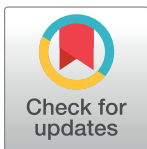


## RESEARCH ARTICLE

# Design and research of music teaching system based on virtual reality system in the context of education informatization

Yan Feng \*

School of Medical Humanities, Xinxiang Medical University, Xinxiang, 453000, China

\* [041019@xxmu.edu.cn](mailto:041019@xxmu.edu.cn) OPEN ACCESS

**Citation:** Feng Y (2023) Design and research of music teaching system based on virtual reality system in the context of education informatization. PLoS ONE 18(10): e0285331. <https://doi.org/10.1371/journal.pone.0285331>

**Editor:** Kathiravan Srinivasan, Vellore Institute of Technology, INDIA

**Received:** February 13, 2023

**Accepted:** April 20, 2023

**Published:** October 5, 2023

**Copyright:** © 2023 Yan Feng. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper and its [Supporting Information](#) files.

**Funding:** This study was supported by Construction of public music curriculum system in colleges and universities based on red culture (No. 2021-XYJG-96). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

Virtual Reality (VR) technology uses computers to simulate the real world comprehensively. VR has been widely used in college teaching and has a huge application prospect. To better apply computer-aided instruction technology in music teaching, a music teaching system based on VR technology is proposed. First, a virtual piano is developed using the HTC Vive kit and the Leap Motion sensor fixed on the helmet as the hardware platform, and using Unity3D, related SteamVR plug-ins, and Leap Motion plug-ins as software platforms. Then, a gesture recognition algorithm is proposed and implemented. Specifically, the Dual Channel Convolutional Neural Network (DCCNN) is adopted to collect the user's gesture command data. The dual-size convolution kernel is applied to extract the feature information in the image and the gesture command in the video, and then the DCCNN recognizes it. After the spatial and temporal information is extracted, Red-Green-Blue (RGB) color pattern images and optical flow images are input into the DCCNN. The prediction results are merged to obtain the final recognition result. The experimental results reveal that the recognition accuracy of DCCNN for the Curwen gesture is as high as 96%, and the recognition accuracy varies with different convolution kernels. By comparison, it is found that the recognition effect of DCCNN is affected by the size of the convolution kernel. Combining convolution kernels of size 5×5 and 7×7 can improve the recognition accuracy to 98%. The research results of this study can be used for music teaching piano and other VR products, with extensive popularization and application value.

## Introduction

With the rapid growth of science and lifestyle change, the way of education and learning is also facing the challenge and opportunity of significant transformation and development. The coronavirus disease 2019 (COVID-19) outbreak has provided a broad space for the development of online music education, but there are many problems with online music teaching [1,2]. Therefore, the study of online music education models with remarkable learning effects and rich learning content has become one of the issues of common concern in society. The goal of music teaching is to cultivate musical talents who have well musical accomplishments

and are applied to the teaching tasks of middle and primary schools. The training process is intensely technical [3,4]. Virtual Reality (VR) is a kind of simulation technology that uses computers, sensors, and various auxiliary devices to create a virtual world to reproduce the real world as much as possible. The most typical feature of VR is immersion, which affects users through sight, hearing, and touch to make them feel immersive. Using VR technology to create realistic virtual scenes and environments is often much cheaper and less difficult than building the real world. Therefore, it has a wide application prospect in many fields such as military training, experimental teaching, industrial design, architectural display, medical anatomy, and virtual games. VR technology has four basic characteristics: existence, interactivity, creativity, and multi-perception. Existence refers to the stereoscopic and verisimilitude of the 3D image scene through computer simulation. Interactivity means the natural interaction between humans and machines, operating things in the virtual environment through a mouse and keyboard or with the help of sensing devices. Creativity refers to creating new vivid physical landscapes such as the water flow in the virtual ocean world according to the laws of physical motion. Multi-perception indicates the use of visual, auditory, and tactile devices to make VR systems multi-perceptual.

At present, there are still many deficiencies in the music teaching system in schools, which can be divided into the following two aspects. 1. The use of modern teaching equipment and the corresponding teaching tools. Music classroom teaching enables students to understand music and through the audiovisual combination to learn music through sound. With the progress of technology and the country's emphasis on education, multimedia devices have gradually entered the classroom of every school. Although multimedia facilities are available in some schools, they are not used to their fullest extent. Teachers need to use multimedia equipment to promote the classroom teaching of music and stimulate students' interest in learning music by projecting pictures, videos, and recordings on multimedia. Students can learn about some performers' classic works and so on. At the same time, to enable students to get closer to and understand music, the school also needs to be equipped with corresponding teaching tools, such as piano, electronic keyboard, cucurbit flute, flute, and other musical instruments. The purpose is to enable students to learn musical instruments in music lessons, learn music theory knowledge, and improve their musical literacy. 2. Innovation of teaching mode and teaching concept. Nowadays, to effectively improve the quality of music teaching in school and make students have good music literacy through music teaching, firstly, the traditional teaching mode and teaching concept should be completely changed. Combining with the requirements of the new curriculum standards, the quality education goal is taken as the guiding direction of students' music teaching mode and concept. The "indoctrination" teaching method in the traditional teaching mode should be transformed, highlighting the main position of students in learning, actively communicating with students, and guiding them to fully realize the importance of music learning to their healthy growth.

In this context, this study developed a virtual piano with an HTC Vive kit and Leap Motion sensor fixed on the helmet as the hardware platform, and Unity3D and related SteamVR plug-in and Leap Motion plug-in as the software platform. The virtual piano is composed of a virtual keyboard with Cube components. The script response function of being approached and pressed or released by Cube components is compiled to record the performance events of the virtual piano keyboard. The sound of the virtual piano is further realized with the help of the Musical Instrument Digital Interface (MIDI). To enhance the performance, a gesture recognition algorithm is proposed and implemented using machine learning (ML) theory. Specifically, Dual Channel Convolutional Neural Network (DCCNN) is used to collect the user's gesture command data. The dual-size convolution kernel is employed to extract the feature information in the image and the gesture instruction in the video. Then the DCCNN recognizes it.

After the spatial and temporal information is extracted, color pattern images and optical flow images of the Red-Green-Blue (RGB) are fed into the DCCNN, and the prediction results are merged to get the final recognition result. The gesture recognition algorithm implemented in this study can be used not only for the property setting of virtual piano but also for other VR products, which have a wide range of promotion and application value.

