J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 2529--2548

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

Combinatorial Press

Research on Virtual Scene Creation and Interactive Learning Model for Music Performance Combining Virtual Reality and Computer Vision Technology

Jingjin Zhang[1,✉]

[1] *Luoyang Weishusheng Middle School, Luoyang, Henan, 471000, China*

ABSTRACT

With the development of virtual reality and computer vision technology, the demand for virtual scenes of music performances is becoming more and more prosperous, which brings new development opportunities for music performances and music teaching. In this paper, we use the beam leveling method to determine the camera parameters in the virtual scene, implement the calibration process and parameter solving for the camera, and implement the segmentation process for the virtual scene image through the GrabCut algorithm, formulate the model constraints and objective function, construct a virtual scene for music performance, and design a virtual scene system for music performance. Based on the virtual scene of music performance, the interactive learning model of music is proposed, and the virtual roaming mode is formulated by combining human-computer interaction technology to realize the interactive learning roaming of music learners in the virtual scene of music performance. The PSNR and SSIM values of the music performance virtual scene constructed by this paper's technology are 25.8291db and 0.9396 respectively, which are higher than those of the virtual scene construction algorithms such as VSRS and JTDI as a comparison. Carrying out music teaching experiments, the experimental class that applies the interactive learning model of this paper for music interactive learning roaming has higher mean values of all dimensions than the control class in both music learning ability and music listening ability, showing significant differences ($P<0.05$).

Keywords: virtual reality technology, beam leveling method, GrabCut algorithm, virtual roaming, virtual scene

1. Introduction

In recent years, with the rapid development of virtual reality technology, it has gradually penetrated into various fields, including dance and music performance. Virtual reality technology provides a new way of artistic expression, bringing dancers and musicians a richer creative space and viewing

✉ Corresponding author.
E-mail address: zhangjingjinyx@126.com (J. Zhang).
Received 30 May 2024; Revised 15 August 2024; Accepted 10 January 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-143

experience [1-4].

The combination of virtual reality (VR) technology and musical performances brings a more immersive and realistic experience to the audience. Through the use of 3D audio, interactive scenes and realistic visual effects, realistic virtual reality music performance creates an artistic space for the audience to participate in. However, it is not easy to realize realistic virtual reality music performance [5-8]. Virtual reality music performance needs to use 3D audio technology to simulate the real sound field, realistic audio experience will make the audience feel the effect of immersive, enhance their participation in the performance [9-12]. To create a realistic virtual scene is the key to realizing realistic virtual reality music performance, in virtual reality music performance, the audience can enter a virtual music world through a head-mounted display or projector, the degree of realism of the visualization effect directly affects the audience's experience of the music performance [13-16]. An important factor in realizing realistic virtual reality music performances is the interactive experience. In traditional music performances, the audience just sits on the seat and listens to the music, and cannot directly participate in it [17-20]. In contrast, in a realistic virtual reality music performance, the audience can interact with the music through controllers, gesture recognition, and even body movements. This interactive experience transforms the audience from passive listeners to active participants, deepening their immersion in the music performance [21-24].

This paper proposes a method to create a virtual scene for music performance, calibrates the camera in the virtual scene for music performance, acquires the scene image, and uses image segmentation method to segment the image target. Based on the image target segmentation results, combined with the user's action parameters, model constraints and objective functions are formulated to realize the construction of the music performance virtual scene. On the basis of the music performance virtual scene construction technology, the music performance virtual scene construction system is built, including the data layer, logic layer and display layer to complete the system architecture design, and the MongoDB database based on distributed file storage is utilized for data management. The functional design of the system is based on user interaction, scene generation, scene rendering, data maintenance and other functional modules. In order to realize interactive learning in the virtual scene of music performance, interactive learning roaming technology is proposed on the basis of human-computer interaction technology, and an interactive learning model based on the virtual scene of music performance is constructed. In the artificial roaming mode based on scene interaction, the input devices and interactive tools are used to change the position of the viewpoint and the direction of the line of sight to achieve the interactive learning roaming purpose. In the automatic roaming approach based on preset paths, the roaming path viewpoints are first set in the scene to roam along the paths in the virtual scene of music performance. Finally, students of S Conservatory of Music are randomly selected as the research subjects to carry out the interactive learning experiment of music performance virtual scene to explore the practical utility of the interactive learning model in music teaching in this paper.

2. Virtual set construction system for music performances

In recent years, with the rapid development of virtual reality and computer vision technology, the construction of virtual scenes has been widely used in music performance and music teaching, playing an increasingly important role. The construction of virtual scenes for music performance will break the limitation of time and space, enrich the music performance mode and provide online music performance practice platform for music learners, and promote the profound change of music classroom teaching and learning environment. In this chapter, we will build a music performance virtual scene construction system, propose the specific construction method of the music performance virtual scene, and determine the execution process of the system functions.

2.1. Music performance virtual scene construction technology

2.1.1. Camera calibration. Implement the calibration process for the camera within the virtual interaction scene.

1) Determining camera parameters. The beam leveling method is essentially a method to minimize the error between the observed and estimated image point values. The beam leveling of the virtual interactive scene space is calculated by setting m virtual interactive scene space and n spatial point-projected images, as shown in the following equation [25]:

$$\min_{p_j,q_i} = \sum_{i=1}^{m}\sum_{j=1}^{n} d(O(p_j,q_i),g_{ij})^2 \tag{1}$$

where g_{ij} is the spatial point projection from the space i to the image j in the scene, p_j is the projection matrix of the image, q_i is the spatial point matrix, $O(p_j,q_i)$ is the projection function, and d is the Euclidean distance between the observation point g and the estimation point u.

The projection matrix determined above is integrated with the spatial point matrix in a parameter set and the spatial point projection is integrated into the observation coefficients as shown in the following equation:

$$\begin{cases} \alpha = \{p_1^T,\cdots,p_n^T,\cdots,q_1^T,\cdots,p_n^T\}^T \\ G = \{p_{11}^T,\cdots,p_{nm}^T\} \end{cases} \tag{2}$$

where p_{ij}^T is the image coordinate at the spatial point i in the image plane coordinates in the virtual scene, G is the observation coefficient, and α is the parameter set.

Based on the above integration results, the initial parameters of the camera are set to α_0, the observation value vector of the camera is estimated $\hat{G}$, and then the beam leveling of the camera is transformed into the minimized Mahalanobis distance using the principle of least squares, and finally the final parameters of the camera as well as the ground coordinates are determined based on the LM algorithm as shown in the following equation:

$$\begin{cases} \hat{G} = f(\alpha) \\ \phi = G - \hat{G} \\ J^T \sum p^{-1} J\beta = J^T \sum p^{-1} J\phi \end{cases} \tag{3}$$

where J^T is the Jacobi matrix and β is the search step for the optimal parameter α.

2) Parameter solving. The parameter solving of the camera is divided into two parts: obtaining the image vanishing point and solving.

