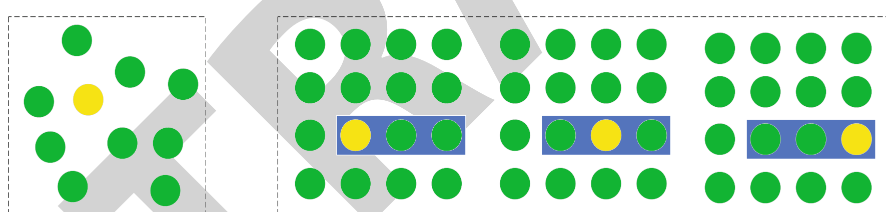


Figure 7: KTH behavior database.



(a) Visual attention region at rest

(b) Visual attention region in motion

Figure 8: An example of visual attention. (a) represents the visual attention area at rest. (b) represents the visual attention area in motion.

RETRACTED

2.7. 3D CNN. Thanks to the core of this structure, convolution in the convolutional layer can extract multiple continuous frame characteristics, features and cube, and previous layers of continuous frame information, so that a period of action can be recorded. 3D CNN is employed to recognize human dance actions in the video, so that the network model can learn the time-series information of human movements. It can avoid the information loss caused by the poor expression ability of time-series information or the use of traditional statistical methods.

3D CNN is a 3D cube in the grid. The advantage of each cube in feature is that it is connected with massive continuous frames in the previous layer, so as to capture information about dance in a period of time. As shown in Figure 6, in the 3D convolution process, the value of a cube is calculated by several frames pressed continuously through a position. The cube eigenvalue of the location is calculated from the local regions of several continuous frames on the convolutional layer.

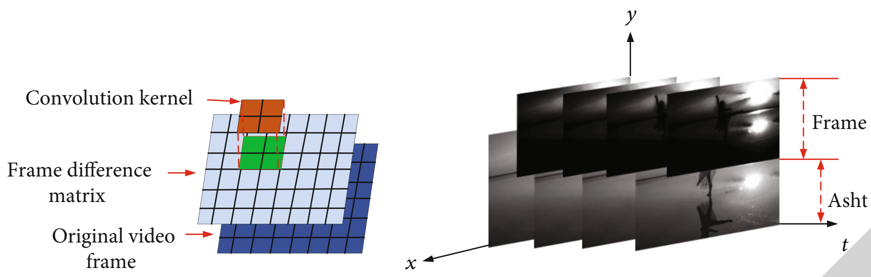
The output of the neurons at the position (i, j, k) of the n-feature cube in the l hidden layer is calculated as follows:

$$x_{lm}^{ijk} = f\left(a_{lm} + \sum \sum_{p=0}^{P_1-1} \sum_{q=0}^{Q_1-1} \sum_{r=0}^{R_1-1} v_{lmn}^{pqr} y_{(l-1)n}^{(i+p)(j+q)(k+r)}\right). \quad (3)$$

y is the input of hidden layer from layer $l-1$ to layer l; x is the output of layer $l\,(i, j, k)$; the size of l-layer convolution kernel is $P_l \times Q_l \times R_l$; $f(*)$ is the activation function; a_{lm} is the common compensation of cube features; n is the pointer to the cube feature, which associates the l layer with the current feature of the cube. v_{lmn}^{pqr} is the weight between neurons of m feature graphs (p, q, r) in layer l and n feature graphs in layer l_1.

In all frame cubes, the same 3D weight ratio and deviation kernel can extract a feature. Multikernel convolution should be employed to extract multiple features in the construction of a 3D neural network model. Moreover, the

RETRACTED



(a) Two channel convolution operation (b) Dual channel input

Figure 9: Schematic diagram of dual channel CNN. (a) denotes dual channel convolution operation. (b) shows the schematic diagram of dual-channel input.

