Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 2538896, 12 pages
https://doi.org/10.1155/2022/2538896

Research Article

The Fusion Application of Deep Learning Biological Image Visualization Technology and Human-Computer Interaction Intelligent Robot in Dance Movements

Nian Jin,[1] Lan Wen,[2,3] and Kun Xie[4]

[1] *College of Music and Dance, Guangzhou University, Guangzhou 510006, China*
[2] *South China Business College, Guangdong University of Foreign Studies, Guangzhou 510545, China*
[3] *Faculty of Education, Guangxi Normal University, Guilin 535400, China*
[4] *Chongqing College of Humanities Science and Technology, Chongqing, China*

Correspondence should be addressed to Kun Xie; 15020240133@xs.hnit.edu.cn

Received 3 August 2022; Accepted 6 September 2022; Published 20 September 2022

Academic Editor: Ning Cao

The paper aims to apply the deep learning-based image visualization technology to extract, recognize, and analyze human skeleton movements and evaluate the effect of the deep learning-based human-computer interaction (HCI) system. Dance education is researched. Firstly, the Visual Geometry Group Network (VGGNet) is optimized using Convolutional Neural Network (CNN). Then, the VGGNet extracts the human skeleton movements in the OpenPose database. Secondly, the Long Short-Term Memory (LSTM) network is optimized and recognizes human skeleton movements. Finally, an HCI system for dance education is designed based on the extraction and recognition methods of human skeleton movements. Results demonstrate that the highest extraction accuracy is 96%, and the average recognition accuracy of different dance movements is stable. The effectiveness of the proposed model is verified. The recognition accuracy of the optimized F-Multiple LSTMs is increased to 88.9%, suitable for recognizing human skeleton movements. The dance education HCI system's interactive accuracy built by deep learning-based visualization technology reaches 92%; the overall response time is distributed between 5.1 s and 5.9 s. Hence, the proposed model has excellent instantaneity. Therefore, the deep learning-based image visualization technology has enormous potential in human movement recognition, and combining deep learning and HCI plays a significant role.

1. Introduction

Modern technologies, such as the Internet and multimedia technology, have developed rapidly. Multimedia systems based on computer information technology have been applied in many fields. The intelligent interactive multimedia is a new platform that develops under the foundation of computer technology [1, 2]. However, the applications of traditional multimedia systems are often independent and mechanized, which are inadequate to meet people's needs. Consequently, the human-computer interaction (HCI) technology emerged. People can interact with the multimedia engine and obtain the required media information quickly and efficiently via HCI. Besides, HCI technology can promote the accurate transmission of information and improve work efficiency [3–5], which has triggered a research boom. In daily life, people can directly express their thoughts or emotions through movements. Therefore, movement recognition and analysis have become a critical direction in the field of HCI and attracted widespread attention, which leads to the wide popularity of human movement-based recognition technology [6, 7].

With the advancement of social informatization, human beings have an increasing requirement in the intelligence level of computers. HCI no longer only depends on the original hardware-based interaction, and some relatively more intelligent interaction methods gradually appear in mass life. The face recognition, gesture recognition, and speech recognition systems constructed by machine learning technology have established a bridge between humans and computers [8]. The emergence of these convenient interaction modes has become a major development trend in the field of HCI. The development of HCI mode aims to enable the computer to serve and adapt to human needs well, so HCI focuses on humans instead of adapting to the computer. Therefore, the friendly interaction between robots and humans is extremely vital in the research of machine learning and HCI. Some scholars focus on the importance of emotional factors related to the interaction between people and computer systems, when exploring the people-centered interaction systems [9]. Motion recognition technology is essentially a classification problem close to machine learning [10].

The above research results imply that the development of the Internet and multimedia technology has made multimedia systems successfully applied to many fields. The friendly interaction between robots and human beings plays an extremely important role in the study of machine learning and HCI. Deep learning shows excellent application potential in function extraction and HCI. A combination of deep learning and HCI is innovatively proposed to extract and identify human skeleton operations to expand the application field of HCI. The ultimate research purpose is to achieve a significant reduction in time costs and dependence on traditional equipment and facilities. The innovative ideas can also achieve the purpose of improving human-computer collaboration and interaction. Moreover, combined with the image visualization technology based on deep learning and HCI system, it is envisaged that the visual geometric group network (VGGNet) and long short-term memory (LSTM) can be optimized. The final HCI system and the research results of the recognition and analysis of human dance provide a reference value.

The contributions based on the extraction and recognition of human dance movements are as follows:

(1) An optimized VGGNet human skeleton movement extraction algorithm is proposed. Its extraction accuracy reaches 96%, which is significantly better than traditional algorithms.

(2) An optimized multiple LSTM human skeleton movement recognition algorithm is proposed. Its recognition accuracy reaches 88.9%, which is significantly better than traditional LSTMs.

(3) An HCI system based on image visualization is designed, and the interaction accuracy rate reaches 92%.

(4) A reference is provided for more in-depth human movement extraction and recognition, and deep learning methods' application range in HCI systems is expanded.

2. Literature Review

2.1. Current Situation of Deep Learning in Dance Education. Dance is an important intangible cultural heritage. Dimitropoulos et al. introduced a research project (i-juries) of intangible cultural heritage, emphasizing the importance of 3D dance interaction [11]. Grammalidis et al. introduced an intangible cultural heritage dataset, i-treasure, including audio and other data information [12]. Doulamis et al. considered that intangible cultural heritage was an important source of cultural diversity, but there were few electronic documents of intangible cultural heritage. According to the "Terpsichore" project funded by the Horizon 2020 of the European Union, they proposed a high-level method based on the digitization of cultural assets [13]. Doulamis et al. discussed the digitization of tangible and intangible cultural heritage and proposed that 3D digital assets would develop into a part of augmented, virtual, and mixed reality experience [14]. Lv studied the application of virtual reality (VR) in 3D environment and HCI system and revealed the excellent performance of VR technology in 3D digitization [15]. The digitization of intangible cultural heritage has become an inevitable development trend, so has dance.

On the recognition and extraction of dance movements, Rallis et al. proposed a dance summarization method based on 3D capture data of the Vicon motion capture system. They analyzed and studied the automatic extraction of dance patterns. This method was a hierarchical scheme based on the temporal and spatial changes of dance characteristics [16]. Aiming at the preservation and dissemination of dance performance, Aristidou et al. proposed a dance action recognition framework based on Laban analysi which used feature space to capture different dance action components and pointed out a new direction for dance evaluation [17]. In terms of editing and synthesis of dance movements, Aristidou et al. used Laban analysis, radial basis function regression, and interpolation methods to map the movement features and emotional features in two directions and realized the stylization of high dynamic dance movements [18]. To sum up, there is a difference between the research of human action recognition and HCI, and there is little research on action recognition in dance education.

