International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 4599-4613

https://doi.org/10.70517/ijhsa463394

Convolutional neural network-based action recognition and human biomechanics modeling for musical instrument performance

Honghe Li[1,*] and Guopeng You[2]
[1] Humanities and Arts Media Department, Changzhi Medical College, Changzhi, Shanxi, 046000, China
[2] Department of Physical Education, Xiamen University of Technology, Xiamen, Fujian, 361000, China
Corresponding authors: (e-mail: lihonghe@czmc.edu.cn).

Abstract With the increasing maturity of deep learning technology, human action recognition based on deep learning has received extensive attention from research scholars. In this paper, based on convolutional neural network and biomechanics theory, the recognition and characterization of musical instrument playing actions are studied. In terms of action recognition, this paper improves the GoogLeNet network structure and constructs a musical instrument playing action recognition model. On this basis, human biomechanical modeling research is carried out. The results show that the average recognition rate of the method proposed in this paper on the publicly available image dataset PPMI is relatively high, reaching 66%, which is better than other comparative methods, confirming the feasibility and effectiveness of the model application system for human action classification. The results of biomechanical modeling analysis show that the reasonable allocation of the time ratio affects the basic rhythm of the movement, and the adjustment of the rhythm of the center of gravity displacement and center of gravity velocity not only affects the basic rhythm of the whole musical instrument playing movement and the quality of the movement, but also the basic rhythm of the movement and the requirements of the movement constrains the allocation ratio of the center of gravity displacement and center of gravity velocity at each stage.

Index Terms convolutional neural network, GoogLeNet model, action recognition, musical instrument performance, biomechanics

I. Introduction

Music has always been an important way that people use to record or convey emotional information [1], [2]. Music can either bring people a feeling of pleasure and relaxation, or make them feel sad and depressed [3]. In the field of music, instrumental performance is a combination of skill and art, which not only requires musical talent and skill, but also requires the performer to have the correct body posture, combining human biomechanics, movement and body language to express the meaning of music [4]-[7]. Through body movements, the performer can better communicate with the audience and convey the emotions and connotations contained in the music, and instrumental movement recognition is of great significance in enhancing musical expression, infectiousness, and personal charisma [8]-[11].

Human movement recognition technology has a wide range of application prospects in the field of computer vision, and human movement recognition technology based on convolutional neural network has become one of the hot spots of research [12], [13]. Convolutional neural network is a kind of deep learning model that can learn feature representation autonomously, with strong abstract feature representation ability and certain invariance to translation, scale, rotation and other transformations, so it can deal with spatial and temporal correlation well in human action recognition [14]-[17]. In musical instrument playing action recognition, convolutional neural network (LSTM) is a more commonly used method [18]. First, data enhancement techniques can be utilized to enhance the diversity and quantity of data, and a pre-trained network model can be used as the initial model for fine-tuning to improve the classification performance [19], [20]. Second, an attention mechanism can be introduced to further mine the key information in the actions and enhance the network's attention to important features [21]-[23]. In addition, we can also use LSTM to process the temporal information to better capture the temporal features of instrument playing actions [24], [25].

In this paper, we study the action recognition and action characterization of musical instrument playing. Firstly, based on the improved GoogLeNet network structure, we realize the construction of action recognition model for musical instrument playing, and experimentally evaluate the performance of the model. Secondly, the biomechanical

joint coordinate system of musical instrument playing is constructed, and human biomechanical modeling is realized. Finally, taking piano playing as a specific object, combining the experimental method with mathematical statistics, the common features and feature differences of the action rhythms of different players were analyzed.

II. Improved musical instrument playing action recognition based on GoogLeNet network

In order to realize the accurate recognition of musical instrument playing movements, this paper improves the GoogLeNet network and constructs the musical instrument playing movement recognition model.

II. A. Convolutional Neural Networks

II. A. 1) Convolutional Neural Network Model Structure

Convolutional neural network [26] model usually includes Convolutional Layer (Conv), Pooling Layer (Pool), Dropout Layer (Dropout), Flatten Layer (Flatten), and Fully Connected Layer (Dense). First, the convolutional layer implements the convolution operation on the input parameter matrix according to the step size, and the process is represented as:

$$s_i = f(w_i \cdot x_{i:i+h} + b) \tag{1}$$

where s_i is the convolution result, $f(\cdot)$ is the activation function, w_i is the convolution kernel weight, h is the width of the convolution kernel, and b is the bias term.

Secondly, the nonlinear transformation is implemented using the activation function ReLU to increase the nonlinear factors of the model, which has the mathematical form:

$$f(x) = \max\{0, x\} \tag{2}$$

The features are then downscaled using a pooling layer to speed up the computation while avoiding the overfitting problem. The pooling method is selected as maximum pooling and the pooling process is represented as:

$$s^l = pool(s^{l-1}) \tag{3}$$

where: s^l denotes the pooling result for the $l-1$ layer and $pool(\cdot)$ denotes the maximum pooling.

Finally, a fully connected layer is used to fit the data distribution and the activation function Softmax is used to compute the probability of each action type to achieve classification. Its mathematical form is:

$$f(x) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{4}$$

where x_i is the output value of the i th node and x_j is the output value of the j th node over nodes of total length n.

The Dropout layer enables the neuron to discard some data with a certain probability, thus preventing overfitting.The Flatten layer takes the multidimensional inputs one-dimensionally and is used for the transition between the convolutional and fully connected layers. The structure of the convolutional neural network model is shown in Figure 1.

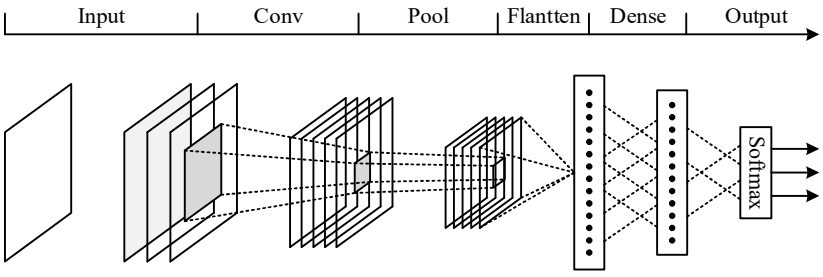


Figure 1: Structure of the convolutional neural network model

II. A. 2) Convolutional neural network model training process

In a convolutional neural network model, the process of obtaining the input matrix features after performing convolution, pooling, and nonlinear operations in the input parameter matrix is called feed forward operation. The last layer of the model calculates the error or loss between the predicted value and the true value, and the error is fed back forward layer by layer from the last layer through the back-propagation algorithm, updating the parameters of each layer, and after updating the parameters the feed-forward operation is carried out at the next time, and so

on until the network model converges, thus achieving the purpose of model training. The training process is as follows:

Step 1: Given a sample set, initialize the weight bias, calculate the state and activation values of each layer through feedforward until the last layer.