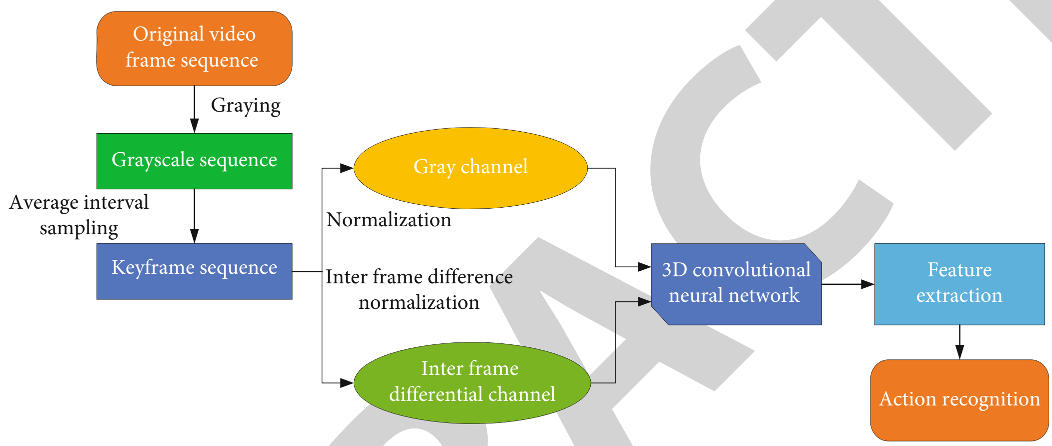


Figure 10: Dual-channel 3D CNN human action recognition process.

number of feature cubes will gradually increase in the input and output direction, so that more and more types of features can be generated to further reduce the features of the cube.

When the video sequence enters 3D CNN, the longest continuous frame must be continuously collected, so as to express the complete dance action information. The data amount will increase dramatically. Besides, information has certain invariance over time. 3D space sampling can reduce the cube volume, reduce the overlap among levels, reduce the difficulty of training, and improve the accuracy of training. For example, the sampling methods often used in 3D CNN include maximum pooling, average pooling, and random pooling. The 3D maximum pooling equation is as follows.

$$x_{i,j,k} = \max_{0\le p\le R_1, 0\le q\le R_2, 0\le r\le R_3,} \left(y_{i\times R_1+p, j\times R_2+q, k\times R_3+r} \right). \quad (4)$$

y is the 3D vector input of the pool layer; x is the output of the pooling layer; R_1, R_2, and R_3 are sampled in three directions continuously. After sampling, the computation time and the strong robustness in time and space are greatly reduced.

2.8. KTH (Kungliga Tekniska Högskolan) Dataset for Human Action Recognition. KTH behavior database [16] is a set of data provided by the KTH Royal Institute of Technology

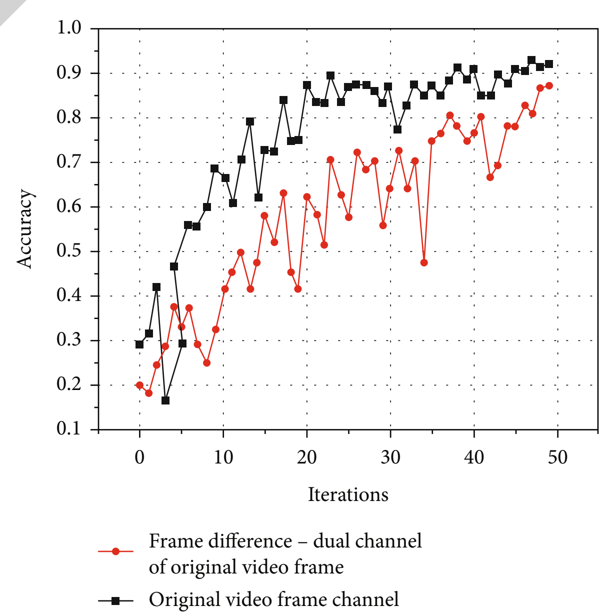


Figure 11: The curve of recognition accuracy of the dual-channel 3D CNN model.

in 2004 to quantify human behavior, as shown in Figure 7. The dataset consists of six actions performed by 25 people in 4 different scenarios: walking, jogging, running, boxing, hand waving, and hand clapping. The first scene is the