(1) Obtaining image vanishing point. Based on the nature of the vanishing point in the interactive image, combine the determined camera parameters to complete the calibration of the camera in the reconstructed scene. Set two vanishing points in the interactive scene image, the main vanishing point is the center of the image, and the image is a parallel line in the projective space, so as to obtain the coordinates of the vanishing point of the image.

(2) Parameter solving. Based on the image vanishing point determined above, set the coordinate system of the camera to $\{x,y,r\}$, and the position of the vanishing point in the camera coordinates is labeled as $\{x_1,y_1,r\}$ and $\{x_2,y_2,r|$, and use the camera calibration method of the vanishing point to determine the focal length of the camera, as shown in the following equation:

$$r = \sqrt{|(x_1-h_0)(x_2-h_0)+(y_1-l_0)(y_2-l_0)|} \tag{4}$$

Where $\{h,l\}$ is the pixel coordinates of the image and r is the camera focal length. Meanwhile, the translation vector of the image is set to η and the scale factor is μ to realize the solution of the camera parameters as shown in the following equation:

$$\mu[h,l,1] = K[Z,\eta](0,0,0,1) = K\eta \tag{5}$$

Where K is a constant.

Based on the above calculation results, the camera parameters are clarified, the calibration of the camera is completed, and the interactive scene-related scene images are acquired.

2.1.2. Scene Reconstruction. When the virtual scene of music performance is constructed, the target object in the scene needs to be decomposed into a number of independent semantic parts, and the corresponding primitives are selected to fit the contours of the parts, and the model construction of the 3D interactive scene is completed through the fitting results.

Set the boundary term of the scene image as $w(\varphi,e)$, and use the GrabCut algorithm to implement segmentation processing on the scene image, as shown in the following equation [26]:

$$M(\gamma,\iota,\lambda,e) = U(\gamma,\iota,\lambda,e) + w(\varphi,e) \tag{6}$$

Where, $M(\gamma,\iota,\lambda,e)$ is the image target segmentation result, where, ι is the number of Gaussian components, γ is the image opacity, λ is the camera parameter, e is the image data, $U(\gamma,\iota,\lambda,e)$, is each region in the scene image.

After the image target segmentation, the target in the scene can be decomposed into a number of semantically independent parts, and the primitive parameters in the scene are divided into two parts: the world coordinate pose as well as the actual shape of the primitive, and the primitive of the scene image is set to be ζ, which establishes the selection conditions of the primitive parameters, as shown in the following equation:

$$\zeta = \sigma(\zeta) \tag{7}$$

Where σ is the primitive constraint factor. Based on the constraints constructed above, the objective function of the 3D scene construction model is formulated by combining the augmented Lagrangian algorithm, and the results are shown in the following equation [27]:

$$f(x) = L_{aug}(x;\kappa,\psi)\sum_{i=1}^{n}\psi_i \tag{8}$$

$$c_{hsi}(x) + \frac{\kappa}{2}\sum_{i=1}^{n}|_{hsi}(x)|^2 \tag{9}$$

Where $f(x)$ is the constructed objective function, $L_{aug}(x;\kappa,\psi)$ is the Lagrange operator, and $c_{hsi}(x)$ is the penalty coefficient.

After determining the primitives based on the above calculation results, the target part contour in the scene image is set to $D = (d_1, d_2, \cdots, d_o)$. Based on the solution results of the camera, the selected primitives are mapped to the world coordinates, and the 3D reconstruction model of the part is constructed through the calculation of the projection contour as shown in the following equation:

$$\begin{cases} \min\rho = \sum_{i=1}^{n}\varpi_2(d_i,\theta(\zeta,N))^2 Q \\ s.t.\sigma(\zeta) = 0 \end{cases} \tag{10}$$

Where, Q is the user's virtual scene response state, $\sigma(\zeta)$ is the constraints, θ is the projection operator of primitives, $\theta(\zeta,N)$ is the projection contour, ϖ is the Euclidean distance in the image space, d_i is the sampling point interval of the part contour, and $\min\rho$ is the constructed 3D reconstruction model.

Finally, the 3D interaction scene is realized by the constructed 3D interaction scene reconstruction model.

2.2. Virtual scene construction system design

In this section, based on the above proposed virtual scene construction technology for music performance, the overall and details of the system realization, from the principle to the execution process of the function, are designed in detail.

2.2.1. System architecture design. The system can be split into back-end processing subsystem and front-end display subsystem. The back-end processing subsystem is mainly responsible for generating and storing 3D scene data, mainly using the nodes and APIs provided by Houdini to develop HDA assets; the front-end display subsystem is mainly developed using Houdini Engine and the existing commercial game engine Unreal Engine. the data generated by HDA is imported to Unreal Engine through Houdini Engine for scene rendering and display. The data generated by HDA is imported into Unreal Engine through Houdini Engine for scene rendering and display. In terms of data preservation, we mainly use the HOM library provided by Houdini for data export, combined with MongoDB for data persistence and version control.
1) Data layer. The data layer is mainly used to save various data in the virtual scene generated by the system, such as the height map of the basic terrain.
2) Logic layer. The logic layer is mainly used to provide functional support for the logical realization of the music performance scene generation function in response to the system's demand analysis and operation.
3) Display layer. The display layer is mainly combined with the graphics engine to realize the interaction between the system and the user. It mainly renders the data generated by the system and roams the virtual scene; at the same time, it provides interfaces for the operation of the system parameters, so that users can adjust the data parameters and view the generation effect.

2.2.2. System database design. In order to realize the persistence of the data and the version control of the generated effects, this system is designed to introduce a database to realize this function. However, due to the huge amount of data stored in this system and the 3D structured model, the conventional relational database is not able to deal with the data management requirements of this system very well. Therefore, this system is designed to use a MongoDB database based on distributed file storage for data management. The database consists of a series of collections, which correspond to a series of files. In contrast to the traditional relational database terminology, the collection here corresponds to the data table, and the file in the collection corresponds to a row of data in the data table.

2.2.3. System functional design. In this section, the system will be divided into functional modules first, and the design will be based on each functional module.
1) User Interaction Module. The user interaction module mainly provides the function of information interaction between the system and the user, which is mainly combined with the graphics engine. It includes the user input sub-module for adjusting the system generation parameters and the data export sub-module for the persistent storage of system data.
2) Scene Generation Module. Music performance venue scene generation module is the cornerstone of other generation modules, other modules need to be generated based on the basic scene data, is the first step in the generation of the entire virtual scene.
3) Scene Rendering Module. Scene rendering function is mainly to provide the generation of 3D rendering results, while providing scene roaming, to facilitate the user to visualize the scene and secondary development.
4) Data Maintenance Module. The data maintenance function facilitates the user to save the generated results, as well as the convenience of subsequent secondary iterative development.

3. Experiments in the construction of virtual scenarios for musical performances

In this chapter, we will examine the virtual scene construction technology for music performance proposed in this paper, discuss its imaging quality performance in virtual scene construction, and analyze the performance of the virtual scene construction system for music performance in this paper.