Step 2: Calculate the error for each layer and feed the error forward from the last layer layer by layer.

Step 3: Calculate the derivatives of the parameters for each layer.

Step 4: Update the parameter matrix.

Step 5: Repeat steps 2 through 4 after updating the parameters, and so on until the model converges, completing the model training.

II. B.GoogLeNet network

II. B. 1) Inception V1 network architecture

GoogLeNet network [27] is also known as Inception V1 network due to the inclusion of the Inception module. The Inception layer uses convolution kernels 1*1, 3*3 and 5*5 for parallel convolution of images to obtain feature images of different dimensions. At the same time, as the network continues to deepen, the high-level network layer can be realized in the network to receive a wider range of domain images. The Inception structure makes GoogLeNet become a classical model of convolutional networks through deeper and wider networks.

II. B. 2) Structural Improvements to Inception V2 Networks

Inception V2 introduces Batch Normalization (BN) on top of Inception V1 and replaces the 5×5 convolution with two 3×3 convolutions.

BN normalization is a normalization network layer used to speed up model convergence and enhance model generalization. The mean and variance are obtained for each channel corresponding to multiple graphs in the input, and then the transformation of the data is done based on the mean and variance so that the input obeys a standard normal distribution.

Assume that the input is:

$$x : B = \{x_1, ..., m\} \tag{5}$$

The mean value B can be derived by calculating as follows:

$$\mu_B = \frac{1}{m}\sum_{i-1}^{m} x_i \tag{6}$$

The variance 2B was calculated by the same formula:

$$\sigma_B^2 = \frac{1}{m}\sum_{i-1}^{m}(x_i - \mu_B)^2 \tag{7}$$

After passing the regularization operation, the mean and variance obtained from Eq. (6) and Eq. (7) will be normalized for each batch of training data to obtain a 0-1 distribution:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{8}$$

The purpose of ε in Equation (8) is to prevent the divisor from being 0. Forcing the inputs to be transformed to data obeying a standard normal distribution would cripple the network. To solve this problem, a scale factor γ and a leveling factor β are added to the normalization formula for scale transformation and bias:

$$y_i = \gamma\hat{x}_i + \beta = BN_{\gamma\cdot\beta}(x_i) \tag{9}$$

II. B. 3) Inception V3 Network Architecture Improvements

Inception V3 builds on Inception V2 by splitting a two-dimensional convolution into two smaller convolutions, splitting a 7×7 convolution into a 1×7 convolution and a 7×1 convolution, which has the advantage of reducing the number of references. It is further argued theoretically that any n×n convolution can be replaced by a 1×n convolution and followed by an n×1 convolution, and that the computational cost savings increase significantly as n grows.

This asymmetric splitting of the convolution structure results in a more pronounced effect than a symmetric splitting into several identical small convolution kernels, allowing for the processing of more and richer spatial features and increased feature diversity.

II. C.GoogLeNet network structure improvement and optimization

II. C. 1) Internal Optimization of the Inception Module

The general structure of GoogLeNet network introduces Inception3 module in the shallow layer, Inception4 module in the middle layer, and Inception5 module in the deep layer, respectively, in which the original network structure is set uniformly for the module structure, and the three modules have the same structure.

In this paper, we first optimize the structure of Inception3 module. The shallow Inception3 removes the 5*5 convolutional kernel and increases the number of channels of the 3*3 convolutional kernel, and at the same time, the 3*3 is also changed into 3*1 and 1*3, which makes the calculation speed faster and quicker.

Next, a bigger structural improvement is made to the Inception4 module in the middle layer. First, the original 5*5 large convolutional kernel structure was replaced by a new small network of convolutional layers composed of two 3*3-sized convolutional kernels connected sequentially. Then two convolutional kernels of size 3*1 and 1*3 are used serially instead of the 3*3 size convolutional kernel.

The next improvement extends the filter banks in the module to make the network structure wider rather than deeper, which reduces the dimensionality to eliminate the representational bottleneck.

For the deeper Inception5 structure, the original structure of the model is retained in this paper.

II. C. 2) Choosing the Leaky-Relu activation function

The Leaky-Relu activation function is given by:

$$f(x)=\begin{cases}\alpha x, x \le 0\\ x, x>0\end{cases} \tag{10}$$

where α is the empirical value.

As opposed to ReLu, the Leaky-Relu function uses a linear function with a small slope at $x<0$. This is equivalent to allowing the gradient corresponding to intervals less than 0 to backpropagate, rather than intercepting it directly.

In the improved Inception network structure, the batch normalization will be performed uniformly after each convolution operation for normalization, the Relu activation function is selected in the shallow network, the Leaky-Relu activation function is selected after the mid-layer, and the two improved shallow Inception3 modules are sequentially accessed through the maximum pooling operation, and the structurally optimized five mid-layer Inception4 modules, and two high-level Inception5 modules with larger convolutional kernels, followed by global average pooling, and finally the feature image is classified for output by Softmax classifier.

II. D.Network training strategy selection

II. D. 1) Dropout strategy

Dropout is used to prevent overfitting problems. For deep neural networks, the goal of training is to make the final output of the network infinitely close to the real Label.

In order to prevent neurons from relying on each other too much, Dropout hides some neurons during the training process.The operation of Dropout during the training process is shown in Figure 2, where the dotted lines indicate the hidden nodes.

Dropout determines whether a neuron is hidden or not according to a set probability, if a neuron is selected to be hidden, then this neuron does not participate in the computation, and the neuron weights are not updated during the back propagation, and the neuron weights will continue to be retained. The network structure adds a Dropout layer after the three fully connected layers and removes the two additional auxiliary classifiers, which reduces the complexity of the network and improves the training speed of the network.

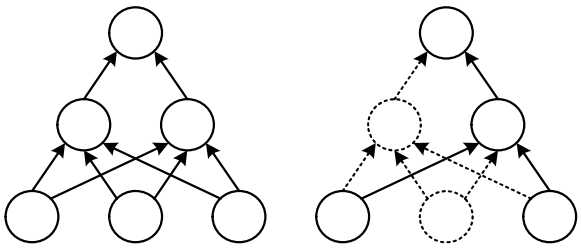


Figure 2: Changes in the network after Dropout

II. D. 2) Adam Optimizer

The adaptive moment estimation (Adam) algorithm combines SGDM first-order momentum and RMSProp second-order momentum, with RMSProp as root-mean-square backpropagation, overcoming the drawbacks of the optimal solution being locally optimal and slow to converge when close to a local minimum solution.