2.2. Research Progress of HCI. Experts and scholars have made great efforts on deep learning and HCI. Bhardwaj et al. applied support vector machine and artificial neural network classifier to fingerprint recognition. By integrating the relevant dynamic information from hundreds of biometric scanning sample datasets, they found that the accuracy of fingerprint dynamic recognition by fusing the deep learning method was improved by 5.3% [19]. Israelsen and Ahmed analyzed the influence of artificial intelligence (AI) agent in HCI and machine learning based on the research of algorithm-guaranteed AI agent and discussed the advantages and disadvantages of different methods [20]. Based on similarity embedding, Spathis et al. proposed an interactive dimension reduction framework (iSP). In this framework, user interaction formed different goals. Gradient descent was used for learning, and an end-to-end composition structure could be trained. By evaluating the framework in two interaction scenarios, they found that the framework could be applied to semisupervised learning, transfer learning, and adaptive learning in interaction field [21]. Using interactive machine learning, Wu et al. studied local decision-making in feature selection of emotion classification task and analyzed the influence of interactive machine learning tools on feature selection results [22]. To improve the performance of multimodal image retrieval by using unmarked and marked multimodal web objects, Xu et al. proposed a semisupervised multiconcept retrieval method based on deep learning (SMRDL). Different from the traditional method of using multiple independent concepts in multiconcept semantic query, the proposed method regarded multiple concepts as a whole scene, which was used for multiconcept scene learning of unimodal retrieval. The comprehensive experimental results on two datasets of MIR flickr2011 and NUS-WIDE indicated that the proposed method was superior to some of the latest methods [23]. Long and Zhao held that intelligent teaching mode overcame the shortcomings of traditional online and offline teaching. However, there were some shortcomings in the real-time feature extraction of teachers and students. In view of this, they used particle swarm image recognition and deep learning technology to process the video teaching image of intelligent classroom. To overcome the shortcomings of premature convergence of standard particle swarm optimization (PSO) algorithm, they proposed an improved multi PSO algorithm strategy. Moreover, to improve the premature problem of PSO in search performance, they combined the algorithm with the useful attributes of other algorithms to improve the diversity of particles in the algorithm, enhance the global search ability of particles, and achieve effective feature extraction [24]. To sum up, there are many research results on the application of deep learning in HCI, but few studies on the combination of the two for dance action extraction.

3. Methods

In computer vision and image processing, movement recognition is a crucial component. However, some problems are found in its research and applications. For example, when extracting and recognizing human skeleton movements, bone modeling is challenging, movement amplitude can affect the extraction results, and feature extraction can be insufficient, increasing the difficulty in analyzing and classifying human movements. Deep learning has developed rapidly. CNN shows excellent performance in feature extraction, while LSTM has significant performance in processing time sequence problems. Therefore, CNN and LSTM are introduced to extract and recognize human skeleton movements. However, traditional CNN models have lots of parameters, using a large convolution kernel to extract features. Traditional LSTM models never consider the connection of multiple different movement times in a long time. Hence, the CNN-based VGGNet is introduced and optimized in parallel. In the meantime, LSTM is improved and optimized before extracting and recognizing human skeleton movements.

3.1. Optimization of VGGNet CNN Model. Cat's visual cortex theory inspires the deep learning-based CNN. Compared with the traditional neural network, CNN extracts the object's local feature information through the convolution layer, a critical CNN component that contains multiple convolution kernels [25]. VGGNet is a typical CNN. Unlike traditional CNNs that employ big convolution kernels to extract features, VGGNet utilizes several $3 * 3$ small convolution kernels for feature extraction. Hence, VGGNet can extract richer features and reduce the calculation amount significantly [26–28].

The features extracted by the convolution layer are integrated to improve the accuracy of VGGNet, i.e., the parallel CNN [29–31].

Extractions of input image features before fusion are as follows:

$$y_1^A = F_1^A(x), \tag{1}$$

$$y_1^B = F_1^B(x). \tag{2}$$

In (1) and (2), F_1^A and F_1^B represent features. The feature information extracted by the two small convolution kernels is fused via the feature fusion module. The convolution operation is denoted as G. The feature map after fusion processing can be written as follows:

$$y_1^c = G_1\left(y_1^A, y_1^B\right) = G_1\left(F_1^A(x), F_1^B(x)\right). \tag{3}$$

The process of fusion of the above feature maps y_1^c, y_1^A, and y_1^B can be expressed as follows:

$$y_1^c = merge\left(y_1^A, y_1^B\right). \tag{4}$$

The above fusion processing can enrich and diversify the extracted features. Graphics Processing Unit (GPU) processing is utilized for training VGGNet to compare the performance of the CNN-based VGGNet before and after optimization. Images in the training set are taken by the Kinect camera and the host computer program. The selected human movements include clapping, slapping, standing, picking up objects, and sitting down.

Movement capture includes the following steps: (1) the demonstrator makes different movements in front of the Kinect camera and (2) Kinect is utilized for evaluating human skeleton changes in real-time. Several demonstrators complete the collection of the entire training set. One thousand images are collected for each movement. Finally, a total of 5,000 human skeleton images under different movements are obtained. The skeleton images affected by the environment are removed, and the remaining human skeleton images are retained. These images train the VGGNet before and after feature fusion. Accuracy and loss rates are taken as evaluation indicators [32, 33]. Parameter settings of the entire training process are shown in Table 1.

3.2. Extraction Algorithm of Human Skeleton Movements. Traditional human pose estimation algorithms extract human skeleton features via the bottom-up manner. Each skeleton extraction object requires a detector, and each

Table 1: VGGNet training parameter settings.

Parameters	Training times	Learning rate	Number of images read	Optimizer
Corresponding value	500	10^{-5}	32	Adam

movement is estimated separately. Therefore, traditional algorithms have many problems, such as false detection, long-running time, and poor instantaneity, which cannot meet the demands. Based on the OpenPose open-source database [34], the optimized VGGNet is the network architecture, and the histogram equalization [35, 36] is introduced to suppress noises, thereby extracting the 2D features of the human skeleton.

OpenPose is an open-source database released in 2017 based on skeleton extraction. Unlike traditional pose estimation algorithms, OpenPose uses a bottom-up method. The joint points of all human body parts are detected first. Then, the nodes are connected to obtain the skeleton, thereby significantly reducing the running time. Also, OpenPose can improve detection accuracy and shorten the running time. Figure 1 illustrates the video information processing by OpenPose.

The unique convolution kernel structure in the CNN can learn spatial information in human actions, and more useful information can be obtained by different convolution kernels. Compared with traditional machine learning methods, CNN is more systematic and comprehensive in task learning with better performances. Unlike traditional CNN models, the VGGNet model extracts features by massive small convolution kernels as a typical CNN model. It can extract more features and reduce calculation amount with satisfactory generalization performance. The optimized VGGNet consists of three parts. The first part processes the image data via the input layer and employs CNN to extract the feature values of body parts. Then, the extracted feature values enter the other two parts for critical point positioning and the body-based 2D vector field positioning. The input to output via the neural network spends a total of k periods, and the information input to the current period is the output feature value obtained through the learning process of $k-1$. The optimized VGGNet's output is formed by a 2D vector field of crucial body parts and a confidence map. As the calculations increase, the candidate human body parts and the corresponding structure division become apparent via this cyclic process. Here, CNN's first convolutional layer is a double convolutional layer, and each contains 64 convolution kernels in the size of $4*4$. Simultaneously, an activation layer and a normalization layer are added after each convolutional layer to process the nonlinear data. A pooling layer is added after the normalization layer to reduce dimensionality and prevent overfitting, located between the two convolutional layers. The Dropout layer comes after the second pooling layer. The Part Affinity Fields (PAFs) [37, 38] are adopted to predict all the human body key points in the images.