3.1. Quality analysis of the construction of virtual scenes for music performances

This experiment uses a dataset from a private dataset obtained through 3DsMax simulation modeling to examine the virtual scene construction effect of this paper's virtual scene construction technique for musical performances through objective imaging quality. The objective imaging quality analysis is mainly evaluated in terms of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). This paper's technology and VSRS, JTDI, PatchMatch, Criminisi and other virtual scene construction algorithms of the peak signal-to-noise ratio of the structure similarity of the comparison, in order to realize the quantitative analysis, the objective evaluation of the index comparison of specific as shown in Table 1. From the comparison results, it can be seen that when the scene structure changes greatly, the PSNR value and SSIM value of this paper's technology are 25.8291db and 0.9396, respectively, which are the highest among all the algorithms, and show the optimal quality of the construction of the virtual scene of music performance.

Table 1. PSNR and SSIM

Algorithm	PSNR(db)	SSIM
VSRS	22.1562	0.8357
JIDI	22.8352	0.8603
PatchMatch	23.2668	0.8954
Criminisi	23.8222	0.8707
Technology of this article	25.8291	0.9396

3.2. Performance Analysis of Virtual Scene Building System for Music Performance

The music performance virtual scene construction system designed in this paper is recorded as experimental group A, and the two groups of traditional virtual reality scene display systems are recorded as experimental group B and experimental group C. Performance comparison experiments are carried out to compare the music performance virtual scene construction time, construction speed, and in-texture display deviation of the three groups of systems.

3.2.1. Comparison of construction time. The average time consumed by Experiment A, B, and C groups to perform virtual scene construction for music performance is shown in Table 2. As can be seen from the table, the average value of the construction time of the virtual scene of music performance in the experimental group A, i.e., the system of this paper, is 0.5487s, while the average construction time of the experimental group B and the experimental group C is 0.8571s and 0.9493s. The average construction time of this paper's system is shortened by 0.3084s and 0.4006s compared to Experiment B and Experiment C. Obviously, this paper's system has a lower average time consumed in the construction of the virtual scene of the musical performance.

Table 2. Average time of virtual scene construction

Number of experiments	Time of virtual scene construction(s)		
	Experimental group A	Experimental group B	Experimental group C
2	0.591	0.819	0.922
4	0.533	0.887	0.968
6	0.539	0.892	0.973
8	0.529	0.854	0.917
10	0.592	0.855	0.928
12	0.534	0.875	0.932
14	0.527	0.816	0.974
16	0.579	0.862	0.958
18	0.534	0.856	0.94
20	0.529	0.855	0.981
Average	0.5487	0.8571	0.9493

3.2.2. Comparison of construction speed. Calculate the speed of experiment A, B, C group to build the virtual scene of music performance multiple experiments to take the average value, the experimental comparison results are shown in Table 3. As can be seen from the table, the average construction speed of the virtual scene in Experiment C group is the lowest, which is only 0.8971 frames·s-1, followed by Experiment B group, where the average construction speed of the virtual scene reaches 1.1185 frames·s-1. The average construction speed of the virtual scene in Experiment A group was 0.5071 frames·s-1 and 0.7285 frames·s-1 more than that in Experiment B and C groups, respectively, reaching 1.6256 frames·s-1, which possessed the fastest construction speed of the virtual scene of music performance.

Table 3. Average speed of virtual scene construction

Number of experiments	Speed of virtual scene construction(frame·s-1)		
	Experimental group A	Experimental group B	Experimental group C
2	1.632	1.174	0.978
4	1.565	1.094	0.927
6	1.596	1.073	0.871
8	1.694	1.044	0.826
10	1.701	1.188	0.853
12	1.582	1.185	0.925
14	1.589	1.12	0.932
16	1.687	1.059	0.87
18	1.669	1.13	0.822
20	1.541	1.118	0.967
Average	1.6256	1.1185	0.8971

3.2.3. Texture presentation bias. The contour textures within the constructed virtual scene of music performance are numbered, and the texture distances of the virtual reality scene and the actual scene are measured respectively to compare the texture deviations of the scene displays of the three groups of systems, and the results of the experimental comparison are shown in Fig. 1. From the figure, it can be seen that the average value of the texture deviation of the virtual scene display constructed by experimental group B and experimental group C is 1.84mm and 5.42mm respectively, while the average value of the texture deviation of the virtual scene display constructed by experimental group A, i.e., the system of this paper, is 6.78mm, which is lower than that of the experimental group B and experimental group C by 4.94mm and 1.36mm respectively, and has a better performance in terms of accuracy.

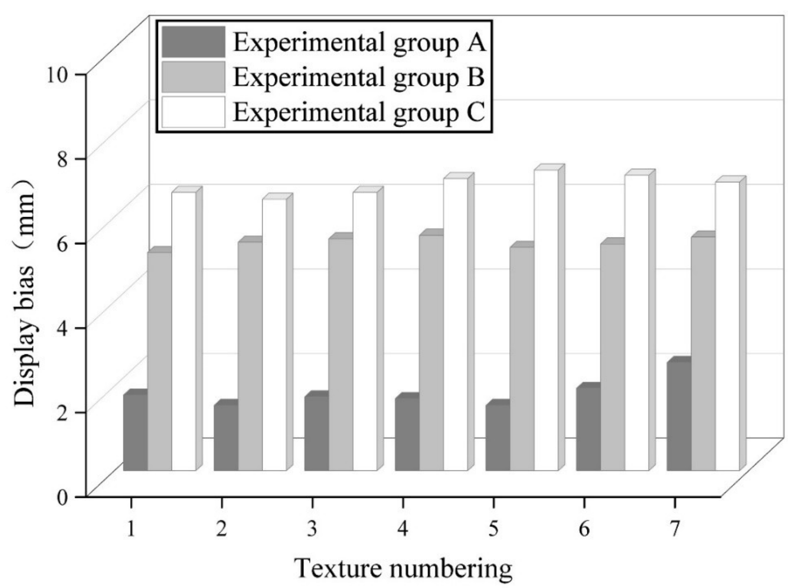


Fig. 1. Comparison results of deviation of the texture display

Overall, the music performance virtual scene construction system designed in this paper reduces the deviation of texture display, improves the accuracy of virtual reality scene display, shortens the construction time of the music performance virtual scene, improves the speed of scene construction, and ensures the construction efficiency of the music performance virtual scene compared with the traditional system.

4. Interactive learning model based on virtual scenarios of musical performances

In the above article, this paper designs a music performance virtual scene construction system to complete the construction of a three-dimensional virtual scene for music performance. How to realize interactive learning in the virtual scene of music performance is an important way to ensure the effective communication between music learners and the virtual scene of music performance, and to guarantee the effect of the virtual scene of music performance. In this chapter, we will construct an interactive learning model on the basis of the music performance virtual scene, and start from the virtual scene roaming technology to explore the methods of realizing interactive learning in the music performance virtual scene.