Denoting the momentum variable v_t by v_t, with $0 \le \beta_1 < 1$ qualifying the hyperparameters, the momentum variable v_t for time step t can be expressed as:

$$v_t \leftarrow \beta_1 v_{t-1} + (1-\beta_1) g_t \tag{11}$$

The stochastic gradient is exponentially weighted by the numerically squared term $g_t \odot g_t$, and subsequently moving averaged to obtain s_t:

$$s_t \leftarrow \beta_2 s_{t-1} + (1-\beta_2) g_t \odot g_t \tag{12}$$

Bias corrections are made to both the variables v_t and s_t by Adam's algorithm:

$$\hat{v}_t \leftarrow \frac{v_t}{1-\beta_1^2} \tag{13}$$

$$\hat{s}_t \leftarrow \frac{s_t}{1-\beta_2^t} \tag{14}$$

Next, Adam's algorithm rescales the learning rate of each element on an element-by-element basis:

$$\overline{g}_t \leftarrow \frac{\hat{\eta}_{v_t}}{\sqrt{\hat{s}_t} + \varepsilon} \tag{15}$$

where the learning rate is η, the constant that maintains numerical stability is denoted by ε, and the independent variables are iterated using g_t:

$$x_t \leftarrow x_{t-1} - g_t \tag{16}$$

In this paper, the Adam optimization algorithm is used in the training of the network model for instrumental movement recognition to ensure that the training of the improved network model has a higher recognition efficiency.

II. E. Experiment and Result Analysis

II. E. 1) Experimental setup

In order to test the validity of the model proposed in this paper, the model is tested by MATLAB2017 processing platform. The hardware configuration of the computer used for the experiment is Intel Core i5-3350P CPU, 3.10 GHz, with 16 GB of RAM.Due to the fact that the training of this network model involves a large number of parameters and a large amount of computation, the model was tested by using Windows 10 GPU computing platform under the operating system.

In the experiments, the experimental dataset is loaded into the improved GoogLeNet for model training to build a GoogLeNet-based human action recognition classification model. In order to ensure the accuracy of action classification, the base learning rate is set to 0.001 during model training.

II. E. 2) Experimental data sets

In this paper, the Musical Instrument Playing (PP-MI) image dataset created by Stanford University is used as the experimental dataset. The dataset consists of 24 different actions corresponding to 12 musical instruments, each instrument contains 2 actions in playing the instrument and holding the instrument. 80% of the dataset was used for the training of the network model and 20% was used for validation. The size of each image is 258×258 pixels.

II. E. 3) Analysis of experimental results

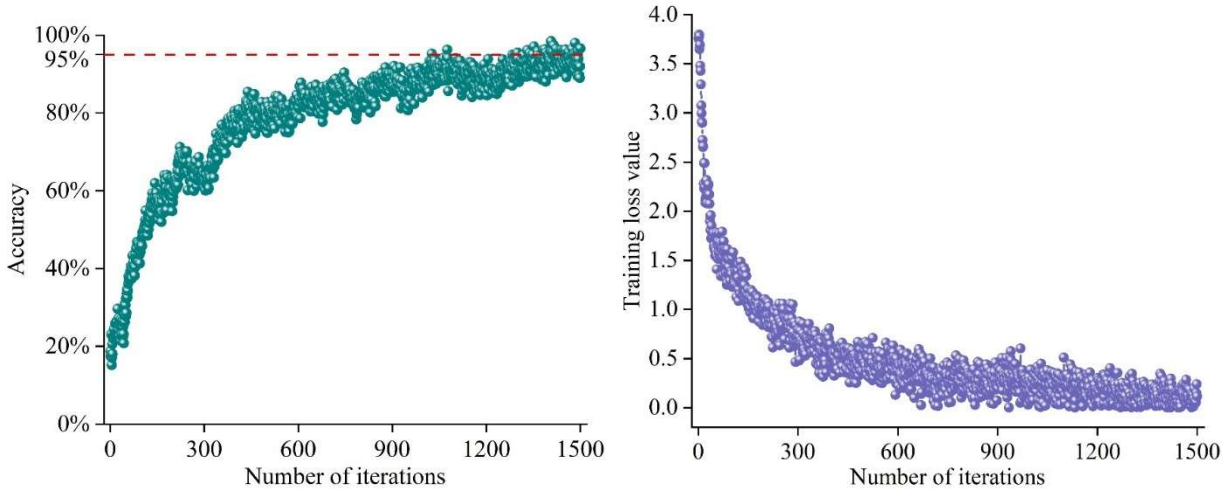
In order to verify the effectiveness of the model, the classification accuracy of the action is used as the evaluation benchmark in this paper. The calculation formula is:

$$A = \frac{n}{N} \times 100\% \tag{17}$$

where: A is the action classification accuracy, n is the number of accurately classified samples, and N is the total number of samples to be tested.

The model training results are shown in Fig. 3. Where (a) is the accuracy of model training, and (b) is the decline in training loss corresponding to the accuracy of model training.

It can be seen that the accuracy is very unsatisfactory at the beginning due to the small number of iterations, but with the increase of the number of iterations, the accuracy gets a significant increase until it finally stabilizes. At the same time, the training loss of the model also gradually decreases and tends to stabilize, and finally close to 0. The training results show that the accuracy of the training of the model floats around 95%, which indicates that this model is able to correctly classify the images of the training dataset, and confirms the feasibility of the present model in the classification and recognition of the human body movements in the process of musical instrument playing.



(a) The accuracy of model training (b) The situation of the decline in training loss

Figure 3: Model training results

The time complexity of this model is the summation of the time complexity of all convolutional layers, i.e., intra-layer concatenation and inter-layer accumulation. Since the proposed model in this paper adopts the Inception module, which greatly reduces the time complexity, this model takes less than 18 min in total during the training of the PPMI dataset, and the average time per image is about 0.58s.

In order to intuitively display the classification results, experiments are conducted on the visualization part of the output layer. The Load Network and Load Image sections of the GUI interface control panel are operated, after which the loaded images are judged and actions are recognized. In addition, in this paper, different backgrounds for the same action and different musical instrument images for the same action are compared. The results of the operation in the case of the same person, the same musical instrument, and different backgrounds show that this model is able to extract the features of the action category during the training period, and the influence of the background can be ignored. And the running results in the case of same action, different instruments show that although the player is playing different instruments with the same action, the model extracts the relevant features of the instruments and can also categorize them.