In summary, extracting human skeleton information includes the following two processes: first, adding the corresponding image data to the input layer of VGGNet and, second, learning the feature value F according to the body parts. The 2D vector field of output corresponding to the human body in the $k=1$ period is

$$S^t = \rho^t\left(F, S^{t-1}, L^{t-1}\right), \forall t \geq 2, \tag{5}$$

$$L^t = \phi^t\left(F, S^{t-1}, L^{t-1}\right), \forall t \geq 2. \tag{6}$$

In (5) and (6), S represents the set of 2D position confidence maps, ρ and ϕ denote the set parameters, t refers to the period corresponding to the feature value, and L signifies the set of 2D vector fields.

The solution to the confidence in the confidence map can be presented as follows:

$$S^*_{j,k}(p) = \exp\left(-\frac{\left\|p - x_{j,k}\right\|_2^2}{\sigma^2}\right), \tag{7}$$

$$S^*_j(p) = \max_k S^*_{j,k}(p). \tag{8}$$

In (7) and (8), S represents the position confidence atlas and p denotes the output image in the corresponding period. Meanwhile, k refers to the number of people in the input image, j stands for the body part's serial number, and σ is a constant.

The joint point position in the 2D vector field is judged according to

$$L^*_{c,k}(p) = \begin{cases} v, \\ 0, \end{cases} \tag{9}$$

$$v = \frac{\left(x_{j2,k} - x_{j1,k}\right)}{\left\|x_{j2,k} - x_{j1,k}\right\|_2}. \tag{10}$$

In (9) and (10), p represents the pixel of the prejudgment part and v denotes the unit vector. On this basis, the average value of the 2D vector field can be written as follows:

$$L^*_c(p) = \frac{1}{n_{c(p)}} \sum_k L^*_{c.k}(p). \tag{11}$$

In (11), $n_{c(p)}$ represents the number of all points of the pixel p on the link c. After testing, candidate positions on PAFs should be determined first. Then, all connected line segments are determined.

The OpenPose open-source library can achieve excellent results of skeleton extraction. However, the image noise limits feature extraction. Therefore, histogram equalization is introduced, which enhances the contrast and reduces the noise by stretching the distribution range of pixel intensity. Videos based on image visualization are processed by Compute Unified Device Architecture (CUDA) to ensure the instantaneity of information extraction. Eighteen key part points are chosen as the input of skeleton movement

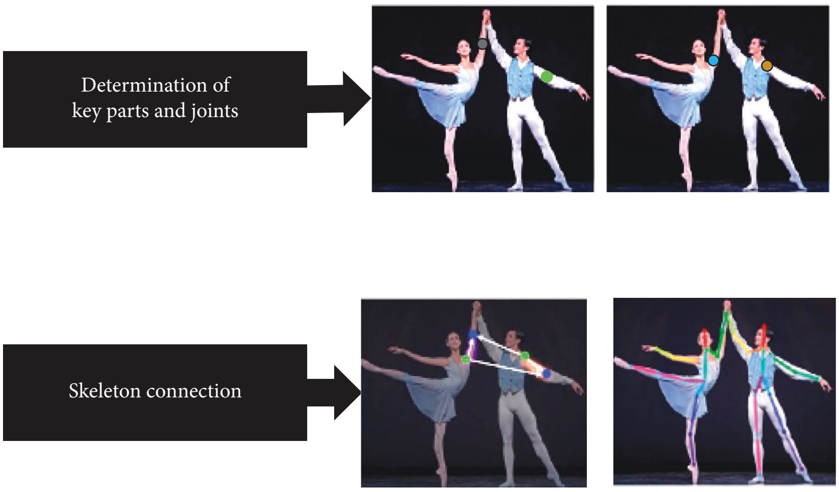

Figure 1: Video information extraction and processing based on open pose (Image URL: https://img-blog.csdnimg.cn/20200910162313905png?x-oss-processimage/watermark, type _ZmFuZ3poZW5naGVpdGk, shadow_10, text_aHR0cHM6Ly9ibG9nLmNzZG4ubmV0L3N1bm55YmxvZ3M, size_16, color_FFFFFF, t_70. Copyright statement URL: https://wwwcsdnnet/company/indexhtml#statement).

extraction while utilizing the OpenPose open-source library. A variable-view movement database containing 40 kinds of aerobic exercises is chosen for analyzing algorithm extraction effects. Eight different movements are chosen for analysis, with the classification accuracy as the primary evaluation indicator.

Here, the optimized 3D CNN (O-3DCNN) algorithm, Spatial-Temporal CNN(ST-CNN) algorithm, and optimized Deformable Part Model CNN (ODPM-CNN) are compared with the optimized VGGNet to prove its effectiveness.

3.3. Skeleton Movement Recognition Based on Optimized LSTM. Traditional neural networks have major limitations in practical application. For example, in time series processing, traditional methods perform well only in short-time series processing. In the separate data processing, the good learning and understanding abilities enable CNN to be applied in practice. However, CNN has limitations in the sequence problem processing related to time correlation. LSTM is a unique Recurrent Neural Network (RNN). LSTM can solve the long-term dependence problem in RNN applications, which has an inseparable relationship with the particular gate structure of LSTM, explicitly referring to input gates, forget gates, and output gates. The input data are calculated according to the following equation:

$$f_t = \sigma\left(w_f\left[h_{t-1}, x_t\right] + b_f\right). \tag{12}$$

In (12), w represents the weight, b corresponds to the deviation, and h_{t-1} denotes the output value corresponding to the time $t-1$. Meanwhile, x_t refers to the input value, σ represents the activation function, and f stands for the forget gate. Moreover, the memory information c_t can be displayed as follows:

$$c_t = f_t c_{t-1} + j_t \tanh\left(w_c \cdot \left[h_{t-1}, x_t\right] + b_c\right). \tag{13}$$

In (13), c_{t-1} represents deciding whether to memorize the information at the time $t-1$ and j_t means the input gate.

Finally, the output gate o_t can be expressed as follows:

$$o_t = \sigma\left(w_o \cdot \left[h_{t-1}, x_t\right] + b_o\right). \tag{14}$$

Although LSTM has many excellent performances, LSTM does not consider the correlation and feature influence between different skeleton movements over a long time. Hence, the LSTM model only depends on the human skeleton joints while recognizing human skeleton movements, resulting in limitations to recognizing human skeleton movements. Therefore, the idea of time integral is introduced. First, the pre-acquired skeleton sequence information is transformed, such as translation and rotation. In this way, all movements can obtain their relative coordinates. If the human skeleton movement has differences due to different times, a multiple LSTM model is used to extract and fuse features [39]. Finally, multiple types of movements are captured by integrating multiple LSTMs. Figure 2 reveals the overall implementation framework of the optimized multi-LSTM human skeleton movement recognition.