4.1. Human-computer interaction technology

People feel the problem of unnatural interaction in the process of interacting with virtual environments, which can be precisely solved by natural language methods such as human gestures, postures, speech, and movements.

4.1.1. Gesture Interaction. Human gesture as one of the main ways to convey information in daily life, gesture recognition has been one of the research hotspots in the field of human-computer interaction, which builds a more convenient communication bridge between machines and people. In the process of human-computer interaction, gesture recognition usually refers to the use of algorithms to recognize human gestures, and users can control or interact with the device by using simple gestures, so that the computer can understand human behavior. In this paper, with the help of KinectV2.0 sensor to obtain the depth image of the human body, human skeletal point binding, and ultimately the human gesture with the virtual environment in the virtual object binding, control virtual objects or with virtual objects for communication.

4.1.2. Posture interaction. In this paper, Kinect sensor is used to track the human skeleton in real time and calculate the distance and angle between each skeletal joint point to determine the human posture. The real-time acquisition of the human skeleton joints and the calculation of the distance and angle between the joints are used to determine the human posture. The Euclidean distance and cosine theorem are used to recognize specific postures, for example, when recognizing the posture of “RaiseRightHand”, which mainly involves the three joints of the right shoulder, the right elbow and the right wrist, the distances and angles between the three joints are judged. When it is within a certain range, it will be recognized successfully. Where A denotes the coordinates of the right wrist joint point, B denotes the coordinates of the right elbow joint point, C denotes the coordinates of the right shoulder joint point, a denotes the distance from the right wrist to the right elbow, b denotes the distance from the right elbow to the right shoulder, and c denotes the distance from the right wrist to the right shoulder. $\angle B$ represents the angle between a and b. When the posture is particularly standard, $\angle B$ is 90°.

The three joint points form a triangle, and the relationship between the three sides can be calculated according to equations (11) to (13):

$$a=\sqrt{[(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2]} \tag{11}$$

$$b=\sqrt{[(x_2-x_3)^2+(y_2-y_3)^2+(z_2-z_3)^2]} \tag{12}$$

$$c=\sqrt{[(x_1-x_3)^2+(y_1-y_3)^2+(z_1-z_3)^2]} \tag{13}$$

The angle of $\angle B$ can be obtained by combining the equations of the three sides. The calculation is as in equation (14):

$$\angle B=\arccos(\frac{a^2+c^2-b^2}{2ac}) \tag{14}$$

4.1.3. Voice Interaction. Speech is the most efficient and direct form of daily communication between human beings, and the use of speech as a way of interaction is more in line with human communication habits than traditional interaction methods such as mouse and keyboard. Therefore, speech recognition is also one of the important ways of human-computer interaction.

Speech recognition technology can convert the user's voice input information into text information, which can be used for intelligent question and answer, which is equivalent to the machine can also hear and understand. Up to now, automatic speech recognition technology has gradually matured, which involves acoustics, analog recognition, artificial intelligence, deep learning and other disciplines, and with the popularization of 5G technology, the application range of voice interaction is becoming more and more extensive. At present, speech recognition technology is mainly used in the field of smart home, chat robot, voice navigation, etc. In addition, the combination of speech recognition technology with machine translation, text matching and other natural language processing technology can be applied in more complex scenes.

4.1.4. Data interaction. Since the virtual learning environment is a mapping of the real space, the virtual learning environment in the process of experimental simulation or intelligent Q&A, often need to associate the data input by the user with the data generated by the system processing, including the inverted pendulum experimental data and the question answer data, through the association of the way to make the user in the parameter change, the system according to the user's input to output Feedback.

4.2. Virtual Scene Roaming Technology for Music Performance

In order to enhance the immersion of the music performance virtual scene and improve the interactivity of music learning, this section will rely on human-computer interaction technologies such as gesture interaction, posture interaction, voice interaction, data interaction, and so on, to realize the interactive learning roaming of the music learners in the music performance virtual scene, and the virtual roaming mode mainly includes two interactive learning modes, namely, manual roaming and automatic roaming [28].

4.2.1. Artificial roaming based on scene interaction. Artificial roaming refers to the user's input device to realize arbitrary roaming in the 3D scene. Through artificial roaming, the user can flexibly and accurately observe the scene in all directions, pick up virtual objects and obtain information. The process of artificial roaming is a process of continuously changing the viewpoint position or view direction and rendering the scene according to the interactive control commands. The interaction control between the user and the scene mainly consists of 10 basic actions: forward, backward, left shift, right shift, look up, look down, turn left, turn right, rise, and fall.

The three-dimensional geometric object is established in the scene coordinate system, but the picture displayed on the screen is only the projection of the three-dimensional scene in the two-dimensional imaging plane perpendicular to the direction of the line of sight under the given viewpoint and line of sight direction. In order to simulate the actual roaming process, it is necessary to establish a viewpoint coordinate system at the viewpoint, and transform the scene from the world coordinate system to the viewpoint coordinate system for processing. The viewpoint coordinate system can be determined by the viewpoint E, the target observation point F and the point G used to determine the viewpoint orientation. Let the unit vector in the direction of x_e axis of the viewpoint coordinate system be N, the unit vector in the direction of y_e axis be M, and the unit vector in the direction of z_e axis be L, then take:

$$
\begin{gathered}
L = \frac{FE}{|EF|} \\
M' = \frac{EG}{|EG|} \\
N = LM' \\
M = NL
\end{gathered}
\tag{15}
$$

1) Forward and Backward. Forward and backward are simply moving the point of view a certain distance along the direction of the line of sight (speed of travel), the direction of the line of sight remains unchanged. Let the coordinates of the point of view before the transformation be $E(E_x, E_y, E_z)$, and the coordinates of the point of view after the transformation be $E'(E'_x, E'_y, E'_z)$, then:

$$E' = E + v_2 L \quad W = L \tag{16}$$

where v_2 is the traveling speed and W is the new line-of-sight unit vector.
2) Left shift and right shift. Left shift and right shift are simply translations of the viewpoint along the x_e-axis direction of the viewpoint coordinate system, with the direction of the line of sight remaining unchanged. i.e:

$$E' = E + v_x L \quad W = L \quad (v_x \text{ is the translation amount}) \tag{17}$$

3) Looking down vs. looking up:

$$E' = E \quad W = LR(\theta) \quad R(\theta) = TR_x(\alpha)R_y(\beta)R_z(\theta)R'_x(\alpha)R'_y(\beta)T' \tag{18}$$