A comparison of the accuracy of action recognition in images using this method with the recognition accuracy of other methods on the PPMI dataset is shown in Fig. 4. Among them, the SPM-LLC method uses a locally constrained linear coding (LLC) model to describe the interaction between the acting individual and the acting object after extracting the image features through the spatial pyramid model (SPM) [28] framework. In the IIBPO-CNN method, the part-based behavioral description method is firstly used, and then, using the detected behavioral object and the behavioral poses of the person which have been obtained. The interaction relationship between the two is described by discriminating information such as the positional relationship between the two and the interaction area. In the GMP-VLAD method, we first start from the collection representation of images, then verify whether different clustering algorithms and the number of clustering centers are applicable to the latest proposed collection representation of images with locally aggregated descriptors, and finally make a study in terms of the role and validity of normalization and pooling on the improved locally aggregated descriptors.

As can be seen in the comparison with other methods, the average recognition accuracy of the proposed method in this paper is higher than that of other methods, but the correct rate is still maintained at about 66%, which still has some room for improvement. The reason for this is that, on the one hand, the difference between different movements of holding the same instrument is not obvious enough, or the difference when playing different instruments with the same movement is not obvious enough. On the other hand, different instruments may be misclassified, such as handheld violin and handheld cello. While the distance or size of the images are different, violin and cello have their similar shape structures, and the difference in human posture is ignored when confusing the 2 instruments, which in turn leads to inappropriate extraction of information about this difference in this model, and therefore incorrect recognition of such movements, which is the focus of the next research.

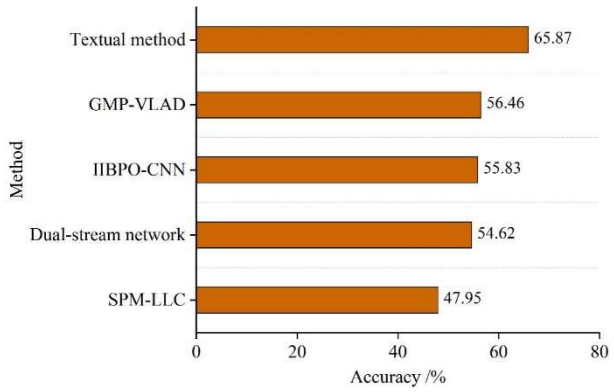


Figure 4: Accuracy rates of different methods on the PPMI dataset

III. Research on human biomechanical modeling for musical instrument performance

In this chapter, based on the accurate recognition of musical instrument playing movements, the human biomechanical modeling of them is further investigated.

III. A. Biomechanical joint coordinate system construction

III. A. 1) Anatomical Characterization Point Definitions

In the definition of skeletal muscle anatomical angle feature points, the motion of the radial-ulnar joint is attributed to the elbow joint and the motion of the tibiofibular joint is attributed to the knee joint. Referring to the human body center of mass related anatomical literature, the main bony feature points of the shoulder, arm, hand of the upper limb and hip, knee and foot of the lower limb of the human body are defined before constructing the joint coordinate system, and the defined limb feature points are shown in Table 1.

Table 1: Bony anatomical Feature points

Feature point	Definition	Feature point	Definition
l	Shoulder lock joint	h	Large rotor
e	Lateral condyle of the humerus	k	Lateral femoral condyle
w	Radial styloid process	a	The lateral condyle of the fibula
t	The lateral terminal of the second metacarpal bone	f	The lateral terminal of the second phalanx

III. A. 2) Definition of joint coordinate system

After determining the main anatomical feature points, the establishment of the global and local coordinate system of the instrumental playing movement begins. The global coordinates are generally defined according to the position of the three basic axes of the human body, i.e., sagittal, frontal, and vertical axes, in the three basic sections, i.e., the transverse, sagittal, and coronal planes.

Sagittal axis of the human body: The axis that extends anteriorly and posteriorly parallel to the transverse plane and perpendicular to the coronal plane is called the sagittal axis, i.e., the x-axis is the sagittal axis.

Coronal axis of the human body: the axis that extends left and right parallel to the transverse plane and perpendicular to the sagittal plane is called the coronal plane, which is also known as the frontal axis, i.e., the y-axis is the coronal axis.

Vertical axis of the human body: the axis parallel to the longitudinal axis of the human body and perpendicular to the transverse plane is called the vertical axis, i.e. the z-axis is the vertical axis.

The local coordinate system and symbols are defined as shown below:

L_i: the ith link.

C_i: center of mass of Bi.

$X_1Y_1Z_1$: coordinate system cemented to the shoulder joint.

C_l: center of mass of the upper arm link.

$X_eY_eZ_e$: coordinate system cemented to the elbow joint.

C_e: center of mass of the forearm link.

$X_wY_wZ_w$: coordinate system cemented to the wrist joint.

C_w : center of mass of the palm link.

$X_hY_hZ_h$: coordinate system cemented to the hip joint.

C_T : center of mass of the thigh link.

$X_kY_kZ_k$: coordinate system cemented to the knee joint.

C_S : center of mass of the calf link.

$X_aY_aZ_a$: coordinate system cemented to the ankle joint.

C_3 : center of mass of the foot link.

T : active force.

P_i : vectorial diameter from C_i to each joint.

Z_1 : is the vertical axis, with the direction pointing down.

X_1 : perpendicular to the normal of the plane formed by the points l and e, with the direction pointing anteriorly.

Y_1 -axis: is the coronal axis, with the direction pointing to the right.

Z_e -axis: the line between e and w, with the direction pointing to w.

X_e -axis: normal to the plane formed by points w and e perpendicularly, direction pointing anteriorly.

Y_e -axis: the coronal axis, with the direction pointing toward the lateral side of the body.

Z_w -axis: the line between w and t, with direction pointing to t.

X_w -axis: normal to the plane formed by the points t and w perpendicular to each other, with the direction pointing toward the anterior side of the body.

Y_w -axis: the coronal axis, with the direction pointing toward the lateral side of the body.

Z_h -axis: the line between h and k, with direction pointing to k.

X_h -axis: normal to the plane formed perpendicular to the points k and h, direction pointing toward the anterior side of the body.

Y_h -axis: the coronal axis, with the direction pointing toward the lateral side of the body.

Z_k -axis: the line between k and a, with direction pointing toward a.