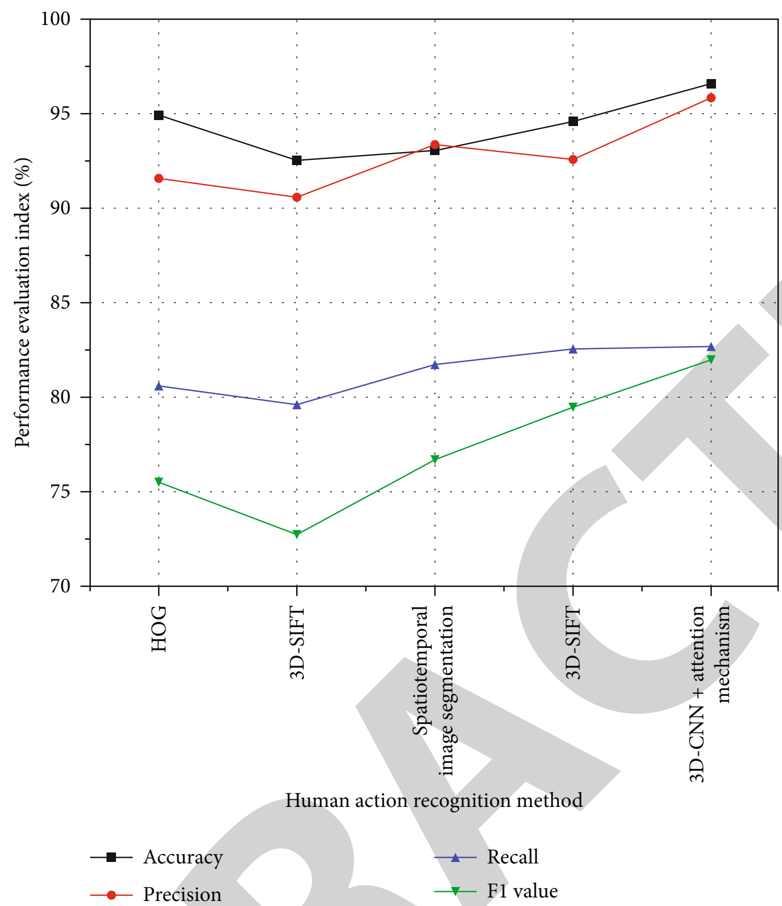


FIGURE 12: Comparison of accuracy, precision, recall, and F1 values of various human behavior recognition algorithms.

normal outdoor fixed angle of view, the second scene is to change the distance and angle outdoors, the third scene is to wear different clothes outdoors, and the fourth scene is indoor. The fixed camera has a 25fps frame rate, a total of 600 videos. Each video has the same behavior 3-4 times, and a total of 2391 video samples can be obtained, including resizing, changing clothes, and lighting changes. 18 people in the dataset are selected as training samples, and 7 people are selected as test samples to test the performance of the algorithm.

2.9. Dual-Channel 3D CNN Human Action Recognition Model. A visual attention mechanism [17] is introduced into the construction of 3D CNN, which can describe the interframe differential channel in the region with obvious changes of human dance action and input into the neural network model together with the original gray video frame to help identify dance action.

The video sequence collected by the camera has the characteristics of continuity. If there is no moving target in the scene, the change of continuous frames is very weak. If there is a moving target, there will be obvious changes between continuous frames. The temporal difference draws on this idea. The target in the scene is moving; so, the image position of the target is different in different image frames. This kind of algorithm performs differential operation on two or three consecutive images in time, subtracts the pixels corresponding to different frames, and judges the absolute value of gray difference. When the absolute value exceeds a certain threshold, it can be judged as a moving target, so as to realize the detection function of moving target.

A crucial advantage of human visual cognition is not to process the whole scene at the same time, but to focus on a part of the visual space, scan the image in a specific order, move from one region to another, and obtain information about the region. The information from these areas is combined with the overall sensory judgment. In a still scene, people's eyes are always focused on different parts of the surrounding environment, as shown in Figure 8(a). Important areas of the still image can be extracted under this principle. Figure 8(b) displays that in the case of motion, the focus of the human visual system is on the part of visual field change, while the static part is ignored, and change is more crucial in judging the current motion.

