Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

AI-enabled enhancement of emotional regulation and musical expression in the art of music conducting

Wenyun Shen[1,] ✉

[1] *Communication University of Zhejiang, Hangzhou, Zhejiang, 310000, China*

ABSTRACT

Music conductors rely on the visual impact of gestures and emotions for the interpretation and expression of musical works. In this paper, we utilize spatio-temporal two-stream convolutional neural network and replace the original VGG-16 network with ResNet-34 network with deeper network structure to construct a conductor recognition model for improving music conductor level. The Dropou optimization is applied in the fully connected layer to reduce the overfitting phenomenon, and the network structure is designed to fuse the temporal and spatial networks in advance with the feature maps, in view of the defects that the network structure of dual-stream convolutional neural network is shallow and the temporal and spatial networks do not learn the temporal and spatial information correlation. After the construction is completed, the model is applied in the teaching of a music college. The spatio-temporal information fusion convolutional neural network proposed in this paper is compared with other existing methods, and it is found that the optimized design helps the convolutional neural network to learn better, and better emotion and action effects can be obtained. It has better recognition accuracy on the dataset and obtained the highest accuracy of 74.3% on the CoST dataset. The results of the dimensions of music perception ability of the conductor students in the experimental class are better than the reference class, and the dimensions of pitch and intensity are more than 20% ahead of the control class, which proves that the model in this paper is more powerful to promote the development of music perception of the conductor students.

Keywords: convolutional neural network, Dropout mechanism, spatio-temporal network, music conducting

1. Introduction

In the art of music conducting, through body language, conducting movements and voice expression, the conductor conveys the understanding of music and emotion directly to the choir members, so that they can accurately understand the conductor's requirements and transform them into voice

✉ Corresponding author.
E-mail address: 15558002187@163.com (W. Shen).
Received 25 March 2024; Revised 05 June 2024; Accepted 15 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-497

expression [1-4]. The conductor's conducting movements and conducting skills can help choir members to better grasp the rhythm, rhyme and pitch, and make the choir more musically expressive. The performance of conducting art can also enhance the confidence and team cohesion of choir members [5-8]. When the conductor can lead by example and infect the choir members with passionate and powerful performances, the members will have trust and respect for the conductor and become more involved in the choral performances [9-10].

Emotional control requires the conductor to have profound musical literacy and keen emotional perception. The conductor needs to accurately grasp the emotions behind each note through the understanding of the composer's creative intention and musical connotation, and pass these emotions to the chorus members through his or her own conducting art [11-14]. Only when the conductor has a deep understanding and experience of the emotions of the musical work, can he truly infect others and make the choir's singing more moving. Emotional regulation not only enables choir members to better understand and feel the music, but also can bring the emotional distance between choir members closer, enhance mutual understanding and team cohesion [15-18]. With the increasing application of artificial intelligence in the field of music, artificial intelligence, with its powerful algorithms and machine learning, will create diverse styles of emotional regulation and musical expression in the art of music conducting [19-20].

In this paper, a music conductor recognition model is constructed based on a two-stream convolutional network. The dual-stream convolutional neural network is designed as a spatial streaming and temporal streaming network, which is trained on a single frame of the video as well as an optical streaming image extracted from consecutive frames, respectively. The two networks extracted the spatial information of the motion as well as the temporal information of the motion respectively, which solved the problem of the shallow structure of the VGG-16 network, resulting in inadequate feature extraction. The recognition results of the two are finally fused using the mean fusion method. The recognition accuracy is compared with the rest of the baseline model, and after case study, it is put into the teaching of conducting profession in a music college.

2. Command recognition model based on two-stream convolutional network

2.1. Requirements for the conductor to integrate the emotion of the work with the movement of the conductor

2.1.1. Fully understand the emotion of the work. In the process of emotional expression, due to the different emotions contained in each musical work, the emotional performance factors presented are not the same, and each work has its own artistic characteristics and styles, the conductor, as one of the performers of the music chorus, should understand the content and emotion of the song itself, and savor the feeling of a certain emotion conveyed by the work, and pass it to the general audience, so that the musical work in the performance to be experience and comprehend.

2.1.2. Mastery of command movements. The article on the choral conductor gesture movement is the basic part of the full set of conductor movement, most of the conductor method in the process of specific use of conventional and special two, even in the same emotional expression can still be found in the details of the content of the difference. Therefore, the conductor needs to repeatedly practice the basic command movement, design the command gesture movement in advance, and use a variety of movement interspersed form to enrich the command technique. In addition, as a kind of specialized musical language, the conductor gesture movement should not be plain and strange, and avoid exaggerating the movement as much as possible, and should comply with the intention of the work on the basis of the appropriate use of the conductor gesture movement for the description of the work.

2.1.3. Developing a good sense of rhythm. The four elements of music are rhythm, melody, harmony and timbre. Rhythm, as one of the elements of conducting, is considered to be the core and skeleton of music conducting. For conductors, whether they have a good sense of rhythm has gradually become an important indicator of their level and ability. Anyone including the conductor's sense of rhythm is slowly cultivated, is in the usual training to cultivate a kind of auditory habits, so emphasize the link of the beat is to cultivate the conductor's sense of rhythm is an effective way to rhythm is no longer all the time in the music, the music is the process of unification of the acoustic rhythm and the body rhythm, so the rhythm of the training of high-quality conductor plays an important and decisive role. People often only pay attention to the command of the colorful, enthusiastic and cheerful external form, but seldom to study the form of the hidden infinite richness of meaning.

2.1.4. Developing flexible hinting skills. From the perspective of broad concept, the definition of implied ability mainly focuses on an expressive ability, which is not based on any language or sound, but purely expressed by self-will. In choral performance, the communication between singers and choristers is mainly based on silent language, and the conductor reminds the singers of the precautions to be taken through changes in gestures and movements. Generally, there are several fixed or temporary movements, which are used by the conductor to guide the singers to complete the performance successfully. The emotion that the writer puts into the composition and the understanding of the musical spirit of the piece by the members of the ensemble are all presented with the help of the conductor's on-the-spot precise judgment and quick reaction. If the conductor himself does not have a strong ability to imply, the conductor's intention can not be conveyed to the whole chorus in the first time, and it is easy to affect the expressive power of the choral performance, and even can not present the actual meaning of the piece in the real sense, and the conductor's actual task can not be completed in accordance with the actual standard successfully. When dealing with the depth of breathing, speed change and voice change in the music, the conductor can fully utilize his own implied ability to express effectively. At present, there are many different ways of cueing.

2.2. Analysis of Convolutional Neural Network Related Technologies

2.2.1. Basic structure and principles of convolutional neural network. The basic structure of convolutional neural network contains convolutional layer, pooling layer and fully connected layer.

1) Convolutional layer. Convolution refers to the process of extracting feature information such as color and texture of an image using a filter, i.e., a convolution kernel, which has the following main features:

(1) Localized connectivity. Unlike traditional artificial neural networks, convolutional neural network neurons are only locally connected to the image, the spatial size of this connection is called the neuron's receptive field, and its size, i.e., the size of the convolution kernel, is a hyperparameter.