Extraction accuracy and loss entropy of various LSTMs are compared to verify the effectiveness of the optimized multi-LSTM human skeleton movement recognition algorithm. Specifically, algorithms selected for comparison include the single-LSTM and double-LSTM. A skeleton sequence input into the optimized F-Multi-LSTM contains 24 frames, among which each frame consists of multiple 2D skeleton points. During analysis, the Adam optimization algorithm is used as the optimization tool, and the initial learning rate is set to 10^{-4}, in an effort to achieve the model's global optimization. The single-LSTM has one input layer, while the double-LSTM has two input layers. The input is assumed as a sentence. In double-LSTM, one side of the input corresponds to the word at the beginning of the sentence and the other side corresponds to the word at the end of the sentence.

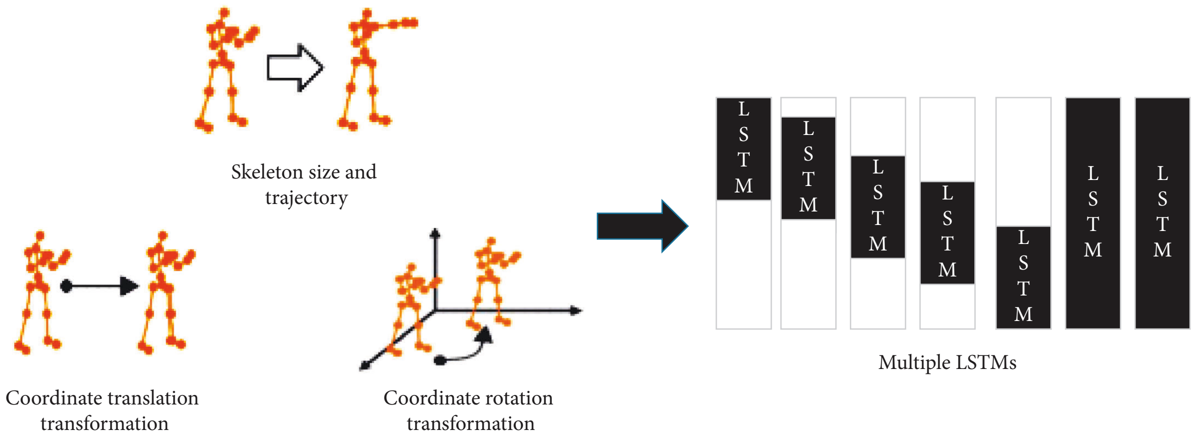
Skeleton size and trajectory
Coordinate translation transformation
Coordinate rotation transformation
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
Multiple LSTMs

Figure 2: Overall implementation framework of multi-LSTM human skeleton movement recognition.

3.4. Design of HCI System Based on Dance Education. Dance education based on physical education helps improve students' physical fitness and transforms traditional sports teaching. According to the above image visualization-based extraction and recognition method of human skeleton movements, the Web3D engine-oriented deep movement recognition system's functional modules are shown in Figure 3.

The system based on dance education and dance movement recognition consists of the front-end interactive function module and the back-end recognition function module. The former is a 3D world built on Web Graphics Library (WebGL) technology, including data processing of video images, 3D processing, and the HCI submodule. The latter consists of two subfunction modules, namely, node recognition and classification of human dance movements.

In this HCI system, the OpenPose open-source database and optimized VGGNet model can estimate facial expressions, positioning of limbs and trunk, and people's feature information. This human skeleton extraction method can identify the critical points of the human body, thereby employing the optimized F-Multi-LSTM skeleton movement recognition network to determine the classification and label attribution of human dance movements. The designed system is based on recognizing and analyzing dance movements. Eight types of dance movements are analyzed and discussed, including stepping and knee lift (S), crouching (C), reaching out and jumping (R), turning and clapping (T), straight punch (B), arm circles (A), jumping (J), and high knee (H).

In the HCI system, the dance pose estimation module and dance movement classification module in the background recognition module are the keys. Accuracy and response time are evaluation indicators to analyze the chosen dance movements, thereby testing the feasibility of the HCI system based on dance education and movement analysis and recognition.

3.5. Data Preprocessing. The image is preprocessed as follows to better meet the needs of behavior recognition: first, the image is uniformly scaled to 432 × 368 based on the center point; second, image denoising. Noises are common in images, in which Gaussian noise is the most common one. The Gaussian filter is used for processing to effectively suppress the Gaussian noise in the image. The one-dimensional Gaussian distribution and two-dimensional Gaussian distribution are shown in (15) and (16), respectively. The Gaussian filter function in open-source computer vision library (Open CV) is used to realize image denoising, and the relevant parameters are optimized.

$$G(\mathrm{x}) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{\mathrm{x}^2}{2\sigma^2}}, \tag{15}$$

$$G(\mathrm{x}, \mathrm{y}) = \frac{1}{\sqrt{2\pi}\sigma^2} \mathrm{e}^{-\frac{\mathrm{x}^2 + y^2}{2\sigma^2}}. \tag{16}$$

4. Results

This section analyzes the optimized VGGNet algorithm's performance through comparison with several human skeleton movement extraction algorithms. The accuracy of the VGGNet algorithm in human skeleton movement extraction is analyzed and optimized on this basis. The effectiveness of the optimized model is verified. Besides, comparative analysis is conducted on the performance of the LSTM model, the single-LSTM model, and the double-LSTM model. Finally, the interaction accuracy and system real-time performance shall prevail to verify the HCI dance education system's performance.

4.1. Performance Comparison of Skeleton Movement Extraction Algorithms. Table 2 presents the comparison result of the extraction accuracy of human movements by several algorithms, including the original and optimized VGGNet.

Table 2 suggests that the optimized VGGNet algorithm presents the best performance in extracting human movements, with the highest accuracy of 98.2%, showing apparent superiority in performance over traditional VGGNet algorithms. The 3D CNN model can only extract a

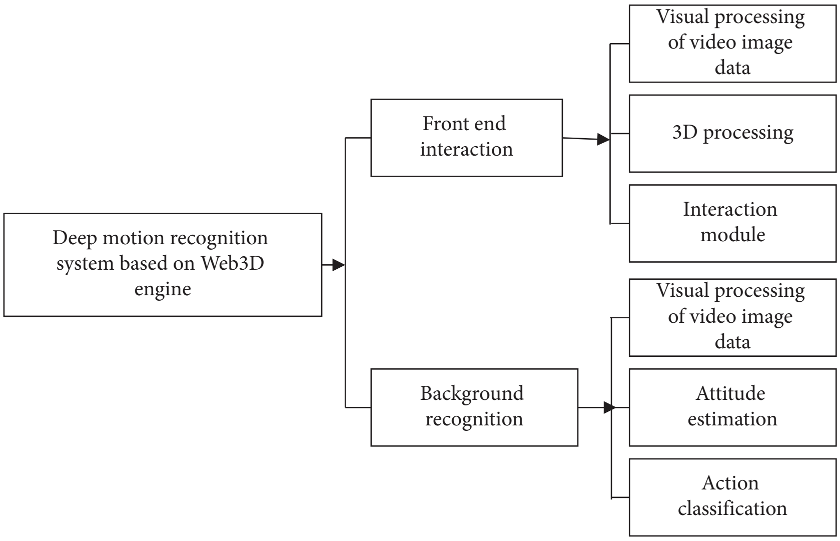
Deep motion recognition system based on Web3D engine
Front end interaction
Visual processing of video image data
3D processing
Interaction module
Background recognition
Visual processing of video image data
Attitude estimation
Action classification

Figure 3: Functional modules of the deep motion recognition system.