X_k -axis: normal to the plane formed by the points h and k perpendicular to each other, direction pointing toward the anterior side of the body.

Y_k -axis: the coronal axis, with the direction pointing toward the lateral side of the body.

Z_a : the line between a and f, with direction pointing to f.

X_a : normal to the plane formed by the points f and a perpendicular to each other, direction pointing toward the anterior side of the body.

Y_a : the coronal axis, with direction pointing toward the lateral side of the body.

M : active moment.

Based on the definitions as above and the definition of the bony anatomical characteristic points in Table 1, a local joint coordinate system can be established.

III. A. 3) Definition of joint angles

Through the definition of anatomical characteristic points, the construction of joint coordinate system and reference to relevant literature, and the physical model of the whole human upper and lower limbs to briefly establish six rigid bodies, 15 angles of rotation, namely, three angles of the shoulder joint: flexion and extension, extension, extension, internal and external rotation, two angles of the elbow joint: flexion, extension, internal and external rotation, one angle of the wrist joint: dorsiflexion, extension, and three angles of hip, knee, and ankle joints: dorsiflexion and extension, internal and external adduction angle, and internal and external rotation angle. Take the shoulder joint as an example, head to hip vertical downward horizontal direction is 0 °, counterclockwise rotation horizontally to the right 90 ° that is perpendicular to the human body, the opposite direction is 270 °, continue to rotate back to the origin for 360 °. Different links around the three basic axes of the human body in the three basic sections to determine the joint angle in this paper, the calculation of the lower limb joint angle of rotation with the same upper limb joints. The specific calculation method is as follows:

Upper arm flexion and extension angle: the angle between the half shift of the upper arm in the sagittal plane and the vertical axis.

Upper arm adduction angle: the angle between the projection of the upper arm on the frontal plane and the vertical axis.

Upper arm rotation inward and outward angle: the angle between the projection of the upper arm on the body's aqueous plane and the frontal axis.

Elbow flexion-extension angle: the angle between the ew line and the el line.

Wrist flexion-extension angle: the angle between the projection of the $\vec{wt}$ vector on the plane perpendicular to the $\vec{w}$ vector and the $\vec{ew}$ vector.

Thigh flexion and extension angle: the angle between the projection of the thigh vector on the sagittal plane of the body and the vertical axis of the body.

Angle of adduction and abduction of the thigh: the angle between the downward vertical axis of the body and the projection of the thigh vector $\vec{hk}$ on the frontal plane of the body.

Thigh rotation inward and outward angle: the angle between the calf longitudinal axis vector $\vec{ka}$ and the vertical axis of the human body in the vertical projection with the thigh.

Knee angle: the angle between the line from k and a to the vertical line.

Ankle flexion-extension angle: the angle between the line from a and f to the vertical line.

III. B. Subjects of the study

According to the biomechanical experiments related to musical instrument playing, most of the sample sizes were selected at 10-20, and the experimental subjects in this study were selected as students majoring in musical instrument playing at the School of Music of the University of H. 10 males and 10 females were selected, with a total of 20 experimental subjects. The experimental subjects have been informed of the content of the experiment and the experimental precautions before conducting the formal experiment, and all of them participated voluntarily.

The inclusion criteria for subjects in this experiment: subjects had participated in regular training for at least three days prior to the test, and those who had not trained for more than a week were not included. Subjects did not have any disease that would affect their ability to complete the instrumental program. Subjects have participated in more than five years of specialized training, have a wealth of theoretical knowledge and practical experience reserves. Subjects understood the content of the test and agreed to participate in the experiment.

III. C. Research methodology

III. C. 1) Experimental methods

(1) Experimental equipment

This experiment adopts simultaneous acquisition of kinetic data and kinematic data, and the test equipment of the overall experiment is divided into three parts: kinetic equipment, kinematic equipment and auxiliary tools.

1) Kinematics equipment

This experiment uses two SONY XCG-CG510C high-speed cameras, three-dimensional motion capture analysis system Simi Motion, matching three-dimensional calibration bar and dongle.

2) Dynamics equipment

A computer, specification of 40 * 60 * 10cm three-dimensional force table 9281EA, acquisition frequency of 1000 Hz. A synchronized trigger that connects the computer and the three-dimensional force measuring table.

3) Auxiliary tools

The diameter of 14 mm reflective sign ball 15, specifications for the 3m * 2m black cloth 8, 3M tape a number of.

(2) Experimental site

The experimental site was the biomechanics laboratory on the second floor of the gymnasium of the University of H. The experimental subject faced the direction of the camera and stood on top of the wooden board 1 to play the instrument in the left and right directions. The direction of the experimental subject's toe is the Y-axis direction, and the direction of the experimental subject's movement is the X-axis direction, and the experimental site is shown in Figure 5. The site is arranged according to the site specifics and equipment requirements.

(3) Experimental process

The site is mainly divided into five parts: students' basic information statistics, site layout, equipment debugging, site personnel preparation, and formal shooting.

1) Basic information statistics: 20 students were registered with basic personal information, including name, gender, years of training, height, weight, and best performance, etc., and were informed of the performance items and requirements of the experimental test.

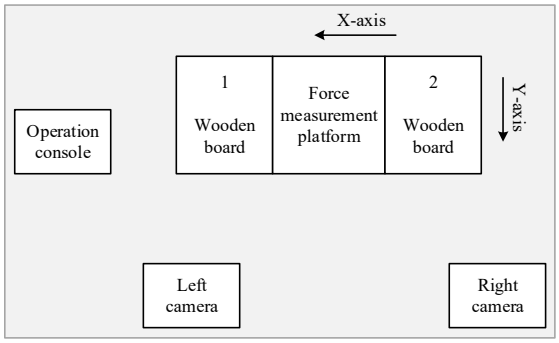


Figure 5: Laboratory site

2) Site arrangement: Due to the laboratory conditions, the subjects were required to stand on the wooden board next to the force measuring table for preparation, so the preparation position tape was put on the board to make sure that the starting position was uniform for all subjects. Since the post-processing of video data by Simi-Motion, a three-dimensional image analysis system, needs to be carried out under dark light, the eight black cloths prepared at the site were arranged and covered.

3) Equipment debugging: including the connection of the computer, the 3D force measuring table and the synchronized trigger three-way, the height of the two camera racks on the site was adjusted to be the same, and at the same time the camera position and angle were fixed, the video frame rate of the two cameras was adjusted to be the same, and the synchronized signal light was placed in the center of the force measuring table in the direction of the subject facing the camera to ensure that the two cameras could shoot the signal light at the same time.