(2) Spatial arrangement. Convolutional neural networks are controlled by three hyperparameters, depth, step size and padding, to control the size of the data body. In this case, depth is the number of channels of output data after convolution operation, which is consistent with the number of convolution kernels [21].

After one convolution operation, the size n_1 of the output feature map can be calculated by the following equation:

$$n_1 = \frac{n + 2p - f}{s} + 1 \tag{1}$$

Since this operation has m convolutional kernel, the depth of the output is m and the size of the output is $[(n + 2p - f)/s + 1]^2 \times m$.

(3) Weight sharing. Weight sharing means that only one convolutional kernel can be used to complete feature extraction at different locations of an image, because the convolutional kernel has the same filtering effect for different locations of the same image, and it is not necessary to use different filters at different locations.

(4) 1×1 convolutional kernel. The traditional convolutional kernel needs to map correlations across channel and spatial dimensions simultaneously, including information in the channel dimension and information in both spatial dimensions.

2) Pooling layer. The pooling layer is also called the downsampling layer, and the common pooling operations in convolutional neural networks are maximum pooling, average pooling, and global average pooling.

(1) Max pooling. Max pooling is the process of selecting the largest feature value of the feature map in the pooling window as the output. When performing max pooling operation, the index position of the largest eigenvalue in the feature map needs to be saved to facilitate back propagation.

(2) Average pooling. Average pooling is the process of averaging all the feature values of the feature map in the pooling window as the output. It can eliminate the error caused by limited neighborhood size to a certain extent, and retain the background information of the image well, which is usually used for scene recognition.

(3) Global Average pooling. In the traditional classification network, the feature maps will be integrated into vector form by the fully connected layer, and then input into the Softmax layer to get the score. Too many fully-connected layers can lead to overfitting, and thus usually need to be paired with a Dropout operation. Instead of full connectivity, global average pooling directly pools the whole feature map on average and then inputs it into the Softmax layer to get the corresponding score.

3) Classifier. Softmax is generally chosen as a classifier in multiclassification problems. It is able to map the inputs of the previous layer to real numbers between 0 and 1 and make them sum to 1. Assuming an input vector of $x \in R^n, z_i = w_i^T x_i + b_i, i = 1, 2, \ldots, k$, then $z = [z_1, z_2, \ldots z_n] \in R^K$ and an output vector of $s \in R^K$ is defined:

$$s = softmax(z) \tag{2}$$

have for the i nd component of vector s:

$$s_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, i = 1, 2, \ldots, K \tag{3}$$

Suppose that a certain sample x corresponds to a vector of labels y and $y = [y_1, y_2, \ldots y_n] \in R^K$, and that sample x passes through the Softmax layer of a certain neural network model and outputs a vector $s \in R^K$. If the cross-entropy loss function is used, then the loss on that sample is:

$$L = -\sum_{k=1}^{K} y_k \ln s_k \tag{4}$$

In the backpropagation process, the partial derivatives $\partial L / \partial w_i$ and $\partial L / \partial b_i$ of the loss function with respect to w_i and b_i need to be found, obviously:

$$\frac{\partial \mathrm{L}}{\partial w_i} = \frac{\partial \mathrm{L}}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_i} \tag{5}$$

$$\frac{\partial \mathrm{L}}{\partial b_i}=\frac{\partial \mathrm{L}}{\partial z_i}\cdot\frac{\partial z_i}{\partial b_i} \tag{6}$$

Easily accessible:

$$\frac{\partial z_i}{\partial w_i}=x \tag{7}$$

$$\frac{\partial z_i}{\partial b_i}=1 \tag{8}$$

Since equation (3) states that any component s_K of s contains all components of z, there is:

$$\frac{\partial L}{\partial z_i}=\sum_{k=1}^{\kappa}\left[\frac{\partial L}{\partial s_k}\cdot\frac{\partial s_k}{\partial z_i}\right] \tag{9}$$

Again from equation (4):

$$\frac{\partial L}{\partial s_k}=\frac{\partial\left(-\sum_{k=1}^{K}y_k\ln s_k\right)}{\partial s_k}=-\frac{y_k}{s_k} \tag{10}$$

The definition of Eq. (4) is then used to find $\partial s_K/\partial z_i$, which is then divided into two cases:
When $k\neq i$, then:

$$\frac{\partial s_k}{\partial z_i}=\frac{\partial\left(\frac{e^{z_k}}{\sum_{j=1}^{K}e^{z_j}}\right)}{\partial z_i}=-s_k s_i \tag{11}$$

When $k=i$, there is:

$$\frac{\partial\left(\frac{e^{z_k}}{\sum_{j=1}^{K}e^{z_j}}\right)}{\partial z_i}=\frac{\partial\left(\frac{e^{z_k}}{\sum_{j=1}^{K}e^{z_j}}\right)}{\partial z_i}=s_i\left(1-s_i\right) \tag{12}$$

brought into Eq. (10), there:

$$\frac{\partial L}{\partial z_i}=\sum_{k=1}^{\kappa}\left[\frac{\partial L}{\partial s_k}\cdot\frac{\partial s_k}{\partial z_i}\right]=-y_i+s_i\sum_{k=1}^{K}y_k \tag{13}$$

Labeled Samples $y=[y_1,y_2,\ldots y_n]\in R^K$ For multicategorization problems usually in the form of One-hot coding, only one element of the vector is 1 and all the remaining elements are 0. Then there are:

$$\sum_{k=1}^{K}y_k=1 \tag{14}$$

So there is:

$$\frac{\partial L}{\partial z_i} = s_i - y_i \tag{15}$$

In summary, under the multiclassification problem:

$$\frac{\partial L}{\partial w_i} = x(s_i - y_i) \tag{16}$$

$$\frac{\partial L}{\partial b_i} = s_i - y_i \tag{17}$$

Use the formula:

$$w_i = w_i - \frac{\partial L}{\partial w_i} \tag{18}$$

$$b_i = b_i - \frac{\partial L}{\partial b_i} \tag{19}$$

The update of the corresponding weights can be accomplished by backpropagation.

4) Backpropagation between convolutional layer and pooling layer

(1) Backpropagation from pooling layer to convolutional layer. Assuming that the pooling layer is the lst layer and its sensitivity is δ^l, and the convolutional layer is the $l-1$rd layer, so that the loss function is L, then:

$$\delta^{l-1} = \frac{\partial L}{\partial z^{l-1}} = \frac{\partial L}{\partial z^l} \cdot \frac{\partial z^l}{\partial a^{l-1}} \cdot \frac{\partial a^{l-1}}{\partial z^{l-1}} \tag{20}$$

Since z^l is the result of downsampling the convolutional layer output $a^l - 1$, $\frac{\partial L}{\partial z^l}\frac{\partial z^l}{\partial a^{l-1}}$ is equivalent to $upsample(\delta^l)$ at this point:

$$\delta^{l-1} = upsample(\delta^l) \cdot \frac{\partial a^{l-1}}{\partial z^{l-1}} = upsample(\delta^l) \cdot \sigma'(z^{l-1}) \tag{21}$$

where $upsample(\delta^l)$ means that the sensitivity δ^l of the pooling layer is first reduced to the size before pooling, and then different backpropagation is performed for different pooling operations.

