Available Online: 27 August 2026 DOI: 10.54254/2753-7080/2026.36382

Research on interactive electronic music creation based on MediaPipe

Nianzhi Xue, Xinyu Wang*

Zhejiang Conservatory of Music, Hangzhou, China

*Corresponding Author. Email: wxy450038600@163.com

Abstract. With the development of computer vision and real-time interaction technologies, interactive music creation based on body gesture recognition has gradually become an important direction in contemporary electronic music practice. As an open-source, lightweight computer vision framework, MediaPipe offers new technical pathways and practical possibilities for interactive music creation due to its features such as rich solution and parameter offerings, high hardware compatibility, flexibility, and extensive support from the community. This study focuses on the application of MediaPipe technology in interactive music creation. It first outlines the core technical characteristics of MediaPipe and, in comparison with currently popular gesture recognition technologies, analyzes its applicable scenarios and features in creative practice. On this basis, three interactive music works based on MediaPipe are selected as case studies. These works are analyzed from the perspectives of sound material, interaction design, and technical implementation, exploring the functional roles and expressive modes of body gesture recognition in different creative contexts. Through the analysis of these works, the study summarizes several creative characteristics of MediaPipe-based interactive music, including the body as a medium for sound generation and control, the body as a tool for visual expression, and the body as a structural and narrative element within the composition. The findings suggest that MediaPipe not only lowers the technical threshold of interactive music creation, but also encourages a transformation of the body from a functional control interface into an aesthetic and structural component of musical works. By combining technical analysis with creative practice, this study aims to provide a practical reference for computer vision–based interactive music creation and to offer methodological insights for future artistic research and development in this field.

Keywords: electronic music, interactive music, MediaPipe

1. Introduction

In fields such as interactive art, virtual reality, and multimedia performance, bodily movement has gradually become an important means of interaction connecting performers with digital media. Through technologies such as sensors and computer vision, performers' bodily actions can be transformed into digital data, which can then be used to generate and control sound, images, and other media elements.

Whether in traditional instrumental performance or live electronic music performance, bodily posture and movement have always constituted an important component of musical expression. Dalcroze argued that the body serves as a link between sound and consciousness, suggesting that musical experience derives not only from auditory perception but is also influenced by bodily movement, spatial relationships, and emotional expression. In contemporary live electronic music, sound, visual elements, and performative actions are increasingly forming a synergistic relationship, in which performers' movements can directly influence the real-time generation and presentation of a musical work.

To establish a real-time connection between bodily actions and electronic music systems, creators have continuously explored different approaches to data acquisition. From early sensor-based interactive installations to contemporary computer-vision-based motion recognition technologies, the acquisition of bodily information has gradually shifted from dedicated hardware toward more open software frameworks. After being processed by interactive systems, such data can be further mapped to sound-generation and sound-processing modules, thereby enabling real-time interaction between the body and electronic music.

MediaPipe is an open-source multimedia vision processing framework developed by Google and released in 2019 [1]. It integrates a range of machine-learning-based visual analysis models that can extract key-point information from targets such as the human body and hands in image and video data and convert this information into data parameters that can be processed by computers. These parameters can subsequently be connected, through programming, to sound-generation, visual-processing, and other systems, providing technical support for the creation of real-time interactive art. In recent years, machine-learning-based vision frameworks such as MediaPipe have gradually been introduced into electronic music and multimedia art practices and have been used to facilitate real-time interactions between bodily movements and sound and visual elements.

Although MediaPipe has been applied in a variety of fields, research on its application to electronic music composition remains relatively limited. Existing studies have focused primarily on technological development, educational applications, and interaction design, while further investigation is still needed into how MediaPipe can be integrated into the electronic music composition process, how bodily data can influence sound generation, and how the corresponding interaction mechanisms can be constructed. Therefore, based on MediaPipe as its technological foundation, this study explores its application in the creation of interactive electronic music.

1.1. Overview of technical approaches

The technical approaches currently used for bodily interaction in live electronic music can be broadly divided into two categories: contact-based and non-contact-based approaches. Contact-based technologies typically collect performers' movement data through wearable sensing devices. Non-contact-based technologies, by contrast, do not require performers to wear or carry any equipment; instead, they obtain information about human movement through spatial sensing or computer vision technologies.

Figure 1. MPU6050, a commonly used IMU sensor

Contact-based bodily interaction approaches include inertial measurement units (IMUs; see Figure 1), which are highly versatile sensing devices that continue to be widely used in live electronic music. By capturing the movement of the wrists and other parts of the body, IMUs can be used to control sound parameters. In addition to IMUs, pressure sensors are also frequently incorporated into interactive installations, in which bodily interaction is achieved through direct physical contact between the performer and the sensing surface. Non-contact approaches include relatively simple and direct technologies such as ultrasonic sensors (see Figure 2) and light-sensitive sensors. These two types of sensing technologies can detect the distance between the human body and the sensor by measuring changes in sound waves or light, thereby enabling bodily interaction. Some non-contact technologies employ infrared spatial sensing. A relatively mature example is Leap Motion (see Figure 2), which uses an infrared optical system to capture the three-dimensional coordinates of the hands, providing new possibilities for controlling sound through gestures. With the development of computer vision and machine-learning-based visual algorithms, motion recognition has gradually evolved toward more open hardware-software solutions. Kinect (see Figure 2) is a relatively mature camera-based interactive device that uses infrared sensing to capture three-dimensional information about the human body and was once widely used in interactive music. However, such devices generally rely on dedicated hardware, and their applications are consequently constrained by factors such as equipment cost and spatial requirements.