Looking up and looking down refers to the process of changing the scene as seen by the user with the head tilted up and the head lowered down. Looking up and looking down do not change the position of the viewpoint, but only the direction of the line of sight changes, so that the direction of the line of sight is rotated by a certain angle around the x_e axis of the viewpoint coordinate system. Set the rotation angle as θ, where T is the transformation matrix that translates the origin of the viewpoint coordinate system to the origin of the world coordinate system, $R_x(\alpha), R_y(\beta)$ is the transformation matrix that rotates the x_e-axis of the viewpoint coordinate system around the x_w-axis by an angle of α to the $x_w o z_w$-axis, and rotates around the y_w-axis by an angle of β to the z_n-axis, and $R'_x(\alpha), R'_r(\beta), T'$ is the transformation matrix that rotates around the x_n-axis and y_n-axis in the opposite direction and translates in the opposite direction, respectively.
4) Left Turn and Right Turn. Turning left and turning right refers to the process of scene change when the user turns left and turns right. Turning left and turning right does not change the position of the viewpoint, but only changes the direction of the line of sight, so that the direction of the line of sight rotates around the y_e axis of the viewpoint coordinate system by a certain angle. Let the angle of rotation be θ, then:

$$E' = E \quad W = LR(\theta) \tag{19}$$

The calculation of $R(\theta)$ is similar to $R(\theta)$ in the overhead and elevation views, except that the rotation around the viewpoint coordinate system x. Axis is replaced by rotation around the viewpoint coordinate system y_e axis.
5) Ascending and descending. Ascending and descending only increase or decrease the height value of

the viewpoint (y coordinate), while the direction of the line of sight remains unchanged. I.e:

$$E' = E + v_y M \quad W = L \quad (v_y \text{ is the change in viewpoint height}) \tag{20}$$

The above cases are the basic ways for users to interact with the scene, and in practical applications, it is often a combination of these ways, thus showing the user a broader perspective.

4.2.2. Automatic roaming based on predefined paths. Auto roaming is the process of pre-setting the roaming path so that the viewpoint travels along the roaming path, thus achieving the purpose of roaming in the 3D scene. The roaming path can be set in the scene plan by clicking the control points with the mouse.

Let the number of frames played per second of the screen be a, P_1, P_2 for the control P_1 point, P for the interpolation point between P_1, P_2, d for the flight distance, t for the flight time, v for the flight speed, and n for the number of interpolation points between P_1, P_2. Then:

$$\begin{gathered} d = \sqrt{(P_{2x} - P_{1x})^2 + (P_{2y} - P_{1y})^2 + (P_{2z} - P_{1z})^2} \\ t = d / v \quad n = ta \\ \begin{aligned} P_x &= P_{1x} + i(P_{2x} - P_{1x}) / (n-1) \\ P_y &= P_{1y} + i(P_{2y} - P_{1y}) / (n-1) \quad (i = 0, 1, \ldots, n-1) \\ P_z &= P_{1z} + i(P_{2z} - P_{1z}) / (n-1) \end{aligned} \end{gathered} \tag{21}$$

In order to make the virtual view roaming process to achieve real-time requirements, generally using pre-processing technology, the above processing process is pre-calculated and stored, and the drawing of the call for this information, you can realize the scene of the fast roaming.

4.2.3. Collision Detection. A common method for collision detection in virtual scenes is to use the wrapping box method, the basic idea of which is to use a simple pediatric body instead of a complex pediatric body. The first rough detection of the object's enclosing box, when the enclosing box intersects its enclosing geometry may intersect; when the enclosing box does not intersect its enclosing geometry must not intersect, so that a large number of impossible intersecting geometry and geometric parts can be excluded, so as to quickly find the intersecting geometric parts.

For the collision detection algorithm, it can be analyzed with the help of a depletion function:

$$T = N_r C_r + N_p C_p + N_u C_u \tag{22}$$

Where, T is the total cost of collision detection, N_v is the number of pairs of enclosing boxes involved in overlap detection, C_v is the cost for a pair of enclosing boxes to do overlap test, N_p is the number of pairs of geometric elements involved in intersection test, C_p is the cost for a pair of geometric elements to do intersection test, N_n is the number of nodes that need to be modified in the enclosing box hierarchy after the object's motion, and C_n is the cost of modifying a node.

The OBB collision detection algorithm between objects can be divided into three parts.

1) Compute an OBB enclosing box for the set of polygons representing the object. The surface of the object is first partitioned into triangular facets, and then the vertices of all the triangular facets that make up the object are counted using first-order and second-order methods to obtain the mean m and covariance matrix C, respectively, and the vertex vectors of the i rd triangle are set to be p^i, q^i and r^i, and the number of triangular facets surrounded by the enclosing box is n. The subscripts j and k represent the (x, y, z) components of the points.

Then the center of the enclosing box is located:

$$m = \frac{1}{3n} \sum_{i=1}^{n} (p^i + q^i + r^i) \tag{23}$$

covariance matrix elements:

$$C_{jk} = \frac{1}{3n}\sum_{i=1}^{n}(\overline{p_j^i p_k^i} + \overline{q_j^i q_k^i} + \overline{r_j^i r_k^i}) \quad 1 \le j,k \le 3 \tag{24}$$

where $\overline{p^i} = p^i - m, \overline{q^i} = q^i - m, \overline{r^i} = r^i - m$, each of which is a 3 × 1 matrix.

C is a real symmetric matrix that can be used as a basis by normalizing its 3 characteristic roots. Find the axial pole vertices along each axis of the basis, specify the size of the OBB based on these pole vertices in each axis upward, and use the axial direction of the basis as the direction of the OBB so that the OBB encloses the pole vertices in each axis upward, so that the constructed enclosing box is more compact with the object.

2) Compose OBB into a tree structure. There are two ways to create a hierarchical structure of an enclosing box: top-down and top-up.The OBB tree is constructed using the top-down method. First, the longest axis of the box is dissected by a plane perpendicular to the axis, and the position of the dissection is the mean position of all vertices in the box. Then the polygons in the box are classified according to which side of the plane the midpoint of the polygon is on, and meters are dissected by this method until they can no longer be divided as leaf nodes of the OBB tree.

3) OBB collision detection. The intersection test between OBBs is based on the separation axis theory. If the projections of two OBBs on an axis (not necessarily a coordinate axis) do not overlap, this axis is called the separation axis. If there is a separation axis between a pair of OBBs, it can be determined that these two OBBs do not intersect.

5. Experiments in interactive learning with virtual scenarios for music performance

This chapter will focus on the practical utility of the interactive learning model based on virtual scenes of music performances proposed in this paper, and analyze its experiential performance in real music interactive learning and its application effect in music teaching.

5.1. Analysis of the Music Interactive Learning Experience

The interactive learning model for music proposed in this paper is centered on artificial roaming based on scene interaction and automatic roaming based on preset paths to realize interactive learning with music virtual performance scenes. In this section, 12 students from S College of Music will be selected as research subjects and invited to participate in interactive learning experiments in a music performance virtual scene to compare the actual effects and interactive learning roaming experiences of the other 2 interactive learning models. The comparison models included LLCM-WIP model and PCT-MCD model. The results of the subject students' interactive learning roaming experience in the music performance virtual scene were mainly reflected in 6 indicators, corresponding to fall, collision, responsiveness, ease of use, error perception, and immersion (in order, I1~I6 are substituted). In this section, analysis of variance (ANOVA) and Tukey post hoc test were used to compare the performance of each model on different indicators and analyze the significance.