(2) Backpropagation from convolutional layer to pooling layer. Assume that the convolutional layer has only one convolutional kernel and that the convolutional layer is layer l with error sensitivity δ^l and the pooling layer is layer $l-1$. Similarly, there are:

$$\delta^{l-1} = \frac{\partial L}{\partial z^{l-1}} = \frac{\partial L}{\partial z^l} \cdot \frac{\partial z^l}{\partial a^{l-1}} \cdot \frac{\partial a^{l-1}}{\partial z^{l-1}} \tag{22}$$

Since there is no activation function for the pooling layer, there is $\frac{\partial z'}{\partial a^{l-1}} = 1$. Therefore, equation (22) can be expressed as follows

$$\delta^{l-1} = \frac{\partial L}{\partial z^{l-1}} = \delta^l \cdot \frac{\partial a^{l-1}}{\partial z^{l-1}} \tag{23}$$

As:

$$z^l = a^{l-1} * w^l = \sigma'(z^{l-1}) * w^l \tag{24}$$

So:

$$\delta^{l-1} = \frac{\partial L}{\partial z^{l-1}} = \delta^l \frac{\partial a^{l-1}}{\partial z^{l-1}} = \delta^l * rot180°(w^l) \cdot \sigma'(z^{l-1}) \tag{25}$$

Take partial derivatives with respect to w^l

$$\frac{\partial L}{\partial w^l} = \frac{\partial L}{\partial z^l} \cdot \frac{\partial z^l}{\partial w^l} = \delta^l \cdot \frac{\partial z^l}{\partial w^l} \tag{26}$$

Is then obtained from $z^l = a^{l-1} * w^l$:

$$\frac{\partial L}{\partial w^l} = \frac{\partial L}{\partial z^l} \cdot \frac{\partial z^l}{\partial w^l} = \delta^l \cdot \frac{\partial z^l}{\partial w^l} = \delta^l \cdot rot180°(a^{l-1})\delta^l = \delta^{l-1} \tag{27}$$

$$w^l = w^l - \frac{\partial L}{\partial w^l} \tag{28}$$

The weight update can be accomplished by back propagation.

2.2.2. Convolutional Neural Network Optimization Methods

1) Migration learning. Migration learning, is the process of taking the structure and weights of a model that has been learned or trained in an old domain and applying them to a new domain, using similarities between data, tasks, or models. There are two ways to use migration learning: the first is fine-tuning, the structure of which is shown in Figure 1. In the case where the existing dataset is small and the data similarity is high. The trained model can be simply modified for new learning tasks. The second one is called fixed feature extractor and the structure is shown in Fig. 2. In the case where the existing dataset is small and the data similarity is not high, the first k layer can be frozen and the second $n-k$ layers can be retrained, i.e., the pre-trained network is used as the feature extractor for the new task.

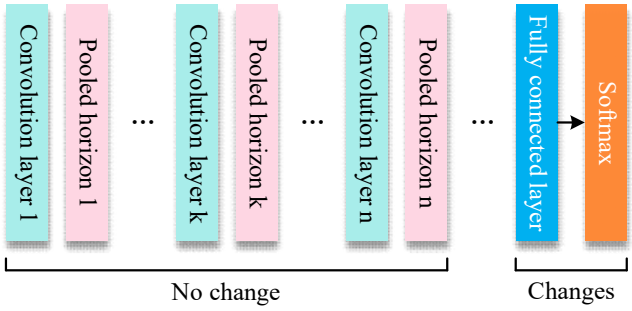

Fig. 1. Fine tuning

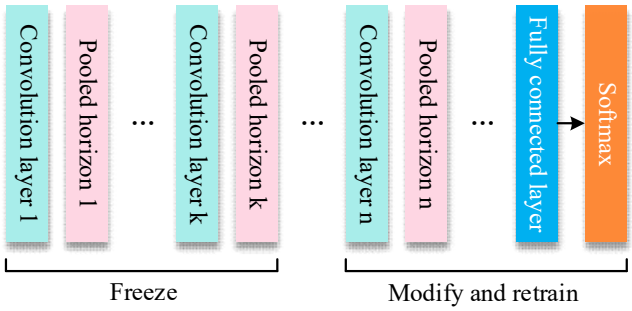

Fig. 2. Feature Extractor

2) Batch normalization. In deep learning, the use of input all the data to train the network has a large memory requirement and the training time for each round is too long, generally a small batch of data mini-batch is used to train the network.

Therefore, batch normalization is proposed on the basis of mini-batch.

Training for a mini-batch. mini-batch is denoted as $B = x\{x_{1....m}\}$, with a total of m training samples, at which point there are:

$$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \tag{29}$$

μ_B is the vector formed by averaging the corresponding features, using which the variance σ_B^2 can be found:

$$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{30}$$

Further, it can be normalized using the following equation:

$$\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{31}$$

However, simply using the above method will change the original data distribution, which will result in the loss of expressive power of the data. Therefore BN introduces two learnable parameters γ & β to linearly transform the normalized data, i.e.:

$$\hat{y}_i = \gamma \hat{x}_i + \beta \tag{32}$$

By bringing in learnable offsets through linear transformations, some of the expressive power of the input data can be recovered.

3) Dropout Mechanism. Dropout is a common technique for optimizing deep neural networks and is usually used in fully connected layers in convolutional neural network models. It can significantly reduce the overfitting phenomenon and improve the generalization ability of the model [22].

The forward propagation of Dropout is calculated as follows:

$$r_j^l = Bernoulli(p) \tag{33}$$

$$y^l = r^l * y^l \tag{34}$$

$$z_i^{l+1} = w_i^{l+1} y^l + b_i^{l+1} \tag{35}$$

$$y_i^{l+1} = f(z_i^{l+1}) \tag{36}$$

where the probability vector r in Eq. (34) is a zero-one vector, which is generated by the Bernoulli function. In the testing phase, the weight parameter of each neural unit needs to be multiplied by the probability p of Dropout to scale the weights. The formula for Dropout in the testing phase is as follows:

$$w_{test}^l = pW^l \tag{37}$$

2.3. Research on command action recognition based on two-stream convolutional network

2.3.1. Spatio-temporal two-stream convolutional neural network modeling. The structure of each part of the network is first described:

1) VGG-16 network. VGG-16 has the following features:(1) The parameters of the convolution kernel can be reduced under the premise of obtaining the same sensory field, which alleviates the overfitting of the network, followed by the increase of the convolution layer along with the nonlinear activation function, which enhances the nonlinearity and generalization ability of the network. (2) A 2×2 pooling kernel was used instead of a 3×3 pooling kernel, and the use of maximum pooling better captures the changes in the image. (3) Convolution is used instead of fully connected testing, which can make the input image arbitrary in size.