Figure 2. Ultrasonic sensor (left), leap motion (center), and azure kinect (right)

Overall, contact-based sensing technologies offer advantages such as stable data acquisition and precise control, but they require performers to wear additional equipment. Non-contact technologies based on depth vision can provide richer spatial information, but they generally depend on dedicated hardware. By comparison, MediaPipe, which combines standard cameras with machine learning models, provides a more lightweight approach to acquiring bodily movement data. Therefore, this study further explores the potential applications of MediaPipe in interactive electronic music composition.

1.2. Related research

In research on interactive music, the body is generally regarded as an important medium connecting the performer, interactive system, and sonic environment. Previous studies have pointed out that embodiment is not simply a matter of using the body to control sound; rather, it involves an ongoing process of coupling among bodily movement, data feedback, musical structure, and the performance context [2, 3]. Building on this perspective, subsequent research has further examined how performers understand interactive relationships through movement, perception, and sonic feedback, and has explored how meaningful movement–sound mappings can be established on the basis of bodily experience and movement structures [4-6].

With the development of computer vision and deep learning technologies, lightweight visual recognition frameworks such as MediaPipe have gradually been applied to interactive music and multimedia art practices. Existing studies in this area have focused primarily on individual artistic projects, technical solution design, and the implementation of interactive systems, with particular attention to technical issues such as pose data acquisition, model application, and movement–sound mapping [7]. These studies have demonstrated the feasibility of MediaPipe for bodily interaction design. However, further practical analysis is still needed regarding how MediaPipe can be integrated into the broader artistic creation process and how technological choices relate to artistic expression under different creative strategies.

Therefore, although existing research provides an important foundation for understanding the role of bodily movement in interactive music, its specific applications in electronic music composition remain open to further exploration. Based on MediaPipe's pose recognition technology and drawing on practices in interactive electronic music, this study investigates how bodily movement can participate in the generation of sound and visual elements and analyzes the characteristics of this technology in creative practice.

2. Technical characteristics of MediaPipe

2.1. Rich and accessible solutions and parameters

For works that employ pose information for interaction, an ideal pose-recognition technology should not only be capable of detecting the presence of the human body within an image, but should also transform pose information into rich, diverse, and easily accessible data parameters, thereby enabling precise control of sound or visual systems. Therefore, a mature pose-recognition framework typically needs to provide both stable and comprehensive recognition models and detailed parameters describing human body posture.

MediaPipe is designed around a series of mature, pretrained solutions. Its official framework integrates a variety of models for human pose recognition. In addition to solutions such as hand recognition (Hands), face recognition (Face), and human pose recognition (Pose), MediaPipe also provides specialized model solutions for functions such as facial landmark tracking and gesture detection. When developing an interactive system, creators can directly invoke these models without undertaking additional deep-learning training, thereby enabling real-time detection and analysis of human poses. This "ready-to-use" design allows pose-recognition technology to be directly applied to interactive music and multimedia creation without requiring users to have a specialized background in deep learning.

At the same time, MediaPipe not only detects the presence of the human body within an image but also further decomposes pose information into multiple quantifiable landmark data points. For example, in the Pose solution (see Appendix), the human skeleton is represented by 33 landmarks, with each landmark output in the form of three-dimensional coordinates (x, y, z) (see Figure 3). Based on these landmark data, creators can further develop parameters describing movement characteristics, such as hand orientation, movement trajectories, and movement speed, and map them to sound parameters or visual controls. The rich and structured API output parameters enable MediaPipe to support a wide range of interaction designs, from simple triggering mechanisms to complex forms of continuous control.

In summary, by providing mature pretrained models and structured landmark-data interfaces, MediaPipe enables creators to obtain detailed human pose information rapidly without having to train their own models and to incorporate this information into real-time interactive systems. This characteristic provides creators with greater flexibility and convenience when developing interactive music and multimedia performance works.