## Literature review

In optimizing the music teaching system, many researchers have conducted in-depth explorations. Loveridge (2020) [5] pointed out that VR technology is a high-tech simulation technology active in many fields such as medicine, aerospace, industry, and games in recent years. VR technology can help students create conditions for lifelong learning through constantly changing novel environments and practical experiences, and improve students' learning interest and learning effect. According to the relevant domestic and international research results of the application of VR technology in music teaching, based on the actual development of VR tools, they used literature analysis, case analysis, and questionnaire to investigate and analyze the use of VR technology in violin teaching, and to explore the limitations and prospects of this technology in music teaching. Soliman et al. (2021) [6] pointed out that due to the particularity of music teaching methods (combination of sensibility and rationality), the existing computer-assisted instruction technology is difficult to be applied in music teaching. This study proposed a VR technology-based music teaching system, which was based on audio processing analysis and brings unique immersion, interactivity, and conceivability through VR technology. This system provided a new model for music teaching preparation, process, and interaction. Gong (2021) [7] proposed a VR-based music teaching and training system. The system includes an instrument audio management module, audio waveform simulation module, instrument type judgment module, and training instrument error correction module, in which the instrument audio management module is employed to transmit the audio signal of the instrument; The simulation module is used to simulate the audio waveform data of the transmitted audio signal. The third module is adopted to automatically judge the instrument type according to the simulated audio waveform. When playing and training a certain instrument, the system will collect the playing audio of the instrument, voluntarily judge the instrument type of the audio, and automatically correct the instrument's tone when playing a certain piece of music, thus improving the correction management during music training and performance. Xu et al. (2022) [8] pointed out that with the continuous reform and innovation of teaching methods in colleges, the curriculum system of students is also constantly enriched and developed. Hence, people's requirements for teaching management and system are increasing. Physical education courses are usually outdoor teaching. Some schools have not established a complete teaching system. To this end, they designed an interactive sports teaching system based on artificial intelligence. Firstly, the hardware of the intelligent sports training interactive system was designed by constructing the total control circuit of the interactive teaching system, determining the corresponding circuit address decoding, improving the audio control circuit, and associating the video with the interactive driver. Secondly, the software design of this interactive system was carried out by creating an intelligent training function module, the design of a training database, and the realization of effective training and teaching of intelligent sports. Finally, the system test verified the corresponding effect, and the relevant system was further improved. Hua et al. (2022) [9] designed an intelligent auxiliary system to enhance the input accuracy and response speed of music training. The architecture is divided into the infrastructure, data, application, and presentation layers. In hardware design, the combination of an Advanced RISC Machine (ARM) and a digital signal processor (DSP) is used to realize data

analysis, human-computer interaction (HCI), and interface interaction. In the software design, the Cepstrum algorithm is utilized to extract cepstrum features of music signals, and the linear smoothing algorithm is used for filtering. The dynamic time normalization method is adopted for pattern matching, and the radial basis function algorithm is employed to output results. Gong (2022) [10] argued that the traditional teaching mode of ethnic vocal music in colleges has such problems as low teaching quality, poor teaching diversity, and low interest of students. Based on this, they studied the innovation of a national vocal music teaching model in colleges based on a Deep Recurrent Neural Network (DRNN) algorithm. They designed a teaching quality evaluation model based on this algorithm. Data and information are collected from students' class status, vocal music test results, classroom interaction, and other aspects. Moreover, the DRNN algorithm is used for comprehensive analysis to realize the multi-analysis and objective evaluation of the whole process of different teaching modes of national vocal music in music colleges.

It can be seen that with the development of digital image technology, the gradual popularization of three-dimensional (3D) panorama is the breakthrough point, and image-based VR technology gradually stands out. 3D panorama has attracted more and more attention due to its low development cost, strong realism, and convenient and quick panorama generation. However, the independent panorama technology lacks audio, so the teaching materials combined with Unity3D will make students feel the real audio-visual world.

## VR piano design based on UNITY 3D

### The construction of a VR hardware platform

In this study, the HTC Vive VR kit is adopted as the hardware platform of the virtual piano. A complete set of HTC Vive VR devices is displayed in Fig 1, consisting of a helmet, two handles, and two positioners [11–13].

The Vive uses Lighting House indoor location technology. Two rows of Light Emitting Diodes (LEDs) inside the locator emit a scanning beam six times per second that alternately scans a location space with 15\*15 feet in horizontal and vertical directions. The HTC Vive has



**Fig 1.** HTC Vive VR devices.

<https://doi.org/10.1371/journal.pone.0285331.g001>

**Table 1. Computer configuration for virtual piano.**

Graphics Processing Unit (GPU)	NVIDIA GeForce GTX 1060
Central Processing Unit (CPU)	Intel (R) Xeon (R) E5-2630
Random Access Memory (RAM)	16.0GB
Video output	16.0GB
USB terminal	USB3.0
Operating system	Windows 7

<https://doi.org/10.1371/journal.pone.0285331.t001>

over 70 photosensitive sensors on the helmet and handles [14,15]. As the beam sweeps through, the helmet begins to count. Once the sensor receives the scanning beam, it uses the relationship between the sensor position and the time it takes to receive the laser to calculate the exact position relative to the locator. With the help of location technology, users can walk around the virtual piano within the detection range and even play the piano on the other side of the keyboard, further increasing the sense of three-dimensionality and immersion. Here, the HTC Vive helmet and Leap Motion sensor are connected to computers through High Definition Multimedia Interface (HDMI) and Universal Serial Bus (USB) interfaces, respectively [16]. Table 1 lists the configuration properties of the computer and the built-in discrete graphics card used in this study.

### The construction of the development platform of the VR piano software and the design of the virtual piano keyboard

Unity3D is a professional game engine developed by Unity Technologies, which enables game developers to easily complete various game creativity and 3D interactive development [17,18]. Unity3D has rendering resources such as physical simulation, normal mapping, screen space environment shading, dynamic shadow, etc. It has two major advantages over other game development tools: an extremely visual workflow and multi-dimensional cross-platform support. By the meaning of visual workflow, scene layout editing, resource binding, and the interactive object's scripting language coding can be easily done. Unity3D spans multiple platforms in terms of deployment targets and can run on 21 platforms, including the current Windows, Mac, Wii, iPhone, WebGL, Windows, and Android [19,20].

The piano comprises 88 black and white keys, black and white keys separated by half a note, from left to right in the bass, middle and high registers. For simplicity, seven notes, do, re, mi, fa, so, la, and si, are set respectively in the bass, middle, and high registers of the developed virtual piano, with 21 white keys and 15 black chromatic keys. Canvas [21–23] is used in the piano interface of Unity3D, which carries the region of all User Interface (UI) elements. This component is bound to the piano key object in the Hierarchy panel, as exhibited in Table 2. There are three rendering modes for Canvas: Screen Space-overlay, World Space, and Screen Space-camera. In Screen space-overlay mode, the canvas fills the entire screen space and puts all UI elements under the canvas on the top layer of the screen, which is not blocked by any

**Table 2. Rendering modes for Canvas.**

Rendering modes	Whether the canvas corresponds to the screen	Pixel correspondence	Cameras
Screen Space-Overlay	Yes	Optional	Not required
Screen Space-Camera	Yes	Optional	Need
World Space	No	Not optional	Need

<https://doi.org/10.1371/journal.pone.0285331.t002>

objects and can be seen without cameras. World Space is the world space mode. In this mode, the Canvas is located in the game scene like a 3D object, and its position and size can be set.