Visual attention is applied to the 3D CNN motion recognition model by constructing interframe difference channel. Frame difference is one of the common methods to detect and segment moving objects. This process includes the following steps. First, the difference between the adjacent images is obtained by subtracting the corresponding pixel value from them. Then, opening and closing, as well as the threshold binarization, are performed on the difference

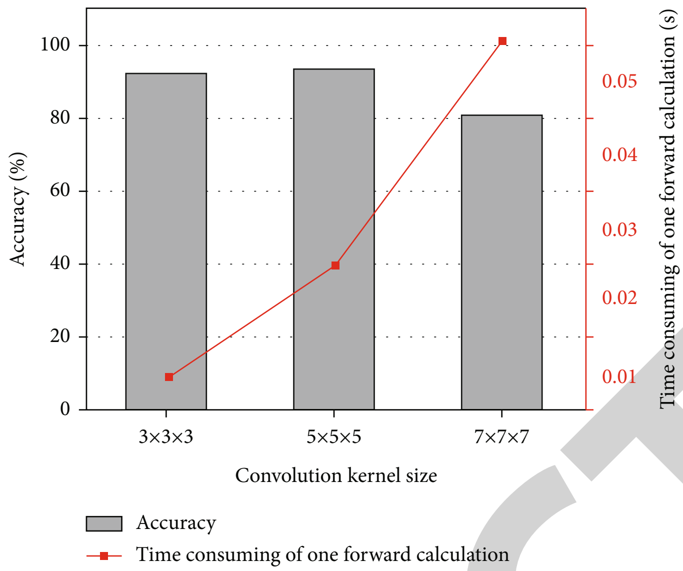


Figure 13: Classification accuracy and time consumption for different convolution kernel sizes.

image. When the microenvironment changes, if the corresponding pixel value change in the region is less than a certain threshold, it is considered as the background region. If the corresponding changed pixel value is greater than the specified threshold in the region, it can be considered that the region is caused by the movement of the object in the image.

The time interval between two adjacent frames is short, and the environment changes little; so, the adjacent image frames can be employed as the background of the current frame model to improve the real-time performance. An interframe differential channel based on visual attention difference is introduced into the 3D CNN of the input layer to expand the neural network model to two channels. The difference matrix is introduced as another channel to input into the neural network model together with the original gray cube video frame, as shown in Figure 9 below.

Figure 10 is a process of human motion recognition using dual channel 3D CNN. The first step is to process the human behavior video in the dataset in gray mode. 18 key images are extracted from the average sampling period as the original video input of 3D CNN. The size is $15 \times 120 \times 160$. The interframe difference channel is calculated based on the extraction of sample frame keys. The key frame is the second channel of 3D CNN, combined with gray video data as training input.

2.10. System Performance Comparison Experiment and Result Analysis. In the system established, a dual-channel 3D CNN model is constructed based on the KTH dataset for dance action recognition. Figure 11 shows the accuracy curve comparison between the constructed 3D CNN model of human action recognition with frame difference channel and only with the original gray data.

Figure 11 shows that the dotted line represents the original video frame channel without adding frame difference channel, and the solid line represents the 3D CNN model with frame difference channel. The network can achieve

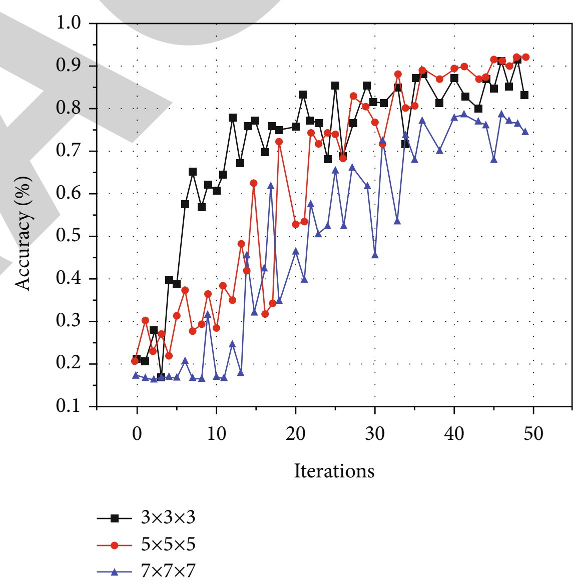


Figure 14: Classification accuracy curves using different convolution kernel sizes.