2) Spatial flow convolutional neural network. Spatial flow convolutional neural network operates on single frame video images, it is used to effectively recognize human movement information within video frames by performing feature extraction on static video frames. The spatial flow convolutional neural network uses the VGG-16 network pre-trained on ImageNet.

3) Time stream convolutional neural network. The input image to the temporal flow convolutional neural network is an optical flow image generated by superimposing the optical flow information of multiple consecutive video frames.

The dense optical flow between two consecutive frames can be viewed as a set of displacement vector fields d_t between frames t and $t+1$, denoted by $d_t(u,v)$ as the feature point (u,v) in the frame t image moving to the corresponding position in the frame $t+1$ image. The horizontal component d_t^x and vertical component d_t^y of d_t in the displacement field can be regarded as different channels of the optical flow image, which is very suitable for the recognition task using convolutional neural network. To better perform the action recognition task, in this paper, the optical flow channels d_i^x, d_i^y of t video optical flow frames are stacked to form the number of input channels of dimension $2L$. Denote the width and height of the image by w and h respectively. For any video frame τ there, its convolutional input can be represented by a video block of dimension $I_\tau \in R^w \times h \times 2L$, and the relevant construction formula is shown in the following equation:

$$\begin{aligned} I_\tau(u,v,2k-1) &= d^x_{\tau+k-1}(u,v) \\ I_\tau(u,v,2k) &= d^x_{\tau+k-1}(u,v) \\ u=[1;w], v&=[1;h], k=[1;L] \end{aligned} \tag{38}$$

For any pixel point (u,v) in the video, its feature map channel $I_\tau(u,v,c), c=[1;2L]$ may encode the horizontal component and the vertical component of the point in the L frames superimposed.

2.3.2. Space-time convergence network design. In this paper, we deepen the network structure and fuse the feature maps extracted from the time-flow spatial-flow network right at the convolutional layer.

1) Residual Neural Network. Residual neural network introduces a constant mapping mechanism in the network, and its network structure is composed of multiple residual blocks. Its formula is shown under Eq.

$$y = F(x,\{W_i\}) + x \tag{39}$$

In Eq. x represents the input vector, y represents the output vector, and $F(x,\{W_i\})$ is the residual mapping to be learned for each residual block, denoted as a multilayer convolution operation in the network [23].

Assuming that F and x have the same dimension, the operation of $F+x$ is to sum the pixel values of each point of F and x, and when the dimensions of the two are not the same, one of the vectors must be processed:

$$y = F(x, \{W_i\}) + W_s x \tag{40}$$

W_s denotes the mapping matrix after the linear mapping of x, which ensures the same W_s dimensionality of F and x after mapping, and then the two matrices come to the summation work to get the output matrix y by Relu function.

ResNet-34 is used to replace the VGG-16 network in the original two-stream network to deepen the network.

In the ResNet-34 network, there are 32 convolutional layers, a fully connected layer and a pooling layer, $Conv1$ to $Conv\_5x$ represent the convolutional layer from the first layer to the fifth convolutional layer, where the number of residual blocks of the second, 3rd, 4th, and fifth convolutional layers is 3, 4, 6, 3, after the first convolutional layer, there is a maximum pooling layer, after the $Conv\_5x$ rd layer, there is an average pooling layer, and immediately after the fifth convolutional layer, there is a fully connected layer, with a total of 1000 neurons.

2) Fusion of network convolutional layers. For the fusion methods of convolutional layers, the following schemes are proposed in this paper: (1) summation fusion method; (2) maximum fusion method: (3) stacking fusion method; and (4) convolutional fusion method. These fusion methods are described below.

Summing fusion method: Summing fusion method calculates the sum of i, j pixel values of two feature maps with the same number of channels corresponding to the same spatial points as the fusion result:

$$y_{i,j,d}^{sum} = x_{i,j,d}^{a} + x_{i,j,d}^{b} \tag{41}$$

Maximum fusion method: Maximum fusion method takes the maximum value of i, j pixel of two feature maps with the same number of channels corresponding to the same spatial point as the fusion result:

$$y_{i,j,d}^{\max} = \max\{x_{i,j,d}^{a}, x_{i,j,d}^{b}\} \tag{42}$$

Stacked fusion approach: two feature maps with the same number of channels d are stacked at the same spatial location i, j:

$$y_{i,j,2d}^{cat} = x_{i,j,d}^{a} \quad y_{i,j,2d-1}^{cat} = x_{i,j,d}^{a} \tag{43}$$

Convolutional fusion method: the convolutional fusion method first stacks two feature maps with the same number of channels at the same spatial location i, j, and next performs a convolution operation using D convolution kernels f of dimension $2D \times 1 \times 1$, plus a bias b of dimension D The specific formula is shown below:

$$y^{conv} = y^{cat} * f + b \tag{44}$$

3) Design of the network model. Combined with what was said in the previous part of this chapter, the improvement strategy for the spatio-temporal dual-flow convolutional neural network is as follows:(1) Deepen the depth of the time-flow as well as spatial-flow convolutional neural network, and replace the original VGG-16 network by using ResNet-34. (2) Improve the fusion strategy, the original dual-stream convolutional neural network only in the spatial flow, the recognition results of the temporal flow network to take the average value of the fusion, the network is proposed to use a different fusion, in the spatial flow, the temporal flow ResNet34's $Conv1, Conv\_2x, Conv\_3x, Com\_4x, Conv\_5x$ after the spatial and temporal features using a different fusion strategy. The optimal network structure was finally derived, and the network structure is shown in Figure 3, which simplifies the network structure of ResNet-34 in order to easily express the network structure, after the network is fused, the spatial

flow, temporal flow network training does not have to be carried out individually, and the error in the training process is back-propagated to the cutoff of the fusion layer.

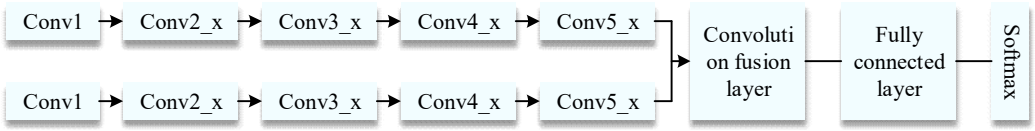

Fig. 3. Network structure proposed in this paper

3. Model performance testing and practical applications

3.1. Data frame rate optimization experiment

According to the characteristics of CNN, the input data must ensure the consistency of dimension as well as size, while the TouchGET dataset does not limit the sampling time in order to ensure the complete expression of emotion in each collection, so the number of sampling frames is not the same for each sample. For this reason, the data must be intercepted or supplemented for each sample to ensure that the dimensionality and size of the input data are consistent. Different lengths of data frames contain different data features, so it is necessary to choose the appropriate number of data frames to ensure that it contains enough data features to distinguish between gesture actions and emotion labels, but also to ensure the real-time nature of the data. This section therefore evaluates the effect of the number of intercepted frames on gesture and emotion recognition. When intercepting the data, for the samples whose individual sample frames are larger than the number of intercepted frames, the samples are segmented according to the specified number of intercepted frames to obtain multiple sub-sample data, and the remaining data with less than the number of intercepted frames will be discarded; for the cases where the individual samples are less than the number of intercepted frames, the data from the last frame is used to populate the samples up to the number of intercepted frames. After data interception, all data samples in the dataset are randomly divided into a training set and a test set, with the training set accounting for 80% of the number of clipped samples and the test set accounting for 20%.