The interactive learning roaming experience of the subject students is specifically shown in Fig. 2.The ANOVA results show that there are significant differences between the models on falling metrics, error perception metrics, and ease of use metrics ($p<0.01$). The mean value of the falling indicator of this paper's model is 1.78, which is significantly lower than that of the LLCM-WIP model ($P<0.05$). During the experiment, when the virtualized body of the subject student is close to the edge of the platform, this paper's model can better and accurately control the virtualized body to move in the corresponding direction, and reduce the falling situation in the virtual scene of music performance. In terms of collision, responsiveness and immersion, the mean values of each model are small and do not show any significant difference ($P>0.05$), and the collision and latency of the avatars in each model are at an acceptable level, but the model in this paper still has the lowest latency and relatively good responsiveness and immersion. In terms of error perception and ease of use, this paper's model achieved the highest mean values of 4.51 and 4.63, respectively, and the mean values of this paper's model were significantly higher than those of the LLCM-WIP model and the PCT-MCD model in terms of error perception and ease of use (P<0.05). Overall, this paper's model provides a relatively better experience in interactive learning roaming, and to a certain extent, it can better meet the learners' needs for interactive music learning.

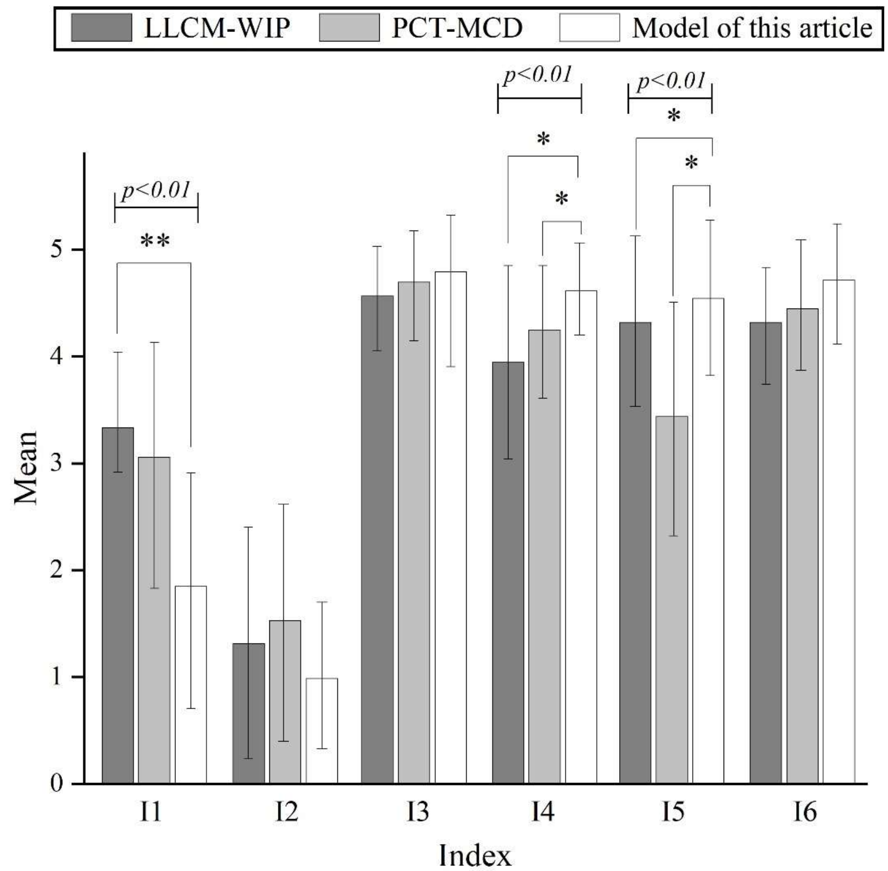


Fig. 2. User study results

5.2. *Analysis of the effect of music teaching application*

In this section, we will further explore the feasibility of the interactive learning model based on the virtual scene of music performance proposed in this paper in music teaching, as well as the impact on students' music learning after application. The location of this experimental study is still the S Conservatory of Music, and the first year class 1 and class 2 of the vocal music major of this university are selected for the control experiment, in which class 1 is the experimental class and class 2 is the control class. The number of students in both classes is 50, and the teachers, teaching materials, teaching mode, music lesson time and teaching schedule are the same, but the experimental class will have interactive learning in the virtual scene of music performance, while the control class will not use it. The duration of the experiment is 60 days.

5.2.1. Music learning ability. At the end of the experiment, the music learning ability of the students in the experimental class and the control class was measured separately, and the results of the measurement are specifically shown in Table 4. It can be seen that the significant values of the experimental class and the control class in the evaluation dimensions of music learning ability are less than 0.05, and there is a significant difference. The mean values of the dimensions in the experimental class are higher than those in the control class, and the difference in the mean values is more than 0.6 points or more, which indicates that the teaching effect of the experimental class that adopts the interactive learning model of this paper and carries out the interactive learning of the virtual scene of music performance is more prominent, and the music learning effect is remarkable.

Table 4. Ability of learning music

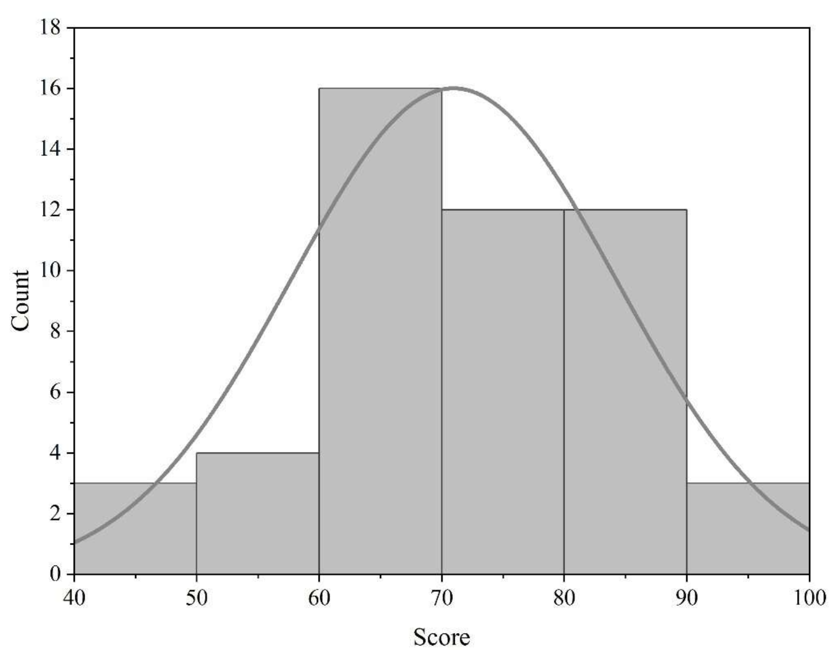
Dimensions	The difference of mean value	Standard deviation	T	P
Singing rhythm	0.839	0.779	7.389	0.000
Singing intonation	0.67	0.811	5.725	0.000
Distinguish the melody trend of the work	0.902	0.777	7.969	0.000
Distinguish the strength of the work	1.176	0.782	10.341	0.000
Distinguish the rhythm of the work	1.25	0.987	8.858	0.000
Emotional feelings of the work	0.719	0.963	5.256	0.000
Distinguish the speed of works	0.633	1.039	4.142	0.000

5.2.2. Music listening skills. After the experiment, the comparison of music listening discrimination ability between the experimental class and the control class is shown in Table 5. From the table, we can see that the mean values of the five dimensions of melody, intensity, rhythm, emotion and tempo of the experimental class are higher than those of the control group, with the difference values of 0.11, 0.16, 0.32, 0.27, 0.18, respectively, and the significant values of p-value are all less than 0.05, which proves that there is a significant difference between the experimental class and the control class, and that the experimental students are more effective in improving the effectiveness of their music learning. The students in the experimental class were more effective in the improvement of music learning.