## MIDI-based virtual piano sound effect and MIDI document management

### Analysis of the working principle of MIDI

During the performance, the MIDI keyboard sends the information by pressing the key to the computer in real-time, and the computer records the info individually. After the music is played, the MIDI format document is kept [24–26]. During replay, the computer interprets the key information one by one from the open MIDI document. With the help of pre-recorded acoustic files of each key or physical modeling, the key sound is synthesized accurately and timely according to each key's playing time and strength, to realize the playback of the whole music. Based on the above basic working principle of MIDI, it is easy to understand that MIDI documents are equivalent to music scores and mainly consist of a series of keystroke events, called MIDI events. According to the MIDI specification, each MIDI Event consists of four parameters, the first parameter is a timestamp, that is, the time interval between the occurrence time of each event and the previous event; The second parameter is the name of the event, such as the press or release of a key; The third parameter is the key name, namely, which key is pressed or released; The fourth parameter is force, that is, the playing force of the key [27–29], which is generally expressed by speed. Table 3 presents the command format of the MIDI Event. In Table 3, the press or loosening of the key is denoted by 9x and 8x, respectively, and the key name and force are expressed by 1 byte, ranging from 1 to 128. Table 3 indicates that in addition to the Note On and Note Off events, there are many other events, such as Aftertouch. Since the notes being played change with time and force [30], using the event needs to be refined according to the length and force of the stay on the keyboard. Besides, the controller is an editor that can be used to control glissando, vibrato, and crescendo in MIDI instruments.

### Property change settings for virtual piano

In Table 3, many MIDI events can change the performance effect in addition to Note On and Note Off events. For example, changing the instrument type can change the timbre. The

**Table 3. MIDI event.**

Category		Parameter (hexadecimal)
Bytes	Implication	
8x	Note off	Note (00-7F); Loosened note; Force: 00-7F
9x	Note on	Note (00-7F); The note pressed; Force: 00-7F
Ax	Key after Touch	Note: 00-7F; Force:00-7F
Bx	Controller	Controller number: 00-7F; Controller parameters: 00-7F
Cx	Program changes	Instrument No.: 00-7F
Dx	Aftertouch	Status: 0xA0-0xAF; Data: nnff
Ex	Portamento	Pitch low: pitch mod 128; Pitch high: pitch div 128
F0	System code	Total system code bytes: length variable; System code: Does not include the beginning F0, but includes the ending F7
FF	Other formats	Type of format: 00-FF
00-7F		Parameters of the last activated format (8x, 9x, Ax, Bx, Cx, Dx, Ex)

<https://doi.org/10.1371/journal.pone.0285331.t003>

```

publicvoid Update Scroll View Items (float xValue, float yValue, float scaleValue)
if (Input.touchCount ==1 && Input.GetTouch(0). phase == TouchPhase.Moved)
float x = Input.GetTouch (0). deltaPosition.x;
iTween.ColorFrom (splatTexture, iTween.Hash ("color", newColor (1, 1,0,0)));
iTween. ScaleFrom (splatTexture, iTween. Hash ("scale", new)}
EventCournt++;
EventData1=0;
EventData2=0XC0~CF;
EventData3=int(x*128);
EventData4= 0;
midiOutShortMsg (lphmo, EventData2<<16+ EventData3<<8+ EventData4);
Event (EventCournt)= [EventData1, EventData2, EventData3, EventData4];

```

**Fig 2.** Code for property change settings.

<https://doi.org/10.1371/journal.pone.0285331.g002>

impact of the instrumental ensemble can be produced by assigning different channels of MIDI to various timbres. In this case, the channel function of MIDI must be used [31–33]. There are 16 channels in total. Additionally, changing the offset value of the keys can change the voice register; the tempo can change the playing speed, and so on. Because the mouse and keyboard cannot be used, it is even inconvenient to use the control handle in the virtual piano implemented here, so it is tough to change the property settings of the virtual piano in real-time. For this reason, this study uses gesture recognition to realize the property change settings of the virtual piano. The principle of gesture recognition will be introduced in detail in the next section. In Fig 2, a sliding bar is set on the upper left of the virtual piano keyboard. Gestures can control the sliding bar's position and color to change the piano's volume and select the instrument's timbre.

### MIDI file archiving for music performance

Each key-playing event of the virtual piano and its property-changing event are recorded in the programmed array. The data recorded in the array is saved to a hard disk file. A MIDI file header is added to the file according to the MIDI specification, which can realize the MIDI file archive of the music played by the virtual piano. MIDI files usually start with a header block followed by one or more track blocks, with header blocks marked by ASCII code MThd and track blocks marked by ASCII code MTrk. The specific composition of the header block of a MIDI file contains information such as header type, header length, and data section. The header length is a variable constant, and the highest bit of each byte is the flag bit. The data section involves format, track number, partition, and a total of 3 16-bit bytes, which need to store high; The track block consists of a series of data streams made up of MIDI messages that are actually repositories of file song content data.

Syntax and format for header blocks:

<Header type><Header length><Format><Number of tracks><Partition>

<4d546864><00000006><ffff><nnnn><dddd>

The ID of the header block and track block is 0x4D546864 and 0x4D54726B; "00000006" is the number of bytes describing the length information of the header block. Currently, in MIDI protocol, this value is fixed at 6. "ffff" indicates a track format, which is generally a synchronous multi-track format; "dddd" specifies the basic time format type, which can be of two types. Type 1 defines the tick number of a quarter note, the tick being the smallest unit of time in MIDI. Type 2 defines the number of Simple Mail Transfer Protocol (SMTPE) frames per second and the tick of each SMTPE frame.

## Gesture recognition algorithm

### Gesture command data acquisition

Solfeggio is the skill of reading music scores using visual, auditory, and other perceptual methods. Solfeggio is helpful for students to quickly master the melody of music, enhance their perception and understanding of music, and improve their performance level. Rapid reflexes from vision to singing are required in this process, which requires intensive training. Some researchers use Convolutional Neural Networks (CNNs) combined with video tracking technology to identify the object. However, how to show the dynamic rhythm of singing and make the teaching process more straightforward and effective remains an open question. Curwen gesture by Hungarian music educator John. Curwen was created and used in music teaching. Thus, applying Curwen gesture to music perception teaching can increase learning interest, improve intonation, and enrich learning content. The main research content of this study is how to use the computer to recognize gestures in music teaching and establish the HCI music perception teaching method. The Curwen gesture is demonstrated in Fig 3.