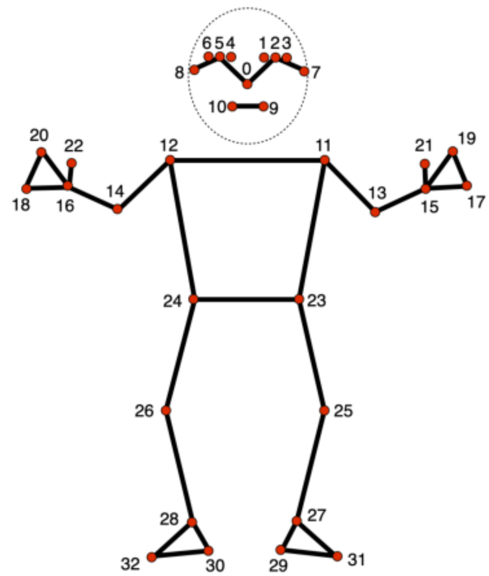


Figure 3. Distribution of keypoints in the MediaPipe pose solution, which provides 33 landmark coordinate values

2.2. Integration with OSC/MIDI technologies and the development of community resources

During interactive music performances, pose-recognition systems often need to exchange data in real time with other modules, such as sound synthesis and visual generation systems, within interactive software environments. Therefore, the ability of a pose-recognition system to integrate effectively with existing cross-platform communication protocols and multimedia development environments is an important factor in determining whether it can be widely adopted by creators.

MediaPipe can be developed and operated in general-purpose programming environments such as Python and C++, making it relatively straightforward to integrate with the Open Sound Control (OSC) and Musical Instrument Digital Interface (MIDI) protocols. In practical applications, the human body landmark data recognized by MediaPipe can be converted into OSC or MIDI messages through scripts and transmitted in real time to other music software or visual systems. With the help of these widely used communication protocols in digital music, pose-recognition data can be transmitted and shared in real time across different software platforms and devices. This enables MediaPipe to be flexibly incorporated into a wide range of electronic music systems and interactive performance environments.

In addition, the development of community resources has played an important role in the wider adoption of MediaPipe. One representative example is the range of software components developed for TouchDesigner. Through these components, creators can directly access MediaPipe's pose-recognition capabilities within TouchDesigner's visual node-based environment and connect and map the resulting landmark data to sound, image, or other control modules. This visual programming approach allows creators to implement complex interaction designs with relatively little conventional coding, thereby further lowering the technical barrier to developing pose-based interactive systems.

In summary, through integration with cross-platform communication protocols such as OSC and MIDI, as well as multimedia development platforms such as TouchDesigner, MediaPipe can establish convenient and efficient connections with a wide range of sound and visual systems. This enables pose-recognition data to flow across different devices and software environments, thereby enhancing MediaPipe's potential for application in the creation of interactive electronic music and multimedia works.

2.3. Software and hardware deployment and compatibility flexibility

In the creative practice of interactive music and multimedia performance, pose-recognition systems often need to be deployed according to different performance environments, such as computer workstations, laptops, or other mobile devices. Therefore, compatibility across software and hardware platforms and flexibility in deployment are also important factors affecting the scope of application of such systems.

Unlike many pose-recognition solutions that rely on dedicated hardware devices, MediaPipe primarily performs human pose analysis using computer vision algorithms and requires only a real-time video input to complete the recognition process. In practical applications, creators can use various types of cameras as input sources, including standard USB webcams (see Figure 4), built-in laptop cameras, and even the camera systems of mobile devices. This openness with regard to video input devices enables MediaPipe to maintain a high degree of adaptability across different hardware environments.

Figure 4. A 1080p fixed-focus USB webcam from UGREEN (priced at RMB 79), which can be used with the MediaPipe framework to acquire pose data (Image source: a Taobao customer)

At the same time, because MediaPipe primarily performs pose recognition and data processing through software, its system deployment is relatively straightforward and does not require the additional configuration of complex sensor networks or dedicated recognition devices. Creators typically need only to run the appropriate recognition scripts in a computer environment and connect the camera's video stream to the system to achieve basic pose-recognition functionality. This lightweight deployment approach enables MediaPipe-based interactive systems to be rapidly established and adjusted in different contexts, including the creative process, rehearsal sessions, and live performances.

Therefore, from the perspective of overall system architecture, MediaPipe supports flexible operation across a range of software and hardware environments through its compatibility with widely available camera devices and its relatively simple software deployment process. This platform compatibility and deployment flexibility give MediaPipe a high degree of adaptability in the creative practice of interactive electronic music.

In the following sections, the author presents two interactive electronic music works created using MediaPipe technology and examines how works of this kind can contribute to artistic innovation in contemporary practice.

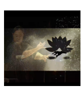
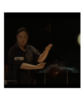
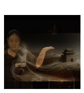
3. Analysis of MediaPipe-based works

3.1. *Flower in Hand, Shadow in Transformation* by Wan Fang and Ziqiao Zeng

Flower in Hand, Shadow in Transformation is an interactive electronic music work composed by Wan Fang, with visual design by Ziqiao Zeng. The work premiered at the closing concert of the 2025 Hangzhou International Electronic Music Festival, titled *Strings Speak, Machines Respond—2025 Hangzhou International Electronic Music Festival Closing Concert*, and has a duration of approximately 4 minutes and 30 seconds. Centered on imagery derived from Eastern aesthetics, the work uses MediaPipe to capture the performer's hand-coordinate data and map it to sound-generation and visual-presentation systems, thereby creating real-time interaction among bodily movement, electronic sound, and visual imagery.