high accuracy in the first few rounds of training after channel difference frames are added, the error can be reduced rapidly, and the convergence can start quickly. The network can quickly capture the characteristics of human behavior under the effect of the frame difference channel. However, the network without frame difference channel must go to the original image to find usable features and gradually learn human behavior features through dozens of rounds of training. Therefore, introducing the visual attention mechanism of the frame difference channel into 3D CNN can improve the accuracy of human behavior recognition.

Figure 12 presents the working model of the proposed human recognition system, which is compared with other methods to determine human behavior, such as HOG

RETRACTED

RETRACTED

(Histogram of Oriented Gradient) [18], 3D-SIFT (Scale-Invariant Feature Transform) [19], and spatiotemporal image segmentation [20] in the 4 evaluation index directions of accuracy, precision, recall, and F1 value. The accuracy of the 3D CNN model is 94.6% on the KTH dataset and 96.6% on the model with frame difference channel. Compared with other values of precision, recall, and F1 value, it can also show the superiority of the human recognition method in the system.

The data in Figure 12 shows that the model with the HOG feature can also achieve high accuracy by using the manual design model. These models need to use massive complex features in the fusion process. They are very difficult and have a large amount of calculation and poor generalization ability. The method proposed does not rely on the characteristics of industrial design. It uses the powerful self-learning ability of the deep neural network to obtain human behavior features from massive self-training samples and adds deep numbers to learn more abstract features at the same time, which can better describe the different properties of human behavior. After the frame difference channel is added, the network can focus on the part of human action change under the influence of the frame difference matrix, ignore the irrelevant background, and obtain the best recognition ability.

The choice of network convolutional kernel size exerts a significant impact on the network performance and classification accuracy of CNN. The large convolution kernel can bring more sensory fields and extract more features, but it also means more parameters and more computation. Figure 13 displays the experimental results of the convolution kernel with different sizes in the process of network construction.

As shown in Figure 13, smaller convolution kernels are more efficient in the calculation, while the $3 \times 3 \times 3$ convolution kernel only needs to be pushed forward by 0.0091 seconds in the calculation. It takes a longer time as the size of the nuclear twist increases. Convolution kernel $7 \times 7 \times 7$ can achieve higher classification accuracy.

Figure 14 suggests that the convergence rate of using a convolution kernel with the size of $3 \times 3 \times 3$ is only a little faster than that of a convolution kernel with the size of $5 \times 5 \times 5$. When a convolution kernel with the size $7 \times 7 \times 7$ is adopted, the recognition accuracy is low, and the calculation time is long. Overall, the best performance of the classification network can be obtained by using the basic convolution with the size of $3 \times 3 \times 3$.

3. Conclusion

With the development of modern information, as one of the most crucial trends in the field of artificial intelligence, human motion recognition technology has broad application prospects and the importance of research theory. However, traditional feature extraction methods usually need more prior knowledge, based on which features are extracted, which has large workload, low strength, weak ductility, and no extensive adaptive features. In particular, some static scenes, such as dance teaching action recognition in interactive music teaching, have many limitations. Therefore, this exploration aims to use deep neural network to improve the shortcomings of traditional methods, reduce physical labor, enhance durability, and improve the accuracy of dance movement recognition.

The proposed dual channel 3D CNN human motion detection system can provide better performance in the human dance motion dataset collected in the simple environment KTH dataset. For example, the model can be used for intelligent static scene monitoring and motion analysis, such as dance motion analysis. The research deficiency is that the model also needs to further study more complex behavior types, such as dancing with teachers and dancing for two people, so as to broaden the application scope of the model. In the future, it is hoped that the model proposed can be further verified in more fields.