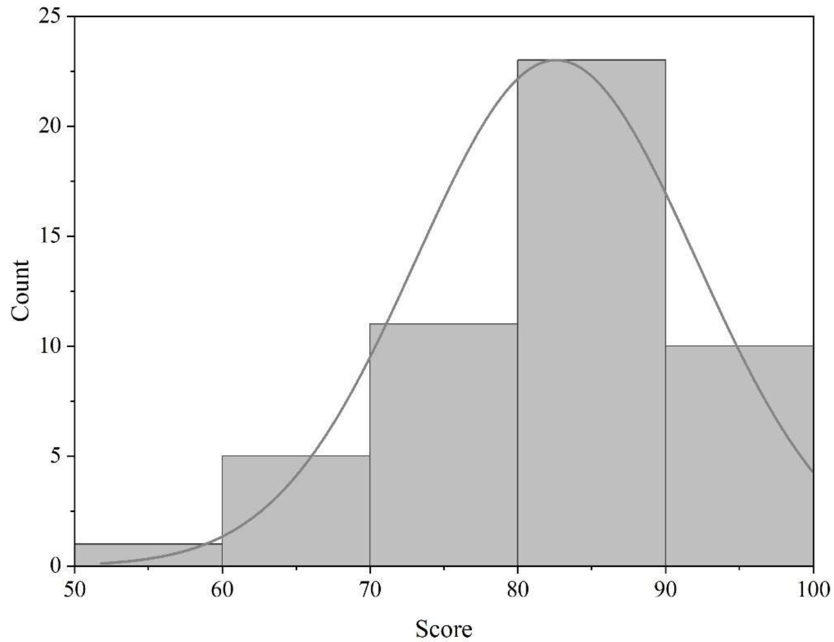
Table 5. Music Listening Ability

Dimensiona	Class	Mean	Standard deviation	T	P
Melody	Experimental class	0.52	0.507	2.018	0.045
	Control class	0.41	0.498		
Strength	Experimental class	0.57	0.478	2.644	0.01
	Control class	0.41	0.483		
Rhythm	Experimental class	0.57	0.503	2.467	0.016
	Control class	0.25	0.46		
Emotions	Experimental class	0.72	0.432	2.536	0.012
	Control class	0.45	0.509		
Speed	Experimental class	0.65	0.489	2.523	0.018
	Control class	0.47	0.505		

5.2.3. Music test scores. In order to get more objective results, students were assessed on their music performance by means of a music test. The music test scores were categorized into four bands: 0-60, 60-80, 80-90, and 80 or above. The music test is divided into four grades, with 0-50 being “failing”, 60-70 being “fair”, 70-80 being “good”, and 80 or more being “Excellent”. By organizing and analyzing the results of students in the experimental class and control class, the music scores of each class were statistically analyzed, and the statistical results are shown in Fig. 3, Figs. (a) and (b) correspond to the control class and the experimental class, respectively. It can be found that the control class has the largest number of students in the score range of 60-80 points, i.e. in the “average” level, with 28 students, while there are only 12 students in the score range of 80-90 points, and only 3 students in the score range of 90 points or above, which indicates that most of the students in the control class have a relatively poor mastery of music knowledge. The largest number of students in the experimental class is in the 80-90 score band, with 23 students, and the majority of students with “good” music test scores, while the number of students in the 90+ and 60-80 score bands is also relatively large, with 10 and 16 students respectively. There was only one student who scored 0-60 points in the music test at the “failing” level, indicating that the students in the experimental class played more stably in the music test, and that their overall level of performance was relatively even, and that they mastered their musical knowledge relatively well.



(a)Control class



(b)Experimental class

Fig. 3. The distribution of music test scores

6. Conclusion

In this paper, virtual reality technology and computer vision technology are applied to the construction of virtual scene of music performance, corresponding to the proposed construction technology of virtual scene of music performance, and build up the corresponding virtual scene construction system. Based on the virtual scene of music performance, an interactive learning model is constructed, and music interactive learning in the virtual scene of music performance is realized with the help of virtual scene roaming technology.

The music performance virtual scene construction experiments are carried out to test the virtual scene construction quality of the music performance virtual scene construction technology in this paper, as well as the performance of the virtual scene construction system. When constructing the virtual scene of music performance, its PSNR value and SSIM value are 25.8291db and 0.9396, respectively, which are higher than the comparative virtual scene construction algorithms such as VSRS, JTDI, PatchMatch, Criminisi, etc., and it has the best virtual scene construction quality. The performance of this paper's virtual scene construction system is explored from three aspects: construction time, construction speed and internal texture display deviation of the virtual scene construction for music performance. In terms of construction time, the average construction time of this paper's system is 0.5487s, which is shortened by 0.3084s and 0.4006s compared with experiment B and experiment C groups, respectively. The system in this paper also has the fastest construction speed, the construction speed reaches 1.6256 frames·s-1, and the average value of the texture deviation of the constructed virtual scene display is only 6.78 mm, which is lower than that of the experimental group B and the experimental group C, 4.94 mm and 1.36 mm. Obviously, the system in this paper has excellent performance, which ensures the efficiency of the construction of the virtual scene of the music performance.

To explore the practical utility of the interactive learning model in this paper, an interactive learning experiment on virtual scenarios for music performance was conducted. The model in this paper has significant differences ($P<0.01$) in the falling indicator, error perception indicator and ease of use indicator, in which the falling indicator achieves the lowest value with an indicator mean of 1.78, while the error perception indicator and the ease of use indicator achieve the highest value, with an indicator mean of 4.51 and 4.63, respectively. The model of this paper can provide better experience in interactive learning roaming. In the music teaching experiment, the experimental class applied this paper's interactive learning model to conduct interactive learning roaming in the virtual scene of music performance, and the mean values of all evaluation dimensions of their music learning ability were higher than those of the control class, and the difference between the mean values of the control class and those of the experimental class was more than 0.6 points or more, which was a significant difference ($P<0.05$). The experimental students' music listening ability also improved significantly, with the mean values of the five dimensions of melody, intensity, rhythm, mood and tempo being higher than those of the control class by 0.11, 0.16, 0.32, 0.27, and 0.18, respectively, and there was also a significant difference ($P<0.05$). The students' learning outcomes were further objectively examined through the music test. While the students in the control class were mainly concentrated in the score band of 60-80 with 28 students, the experimental class had the highest number of students in the score band of 80-90 with 23 students. In addition, the experimental class also has more students in the score bands of 90 and above and 60-80 than the control class, with 10 and 16 students respectively. It can be seen that the interactive learning model in this paper plays a positive role in promoting music teaching.