4) Site personnel preparation: including staff and subject preparation. Subjects were given a 10-minute warm-up and a 20-minute practice of technical movements in the mirror before the test. On-site staff: The master control assigns the staff to tasks, including lighting control, professional camera staff, and joint position training for those posting reflective spots.

5) Formal shooting: first, the field staff were assigned tasks, then the venue was calibrated for the session, and the subjects were given 15 marking points to post. Next, the subjects were tested on their instrument playing movements, and the best video of one performance was selected as the final video for analysis. Finally, after all subjects were tested, the person in charge of the computer organized and classified the data, as well as conducting and organizing the videos from the two cameras, while other staff members carried out the site cleaning work.

III. C. 2) Mathematical and statistical methods

(1) Kinematic data analysis

The three-dimensional motion analysis system Simi-Motion was used to analyze the filmed left and right camera videos, the Hanavan human body center of gravity model was used for digital calculation and processing, manual punching and automatic tracking were used to determine the joint positions of the marker points in the TLD spatially calibrated three-dimensional coordinate system, and the Butterworth digital filtering that comes with the system was used to carry out the data Smoothing processing, after which the kinematic data were exported according to the indexes they selected. The specific operation process is as follows:

1) Definition of research index

Indicators of this study: time of the center of gravity transfer phase, center of gravity displacement, center of gravity displacement speed, hip joint angle, knee joint angle, ankle joint angle, and joint displacement.

Center of gravity position change: the position change of the center of gravity of the human body in the three-dimensional space coordinate axis during the center of gravity transfer phase.

Hip angle: the angle between the line between the shoulder joint and the hip joint and the line between the hip joint and the knee joint.

Knee angle: the angle between the line between the hip joint and the knee joint and the line between the knee joint and the ankle joint.

Ankle angle: the angle between the line between the knee and ankle joints and the line between the ankle joint and the third toe bone.

2) Establishing the human body model

In Simi-Motion system, according to the Hanavan human body model for built-in point building, the calibration video and motion video import, using manual pointing and automatic tracking to determine the joint position of the

marker point after the use of the system comes with Butterworth digital filtering of the data for smoothing the data after the presentation of its original data model.

3) Original data extraction

According to the original 3D data, the index data are extracted according to the definition of the research index.

4) Raw Data Export

Export the extracted raw data.

(2) Dynamic data analysis

Collect the mechanical data of the center of gravity transfer stage by using the three-dimensional force measuring table, obtain the COP values of ground reaction, moment and center of pressure over time, import the data into excel software for data arrangement, and then use Matlab software for programming to calculate the peak value of the ground reaction force, impulse and the center of pressure trajectory.

(3) Data statistics

The obtained kinetic data and kinematic data were organized in Excel to obtain the average value of the data, and then IBM SPSS Statistics 21 was used to carry out hypothesis testing on the data, after which the mean and standard deviation were calculated. The differences between the data were compared by paired samples test and independent samples test. Using Matlab software programming, the peak ground reaction force, the coordinates of the center of pressure, and the impulse of force were calculated, and finally the data were standardized.

III. D. Findings and analysis

In this section, based on the biomechanical modeling method, the experimental method and mathematical statistics were used to analyze the rhythm of the movements of three randomly selected students from 20 subjects when they performed piano playing.

III. D. 1) Common Characterization of Movement Rhythms

The basic profile of the subjects selected for experimental analysis is shown in Table 2.

Table 2: Basic Information of the experimental subjects

Serial number	Gender	Height /cm	Weight /kg	Age	Training time /year
1	Male	181	64	24	13
2	Male	182	71	23	12
3	Male	179	62	22	12

(1) Common characterization of body center of gravity displacement

The distribution of the center of gravity in the Z-axis curve of the three piano players is shown in Figure 6. According to the test results, it is reflected that in completing the piano playing movement, the three subjects' body center of gravity displacement in the Z-axis, i.e., the vertical direction, varied greatly in amplitude, and showed an irregular curve state in the Z-axis direction. The movement of the subjects' body center of gravity displacement in the Z-axis direction has both rise and fall, forming a prolonged undulating and smooth fluctuation state, which shows that the subjects' body center of gravity in the Z-axis direction changes a lot in the performance of the piano playing movement, from a straight line to a high point and then to a low point.

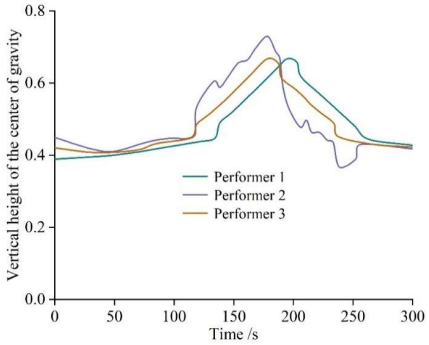


Figure 6: The center of gravity is distributed along the Z-axis curve

Displacement characteristics of body center of gravity displacement in the horizontal direction: In the process of piano playing, the body displacement in the X-axis, i.e., front and back direction, is linear over time. The size of body mobility can show the coordination ability of the player to a certain extent.

(2) Common Characterization of the Velocity of the Center of Gravity of the Body

During piano playing, the magnitude of the body center of gravity velocity of the three different subjects was extremely small in the Y-axis, i.e., left-right direction, and faster in the X- and Z-axis directions. The small amplitude of the body center of gravity velocity in the Y-axis direction indicates that the body center of gravity velocity varies very little in the left and right directions during piano playing, which is due to the fact that the seated posture is always maintained during the playing process and the fluidity of the movement is better. The speed of the center of gravity in the Z-axis varies from slow to fast, which means that the center of gravity maintains a horizontal height at the beginning, and then fluctuates in the later stages of the movement due to the fast changes in the playing gestures, which results in a relatively fast speed of the center of gravity in the vertical direction. The faster speed of the body's center of gravity in the X-axis direction indicates that the speed of the body's center of gravity is faster in the front-back direction during piano playing, indicating that the body fluctuates back and forth at a faster rate.

III. D. 2) Analysis of differences in movement rhythm characteristics

As different players play piano movements under the same repertoire, the total time is the same, the tempo is the same and the movements are the same, but because of the differences in speed and time value and tempo processing on the movement characteristics of the differences presented.

(1) Analysis of differences in time and rhythm characteristics

The time data of different subjects in each stage are shown in Table 3. It can be seen that in the beginning stage, No.1~3 players took 120.50s, 99.50s and 97.40s respectively, with an average of 105.80s and a standard deviation of 12.77s. No.1 player took the longest time and No.3 player took the shortest time. In the transition stage and the ending stage, it was player No. 1 who took the shortest time and player No. 3 who took the longest time.