Table 2: Comparison of several algorithms' extraction accuracy.

Algorithms	Original VGGNet	Optimized VGGNet	O-3DCNN	ST-CNN	ODPM-CNN
Extraction accuracy (%)	96.9	98.2	91.2	90.5	97.08

type of features from a three-dimensional space because the weights of the convolution kernel are the same in the whole space; that is, the weights are shared by the same convolution kernel, so the extraction accuracy of 3D CNN is only 91.2%. The spatial invariance of ST-CNN refers to the invariance of spatial transformation of images such as rotation, translation, and scaling. Even if the input is transformed or slightly modified, the model can recognize and extract features. ST-CNN is the most time-consuming and error-prone place in debugging interpolation and image index, so the extraction accuracy of ST-CNN is only 90.5%. ODPM-CNN model is a variability network and ODPM-CNN just the opposite, and its recognition accuracy reached 97.08%. The optimized VGGNet is also superior to other human movement extraction algorithms. In this way, the effectiveness of the proposed skeleton extraction algorithm is verified preliminarily.

4.2. Extraction Results of Human Skeleton Movements. The accuracy distribution of the eight human skeleton movements' extraction results by optimized VGGNet on OpenPose open-source database is shown in Figure 4.

This collection of 100 dance pictures is seen as a total sample, and each picture contains eight parts of the action changes. *S* represents the step and knee lifting head, shoulders, elbows, wrists, hips, knees, ankle bone node extraction accuracy; other *C*, *R*, *T*, *B*, *A*, *J*, and *H* dataset content for the above eight parts of the extraction accuracy changes under the action of the title annotation. The extraction accuracy of the head is the highest, reaching 96%, and 100 images are correctly extracted. The extraction accuracy of the shoulder reaches 84.8%, with 90 pictures extracted correctly. The extraction accuracy of the elbow reaches 92.6%, with 89 pictures extracted correctly. The extraction accuracy of the wrist reaches 87.6%, with 86 pictures correctly extracted. The extraction accuracy of the hip reaches 91.0%, with 100 pictures extracted correctly. The extraction accuracy of the knee reaches 95.8%, with 90 pictures extracted correctly. The extraction accuracy of the ankle reaches 86.7%, with 88 pictures extracted correctly. Figure 4 signifies that the extraction accuracy of bone nodes in eight body parts is different, and the proportion of sample number is also different. Moreover, Figure 4 implies that the

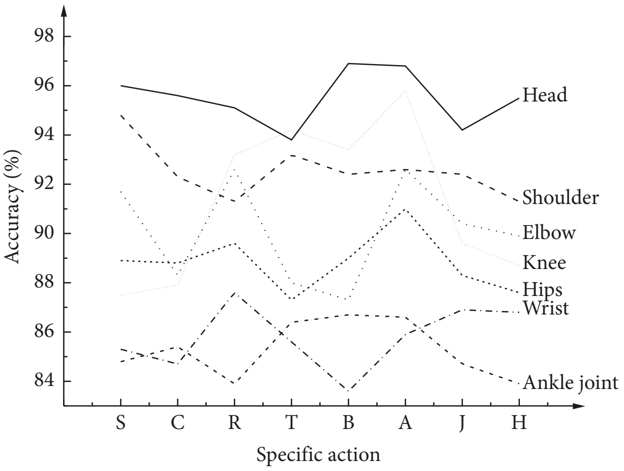

Figure 4: Accuracy distribution of movement extraction results of the human skeleton. (Labels on the *x*-axis represent dance movements. *S* represents stepping and knee lift, *C* represents crouching, *R* stands for reaching out and jumping, *T* stands for turning and clapping, *B* denotes straight punch, *A* denotes arm circles, *J* refers to jumping, and *H* refers to high knee).

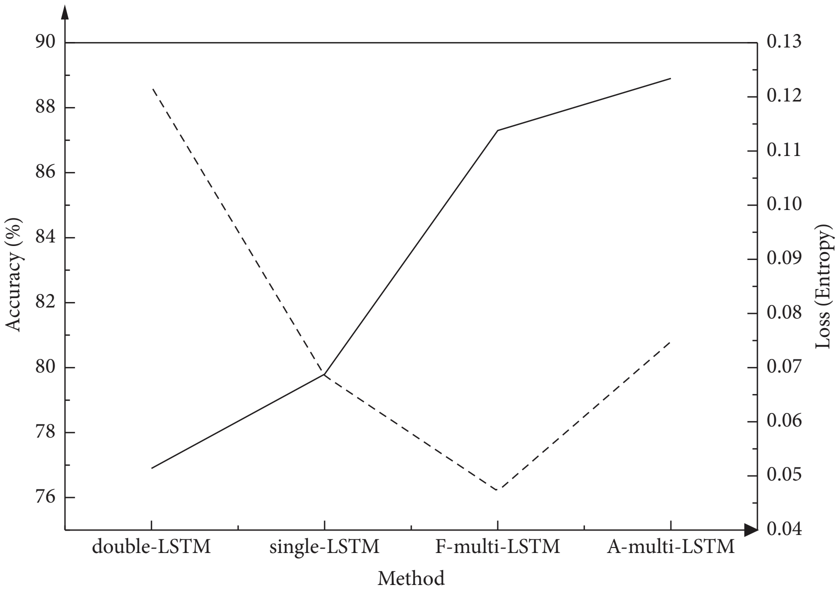

Figure 5: Comparison results of LSTM models.

proportion of accurate number extracted from the large part of the space occupied by the body parts will be significantly higher.

4.3. Skeleton Movement Recognition Results of Multiple LSTMs.

The single-LSTM, double-LSTM, F-multi-LSTM, and A-multi-LSTM are compared. The results are shown in Figure 5.