About the Author

Jingjin Zhang was born in 1986 in Henan Province, People's Republic of China. Obtained a Master's degree in Art from the Belarusian State University of Culture and Arts in 2010. Obtained a DFA in Musicology from Krirk University in Thailand in 2024. Currently teaching at WeiShusheng Middle School in Luoyang. My main research areas are music education and aesthetic education.

References

[1] Smith, S. (2018). Dance performance and virtual reality: an investigation of current practice and a suggested tool for analysis. International Journal of Performance Arts and Digital Media, 14(2), 199-214.

[2] Piitulainen, R., Hämäläinen, P., & Mekler, E. D. (2022, April). Vibing together: Dance experiences in social virtual reality. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-18).

[3] Loveridge, B. (2020). Networked music performance in virtual reality: current perspectives. Journal of Network Music and Arts, 2(1), 2.

[4] Bellinger, D., Wehrmann, K., Rohde, A., Schuppert, M., Störk, S., Flohr-Jost, M., ... & Erhardt-Lehmann, A. (2023). The application of virtual reality exposure versus relaxation training in music performance anxiety: a randomized controlled study. Bmc Psychiatry, 23(1), 555.

[5] Zyl, M. V. (2020). The effects of virtual reality on music performance anxiety among university-level music majors. Visions of Research in Music Education, 35(1), 15.

[6] Petrović, N. (2020, September). Augmented and virtual reality web applications for music stage performance. In 2020 55th international scientific conference on information, communication and energy systems and technologies (ICEST) (pp. 33-36). IEEE.

[7] Charron, J. P. (2017). Music audiences 3.0: Concert-goers' psychological motivations at the dawn of virtual reality. Frontiers in psychology, 8, 800.

[8] Van Kerrebroeck, B., Caruso, G., & Maes, P. J. (2021). A methodological framework for assessing social presence in music interactions in virtual reality. Frontiers in Psychology, 12, 663725.

[9] Bissonnette, J., Dubé, F., Provencher, M. D., & Moreno Sala, M. T. (2016). Evolution of music performance anxiety and quality of performance during virtual reality exposure training. Virtual Reality, 20, 71-81.

[10] Li, L., & Wang, G. (2022). Design and application of interactive music equipment based on wireless wearable sensors. Scientific Programming, 2022(1), 4719884.

[11] Loveridge, B. (2023). An overview of immersive virtual reality music experiences in online platforms. Journal of Network Music and Arts, 5(1), 5.

[12] Dang, K., & Lee, S. (2024, October). Musical Performances in Virtual Reality with Spatial and View-Dependent Audio Descriptions for Blind and Low-Vision Users. In Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (pp. 1-5).

[13] Wolf, K. E., & Fiebrink, R. (2019). Personalised interactive sonification of musical performance data. Journal on Multimodal User Interfaces, 13, 245-265.

[14] Delle Monache, S., Comanducci, L., Buccoli, M., Zanoni, M., Sarti, A., Pietrocola, E., ... & Cospito, G. (2019). A Presence‐and Performance‐Driven Framework to Investigate Interactive Networked Music Learning Scenarios. Wireless Communications and Mobile Computing, 2019(1), 4593853.

[15] Sra, M., Maes, P., Vijayaraghavan, P., & Roy, D. (2017, November). Auris: creating affective virtual spaces from music. In Proceedings of the 23rd ACM symposium on virtual reality software and technology (pp. 1-11).

[16] Gómez-Sirvent, J. L., Fernández-Sotos, A., Fernández-Caballero, A., & Fernández-Sotos, D. (2024). Assessment of music performance anxiety in a virtual auditorium through the study of ambient lighting and audience distance. Virtual Reality, 28(2), 71.

[17] Chakraborty, S., Dutta, S., & Timoney, J. (2021). The Cyborg Philharmonic: Synchronizing interactive musical performances between humans and machines. Humanities and Social Sciences Communications, 8(1), 1-9.

[18] Cao, Y., & Park, J. (2023). Application of the Interactive Art of New Media Installation in Concert Stage Performance. Harmony, 5(3), 51-55.

[19] Malloch, J., Garcia, J., Wanderley, M. M., Mackay, W. E., Beaudouin-Lafon, M., & Huot, S. (2019). A design WorkBench for interactive music systems. New Directions in Music and Human-Computer Interaction, 23-40.

[20] Sales, C., Wang, P., & Jiang, Y. (2023, August). An Interactive Tool for Exploring Score-Aligned Performances: Opportunities for Enhanced Music Engagement. In Proceedings of the 18th International Audio Mostly Conference (pp. 30-33).

[21] Bluff, A. (2017). Interactive art, immersive technology and live performance. University of Technology Sydney (Australia).

[22] Cheng, D. (2024). Exploring the Relationship Between Interactive Technology and Interactive Music Creation. Journal of Humanities, Arts and Social Science, 8(8).

[23] Canazza, S., De Poli, G., & Rodà, A. (2015). Caro 2.0: an interactive system for expressive music rendering. Advances in Human‐Computer Interaction, 2015(1), 850474.

[24] Lindetorp, H. (2019). Immersive and interactive music for everyone. In Proceedings of the Nordic Sound and Music Computing Conference (pp. 16-20). SMC Sound and Music Computing Network.

[25] Huang Huaiyang,Sun Yuxiang,Wu Jin,Jiao Jianhao,Hu Xiangcheng,Zheng Linwei... & Liu Ming. (2021). On Bundle Adjustment for Multiview Point Cloud Registration. IEEE ROBOTICS AND AUTOMATION LETTERS(4),8269-8276.

[26] Zha Jiale,Chen Huaixin,Ren Chengjie,Wang Chenggang & Li Siqi. (2020). A novel method of extracting geometric features of ships based on GrabCut algorithm. Journal of Physics: Conference Series(1),012092-.

[27] Torrealba E.M.R.,Silva J.G.,Matioli L.C.,Kolossoski O. & Santos P.S.M. (2022). Augmented Lagrangian algorithms for solving the continuous nonlinear resource allocation problem. European Journal of Operational Research(1),46-59.

[28] JunzheChen,XingShe,YuanxinFan & WenwenShao. (2024). Virtual Roaming of Cultural Heritage Based on Image Processing. Computer Animation and Virtual Worlds(6),e70000-e70000.