The ratios of the time spent by players No. 1 to No. 3 in these three phases were 1.83:1.68:1, 1.43:1.84:1, and 1.14:1.50:1, respectively, with the ratio of the time spent by player No. 1 in the order of beginning > transition > end, and the ratio of the time spent by players No. 2 and No. 3 in the order of transition > beginning > end.1 Player 1 The treatment of the time rhythm of the piano playing movements is more in line with the musical rhythm characteristics of the piano playing itself and the technical requirements of the movements. To summarize, the time ratio of the beginning, transition and end stages of piano performance should be from the largest to the smallest. Therefore, the ratio of time rhythm should be grasped during the piano performance training.

Table 3: Time data of different subjects at each stage

Serial number	The beginning stage /s	Transitional stage /s	Conclusion stage /s	Total /s	Ratios of each stage
1	120.50	110.71	65.86	297.07	1.83:1.68:1
2	99.50	127.55	69.34	296.39	1.43:1.84:1
3	97.40	128.55	85.47	311.42	1.14:1.50:1
$\overline{X} \pm SD$	105.80	122.27	73.56	301.63	
	±12.77	±10.02	±10.46	±8.49	

(2) Analysis of the differences in the speed and rhythm characteristics of the center of gravity in each stage

The speed rhythm of the change of the center of gravity in each stage of the piano playing action accomplished by different players refers to the change of the center of gravity of the body in the speed in the three stages of the beginning stage, the transition stage and the end stage, including the horizontal speed of the center of gravity of the body, the vertical speed of the center of gravity of the body, and the left and right speeds of the center of gravity of the body, respectively, as shown in Tables 4~6.

The center of gravity level velocity shows that player #1 was the slowest in the beginning stage at 0.033/m/s, while player #3 was the fastest. According to the tempo and movement requirements of the selected piano program, the relatively slowest speed among the 3 is preferred. The transition and ending phases should be relatively fast compared to the beginning phase, with player #1 being relatively the fastest and player #3 being the slowest in the ending phase. Player #1 and #2 ranked the 3 stage ratios as Beginning Stage < Transition Stage < End Stage, while Player #3 ranked Beginning Stage < End Stage < Transition Stage. Player #3 does not quite match the rhythmic speed of the movement, and player #1 is relatively slow in the transition and ending phases and relatively fast at the beginning. The horizontal speed of the center of gravity affects the proportion of the rhythmic distribution of the whole piano playing movement, so the speed of the center of gravity should be relatively slow at the beginning, and relatively fast at the transition and the end.

The vertical velocity of the center of gravity shows that: the speed of the three people in the beginning stage is small, in which the No.1 player has the smallest speed, and the standard deviation is 0.002m/s. The vertical velocity of the center of gravity in the beginning stage is relatively slower, which indicates that the movement undulation in the beginning stage is small and slow in the piano playing movement, in which the center of gravity speed of the No.1 player is the smallest. The vertical velocity of the center of gravity in the transition stage and the end stage is relatively fast, and the vertical velocity of the center of gravity in the transition stage should be smaller than that in the end stage. The ratio shows that the vertical speed of the center of gravity of the three players in the three phases are as follows: the beginning phase, the transition phase, and the end phase. However, the difference between each stage is more obvious for player No. 1, which indicates that the undulating rhythm in the whole piano playing movement is from slow to fast, which is the most obvious.

The center of gravity velocities showed that the difference between the three stages was small, and the order of the ratios was: beginning stage < transition stage < end stage. Analyzing the rhythm of the center of gravity speed of the three subjects, it was concluded that the horizontal speed of the center of gravity affects the proportion of the rhythmic distribution of the whole piano playing movement, so the center of gravity speed should be relatively slower in the beginning stage, and relatively faster in the transition stage and the end stage. Different subjects presented different allocation ratios of center of gravity speed, and the allocation ratio of the No.1 player was the most in line with the basic requirements of movement rhythm, and the ratio of center of gravity speed in the three stages was in the order of beginning stage<transition stage<end stage.

The above analysis shows that the distribution ratio of the center of gravity speed in the 3 stages should be grasped, which affects the whole piano playing action rhythm and action requirements, and at the same time, the basic rhythm and requirements of the action constrain the distribution ratio of the center of gravity speed in the 3 stages.

Table 4: Horizontal speed of body center of gravity

Serial number	The beginning stage /m·s-1	Transitional stage / m·s-1	Conclusion stage / m·s-1	Ratios of each stage
1	0.033	0.046	0.058	1:1.39:1.76
2	0.035	0.043	0.055	1:1.23:1.57
3	0.036	0.042	0.056	1:1.17:1.56
$\overline{X} \pm SD$	0.035	0.044	0.056	
	±0.002	±0.002	±0.002	

Table 5: Vertical velocity of center of gravity

Serial number	The beginning stage /m·s-1	Transitional stage / m·s-1	Conclusion stage / m·s-1	Ratios of each stage
1	0.005	0.019	0.031	1:3.8:6.2
2	0.009	0.013	0.027	1:1.44:3
3	0.008	0.011	0.029	1:1.38:3.63
$\overline{X} \pm SD$	0.007	0.014	0.029	
	±0.002	±0.004	±0.002	

Table 6: Left-right velocity of center of gravity

Serial number	The beginning stage /m·s-1	Transitional stage / m·s-1	Conclusion stage / m·s-1	Ratios of each stage
1	0.039	0.047	0.051	1:1.21:1.31
2	0.045	0.049	0.05	1:1.09:1.11
3	0.042	0.048	0.049	1:1.14:1.17
$\overline{X} \pm SD$	0.042	0.048	0.050	
	±0.003	±0.001	±0.001	

IV. Conclusion

In this paper, we constructed a musical instrument playing action recognition model based on improved GoogLeNet, and performed human biomechanical modeling of musical instrument playing actions, and explored the similarities and differences in the rhythmic features of different players' actions.

The accuracy of the trained action recognition model in this paper fluctuates around 95%, and the training loss is close to 0, which proves the feasibility of this model in classifying and recognizing musical instrument playing actions. Compared with other action recognition methods, the average recognition accuracy of this paper's method achieves the highest value, but the correct rate still remains around 66%, which still has some room for improvement and can be the focus of the next research.