The title *Flower in Hand, Shadow in Transformation* derives from the Zen Buddhist anecdote "Picking Up a Flower and Smiling". In this context, "picking up a flower" symbolizes the subtle communication of meaning through the hands, while "shadow in transformation" refers to the dynamic process through which visual imagery is generated and transformed in response to bodily movement. Accordingly, hand movement in the work serves not only as a means of interactive control but also as an integral component of its artistic concept. Through MediaPipe-based hand-pose recognition, the performer's bodily actions are transformed into computer-processable data, which subsequently participates in the generation of both sound and visual elements.

Table 1. Analysis of the main pose-based interactions in *Flower in Hand, Shadow in Transformation*

Section	Image	Movement Description	Main Sounds	Movement Control
A		One hand forms a lotus-like gesture.	1. Continuous pitched percussion sounds; 2. Multiple ambient sounds.	When the linear coordinate difference between the thumb and index finger approaches zero, a "lotus" visual is triggered, together with pitched percussion (bell) sounds.
B		This section is triggered when the right hand enters the frame; both arms repeatedly open and close.	Pre-recorded sound clips featuring a triple-meter rhythmic pattern, ethnic percussion, and pitched percussion.	1. When the other hand enters the camera frame, changes in the visual structure are triggered and low-frequency sounds are introduced. 2. The coordinate difference at the base of the palms controls particle-flow speed and percussion pitch.
C		The hand moves continuously within the camera frame.	Same as Section B.	The hand's y-coordinate affects the brightness of the Synth Pad background sound, primarily by controlling parameters such as the filter cutoff frequency.

The interactive design of *Flower in Hand, Shadow in Transformation* follows a structural logic of "from simplicity to complexity". The work employs progressively sophisticated forms of bodily control across its different sections, allowing the performer's movements to gradually develop from triggering individual sounds to controlling multiple dimensions of sound and visual elements (see Table 1).

Section A centers on the "flower-picking" gesture as its primary form of interaction. The system uses MediaPipe to identify the positions of hand landmarks and calculates the distance between the thumb and index finger. When the distance between the two fingers falls below a predefined threshold, the system triggers a bell-like sound while simultaneously generating a corresponding lotus visual effect. In this context, the gesture functions not merely as a signal for triggering sound but also as a connection between sound, visual imagery, and the conceptual imagery of the work, transforming the act of "picking up a flower" into an artistically symbolic gesture.

Section B further expands the dimensions of bodily control. Compared with the relatively discrete trigger-based interaction in Section A, Section B uses the spatial distance between the two hands to control the flow rate of particle-based visuals and changes in percussion pitch, transforming bodily movement from a "trigger" into a medium for continuous parameter control. By varying the distance between the hands, the performer adjusts the rate of change in both sound and visual elements, allowing bodily movement to directly participate in the development of the musical structure.

Section C continues the sound materials introduced in Section B and incorporates AIGC-generated imagery as a visual climax (see Figure 5). At this stage, changes in the height of the performer's hand within the camera's field of view affect the perceived brightness of the Synth Pad timbre, maintaining a synchronized relationship between sonic variation and visual generation. Bodily movement at this stage no longer controls merely an individual sound parameter; instead, it becomes an important factor in organizing sound, imagery, and the overall spatial experience.

Figure 5. AI-generated visuals in Section C of *Flower in Hand, Shadow in Transformation*

The interactive system of *Flower in Hand, Shadow in Transformation* was primarily implemented using the TouchDesigner platform, with MediaPipe serving as its visual recognition module and providing the capability to acquire hand-landmark data. Compared with interaction methods that rely on dedicated sensing devices, the camera- and machine-learning-based approach of MediaPipe allows performers to engage in bodily interaction without wearing additional equipment, thereby reducing physical constraints during performance.

It is noteworthy that the work does not extensively employ all of the pose data output by MediaPipe. Instead, it selects data with clear semantic relationships according to the artistic requirements of each section. For example, Section A uses the distance between the thumb and index finger to correspond to the "flower-picking" gesture; Section B uses the spatial distance between the two hands to control changes in sound and visual elements; and Section C uses variations in hand height to influence timbral parameters. This selective use of a small number of key data points demonstrates that the richness of available data is not the sole determinant of expressive potential in interactive music. The relationship between the selection of movements and their artistic semantics is equally important.

Therefore, in *Flower in Hand, Shadow in Transformation*, the role of MediaPipe extends beyond providing pose-recognition functionality. While lowering the technical barrier to incorporating bodily behavior into digital music systems, it also provides creators with a means of transforming bodily movement into an artistic language.