Figures 4 and 5 show the gesture and emotion classification results, respectively. The horizontal axis of the figure represents the number of frames intercepted of different lengths, and the vertical axis represents the classification accuracy.

When the number of intercepted frames is from 8 to 32, the accuracy of gesture classification increases with the increase of the number of intercepted frames; when the number of intercepted frames is larger than 32, the accuracy of gesture classification basically remains unchanged, and the accuracy of gesture classification is highest when the length of intercepted frames is 48.

When the number of intercepted frames is from 8 to 56, the accuracy of sentiment classification increases with the increase of the number of intercepted frames; when the number of intercepted frames is larger than 56, the accuracy of sentiment classification slightly decreases and tends to stabilize. It can be seen that when the number of frames is short, the haptic data contains insufficient spatial information over time to extract effective haptic gesture and emotion information from the data, while with the increase of sampling time, the expression of gesture and emotion features in the haptic data is richer, and the classification accuracy increases, but stabilizes after a certain degree of increase, which is due to the fact that the redundant haptic data cannot provide more haptic gesture and emotion feature information. Therefore, the haptic data can basically cover most of the feature information of gesture movements when the interception length is 56 frames, and 56 frames of data are sufficient to express the emotional information in haptics.

The algorithm in this paper uses the shortest possible sampling time as the network input while maximizing the classification accuracy, which ensures the function of analyzing the data in real time in human-computer interaction.

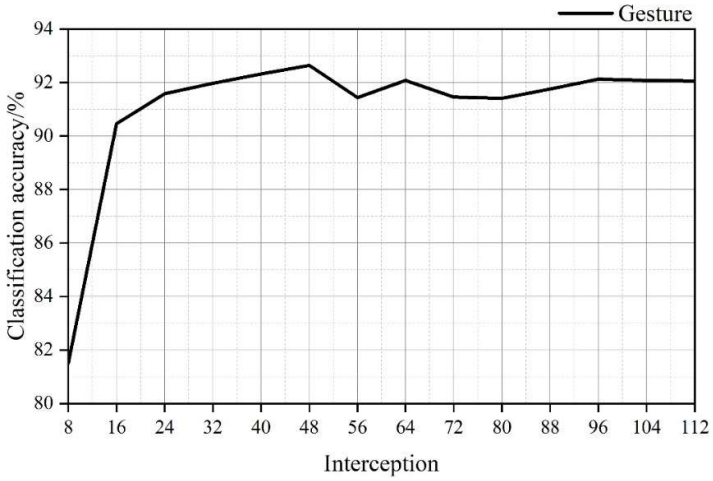

Fig. 4. Gesture classification

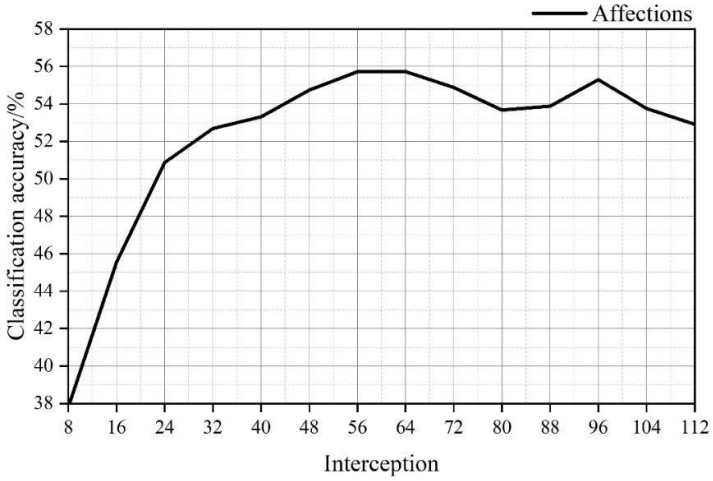

Fig. 5. Affective classification

3.2. Cross-subject experimental results

To build a cross-subject recognition model, this section uses the leave-one-subject method to train and test the model. The tactile data of one subject is used as the test dataset and the data of the remaining subjects is used as the training set to obtain the test results of this subject.

Fig. 6 and Fig. 7 show the training loss optimization process with the subject as the test set. Figure 6 shows the training and testing loss variation for the tactile gesture recognition model and Figure 7 shows the training and testing loss variation for the tactile emotion recognition model. As can be seen from the figure, the training loss of the cross-subject recognition model eventually stabilized under both tasks, but the testing loss first decreased rapidly and then gradually increased. In the early stage of training, the decrease in testing loss is due to the fact that the features learned by the model in the training data are partially applicable to the test samples; as the training proceeds, the model keeps fitting to the data samples in the training set, and the variability of the features of the training set and the test set in the model's output increases, and thus the testing loss gradually rises.

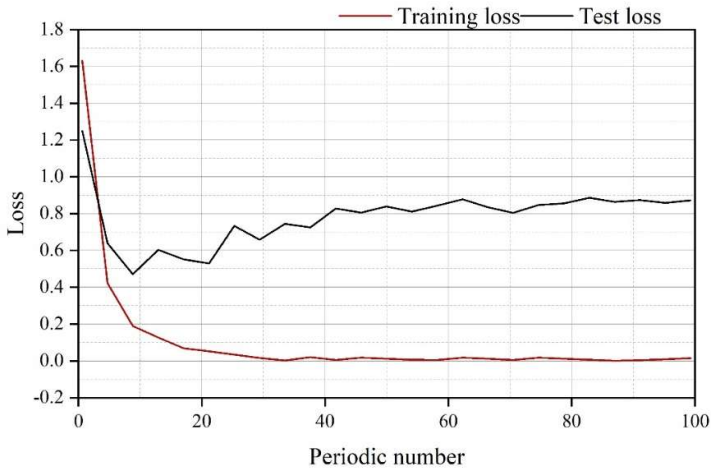

Fig. 6. Tactile gesture recognition model training and test loss change

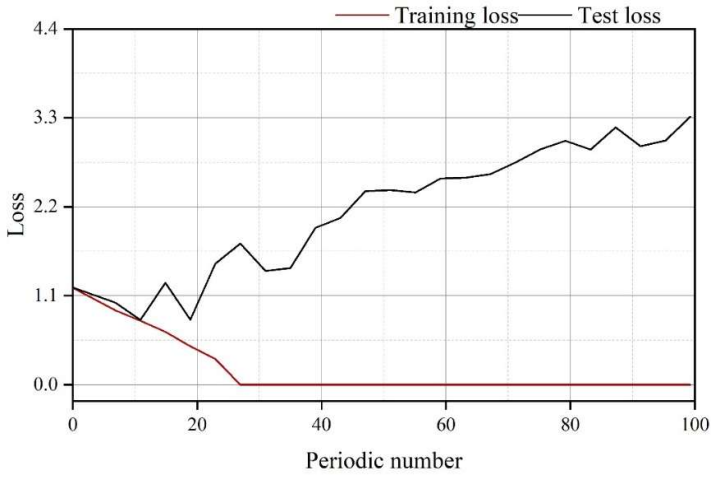

Fig. 7. The tactile emotion recognition model training and test loss change

The cross-subject gesture classification results are shown in Table 1. From the classification results, it can be seen that subject #6 had the highest gesture recognition accuracy of 96.22 %, subject #14 had the lowest gesture recognition accuracy of 74.02%, and the average gesture recognition accuracy of the 15 subjects was 84.11%.