The parameters represented by the abscissa in Figure 5 are different neural network models. The corresponding left-axis variables refer to the accuracy, and the corresponding right-axis variables stand for loss rates. Single-LSTM is a sequence that supports one-way variable input and output, while double-LSTM is a sequence that supports two-way input and output. Multi-LSTM is a multidimensional LSTM for high-frequency time series, which supports multiple parallel input sequences with multiple inputs, rather than the planar structure of multiple inputs in other models. F-Multi-LSTM is an optimized multidimensional LSTM, and A-Multi-LSTM is expressed as a pair of optimized multidimensional LSTM. The double-LSTM has higher accuracy than the single-LSTM according to the comparison results of loss rate and accuracy of single-LSTM and multi-LSTM. The recognition accuracy reaches 79.8%, and the loss rate is 0.0685. Compared with the single-LSTM model, the difference is 43.8%; overall, the recognition accuracy and loss rate of the proposed multi-LSTM model are the best. Specifically, the single-LSTM model's recognition accuracy reaches 88.9%, and the loss rate is 0.0748, which is the best among the comparative algorithms. Compared with the traditional LSTM model before improvement, the optimized LSTM model has higher recognition accuracy. The optimized LSTM model has the best applicability in recognizing human skeleton movements.

4.4. HCI System Performance Based on Dance Education and Movement Analysis.

The eight dance movements are chosen as the benchmark. According to the indicators of interaction accuracy and system instantaneity, the HCI system's performance for dance education is shown in Figure 6.

In the dance education HCI system, the eight dance movements' overall interaction accuracy is above 70%. The interaction accuracy of movement B is the highest, reaching 92%. The overall accuracy of interactive recognition is distributed in the range of 72%–92%, with a large span. The overall response time corresponding to the eight dance movements is distributed within 5.1 seconds to 5.9 seconds, showing that the dance education HCI system has a high instantaneity.

5. Discussion

The above results indicate changes in the OpenPose open-source database's recognition accuracy and the optimized VGGNet model. The reason is that the head has almost no changes in coordinates or rotation angle. Besides, the movement range of the head is small. Therefore, the accuracy of classification and recognition of the head is the highest. In contrast, the shoulders are greatly affected by external factors, such as rotation angle and abscissa among different movements. Hence, classification and recognition accuracy of the shoulders are relatively low. The elbow movements and the wrist movements are affected by changes in moving speed and longitudinal coordinates. If the human body's moving speed is slow and the position between the arm and the camera is not parallel, the classification and recognition accuracy will be high. The hips are easily affected by changes in the leg movements. The overall accuracy of classification and recognition corresponding to the knees is high, but movements with large fluctuations, such as movement *H*, can significantly affect classification and recognition accuracy. Therefore, the accuracy is low. The ankles and other

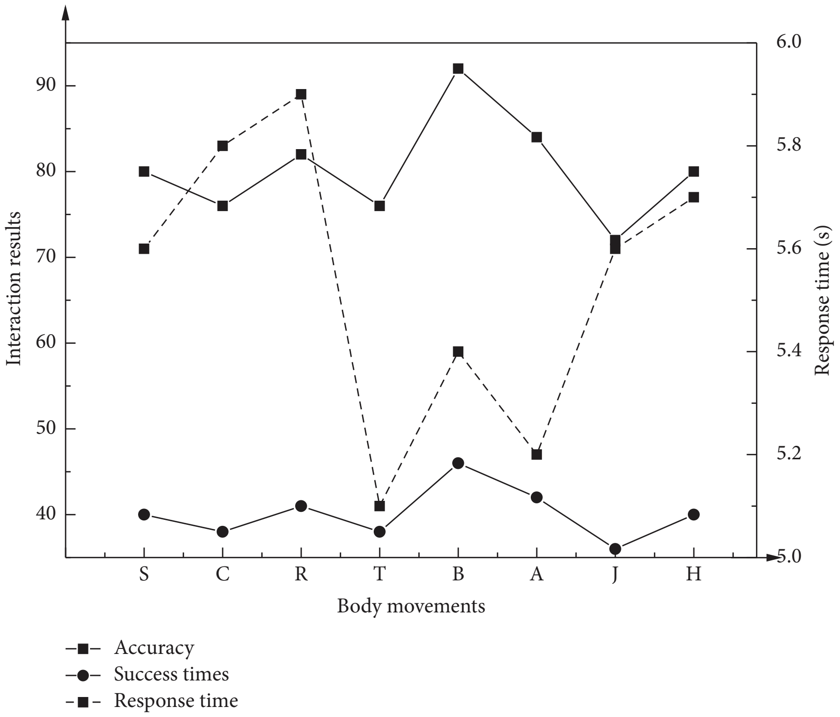

Figure 6: HCI system performance in terms of accuracy and instantaneity.

parts' classification and recognition accuracies are low, probably because of external factors such as clothes and shoes. Although the classification and recognition accuracy of different dance movements are mainly different, the average accuracy is high, confirming the proposed algorithm's effectiveness.

The multiple LSTM model is also advantageous in skeleton movement recognition. Because the optimized LSTM model is robust, its learning and classification abilities are increased, thereby increasing its accuracy in recognizing different dance movements. The distribution changes of the interaction accuracy corresponding to the dance education HCI system reveal that the interaction accuracy corresponding to different movements has a large span. Compared with the model training process, the actual interaction will be affected by sophisticated environmental conditions, such as different lighting, the restraint between different dance movements, and the conversion frequency of various dance movements. Under sophisticated environmental conditions, the interaction accuracy of the HCI system drops. Hence, attention should also be paid to improve datasets in actual HCI applications.

Meanwhile, the proposed algorithm is compared with the methods proposed by other scholars [40–49] to verify its superiority. For the training, the input image size is set to 432 × 368, the number of cycles is set to 50, the batch size is set to 16, and the initial learning rate is set to 0.001. Table 3 reflects the results. Table 3 demonstrates that the proposed multi-LSTM model has the highest accuracy in bone motion recognition, and the recognition accuracy has been improved by 27.79%, 17.69%, and 27.62%, respectively, compared with the comparative methods.

Table 3: Comparison of experimental results of different models.

Algorithms	Accuracy (%)
Method 1 [40]	61.11
Method 2 [41]	71.21
Method 3 [42]	61.28
Multi-LSTM model	88.9

6. Conclusions

For the dance education HCI system, the CNN-based VGGNet model is optimized and applied to extract human skeleton movements based on the OpenPose open-source database and histogram equalization. The proposed extraction algorithm for human skeleton movements shows intentional performance in extracting eight different dance movements, with the highest accuracy rate reaching 96%. From the comparison results of loss value and accuracy between a single-LSTM model and a multi-LSTM model, the accuracy of bone motion recognition by the multi-LSTM model is 79.8%, which is higher than that by a single-LSTM model. The optimized multi-LSTM model has higher accuracy in recognizing human skeleton movements than the traditional LSTM models. The constructed HCI system has an interaction accuracy of 92%. This work achieves the extension of application range of deep learning in skeleton movement recognition and the organic combination of deep learning and HCI.

The contributions based on the extraction and recognition of human dance movements are as follows:

(1) An optimized VGGNet human skeleton movement extraction algorithm is proposed, which achieves a

better extraction accuracy than traditional algorithms, attaining 96%.

(2) An optimized multiple LSTM human skeleton movement recognition algorithm is proposed. Its recognition accuracy reaches 88.9%, which is significantly better than traditional LSTMs.

(3) A HCI system based on image visualization is designed, with the interaction accuracy rate of 92%.

(4) A reference is provided for more in-depth human movement extraction and recognition, and deep learning strengthens the applicability to the HCI system.