Overall, *Flower in Hand, Shadow in Transformation* demonstrates a representative application of MediaPipe in interactive electronic music: bodily movement is no longer merely an input signal for controlling sound parameters, but becomes a central medium connecting sound, visual elements, and the conceptual framework of the work. From using the "flower-picking" gesture to trigger sound, to using two-handed movement to control changes in sound and imagery, and finally to using bodily posture to influence visual generation, the work progressively establishes the body as an instrument, a central component of the interactive system, and an artistic image. This case demonstrates that in vision-based interactive music creation, the value of a technological system lies not only in its ability to capture movement, but also in how bodily behavior is transformed into artistically meaningful interactive relationships (see Figure 6).

Figure 6. Performance scene from *Flower in Hand, Shadow in Transformation* (2025 Hangzhou international electronic music festival closing concert)

3.2. Xinyu Wang — *Lunar Tune*

Lunar Tune is an interactive electronic music work incorporating pose-based interaction technology. It was composed and visually designed by Xinyu Wang, with percussionist Meng Liu performing the digital percussion and pose-based interactive components. The author contributed the MediaPipe-based pose interaction design for the work. The piece has a duration of approximately six minutes and premiered successfully in 2025 at the concert Algorithmic Concerto: 2025 Intelligent Music Manifesto, part of the Hangzhou International Electronic Music Festival. It was subsequently performed again at the festival's closing concert.

The work was inspired by an important event in the history of human space exploration: the successful soft landing of the Chang'e-4 probe on the far side of the Moon in 2019. The central creative material consists of lunar far-side spectral data published by the Chang'e-4 research team in *Nature*. Through intelligent technologies, these scientific data are transformed into parameters for sound generation, visual presentation, and interactive control, thereby integrating scientific data with artistic expression.

The title *Lunar Tune* is a homophone of the Chinese term for "musical sound" (*yueyin*), intended to construct a symbolic representation of the "sound of the Moon" through technological and artistic means. The mathematical characteristics of the spectral data are transformed into electronic sound materials, while the performer's bodily movements further participate in the process of sound modulation. Natural data, digital systems, and bodily performance thus collectively constitute the work's multidimensional mode of expression.

The sonic structure of *Lunar Tune* revolves around electronic sounds generated from lunar data and digital percussion performance. Overall, the work follows a structural progression from spatialized atmospheric textures toward increasingly rhythmic and kinetic activity (see Table 2).

Section A primarily creates a deep and expansive sonic space. Some synthesized sounds are generated from spectral data collected from the far side of the Moon. By mapping data parameters to sound parameters such as wavetable sampling positions and amplitude, the work produces electronic timbres with a distinctive data-driven origin. At the same time, the transient sounds produced by the digital percussion performance provide rhythmic support within the overall sonic space.

Section B shifts toward a more clearly defined rhythmic structure, constructing a dynamic sonic environment through continuous rhythmic patterns and ethnic percussion sounds. During this stage, sound materials generated from spectral data continue to be employed, creating a dialogue between electronic sounds and acoustic percussion and, more broadly, between traditional performance and data-driven sound generation.

Pose-based interaction in *Lunar Tune* is concentrated primarily in Section A. Its purpose is not to replace traditional percussion performance but to expand the dimensions of sound control within the existing performance approach. The work uses the MediaPipe Pose solution to acquire the performer's bodily pose data, enabling the performer to control electronic sound parameters through arm movements while simultaneously performing digital percussion.

Table 2. Analysis of the main interactive components in *Lunar Tune*

Section	Image	Movement Description	Main Sounds	Movement Control
A		Two alternating actions: 1. Striking the digital percussion instrument with drumsticks; 2. Raising the arms (using MediaPipe).	Various synthesized Impact and Lead sounds generated from spectral data used as wavetables.	1. Striking the percussion instrument triggers various transient (Impulse) sounds; 2. Raising the arms high changes the frequency of sustained, long-line sounds.
B		MediaPipe is not used for pose tracking. The digital percussion instrument is struck rapidly with drumsticks.	Rhythmic, intense synthesized sounds with bright timbral characteristics, together with ethnic percussion sounds.	Striking the digital percussion instrument triggers the corresponding sounds.

At the overall level, the interactive design of *Lunar Tune* combines contact-based performance with non-contact pose-based control. Digital percussion performance constitutes a discrete form of interaction, in which striking actions trigger sounds with transient characteristics, whereas the arm movements captured by MediaPipe constitute a continuous form of interaction, in which changes in bodily position and spatial movement control the parameters of sustained electronic sounds. These two forms of interaction create a contrast between "points" and "lines", allowing traditional instrumental performance and bodily pose-based control to complement each other within the same system.

At the technical level, the system uses Python to invoke the MediaPipe model, acquire the coordinates of the left and right wrist landmarks, and convert them into sound-control parameters. The wrists are selected as control points primarily for reasons of stability: their positions reflect the overall movement of the arms while offering more reliable recognition than finer-grained hand landmarks, thereby reducing unintended triggers caused by subtle movements during performance.