Table 1. The accuracy of the gesture classification (%)

Subjects	Classification accuracy	Subjects	Classification accuracy	Subjects	Classification accuracy
1	93.82	6	96.22	11	96.17
2	76.16	7	87.87	12	82.19
3	81.43	8	74.66	13	91.4
4	95.25	9	77.98	14	74.02
5	80.9	10	74.46	15	79.1
AVE				84.11	

The gesture categorization results of the 15 subjects were summarized to obtain the cross-subject test evidence gesture categorization identification confusion matrix shown in Table 2. It can be seen that the most confusing gestures include 'scratching' and 'tickling' (total misclassification of 212 data), 'pushing' and 'pressing' (a total of 192 data were misclassified). The misclassification is mainly due to the similarity of the gestures, which may also be misrecognized by humans during social encounters.

Table 2. Identification confusion matrix

	Hit	Caress	Knead	Flap	Poke	Rap	Push	Press	Torsion	Tickle
Hit	635	0	0	11	44	19	7	0	0	0
Caress	0	911	37	36	0	3	3	3	21	4
Knead	0	18	524	1	0	0	0	0	3	17
Flap	18	48	0	901	1	20	15	26	2	1
Poke	36	0	0	0	300	31	0	0	0	0
Rap	14	4	0	19	38	427	2	0	4	13
Push	0	7	0	3	0	1	395	84	1	0
Press	0	0	0	11	0	0	108	439	1	0
Torsion	0	26	10	2	0	9	0	0	265	95
Tickle	0	12	69	2	0	30	0	0	117	272

The experimental results are compared with other models and the results are obtained as shown in Table 3. It can be seen that the 2D CNN and 3D CNN networks have higher parameter counts and the classification accuracy (72.38% and 75.26%, respectively) is much lower than the 87.46% of the model designed in this paper.2D CNNs are theoretically incapable of extracting features in the temporal dimension, whereas 3D CNNs use a larger time-step and a smaller number of channels. As a result, the existing networks used for tactile gesture recognition are unable to extract gesture features efficiently. The algorithm in this paper increases the network nonlinearity and network depth, which can effectively extract the temporal and spatial features of gesture signals with high classification accuracy.

Table 3. Comparison with the existing method of gesture classification

Model	Verification method	Accuracy rate	Model size (mb))	Flops (g)
3D CNN	Cross test certificate	69.15%	4.72	5.29
2D CNN	Cross test certificate	72.38%	2.69	5.74
3D CNN	Cross test certificate	75.26%	3.54	3.26
This model	Cross test certificate	87.46%	0.89	0.71

Tests were performed on the CoST dataset and the test results are shown in Table 4. The test results of other research methods on CoST dataset are also given in Table 4. The experimental results show that the algorithm of this paper currently obtains the highest accuracy of 74.3% on the CoST dataset.

Table 3 and Table 4 show that the music conductor recognition model in this paper can achieve better classification results for different gesture datasets.

Table 4. Comparison of identification results with existing methods

Classification method	Accuracy rate
Beeses classifier	52%
Deep code	56%
Random forest	60.05%
CNN-RNN	52.37%
CNN	63.29%
This model	74.30%

The four emotions 'joy', 'anger', 'sadness' and 'relaxation' can clearly represent the dimensional emotion model's four quadrants, so these four emotions were chosen for cross-subject emotion recognition.

Table 5 shows the results of cross-subject emotion categorization. In cross-subject categorization, the average emotion categorization accuracy of the 15 subjects was 60.69%, with the highest categorization accuracy of 76.11% for subject #1 and the lowest categorization accuracy of 37.16% for subject #13.

Table 5. The accuracy of cross-induced emotional classification (%)

Subjects	Classification accuracy	Subjects	Classification accuracy	Subjects	Classification accuracy
1	76.11	6	54.39	11	68.23
2	59.27	7	69.08	12	66.12
3	66.54	8	55.7	13	37.16
4	65.7	9	71.54	14	47.19
5	52.04	10	50.98	15	70.39
AVE				60.69	

The classification results of the 15 subjects were summarized to obtain the confusion matrix for cross-subject emotion classification as shown in Table 6. From the table, it can be seen that 'anger' has the best emotion categorization with a classification accuracy of 75.87% (371/489), and 'joy' has the lowest categorization accuracy of 41.58% (195/469).

Table 6. The confusion matrix of the test affective classification identification

	Joy	Anger	Sadness	Relax
Joy	195	69	49	48
Anger	84	371	16	8
Sadness	68	34	197	73
Relax	122	15	123	319
Tot	469	489	385	448

3.3. Experimenting with Strategies for Inspiring and Commanding Emotional Expression

3.3.1. Interaction links. In the conductor of musical repertoire, it is not possible to rely only on individual body language to express the emotion of the work, in order to better show the choral music, you need to pay attention to the connection between the gestures of each part. This connection is the connection between “loose” and “tight”, “dynamic” and “static”.

1) “Loosening” and “Tightening”. “Looseness” and “tightness” require the conductor's muscle activity to be relaxed and tightened in different situations. Loose, for the conductor is to achieve controlled relaxation, the muscles should be relaxed, the movement should be natural, but not completely relaxed without form, so that the conductor's movements can be more flexible. Tight, the conductor needs to mobilize his muscles, but not clumsy, to keep the whole limb loose and tight consistent, to achieve perfect control and management of the limb effect.

2) “Movement” and “Static”. The “movement” and “static” require the conductor to keep a reasonable standing posture in the conductor to ensure that the torso presents a sense of static, but the hand should move. Among them, the left hand is mainly responsible for suggestive gestures, so that the emotions of the chorus members can be reasonably guided to better express the emotions of the work; the right hand is mainly responsible for controlling the rhythm and speed of the melody, and changing the language of gestures according to the music melody. The right hand is mainly responsible for controlling the rhythm and tempo of the melody and changing the gestures according to the music melody.

3.3.2. Command gesture design. What plays a decisive role in the design of the conductor's language for choral works is the main style and content of the work. For example, “On the Taihang Mountains” combines lyrical and marching styles and can be divided into two parts, the first part with a mixed motion two-beat figure and the second part with a linear motion. There are many others similar to this, so it can be seen that appropriate conducting gestures should be devised for different stylistic types of works. Different styles of works determine the conductor gestures, and the main styles of works are march style, lyrical, and combat.