CNN is a feedforward neural network that includes convolution computation. It can perform supervised and unsupervised learning by mimicking the visual mechanism of the human eye. The convolution kernel parameters in the hidden layer have the characteristics of weight sharing and interlayer connection, which makes it possible to learn the features of grid-like

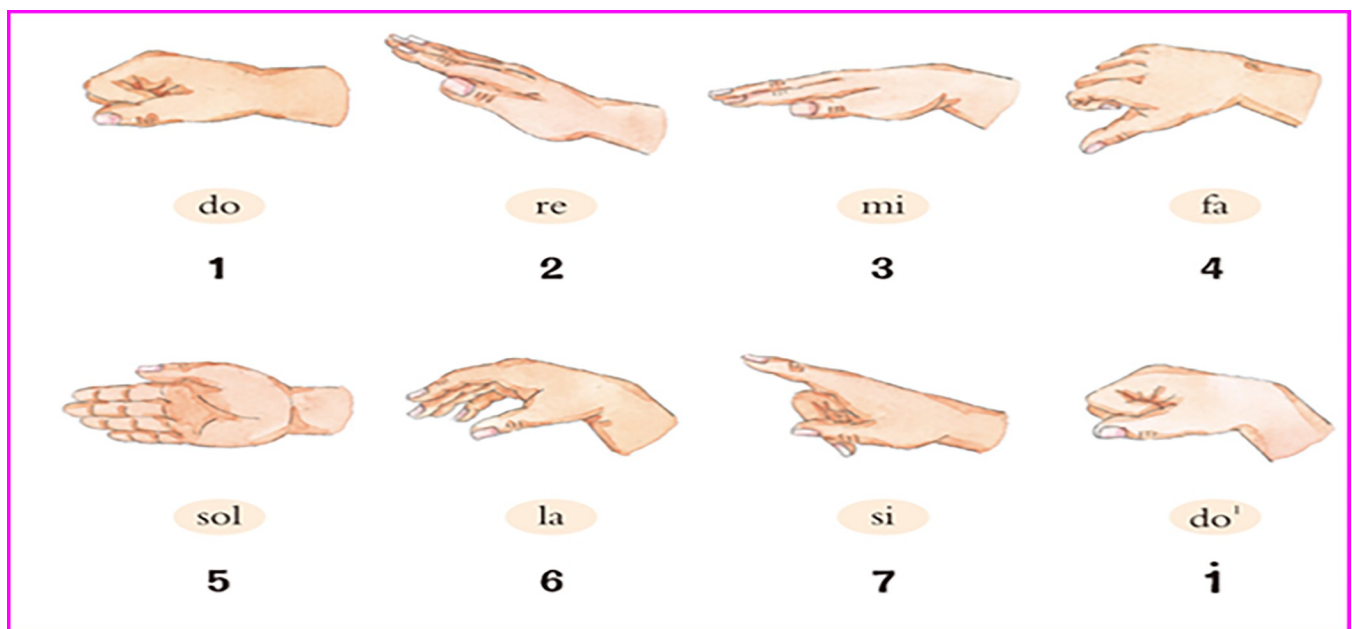
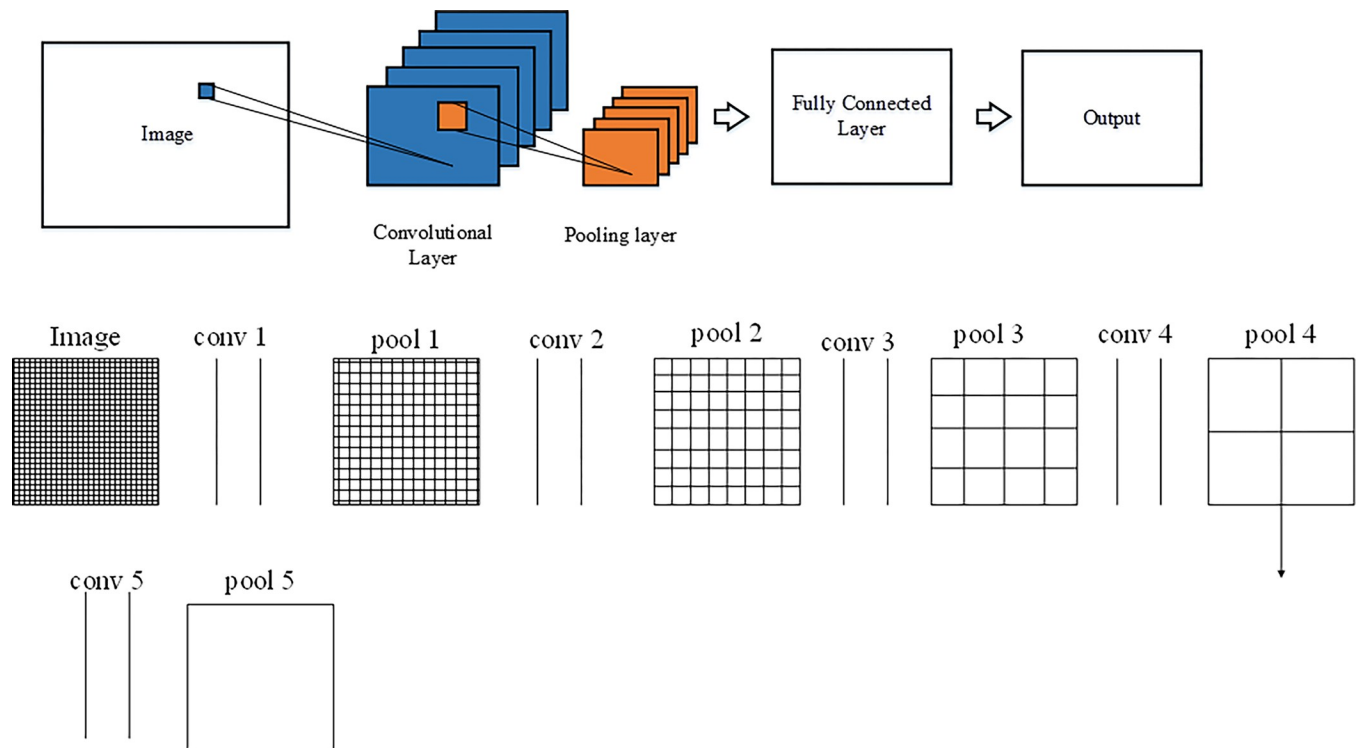


Fig 3. Curwen gesture.

<https://doi.org/10.1371/journal.pone.0285331.g003>



**Fig 4. CNN structure.**

<https://doi.org/10.1371/journal.pone.0285331.g004>

topology with reduced computation. As a result, CNN is often used to process images and speech. The structure of CNN consists of input, hidden, and output layers. The input layer can handle multidimensional data and must standardize input characteristics. The output layer can use logical or normalized functions to output the category tags. The hidden layer usually contains three common layers: convolutional layer, pooling layer, and fully connected (FC) layer, among which the convolutional and pooling layers are the unique structure of CNN. The convolutional layer contains the weight coefficient, while the pooling layer does not. The specific CNN structure is shown in Fig 4.

DCCNN is employed to recognize static gestures and extract advanced features to realize advanced recognition of multiple gestures, avoiding the preprocessing of traditional gesture recognition. DCCNN can obtain advanced features of different granularity and enhance gesture recognition.

CNN can extract the required features after training. In general, CNN has multiple convolutional and pooling layers. There are many convolution kernels in the convolutional layer to extract local features, and each convolution kernel can map a featured image. Under the same conditions, a small convolution kernel can increase the feature information, reduce the number of parameters, and improve the recognition effect. Therefore, the recognition effect of multiple small convolution kernels is higher than that of large convolution kernels. If the convolution kernels are too small, image features may not be extracted. Traditional CNN uses a fixed-size convolution kernel, so the image granularity is also fixed. Thereby, some features will be lost during the learning process, thus reducing the accuracy of network recognition. DCCNN with dual size convolution kernel has two convolution layers, two pooling layers, and one FC layer, as plotted in Fig 5. The FC layers of the two cellular neural networks are combined using FC mapping and are finally input to a classifier for feature classification.