Through the analysis of the spatial rhythm of the center of gravity change in each stage of three different subjects, it can be seen that the overall characteristics of the horizontal displacement of the center of gravity, vertical displacement, and left-right displacement of the piano playing action under the same repertoire of different players are the same, but there is a certain degree of variability of the individual in each stage, which is manifested in:

(1) The allocation of time ratio affects the basic rhythm of the movement. In the beginning stage, the horizontal displacement of the center of gravity should be as large as possible, the vertical displacement should be as small as possible, and the left and right displacements should be as small as possible. In the transition phase the horizontal displacement should be close to the beginning phase, the vertical displacement distance and amplitude is larger, and the left-right displacement is very small. In the end stage, the horizontal displacement is the smallest relative to the first two stages, and the center of gravity horizontal velocity in the three stages is the best order of ratio: start stage < transition stage < end stage.

(2) The adjustment of the rhythm of center of gravity displacement and center of gravity speed not only affects the basic rhythm and requirements of the whole piano playing movement, but also restricts the distribution ratio of center of gravity displacement and center of gravity speed in each stage.

References

[1] Reybrouck, M., & Eerola, T. (2017). Music and its inductive power: a psychobiological and evolutionary approach to musical emotions. Frontiers in psychology, 8, 494.

[2] Taruffi, L., Allen, R., Downing, J., & Heaton, P. (2017). Individual differences in music-perceived emotions: The influence of externally oriented thinking. Music Perception: An Interdisciplinary Journal, 34(3), 253-266.

[3] Ribeiro, F. S., Santos, F. H., Albuquerque, P. B., & Oliveira-Silva, P. (2019). Emotional induction through music: Measuring cardiac and electrodermal responses of emotional states and their persistence. Frontiers in psychology, 10, 451.

[4] Doğantan-Dack, M. (2016). The role of the musical instrument in performance as research: The piano as a research tool. In Artistic practice as research in music: Theory, criticism, practice (pp. 169-202). Routledge.

[5] Mazur, Z., & Laguna, M. (2019). The role of affect in practicing a musical instrument: A systematic review of the literature. Psychology of Music, 47(6), 848-863.

[6] TURSINOVICH, N. D. (2021). Uzbek National Musical Instrument Performance. JournalNX, 7(1), 315-318.

[7] Kim, H. S., & Kim, H. S. (2018). Effect of a musical instrument performance program on emotional intelligence, anxiety, and aggression in Korean elementary school children. Psychology of Music, 46(3), 440-453.

[8] Girgin, D. (2017). The relations among musical instrument performance self-efficacy, self-esteem and music performance anxiety in pre-service music teachers. Educational Research and Reviews, 12(11), 611-616.

[9] Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2020). This time with feeling: Learning expressive musical performance. Neural Computing and Applications, 32, 955-967.

[10] Visi, F., Coorevits, E., Schramm, R., & Miranda, E. R. (2017). Musical instruments, body movement, space, and motion data: music as an emergent multimodal choreography. Human technology, 13(1), 58.

[11] Manitsaris, S., Tsagaris, A., Dimitropoulos, K., & Manitsaris, A. (2015). Finger musical gesture recognition in 3D space without any tangible instrument for performing arts. International Journal of Arts and Technology, 8(1), 11-29.

[12] Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., & Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. Sensors, 19(5), 1005.

[13] Al-Faris, M., Chiverton, J., Ndzi, D., & Ahmed, A. I. (2020). A review on computer vision-based methods for human action recognition. Journal of imaging, 6(6), 46.

[14] Kamath, U., Liu, J., Whitaker, J., Kamath, U., Liu, J., & Whitaker, J. (2019). Convolutional neural networks. Deep learning for NLP and speech recognition, 263-314.

[15] Vonder Haar, L., Elvira, T., & Ochoa, O. (2023). An analysis of explainability methods for convolutional neural networks. Engineering Applications of Artificial Intelligence, 117, 105606.

[16] Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., & De, D. (2020). Fundamental concepts of convolutional neural network. Recent trends and advances in artificial intelligence and Internet of Things, 519-567.

[17] Liu, M., Shi, J., Li, Z., Li, C., Zhu, J., & Liu, S. (2016). Towards better analysis of deep convolutional neural networks. IEEE transactions on visualization and computer graphics, 23(1), 91-100.

[18] Costa, Y. M., Oliveira, L. S., & Silla Jr, C. N. (2017). An evaluation of convolutional neural networks for music classification using spectrograms. Applied soft computing, 52, 28-38.

[19] Solanki, A., & Pandey, S. (2022). Music instrument recognition using deep convolutional neural networks. International Journal of Information Technology, 14(3), 1659-1668.

[20] Li, C., Zhong, Q., Xie, D., & Pu, S. (2017, July). Skeleton-based action recognition with convolutional neural networks. In 2017 IEEE international conference on multimedia & expo workshops (ICMEW) (pp. 597-600). IEEE.

[21] Huang, W., Zhang, L., Wu, H., Min, F., & Song, A. (2022). Channel-equalization-HAR: A light-weight convolutional neural network for wearable sensor based human activity recognition. IEEE Transactions on Mobile Computing, 22(9), 5064-5077.

[22] Wang, P., Li, W., Li, C., & Hou, Y. (2018). Action recognition based on joint trajectory maps with convolutional neural networks. Knowledge-Based Systems, 158, 43-53.
[23] Ijjina, E. P., & Chalavadi, K. M. (2016). Human action recognition using genetic algorithms and convolutional neural networks. Pattern recognition, 59, 199-212.
[24] Hou, Y., Li, Z., Wang, P., & Li, W. (2016). Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology, 28(3), 807-811.
[25] Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., & Feng, D. D. (2018). Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(9), 1806-1819.
[26] Amirhossein Farajollahi & Mir Masoud Seyyed Fakhrabadi. (2025). Convolutional neural networks to predict dispersion surfaces-based properties of acoustic metamaterials with arbitrary-shaped unit cells. Results in Engineering,26,104905-104905.
[27] Jinli Wang, Jin Tong, Jun Li, Chunli Cao, Sirui Wang, Tianyu Bi... & Xinwu Cui. (2024). Using the GoogLeNet deep-learning model to distinguish between benign and malignant breast masses based on conventional ultrasound: a systematic review and meta-analysis..Quantitative imaging in medicine and surgery,14(10),7111-7127.
[28] Fulong Liu, Gang Li, Shuqiang Yang, Wenjuan Yan, Guoquan He & Ling Lin. (2020). Detection of heterogeneity in multi-spectral transmission image based on spatial pyramid matching model and deep learning. Optics and Lasers in Engineering,134.