The main feature of the march style is the strong sense of rhythm, such as the “March of the Volunteers”, which has a great sense of shock, and its music tone is also very loud and clear, and there is a sense of urging people to move forward, and its emotional expression is also more stirring, the use of the rhythm of the melody to cause the listener's resonance. For this kind of work, in the design of the command gesture action, need to have a strong sense of power, each action should be sharp and bold, the action should be full of power, fast and decisive, and pay attention to the beat and the beat, close to the beat, contains full of emotion. In the conductor's way, the pivot point movement is also the main style, reflecting the atmospheric emotion of the work.

The lyrical style of the work contains a variety of content, including not only love songs, sadness, and praise and so on are the content of its performance. In the lyrical choral program with a softer rhythm, the conductor's gestures are slower, on the contrary, in the praise or cheerful choral program, the conductor's movements will be accelerated with the rhythm. Different swing speeds and strengths show the different emotions of the work, and in the conductor to curve-based, which is also another conductor gestures to show the emotional expression of the lyrical repertoire.

Combat style works have outstanding characteristics, the conductor of the process of swinging the beat to have a kind of internal resistance, this internal resistance in the strength of the obvious changes, with special emphasis on the beat point, through the command gestures on the beat point to show the intensity of the battle and the spirit of resistance. On the Taihang Mountains” has the combat nature, so in its conductor action, it should be mainly linear, through the posture action in the conductor, such as unfolding the arms and the expression of the firm conductor to lead the chorus members to show the different emotions and the combat nature of the work.

3.4. Effectiveness of model application

A one-semester teaching experiment was conducted in the conducting department of a conservatory, and two classes were selected as the experimental class and the reference class, respectively. The experimental class was taught using the model designed in this paper, and the control class was taught according to the original teaching plan. Conducting level tests were conducted before the beginning and after the end of the semester in the form of an orchestra performance.

Excel and SPSS were used to process the data from the pre-test and post-test to compare the data differences between the reference class and the experimental class.

3.4.1. Pre-test results and statistical analysis. The four components of musical ability of the experimental class and the reference class, i.e., musical perception, musical expression, musical creativity and musical comprehension, were counted to analyze whether the situation of the two classes in the indicators of musical ability was at a similar level and to judge whether the current situation of the musical ability of the two classes was suitable for the experiment. The results are shown in Table 7. Carrying out the chi-square test on the experimental data, the data show that there is no significant difference between the two groups of conducting students in various aspects of musical ability and overall musical ability, and the data are homogeneous and good enough to carry out the experiment.

Table 7. The previous test of music ability

Project		Mean	MD	SD	t	Sig.(Double tail)
perceptiveness	Laboratory orchestra	15.385	0.238	3.51999	0.3774	0.72828
	Reference orchestra	15.147				
expressiveness	Laboratory orchestra	13.974	-0.374	4.96442	-0.42126	0.69666
	Reference orchestra	14.348				
creativity	Laboratory orchestra	15.98	0.204	3.95861	0.28764	0.7956
	Reference orchestra	15.776				
understanding	Laboratory orchestra	18.938	0.68	3.08142	1.23318	0.24174
	Reference orchestra	18.258				

3.4.2. Post-test Results and Statistical Analysis of Experimental and Reference Classes. The results of the posttests of the experimental and reference classes of conducting students in the four dimensions of music perception, music expression, music creativity and music comprehension were tabulated. The results of the post-test are shown in Table 8.

The data show that there is a significant difference between the experimental class and the reference class of conducting students in terms of musical expressiveness, and a highly significant difference between the experimental class and the reference class of conducting students in terms of musical perception, musical creativity and musical understanding. The experimental class is 2.596, 2.492, 3.6783 and 3.669 points higher than the reference class in the four dimensions respectively, and the results of the dimensions of the students' musical ability are better than those of the reference class, which indicates that under the training of the recognition model of this paper, the interoperability of all the senses is mobilized to promote the development of the conductor's students' musical ability more strongly, which is of great significance for the development of the conductor's students' musical ability.

Table 8. The results are counted

Project		Mean	MD	SD	t	Sig.(Double tail)
perceptiveness	Laboratory orchestra	18.419	2.56	3.46	4.08	0.000**
	Reference orchestra	15.823				
expressiveness	Laboratory orchestra	15.752	2.18	5.28	2.15	0.035*
	Reference orchestra	13.26				
creativity	Laboratory orchestra	17.832	2.82	5.19	2.94	0.007**
	Reference orchestra	14.1537				
understanding	Laboratory orchestra	20.795	2.89	3.29	4.75	0.000**
	Reference orchestra	17.126				

3.4.3. Pre- and post-test analysis of experimental classes. Table 9 shows the statistics of the pre and post-test results of the experimental class of conducting students in the four dimensions of musical perception, musical expression, musical creativity and musical understanding. It can be seen that the post-test results of the dimensions of musical ability of the experimental class of conducting students are better than the pre-test results, and all these measurements show that the music activities of the conducting students based on artistic generalization can promote the development of the musical ability of the conducting students more strongly through the mobilization of the interoperability of all the senses, and that the music activities of the conducting students based on artistic generalization are of great significance to the development of the musical ability of the conducting students.

Table 9. Laboratory class survey

Project	Mean	MD	SD	t	Sig.(Double tail)

perceptiveness	Pretest	15.385	-3.38	3.85	-4.62	0.000**
	Posttest	18.419				
expressiveness	Pretest	13.974	-2.26	5.84	-2.11	0.45*
	Posttest	15.752				
creativity	Pretest	15.98	-2.27	5.59	-2.27	0.036*
	Posttest	17.832				
understanding	Pretest	18.938	-11.18	2.43	-4.54	0.000**
	Posttest	20.795				

3.4.4. Results and statistical analysis before and after reference classes. Table 10 shows the results of the pre- and post-test for the reference class. The data show that there is a significant difference between the pre- and post-test results of musical perception of the conductor students in the reference class, and there is no significant difference between the pre- and post-test results of musical expression, musical creativity and musical comprehension in the reference class. The results of the post-test and pre-test of the dimensions of musical ability of the conductor students in the reference class did not show any significant improvement, indicating that the traditional music teaching of the conductor students did not strongly promote the development of musical ability of the conductor students.

Table 10. Control class survey

Project		Mean	MD	SD	t	sig.(Double tail)
perceptiveness	Pretest	15.147	-1.05	2.1534	-2.578	0.015*
	Posttest	15.823				
expressiveness	Pretest	14.348	0.208	4.0729	0.271	0.792
	Posttest	13.26				
creativity	Pretest	15.776	0.64	4.5817	0.813	0.431
	Posttest	14.1537				
understanding	Pretest	18.258	0.109	2.9086	0.192	0.853
	Posttest	17.126				

3.4.5. Analysis of the Dimensions of Musical Emotional Expression. The following is a comparison of the posttest results of the emotional expression dimensions of pitch, intensity, rhythm, tempo, and melody of the orchestra under the direction of the experimental class and the reference class. Table 11 shows the results of each dimension.