Due to computational resource limitations, other larger and more complex datasets are considered in this experiment [50–55]. In addition, the algorithm can meet the real-time requirements, the recognition speed is still very slow. In view of the above problems, it is worth further expanding the datasets in complex scenes in the subsequent work and further optimizing the model to improve the detection speed [56].

Limited by the computing resources, other larger and more complex datasets are not explored [57, 58]. In addition, the recognition speed of the algorithm is slow although it can meet the real-time performance [59–61]. In view of the above problems, the dataset will be further expanded, especially in complex scenarios, which further optimizes the model to improve the speed of detection.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent

Informed consent was obtained from all individual participants included in the study.

Conflicts of Interest

All authors declare that they have no conflicts of interest.

Authors' Contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

References

[1] S. Kohek, D. Strnad, B. Zalik, and S. Kolmanic, "Interactive synthesis and visualization of self-organizing trees for large-scale forest succession simulation," *Multimedia Systems*, vol. 25, no. 3, pp. 213–227, 2019.

[2] A. Canovas, J. M. Jimenez, O. Romero, and J. Lloret, "Multimedia data flow traffic classification using intelligent models based on traffic patterns," *IEEE Network*, vol. 32, no. 6, pp. 100–107, 2018.

[3] A. Rapp, "Design fictions for learning: a method for supporting students in reflecting on technology in Human-Computer Interaction courses," *Computers & Education*, vol. 145, Article ID 103725, 2020.

[4] M. Hibbeln, J. L. Jenkins, C. Schneider, J. S. Valacich, and M. Weinmann, "How is your user feeling? Inferring emotion through human-computer interaction devices," *MIS Quarterly*, vol. 41, no. 1, pp. 1–21, 2017.

[5] J. Bergstrm and K. Hornbk, "Human--Computer interaction on the skin," *ACM Computing Surveys*, vol. 52, no. 4, pp. 1–14, 2019.

[6] M. R. Malgireddy, I. Nwogu, and V. Govindaraju, "Language-motivated approaches to action recognition," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2189–2212, 2017.

[7] S. Singh, C. Arora, and C. V. Jawahar, "Trajectory aligned features for first person action recognition," *Pattern Recognition*, vol. 62, pp. 45–55, 2017.

[8] D. J. Kim, W. K. Song, J. S. Han, and Z. Z. Bien, "Soft computing based intention reading techniques as a means of human-robot interaction for human centered system," *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol. 7, no. 3, pp. 160–166, 2003.

[9] L. Y. Mano, B. S. Faiçal, V. P. Gonçalves et al., "An intelligent and generic approach for detecting human emotions: a case study with facial expressions," *Soft Computing*, vol. 24, no. 11, pp. 8467–8479, 2020.

[10] U. Erkan, "A Precise and Stable Machine Learning Algorithm: Eigenvalue Classification (EigenClass)," *Neural Computing and Applications*, vol. 33, no. 10, pp. 5381–5392, 2020.

[11] K. Dimitropoulos, S. Manitsaris, F. Tsalakanidou, and N Spiros, "Capturing the Intangible: An Introduction to the I-Treasures Project," in *Proceedings of the 9th International Conference on Computer Vision Theory and Applications (VISAPP2014)*, IEEE, Lisbon, Portugal, October 2014.

[12] N. Grammalidis, K. Dimitropoulos, F. Tsalakanidou, and A Kitsikidis, "The I-Treasures Intangible Cultural Heritage Dataset," in *Proceedings of the 3rd International Symposium on Movement and Computing*, July 2016.

[13] A. Doulamis, A. Voulodimos, N. Doulamis, and S Soile, "Transforming Intangible Folkloric Performing Arts into Tangible Choreographic Digital Objects: The Terpsichore Approach," in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP) 2017, Special Session on Computer Vision, Imaging and Computer Graphics for Cultural Applications*, Porto, Portugal, February 2017.

[14] N. Doulamis, A. Doulamis, C. Ioannidis, and M Klein, "Modelling of Static and Moving Objects: Digitizing Tangible and Intangible Cultural Heritage," *Mixed Reality and Gamification for Cultural Heritage*, Springer International Publishing, Berlin, Germany, 2017.

[15] Z. Lv, "Virtual reality in the context of Internet of things," *Neural Computing & Applications*, vol. 32, no. 13, pp. 9593–9602, 2019.

[16] I. Rallis, N. Doulamis, A. Doulamis, A. Voulodimos, and V. Vescoukis, "Spatio-temporal summarization of dance choreographies," *Computers & Graphics*, vol. 73, pp. 88–101, 2018.

[17] A. Aristidou, E. Stavrakis, P. Charalambous, Y. Chrysanthou, and S. L. Himona, "Folk dance evaluation using laban movement analysis," *Journal on Computing and Cultural Heritage*, vol. 8, no. 4, pp. 1–19, 2015.

[18] A. Aristidou, Q. Zeng, E. Stavrakis, and K Yin, "Emotion control of unstructured dance movements," in *Proceedings of the Acm Siggraph*, pp. 1–10, ACM, Los Angeles, California, July 2017.

[19] I. Bhardwaj, N. D. Londhe, and S. K. Kopparapu, "Performance evaluation of fingerprint dynamics in machine learning and score level fusion framework," *IETE Technical Review*, vol. 36, no. 2, pp. 178–189, 2019.

[20] B. W. Israelsen and N. R. Ahmed, "Dave. I can assure you. That it's going to be all right A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–37, 2019.

[21] D. Spathis, N. Passalis, and A. Tefas, "Interactive dimensionality reduction using similarity projections," *Knowledge-Based Systems*, vol. 165, pp. 77–91, 2019.

[22] T. Wu, D. S. Weld, and J. Heer, "Local decision pitfalls in interactive machine learning: an investigation into feature selection in sentiment analysis," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 4, pp. 1–27, 2019.

[23] H. Xu, C. Huang, and D. Wang, "Enhancing semantic image retrieval with limited labeled examples via deep learning," *Knowledge-Based Systems*, vol. 163, no. JAN.1, pp. 252–266, 2019.

[24] S. Long and X. Zhao, "Smart teaching mode based on particle swarm image recognition and human-computer interaction deep learning," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 4, pp. 5699–5711, 2020.

[25] F. Zhang, N. Cai, J. Wu, G. Cen, H. Wang, and X. Chen, "Image denoising method based on a deep convolution neural network," *IET Image Processing*, vol. 12, no. 4, pp. 485–493, 2018.

[26] D. Singh, M Erinc, H Sten, and K Johannes, "Convolutional and recurrent neural networks for activity recognition in smart environment," in *Towards Integrative Machine Learning and Knowledge Extraction: BIRS Workshop, Banff, AB, Canada, July 24-26, 2015, Revised Selected Papers*, A. Holzinger, R. Goebel, M. Ferri, and V. Palade, Eds., Springer International Publishing, Berlin, Germany, pp. 194–205, 2017.