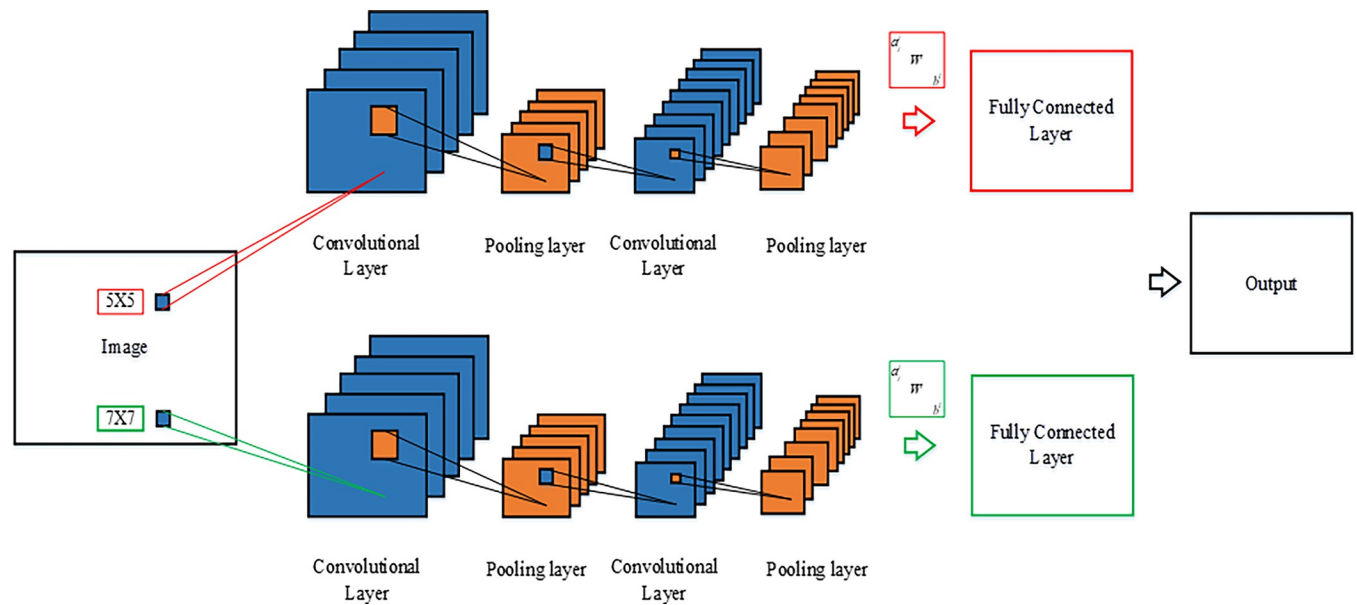


Fig 5. DCCNN structure.

<https://doi.org/10.1371/journal.pone.0285331.g005>

## Gesture command recognition network

Dynamic gestures contain more information about motion than static gestures. Therefore, how to recognize the gesture is the focus of this section. The optical flow characteristics of the video are analyzed, and the motion information of the video is identified by DCCNN. Video information is divided into spatial information and temporal information. Spatial information refers to the surface information of video frames, while temporal information refers to the optical flow between frames. Thus, spatial and temporal information is extracted from RGB and optical flow images, and DCCNN is applied to identify motion information. The two images are input into the network, respectively, and the predicted results of each network are combined to form the final recognition result. DCCNN consists of two dimensions, space and time, respectively processing spatial and temporal information. One is single-frame CNN, where videos are input via single-frame RGB images to describe the spatial characteristics of the images; Another CNN inputs the video through dense optical streams of multiple frames to describe the temporal properties of the video, combining the two networks to identify the video. DCCNN can acquire motion information of video data effectively. CNN is used to convolve, merge and fully connect spatial feature images and temporal images, and recognize video by combining spatial and temporal features.

The constructed DCCNN structure uses a convolution kernel in the convolutional layer to convolve single-frame images and multi-frame optical flow images. Its structure is portrayed in Fig 6. The final result is obtained by the weighted sum of the output.

## Results and analysis

### Static gesture command recognition result

To verify the influence of the size of the DCCNN kernel on the recognition effect, two parallel experiments are conducted using convolution kernels of different sizes, and the results are denoted in Fig 7.

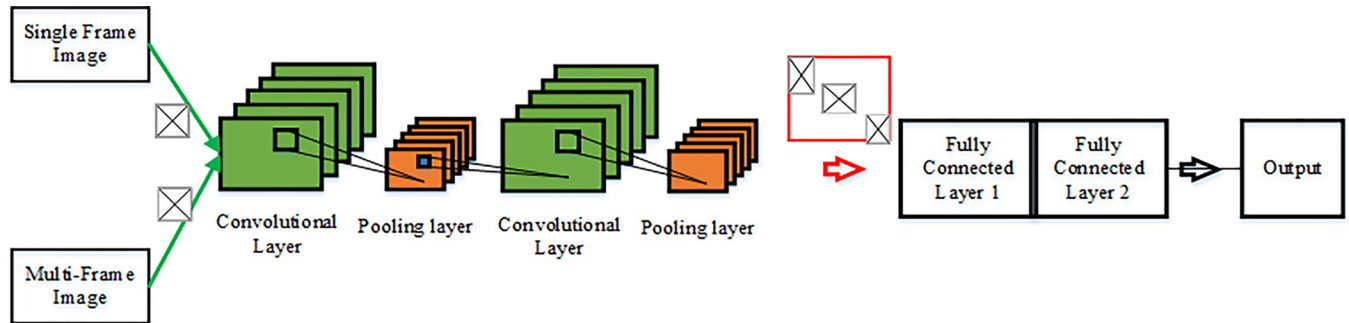


Fig 6. Early merger strategy adopted by DCCNN.

<https://doi.org/10.1371/journal.pone.0285331.g006>

Fig 7 signifies that the recognition accuracy of DCCNN can be as high as 96%, and the recognition accuracy varies with different convolution kernels. By comparison, it is found that the recognition effect of DCCNN is affected by the size of the convolution kernel. Combining convolution kernels of size 5×5 and 7×7 can improve the recognition accuracy to 98%. The experiment also shows that DCCNN can recognize static gesture images of different scales, and combining the information of these images can obtain richer feature information and better recognition results.

The learning curve of the neural network is suggested in Fig 8.

The learning curve is to display the performance of the network against new data by calculating the accuracy and cross-validation of the training set under different training set sizes, to determine whether the variance or deviation of the network is too high and whether increasing the training set can reduce the overfitting phenomenon of CNN. Fig 8 indicates that there is no overfitting phenomenon in this algorithm.

### Dynamic gesture command recognition result

The three types of cellular neural networks involved were tested using training from scratch, pre-training spatial network, and cross-input pre-training strategies, and the results are illustrated in Fig 9.