Figure 7. Camera capture of MediaPipe in operation during the performance of *Lunar Tune*

In the performance system, MediaPipe is primarily responsible for human body landmark recognition (see Figure 7), while subsequent data processing and parameter mapping are implemented independently within the Max/MSP environment. Data transmission between the pose-recognition module and the sound-generation module is achieved via the UDP protocol, allowing the two systems to remain relatively independent and thereby improving the stability of the live performance system.

Figure 8. Max patch interface for manually adjusting and monitoring the mapping range of MediaPipe values in the work

Because the data output by MediaPipe consists of normalized coordinates, the performer's range of movement during a performance can be affected by factors such as the stage space, camera position, and individual movement habits. Therefore, the author specifically designed a dynamic range-calibration mechanism. During rehearsals or performance preparation, the system can record the performer's actual range of movement and adjust the coordinate-mapping intervals based on the collected data, enabling pose data to be consistently converted into sound-control parameters across different performance environments (see Figure 8).

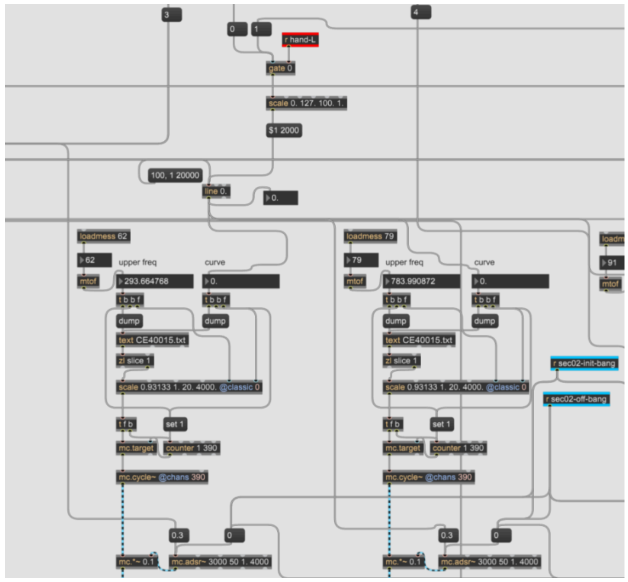


Figure 9. Part of the main max patch in *Lunar Tune* responsible for receiving MediaPipe coordinate data and controlling specific sound parameters

In the main patch responsible for receiving all data for sound generation (see Figure 9), the composer uses the receive hand-L object at the top to receive the spatial coordinate data of the performer's left hand identified by MediaPipe. After the hand-coordinate data undergoes range mapping and smoothing, it is used to control the exponential base value of the scale object below. The data processed by the scale object originates from lunar far-side spectral data containing 390 data points, which are loaded through a text object. As the exponential base value changes in response to real-time pose variations, the program dynamically modifies the mapping distribution of all spectral data points. These changes are then sonified through 390 independently operating sine-wave oscillators built into the mc.cycle~ object, enabling the large-scale spectral dataset to be transformed into sound in real time. In this computational approach, MediaPipe facilitates efficient interactive sound generation based on a large volume of data.

The practical implementation of *Lunar Tune* demonstrates the lightweight deployment and flexibility of MediaPipe in live interactive electronic music. Compared with pose-recognition solutions that rely on dedicated depth-sensing cameras, this system requires only a standard camera and a computer running the MediaPipe environment to achieve real-time pose recognition, allowing the interactive module to adapt relatively quickly to different performance environments. In addition, the work adopts a modular system architecture, with pose recognition, data processing, and sound generation implemented in separate software environments (see Figure 10) and connected through open communication protocols. This design reduces the impact that a failure in any single program may have on the overall system while also improving the flexibility of system debugging and maintenance.

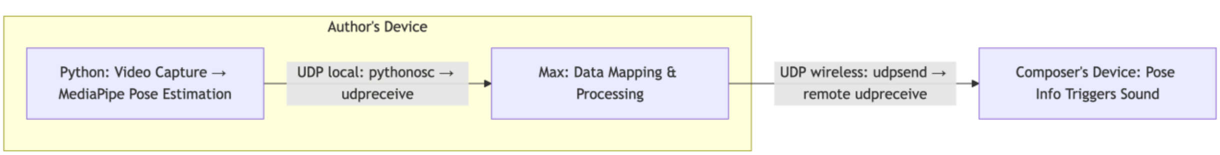


Figure 10. Pose-interaction data flow in *Lunar Tune*

From a practical perspective, *Lunar Tune* demonstrates another approach to using MediaPipe as a tool for interactive music creation. Its value lies not merely in providing human body landmark data, but in serving as an intermediary layer connecting the performer's bodily behavior with a complex electronic music system, allowing bodily movement to enter the sound-generation process in a more natural and direct manner.

From a creative perspective, *Lunar Tune* demonstrates the potential of the body as an extended medium within an interactive system. Bodily movement is no longer merely an auxiliary means of control beyond conventional instrumental performance; rather, it becomes an important link connecting performance actions, data processing, and electronic sound generation. At the same time, the work's modular system architecture and dynamic mapping mechanism demonstrate that MediaPipe can provide interactive electronic music with a low-hardware-dependency and easily deployable approach to acquiring bodily data (see Figure 11).