The data show that there were significant differences in the posttest results of the dimensions of pitch, intensity, tempo, and melody, and no significant differences in the tempo dimension. The results of the dimensions of music perception ability of the conductor students in the experimental class were better than those of the reference class, and all of them were more than 0.48 points higher than those of the reference class. It proves that the teaching based on the command recognition model promotes the development of music perception of the conducting students more strongly by mobilizing the interoperability of all senses.

Table 11. The statistics of all dimensions of music

Project		Mean	MD	SD	t	sig.(Double tail)
Pitch	Laboratory orchestra	2.63	0.5096	1.26	2.15	0.035*
	Reference orchestra	2.15				
Intensity	Laboratory orchestra	4.79	0.6799	1.58	2.284	0.031*
	Reference orchestra	3.84				

Rhythm	Laboratory orchestra	3.52	0.76416	1.89	2.195	0.036*
	Reference orchestra	2.27				
Speed	Laboratory orchestra	4.46	0.1686	1.56	0.573	0.575
	Reference orchestra	3.83				
Melody	Laboratory orchestra	3.95	0.47452	1.14	2.243	0.033*
	Reference orchestra	3.21				

4. Conclusion

This paper constructs a music conductor recognition model based on two-stream convolutional network to realize the accurate capture and recognition of music conductor's movements and emotions. The data are intercepted for experiments to test the performance of this paper's model, and are applied to the teaching of conducting majors in a music college. It is found that the algorithm in this paper uses the shortest possible sampling time as the network input while maximizing the classification accuracy, which ensures the function of real-time data analysis in human-computer interaction. The classification accuracies of the two baseline models were 72.38% and 75.26%, which were much lower than the 87.46% of the model designed in this paper. The results of the dimensions of musical ability of the conductor students in the experimental class are better than those of the reference class, which indicates that under the training of the recognition model in this paper, the interoperability of all the senses can be mobilized to promote the development of the musical ability of the conductor students more powerfully, which is of great significance to the development of the conductor students' musical ability.

References

[1] Nozimov, A. (2023). FORMATION STAGES OF THE HISTORY OF THE ART OF CONDUCTING. Development and innovations in science, 2(1), 20-23.

[2] KANIOWSKA, M. (2018). The Art of Conducting from a Historical Perspective. Music Science Today: The Permanent & the Changeable/Muzikas Zinatne Sodien: Pastavigais un Mainigais, 10.

[3] Schmidt, S., Längler, M., Altenbuchner, A., Kobl, L., & Gruber, H. (2021). Acquiring the art of conducting: Deliberate practice as part of professional learning. Journal of Advanced Academics, 32(3), 354-379.

[4] Rucsanda, M. D. (2020). The role of the Conductor of children choir in the improvement of Interpersonal Communication through Music. Bulletin of the Transilvania University of Braşov, Series VIII: Performing Arts, 13(2-Suppl), 257-266.

[5] Muraru, A. (2021). THE ROLE OF FACIAL EXPRESSIONS IN THE ART OF MUSICAL CONDUCTING. Studia Universitatis Babes-Bolyai-Musica, 66(2), 247-253.

[6] Savchenko, R., & Savchenko, Y. (2019). Choral conducting competence as a pedagogical system. Journal of history culture and art research, (8), 382-390.

[7] Odusanya, O. S., & Idọlọr, E. G. (2023). Interpretive Approach to Conducting African Choral Works: A Conductors Task. CENTRAL ASIAN JOURNAL OF ARTS AND DESIGN, 4(10), 36-48.

[8] Mirea, R. (2018). Emotions and cognition in the music service. Învăţământ, Cercetare, Creaţie, 4(1), 107-118.

[9] Terasawa, H., Hoshi-Shiba, R., & Furukawa, K. (2019). Embodiment and interaction as common ground for emotional experience in music. In Proceedings of the 14th International Symposium on CMMR, eds Aramaki M., Derrien O., Kornland-Martinet R., Ystad S.(Marseille:) (pp. 777-788).

[10] Crescenzi, D. (2024). CONDUCTING GESTURE. CONDUCTING TECHNIQUES. Review of Artistic Education, (28), 337-348.

[11] Zhang, Y. (2024, September). Analysis of the Dual Role of Choral Conductors in Music Education. In 2024 3rd International Conference on Science Education and Art Appreciation (SEAA 2024) (pp. 110-116). Atlantis Press.

[12] Cook, T., Roy, A. R., & Welker, K. M. (2019). Music as an emotion regulation strategy: An examination of genres of music and their roles in emotion regulation. Psychology of Music, 47(1), 144-154.

[13] Xu, Y. (2023, October). The Role of Body Language in Orchestra Conducting. In 2023 2nd International Conference on Public Culture and Social Services (PCSS 2023) (pp. 127-133). Atlantis Press.

[14] Akbarova, G. N., Dyganova, E. A., Shirieva, N. V., & Adamyan, A. Z. (2017). THE ROLE OF THE CONDUCTOR'S PROFESSIONAL COMMUNICATIONS IN INTERACTION WITH A CHOIR. The Turkish online journal of design, art, and communication, 1510-1515.

[15] Jansson, D., Elstad, B., & Døving, E. (2021). Choral conducting competences: Perceptions and priorities. Research Studies in Music Education, 43(1), 3-21.

[16] Durrant, C. (2017). The framing of choirs and their conductors. The Oxford handbook of choral pedagogy, 221-236.

[17] Claraco, A. V. (2015). The process of nonverbal communication between choir and conductor. Revista Electrónica de LEEME, 36, 49-70.

[18] Corbalán, M., Pérez-Echeverría, M. P., Pozo, J. I., & Casas-Mas, A. (2019). Choral conductors to stage! What kind of learning do they claim to promote during choir rehearsal?. International Journal of Music Education, 37(1), 91-106.

[19] Soo, V. W., Huang, C. F., Su, Y. H., & Su, M. J. (2018). AI applications on music technology for edutainment. In Innovative Technologies and Learning: First International Conference, ICITL 2018, Portoroz, Slovenia, August 27–30, 2018, Proceedings 1 (pp. 594-599). Springer International Publishing.

[20] Li, Y. (2024). Innovative Cultivation of Choral Talents in Primary and Secondary Schools. Advances in Educational Technology and Psychology, 8(4), 32-38.

[21] Yongda Lin, Shiyu Xia, Lingxiao Wang, Baiyu Qiao, Hu Han, Linhui Wang... & Yajia Liu. (2025). Multi-task deep convolutional neural network for weed detection and navigation path extraction. Computers and Electronics in Agriculture109776-109776.

[22] Huma Israr, Safdar Khan, Muhammad Tahir, Muhammad Shahzad, Muneer Ahmad & Jasni Zain. (2023). Neural Machine Translation Models with Attention-Based Dropout Layer. Computers, Materials & Continua(2),2981-3009.

[23] Lina Song, Yousheng Tan, Fajun Yu, Yangcheng Luo & Jingjing Zheng. (2024). Optimal approximations for the free boundary problems of the space-time fractional Black-Scholes equations using a combined physics-informed neural network. Scientific Reports(1),25289-25289.