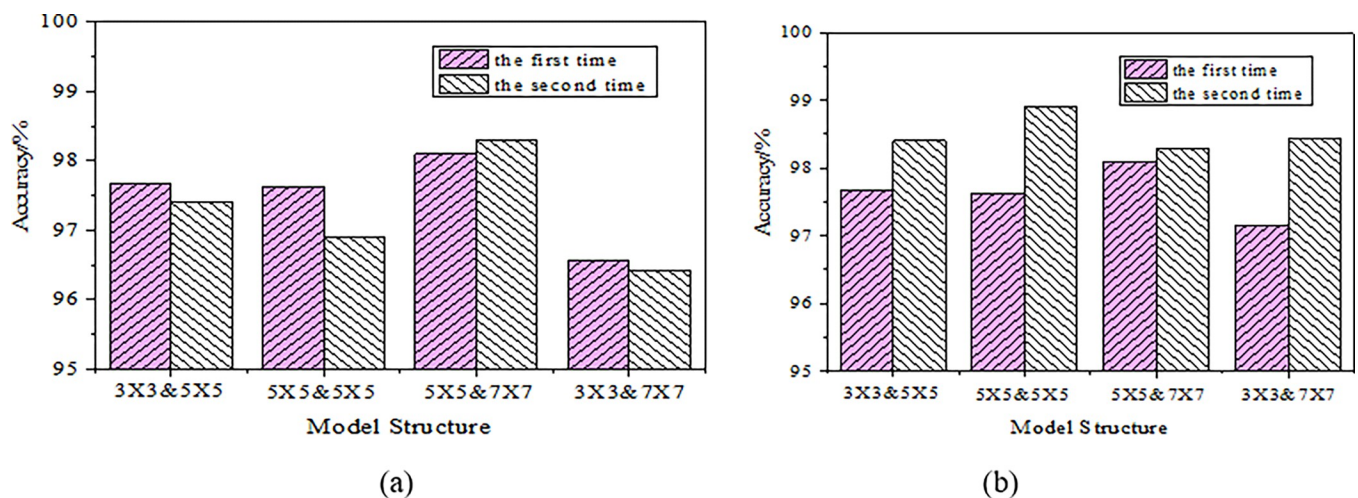


Fig 7. Recognition accuracy of cellular neural networks with different kernel sizes (a.lr = 0.01; b.lr = 0.001).

<https://doi.org/10.1371/journal.pone.0285331.g007>

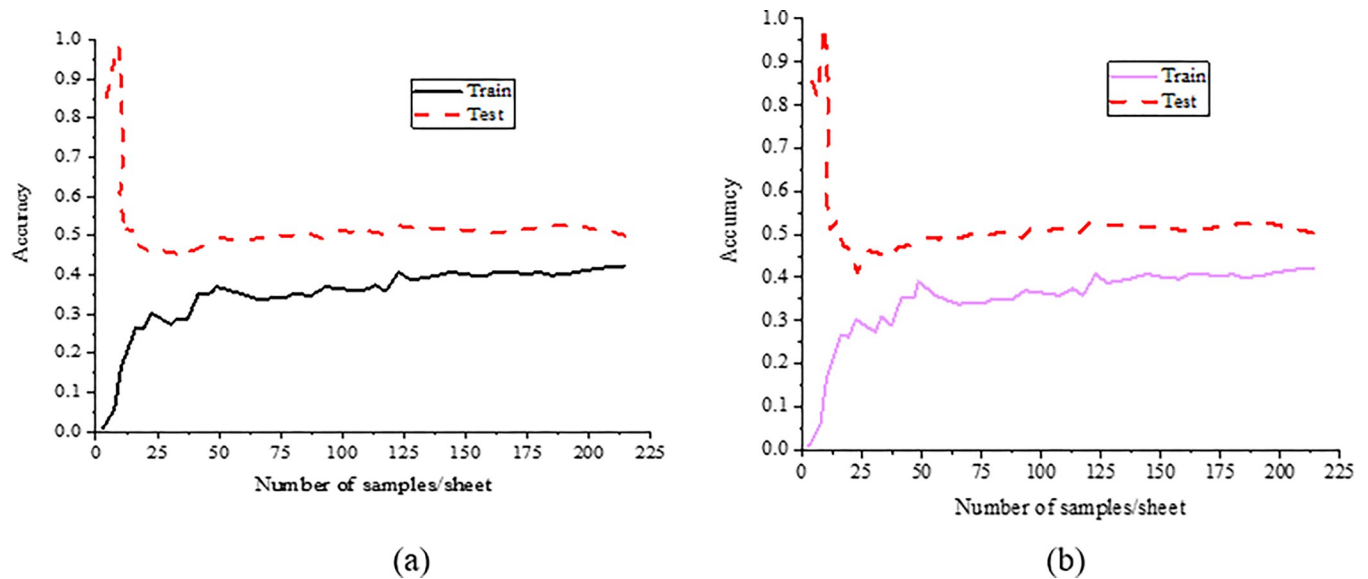


Fig 8. The learning curve (a.lr = 0.01; b.lr = 0.001).

<https://doi.org/10.1371/journal.pone.0285331.g008>

Fig 9 expresses that both spatial CNN and temporal CNN have universal effects in dynamic gesture recognition, and the accuracy rate is lower than 90%. The accuracy rate of DCCNN is about 90%, which is higher than the other two networks. In addition, if the same CNN is trained using different algorithms, the recognition results are variable. Among them, the recognition effect of cross-input pre-training is the best among the three, and the recognition accuracy is 91%. The reason may be that this training way can effectively reduce the overfitting phenomenon of CNN.

### Test results of system pressure

According to the actual demand, the constructed music teaching system based on the VR system has been subjected to two stress tests, and the test results are exhibited in Table 4.

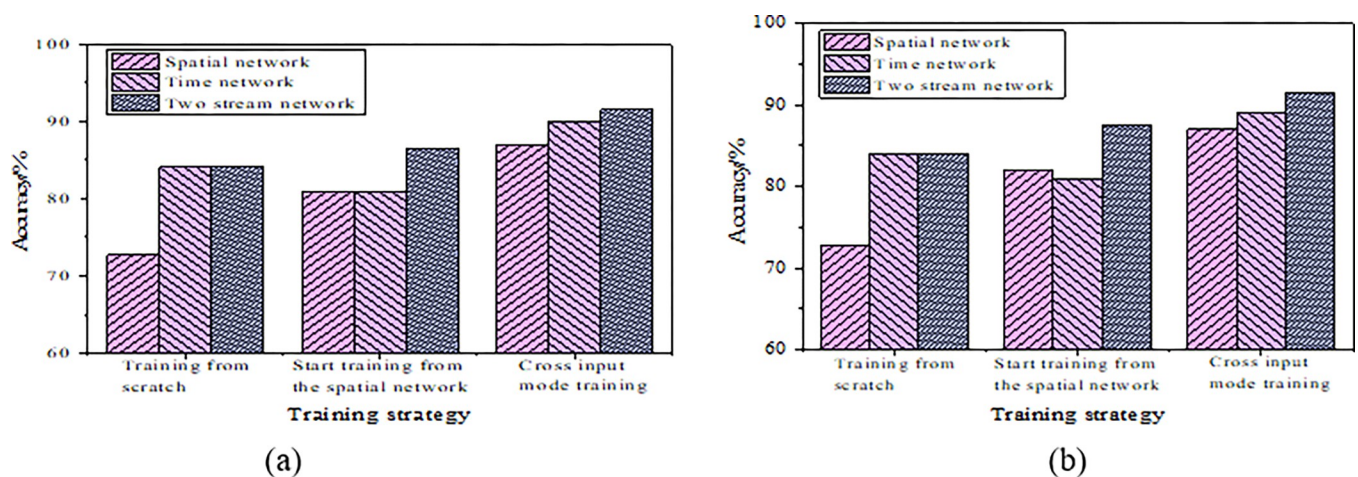


Fig 9. CNN test results under three different training methods (a.lr = 0.01; b.lr = 0.001).