Figure 11. Performance scene from *Lunar Tune* (2025 Hangzhou international electronic music festival closing concert)

4. Characteristics of MediaPipe-based interactive music creation

4.1. The body as instrument

In pose-based interactive music, "the body as instrument" does not simply mean that bodily movements participate in sound control. Rather, it refers to a performance paradigm in which the performer's body directly enters the processes of sound generation and musical performance through a real-time interactive system, making bodily movement itself part of the musical structure. In contrast to traditional musical performance, which relies on physical acoustic instruments such as pianos and string instruments or interfaces such as MIDI controllers for sound production, interactive music systems based on technologies such as MediaPipe transform the performer's movements, spatial position, and other bodily behaviors into sound-control parameters in real time, thereby establishing a performance mode centered on bodily behavior.

According to relevant research [8], in performance paradigms modeled on traditional instrumental ensemble performance, interactive performance systems composed around "software instruments", "virtual performers", and "composed instruments" closely associate the instrument's control interface (such as keys and strings) with its sound-production mechanism. Physical instruments depend on the application of physical energy and provide proprioceptive and tactile feedback, which are considered essential to achieving precise control and expressive performance. Performers must acquire specific techniques through long-term training and develop performance experience through the causal relationship between bodily movements and sonic feedback. However, with the development of digital interactive music, the concept of the "instrument" is no longer restricted to sound-producing devices with fixed physical forms. Over the past several decades of interactive electronic music, it has gradually expanded into an interactive behavioral system—often conceptualized as a *behavioural object*—composed of movement, mapping, and feedback. In this context, the characteristics associated with an "instrument" in interactive music no longer derive entirely from a paradigm in which physical structure is the defining feature. Instead, they are increasingly reflected in whether a system can establish stable relationships between movement and feedback and whether performers can develop learnable and repeatable performance experiences through interaction.

Within this new paradigm of interactive music, whereas traditional instrumentalists control a physical instrument interface and generate sound through parameter mapping and software instruments, pose-based interactive systems allow bodily movements to participate directly in the logic of sound generation. The performer's movements are captured by a camera, transformed into real-time data, and subsequently mapped to sound parameters such as pitch, timbre, spatialization, and granular density. Rather than relying on the "physical energy" and "tactile feedback" associated with conventional physical instruments, pose-based performance replaces these mechanisms with the macroscopic physical feedback of the body "acting against gravity and within space". The body itself becomes an interface for energy conversion, thereby establishing a continuous "movement–feedback" loop. Through repeated interaction, performers can gradually develop a sense of system control, habitual movement patterns, and bodily memory, allowing the system to acquire a certain degree of *playability* and *learnability*. Even without the fixed physical structure of a traditional instrument, such systems can still establish a behavioral logic analogous to instrumental performance.

The works discussed in this study provide clear examples of how this concept can be realized through different creative approaches. For example, in *Stillness · Reading the Air*, the performer adopts different poses in front of the camera and moves the hands to different areas within the camera frame to trigger different types of sonic changes, enabling real-time control over dynamics, frequency, and timbre. By contrast, *Flower in Hand, Shadow in Transformation* and *Lunar Tune* primarily employ continuous tracking of the coordinates of individual or multiple bodily landmarks. By analyzing landmark positions, distances, and movement trajectories, these works establish mappings with backend sound-generation programs, thereby producing more continuous and nuanced sonic transformations. This data-driven mapping of bodily behavior gives the performance a stronger sense of real-time responsiveness and instrumental agency.

The various MediaPipe models provide extensive parameters for body-joint and landmark coordinates, and different solutions can be used in combination. Compared with interactive approaches based on contact sensors or hardware controllers, this type of vision-based pose-recognition technology can provide more continuous and unconstrained movement input. At the same time, through communication protocols such as MIDI and OSC, bodily movement data can be mapped in real time to different parameters in digital audio workstations, Max/MSP, or sound-synthesis systems, enabling multidimensional sound control. Continuous, high-performance transmission of real-time coordinate data allows performers to establish more stable feedback relationships between movement and sound, thereby enhancing the real-time performability of interactive systems.

Therefore, "the body as instrument" represents not merely a technological approach to sound control but also a reconceptualization of the notion of the "instrument" within the context of interactive music. In pose-based interactive systems, the body, data mapping, and sonic feedback collectively constitute a new performance structure. The body is no longer merely an auxiliary visual element in musical performance; instead, it becomes part of the mechanism through which sound itself is generated. This interaction centered on bodily behavior further expands the expressive boundaries of the relationship among performance, instruments, and the body in digital music.