[27] C. C. Wong, Y. Gan, and C. M. Vong, "Efficient outdoor video semantic segmentation using feedback-based fully convolution neural network," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5128–5136, 2020.

[28] M. Tahir, H. Tayara, and K. T. Chong, "iRNA-PseKNC(2-methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components," *Journal of Theoretical Biology*, vol. 465, pp. 1–6, 2019.

[29] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "Multi-class Arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network," *Information Fusion*, vol. 53, pp. 174–182, 2020.

[30] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *Journal of Digital Imaging*, vol. 30, no. 2, pp. 234–243, 2017.

[31] A. Ardakani, C. Condo, and W. J. Gross, "Fast and efficient convolutional accelerator for edge computing," *IEEE Transactions on Computers*, vol. 69, no. 1, pp. 138–152, 2020.

[32] M. Hammad and K. Wang, "Parallel score fusion of ECG and fingerprint for human authentication based on convolution neural network," *Computers & Security*, vol. 81, pp. 107–122, 2019.

[33] P. Yao, H. Wu, B. Gao et al., "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, no. 7792, pp. 641–646, 2020.

[34] Z. Dong, X. Du, and Y. Liu, "Automatic segmentation of left ventricle using parallel end-end deep convolutional neural networks framework," *Knowledge-Based Systems*, vol. 204, Article ID 106210, 2020.

[35] A. Yasoubi, R. Hojabr, and M. Modarressi, "Power-efficient accelerator design for neural networks using computation reuse," *IEEE Computer Architecture Letters*, vol. 16, no. 1, pp. 72–75, 2017.

[36] A. Sellami and H. Hwang, "A robust deep convolutional neural network with batch-weighted loss for heartbeat classification," *Expert Systems with Applications*, vol. 122, pp. 75–84, 2019.

[37] A. Holzinger, R. Goebel, and M. Ferri, "Towards integrative machine learning and knowledge extraction: BIRS workshop, banff, AB, Canada, july 24-26, 2015, revised selected papers," *Lecture Notes in Computer Science*, Berlin, Germany, 2017.

[38] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: event-based data for pose estimation, visual odometry, and SLAM," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.

[39] M. Shakeri, M. H. Dezfoulian, H. Khotanlou, A. Barati, and Y. Masoumi, "Image contrast enhancement using fuzzy clustering with adaptive cluster parameter and sub-histogram equalization," *Digital Signal Processing*, vol. 62, pp. 224–237, 2017.

[40] Y. Shi, Y. Wei, D. Pan et al., "Student body gesture recognition based on Fisher broad learning system," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 17, no. 01, Article ID 1950001, 2019.

[41] W. Huang, N. Li, Z. ., J. Qiu, N. Jiang, B. Wu, and B. Liu, "An automatic recognition method for students' classroom behaviors based on image processing," *Traitement du Signal*, vol. 37, no. 3, pp. 503–509, 2020.

[42] Y. Y. Cheng, Z. J. Dai, Y. Ji, and L Simin, "Student action recognition based on deep convolutional generative adversarial network," in *Proceedings of the 2020 Chinese control and decision conference (CCDC)*, vol. 35, no. 6, pp. 128–133, IEEE, Hefei, China, August 2020.

[43] S. Memiş, S. Enginoğlu, and U. Erkan, "Numerical data classification via distance-based similarity measures of fuzzy parameterized fuzzy soft matrices," *IEEE Access*, vol. 9, pp. 88583–88601, 2021.

[44] R. Liu, X. Wang, H. Lu et al., "SCCGAN: style and characters inpainting based on CGAN," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 3–12, 2021.

[45] W. Zhou, J. Liu, J. Lei, L. Yu, and J. N. Hwang, "GMNet: graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.

[46] G. Sun, Y. Cong, Q. Wang, B. Zhong, and Y. Fu, "Representative task self-selection for flexible clustered lifelong learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 4, 1481 pages, 2020.

[47] F. Liu, G. Zhang, and J. Lu, "Multisource heterogeneous unsupervised domain adaptation via fuzzy relation neural

networks," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 11, pp. 3308–3322, 2021.

[48] W. Yang, X. Chen, Z. Xiong, Z. Xu, G. Liu, and X. Zhang, "A privacy-preserving aggregation scheme based on negative survey for vehicle fuel consumption data," *Information sciences*, vol. 570, pp. 526–544, 2021.

[49] D. Li, S. S. Ge, and T. H. Lee, "Simultaneous arrival to origin convergence: sliding-mode control through the norm-normalized sign function," *IEEE Transactions on Automatic Control*, vol. 67, no. 4, pp. 1966–1972, 2022.

[50] B. Zhu, Q. Zhong, Y. Chen et al., "A novel reconstruction method for temperature distribution measurement based on ultrasonic tomography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 7, pp. 2352–2370, 2022.

[51] W. Zheng, X. Tian, B. Yang et al., "A few shot classification methods based on multiscale relational networks," *Applied Sciences*, vol. 12, no. 8, p. 4059, 2022.

[52] X. Wu, W. Zheng, X. Chen, Y. Zhao, T. Yu, and D. Mu, "Improving high-impact bug report prediction with combination of interactive machine learning and active learning," *Information and Software Technology*, vol. 133, Article ID 106530, 2021.

[53] Y. Wang, H. Wang, B. Zhou, and H. Fu, "Multi-dimensional prediction method based on Bi-LSTMC for ship roll," *Ocean Engineering*, vol. 242, Article ID 110106, 2021.

[54] Y. Ban, M. Liu, P. Wu et al., "Depth estimation method for monocular camera defocus images in microscopic scenes," *Electronics*, vol. 11, no. 13, p. 2012, 2022.

[55] W. Zheng, X. Liu, and L. Yin, "Research on image classification method based on improved multi-scale relational network," *PeerJ Computer Science*, vol. 7, p. e613, 2021.

[56] J. Wang, J. Tian, X. Zhang et al., "Control of time delay force feedback teleoperation system with finite time convergence," *Frontiers in Neurorobotics*, vol. 16, Article ID 877069, 2022.

[57] J. Li, K. Xu, S. Chaudhuri, E. Yumer, H. Zhang, and L. Guibas, "Grass: generative recursive autoencoders for shape structures," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, 2017.

[58] M. Qi, S. Cui, X. Chang et al., "Multi-region Nonuniform Brightness Correction Algorithm Based on L-Channel Gamma Transform," *Security and Communication Networks*, vol. 2022, Article ID 2675950, 2022.

[59] W. Zheng and L. Yin, "Characterization inference based on joint-optimization of multi-layer semantics and deep fusion matching network," *PeerJ Computer Science*, vol. 8, p. e908, 2022.

[60] W. Wang, Z. Chen, and X. Yuan, "Simple low-light image enhancement based on Weber-Fechner law in logarithmic space," *Signal Processing: Image Communication*, vol. 106, p. 116742, 2022.

[61] W. Zhou, L. Yu, Y. Zhou, W. Qiu, M. W. Wu, and T. Luo, "Local and global feature learning for blind quality evaluation of screen content and natural scene images," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2086–2095, 2018.