<https://doi.org/10.1371/journal.pone.0285331.g009>

**Table 4. Test results of system pressure.**

Number of parallel tasks in the system (PCS)	10	20	40	60	80	100	120	150	200
Response time of the first test system(s)	1.6	1.9	1.9	2.2	3.1	3.6	3.8	4	5.5
Response time of the second test system(s)	1.8	2.1	2.1	2.1	3.3	4.1	4.3	4.5	6

<https://doi.org/10.1371/journal.pone.0285331.t004>

As can be seen from the experimental results of the system pressure test in [Table 4](#), when there are 10, 100, and 200 parallel tasks in the system, the response time of the system in the first experiment is 1.6s, and 3.6s, 5.5s, and the response time of the system in the second experiment is 1.8s, 4.1s, and 6s. From the overall trend of data change, when the number of test tasks reaches the peak, the system's response time is 6s, and no abnormal problems occur. It can be found that the designed system has an excellent anti-pressure ability.

The results of the system comparison test are outlined in [Table 5](#).

In [Table 5](#), when the number of video playback devices is 10 and 100, the playback fluency of the teaching video of the previous music teaching system is 87% and 82%, while that of the constructed VR-based music teaching system is 90% and 90%. From the overall change trend of the data, it can be seen that with the increasing number of video playback devices, the fluency of system video playback will fluctuate to a certain extent, but the overall trend is declining. Compared with the previous music teaching system, the fluency of teaching video playback of the VR-based music teaching system constructed here has been slightly improved. Thereupon, it means that the proposed system is more suitable for the daily teaching work of colleges and universities.

## Discussion

This study uses two data sets from ChaLearn Gesture Dataset (CGD)2011 and Jester for training. The recognition accuracy of the two-stream CNN proposed by Solanki and Pandey (2019) [34], the two-stream 3D Convolution (C3D) proposed by Du et al. (2019) [35], and the dual-channel algorithm proposed in this study are compared, as revealed in [Fig 10](#).

[Fig 10](#) details that after training with CGD2011 and Jester data sets, the recognition accuracy of dual-channel is 91% and 88%, respectively, higher than that of C3D and two-stream CNN. Hence, the designed algorithm has a good video recognition effect and can avoid the loss of video information in the process of video recognition.

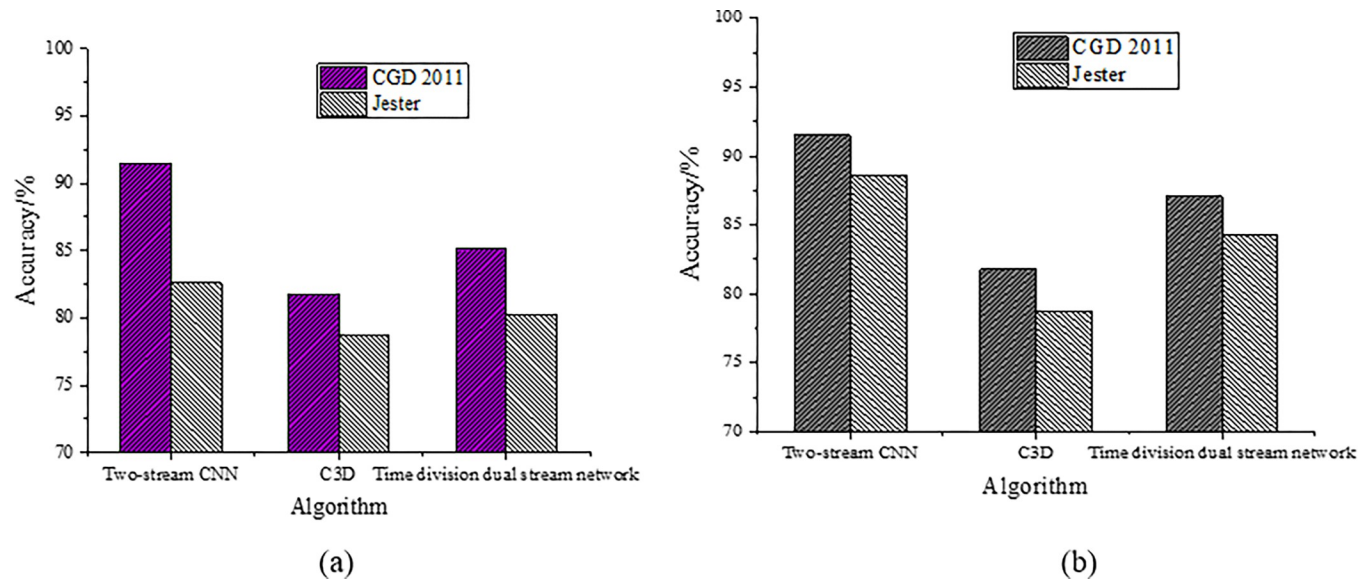
## Conclusion

With the popularity of modern computers, smartphones, and other smart terminals, VR technology will be widely used and further challenge the traditional music teaching mode. Therefore, the application of VR technology in music teaching will be more extensive, such as rhythm skills training. VR technology can simulate related music equipment and construct related equipment simulation training. This study studies the music teaching system based on a virtual simulation system. A virtual piano is developed with Unity3D and related SteamVR plug-in and Leap Motion plug-in as the software platform and HTC Vive kit and Leap Motion sensor fixed on the helmet as the hardware platform. Based on ML theory, a gesture

**Table 5. Test results for the smoothness of video playback.**

Number of video playback devices (PCS)	10	20	40	60	80	100	120	150	200
The playback fluency of the designed system (%)	90	91	92	90	92	90	94	92	91
The playback fluency of the original system (%)	87	80	81	89	80	82	86	85	81

<https://doi.org/10.1371/journal.pone.0285331.t005>



**Fig 10. Dynamic gesture recognition results (a.lr = 0.01; b.lr = 0.001).**

<https://doi.org/10.1371/journal.pone.0285331.g010>

recognition algorithm is proposed and implemented. Specifically, DCCNN is adopted to collect gesture command data of users, and a dual-size convolution kernel is applied to extract feature information in images and gesture commands in videos. The DCCNN then recognizes it. Future work will focus on the above two aspects. (1) The model's compatibility is further improved, and the code is optimized. (2) Relevant data will be further collected to construct reasonable datasets to understand the actual situation and optimize the model.