Data Availability

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

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] F. M. Narita, “Informal learning in action: the domains of music teaching and their pedagogic modes,” *Music Education Research*, vol. 19, no. 1, pp. 29–41, 2017.

[2] R. Song, “Research on the application of computer multimedia music system in college music teaching,” *Journal of Physics: Conference Series*, vol. 1744, no. 3, pp. 1–1, 2021.

[3] E. K. Raheb, M. Stergiou, A. Katifori, and Y. Ioannidis, “Dance interactive learning Systems,” *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–37, 2019.

[4] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, “Exploiting deep residual networks for human action recognition from skeletal data,” *Computer Vision and Image Understanding*, vol. 170, pp. 51–66, 2018.

[5] H. B. Zhang, Y. X. Zhang, B. Zhong et al., “A comprehensive survey of vision-based human action recognition methods,” *Sensors*, vol. 19, no. 5, article 1005, 2019.

[6] C. Chen, K. Liu, and N. Kehtarnavaz, “Real-time human action recognition based on depth motion maps,” *Journal of Real-Time Image Processing*, vol. 12, no. 1, pp. 155–163, 2016.

[7] M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition,” *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[8] M. A. Khan, M. Sharif, T. Akram, M. Raza, T. Saba, and A. Rehman, “Hand-crafted and deep convolutional neural network features fusion and selection strategy: an application to intelligent human action recognition,” *Applied Soft Computing*, vol. 87, pp. 105986–105986, 2020.

[9] S. al-Obaidi, H. al-Khafaji, and C. Abhayaratne, “Modeling temporal visual salience for human action recognition enabled visual anonymity preservation,” *IEEE Access*, vol. 8, pp. 213806–213824, 2020.

[10] P. Wang, H. Liu, L. Wang, and R. X. Gao, “Deep learning-based human motion recognition for predictive context-

aware human-robot collaboration," *CIRP Annals*, vol. 67, no. 1, pp. 17–20, 2018.

[11] D. M. Córdova-Esparza, J. R. Terven, H. Jiménez-Hernández, and A. M. Herrera-Navarro, "A multiple camera calibration and point cloud fusion tool for Kinect V2," *Science of Computer Programming*, vol. 143, pp. 1–8, 2017.

[12] Y. Li, D. Ma, Y. Yu, G. Wei, and Y. Zhou, "Compact joints encoding for skeleton-based dynamic hand gesture recognition," *Computers & Graphics*, vol. 97, pp. 191–199, 2021.

[13] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 1044–1048, 2018.

[14] J. Yang, F. Wang, and J. Yang, "A review of action recognition based on convolutional neural network," *Journal of Physics: Conference Series*, vol. 1827, no. 1, pp. 012–138, 2021.

[15] W. Shin, S. J. Bu, and S. B. Cho, "3D-convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance," *International Journal of Neural Systems*, vol. 30, no. 6, article 2050034, 2020.

[16] C. Wiezorek, A. Parisio, T. Kyntäjä et al., "Multi-location virtual smart grid laboratory with testbed for analysis of secure communication and remote co-simulation: concept and application to integration of Berlin, Stockholm, Helsinki," *IET Generation, Transmission & Distribution*, vol. 11, no. 12, pp. 3134–3143, 2017.

[17] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22081–22091, 2017.

[18] C. Hu, C. Fan, and B. Liu, "Nearest atom of local spatiotemporal features for action recognition," *The Journal of Computer Information Systems*, vol. 11, no. 1, pp. 341–348, 2015.

[19] F. Xie, S. Gong, C. Liu, and Y. Ji, "Human behavior recognition based on local and global features of visual words," *Computer Science*, vol. 42, no. 11, pp. 293–298, 2015.

[20] J. Wang, C. Li, and Y. Li, "Human behavior recognition method based on regional Hof and dictionary learning," *Computer Application Research*, vol. 9, pp. 309–312, 2017.

RETRACTED