## Supporting information

### S1 Data.

(ZIP)

## Author Contributions

**Data curation:** Yan Feng.

**Methodology:** Yan Feng.

**Writing – original draft:** Yan Feng.

## References

1. Anglada-Tort M, Sanfilippo K R M. Visualizing Music Psychology: A Bibliometric Analysis of Psychology of Music, Music Perception, and Musicae Scientiae from 1973 to 2017. *Music & Science*, 2019; 2:205.
2. Burak S. Self-efficacy of pre-school and primary school pre-service teachers in musical ability and music teaching. *International INT J MUSIC EDUC*, 2019; 37(2):257–271.
3. Rickels D A, Hoffman E C III, Fredrickson W E. A comparative analysis of influences on choosing a music teaching occupation. *J HIST RES MUSIC EDU*, 2019; 67(3): 286–303.
4. Schiavio A, Biasutti M, van der Schyff D. A matter of presence: A qualitative study on teaching individual and collective music classes. *MUSIC SCI*, 2020; 24(3): 356–376.
5. Loveridge B. Networked music performance in virtual reality: current perspectives. *Journal of Network Music and Arts*, 2020; 2(1):2.

6. Soliman M, Pesyridis A, Dalaymani-Zad D. The application of virtual reality in engineering education. *J APPL ENG SCI*, 2021; 11(6):2879.
7. Gong Y. Application of virtual reality teaching method and artificial intelligence technology in digital media art creation. *ECOL INFORM*, 2021; 63:101304.
8. Xu M, Liu D A, Zhang Y. Design of Interactive Teaching System of Physical Training Based on Artificial Intelligence. *Journal of Information & Knowledge Management*, 2022; 21: 1–16
9. Hua Z, Liang Y, Tao J. Design of music training assistant system based on artificial intelligence. *ICST Transactions on Scalable Information Systems*, 2022; 1: 173450.
10. Gong X. Research on Discrete Dynamic System Modeling of Vocal Performance Teaching Platform Based on Big Data Environment. *Discrete Dynamics in Nature and Society*, 2022; 2: 34–37.
11. Vasil M, Weiss L, Powell B. Popular music pedagogies: An approach to teaching 21st-century skills. *J MUSIC TEACH EDUC*, 2019; 28(3):85–95.
12. Metić M, Svalina V. Potential use of music therapy elements in Music teaching. *Metodički ogleđi: časopis za filozofiju odgoja*, 2020, 27(1): 149–173.
13. Öztuğ E K, Saldun O. Examination of Studies Structured with Pre-School Music Teaching Using Content Analysis Method. *Near East University Online Journal of Education*, 2020; 3(2):116–128.
14. Julia J, Supriyadi T, Iswara P D. Training the Non-Specialist Music Teacher: Insights from an Indonesian Action Research Study. *Universal Journal of Educational Research*, 2020; 8(2):547–558.
15. Kim J. Implementing a Co-Teaching Model in Music Student Teaching: A Literature Review. *Applications of Research in Music Education*, 2019; 38(1):18–24.
16. Bayley J G, Waldron J. “It’s never too late”: Adult students and music learning in one online and offline convergent community music school. *INT J MUSIC EDUC*, 2020; 38(1):36–51.
17. Ren D. Application and Research on Computer Aided Technology in the New Music Teaching Method. *J ENG DESIGN*, 2020; 2:1–9.
18. Cho E. Sensitive periods for music training from a cognitive neuroscience perspective: A review of the literature with implications for teaching practice. *International Journal of Music in Early Childhood*, 2019; 14(1):17–33.
19. Palkki J, Sauerland W. Considering gender complexity in music teacher education. *J MUSIC TEACH EDUC*, 2019; 28(3):72–84.
20. Zhang J D, Susino M, McPherson G E, Schubert E. The definition of a musician in music psychology: A literature review and the six-year rule. *PSYCHOL MUSIC*, 2020; 48(3):389–409.
21. Savage P E, Loui P, Tarr B, Schachner A, Glowacki L, Mithen S. Music as a coevolved system for social bonding. *BEHAV BRAIN SCI*, 2020; 5:1–36. <https://doi.org/10.1017/S0140525X20000333> PMID: 32814608
22. Freer E, Evans P. Choosing to study music in high school: Teacher support, psychological needs satisfaction, and elective music intentions. *PSYCHOL MUSIC Psychology of Music*, 2019; 47(6):781–799.
23. Ros-Morente A, Oriola-Requena S, Gustems-Carnicer J, Gustems-Carnicer J, Guiu G F. Beyond music: Emotional skills and its development in young adults in choirs and bands. *INT J MUSIC EDUC*, 2019; 37(4):536–546.
24. Kellems R O, Cacciatore G, Osborne K. Using an augmented reality–based teaching strategy to teach mathematics to secondary students with disabilities. *CAREER DEV TRANSIT E*, 2019; 42(4):253–258.
25. Bertka C M, Pobiner B, Beardsley P. Acknowledging students’ concerns about evolution: a proactive teaching strategy. *Evolution: Education and Outreach*, 2019; 12(1):3.
26. Vasil M. Nonformal Teaching and Informal Learning: Popular Music Education and Orff Schulwerk. *The Bloomsbury Handbook of Popular Music Education: Perspectives and Practices*, 2019; 4:249.
27. Zheng H T, Chen J Y, Liang N. A deep temporal neural music recommendation model utilizing music and user metadata. *Applied Sciences*, 2019; 9(4):703.
28. Schedl M. Deep Learning in Music Recommendation Systems. *FRONT APPL MATH STAT*, 2019, 5: 44.
29. Chen X. The application of neural network with convolution algorithm in Western music recommendation practice. *J AMB INTEL HUM COMP*, 2020; 4:1–11.
30. Gunawan A A S, Suhartono D. Music Recommender System Based on Genre using Convolutional Recurrent Neural Networks. *Procedia Computer Science*, 2019; 157:99–109.
31. Wang R, Ma X, Jiang C, Ye Y, Zhang Y. Heterogeneous information network-based music recommendation system in mobile networks. *J COMPUT NETW COMMUN*, 2020; 150:429–437.
32. Zhang M, Yang Z. GACoforRec: Session-based graph convolutional neural networks recommendation model. *IEEE ACCESS*, 2019; 7:114077–114085.

33. Wen X. Using deep learning approach and IoT architecture to build the intelligent music recommendation system. *SOFT COMPUT*, 2020; 7:1–10.
34. Solanki A, Pandey S. Music instrument recognition using deep convolutional neural networks. *INT J INF TECH DECIS*, 2019; 30:1–10.
35. Du X, He X, Yuan F, Tang J, Chua T S. Modeling Embedding Dimension Correlations via Convolutional Neural Collaborative Filtering. *ACM TRANS MANAG INF*, 2019; 37(4):1–22.