4.2. The body as the center of the system

In pose-recognition-based interactive music performance, the body not only serves the function of controlling sound but also increasingly becomes the central element connecting sound and visual media. In contrast to traditional multimedia electronic music performances, in which visual content often functions as an auxiliary layer accompanying the sound, multimodal interactive systems frequently share the same movement data, temporal structures, and generative logic across sound and visual media. As a result, the performer's bodily movements can simultaneously drive changes in different media. Within this model, the body is no longer merely a single "performance tool" but becomes a unified interaction source that organizes the operation of the entire audiovisual system.

Relevant creative research has indicated that in interactive music performance systems integrating sound and visual elements, visual materials have gradually moved beyond the constraint of being merely "auxiliary to sound". Through real-time mapping and parameter sharing, visual and sonic elements establish interconnected generative relationships, allowing the work to develop an "audiovisual isomorphic" system structure [9]. Within this context, sound and visual media are no longer independent mediums but synchronized manifestations of the same interactive behavior across different perceptual dimensions.

The works examined in this study clearly demonstrate this "body as the center of the system" structure. For example, in *Flower in Hand, Shadow in Transformation*, MediaPipe continuously tracks changes in the coordinates of the performer's hand landmarks and maps the movement data simultaneously to sound generation, visual feedback, and spatial movement structures. A particular hand gesture, for instance, can simultaneously trigger a lotus visual and a bell-like pitched percussion sound, while changes in the distance between the two hands can further influence particle-flow speed, percussion pitch, and the introduction of a low-frequency sound layer. In this process, bodily movement does not separately control sound and visual elements. Rather, it serves as a unified data source that simultaneously drives parameters across multiple media, establishing a real-time generative structure in which sound, visual elements, and movement share a common logic. The performer's body consequently becomes a central interactive node connecting sound, visual media, and feedback from the performance space.

The development of this creative approach is closely related to the increasing maturity of real-time interactive workflows in recent years. The widespread adoption of MIDI and OSC protocols has made it increasingly convenient to transmit movement data between different software environments in real time. In particular, the development of node-based real-time visual platforms such as TouchDesigner has enabled greater scalability for integrated sound–image–performance interactive systems. Within such platforms, creators can not only access community-developed MediaPipe components designed for platform users but also utilize extensive libraries of CHOP, TOP, and SOP operators to receive data from MediaPipe or other sources and perform image generation, particle simulation, spatial transformation, and audiovisual synchronization in real time. This modular workflow allows performers to establish mappings between bodily movement and visual feedback with greater flexibility while also enhancing system stability and adjustability during live performances.

In summary, under the model of "the body as the center of the system", sound, visual media, and space no longer exist as separate and independent media. Instead, they form a shared structure and real-time network of relationships driven by bodily movement. The body thus shifts from being an auxiliary performance element in traditional performance toward becoming the central node that organizes the operation of the entire multimodal interactive system.

4.3. The body as image

In most interactive music performances, the performer's bodily movements themselves constitute a core component of the performance, jointly shaping the overall stage presentation with sound and visual elements. The body is not merely a "controller" that triggers and modifies sound and visual elements; it is also an important component of the performance as a whole, contributing alongside sound and visual media to the development of the work. Consequently, in the creation of pose-based interactive music, creators must not only design the sound, visual elements, and interactive programs but also carefully consider and choreograph the physical performance itself.

In many traditional electronic music performances, sound generation is largely accomplished through computers, controllers, and backend software, while the performer's body is often concealed behind equipment, consoles, or other stage structures. Compared with the clearly visible relationship between physical movement and sonic feedback in traditional instrumental performance, bodily behavior in electronic music performance often lacks sufficient visibility. Audiences may find it difficult to determine directly which actions trigger particular sounds or to perceive changes in musical structure and emotional expression through the performer's bodily state. In some performances centered primarily on computer-based audio programs, performers may remain in a relatively fixed position for extended periods, operating only keyboards, mice, or controllers. This can produce a pronounced tendency toward "de-embodiment" in the stage presentation. Although sound continues to be generated in the performance space, the sense of human presence may become comparatively diminished. Electronic music performance has consequently long been associated with discussions concerning the invisibility of performance actions and the lack of a sufficiently strong sense of liveness.

Relevant scholarship has pointed out that musical perception does not originate solely from auditory experience but is instead constituted through an embodied process involving bodily movement, visual observation, and perceptual experience [10]. The introduction of pose-based interactive technologies has, to some extent, re-emphasized the central role of the body in electronic music performance. By incorporating bodily movement itself into the generation of sound and visual elements, these technologies make the performer's movement trajectories, spatial postures, and changes in physical actions important components of stage expression once again. Audiences can not only "hear" and "see" changes in sound and visual elements but also perceive the real-time relationship between performance actions and system feedback through the performer's visible bodily movements. The body therefore returns to the center of electronic music performance, shifting electronic music from a relatively "de-embodied" control paradigm toward a more embodied performance structure emphasizing the visibility of movement, real-time feedback, and bodily participation.

The three works analyzed in this study all demonstrate the creators' application of the body as an artistic image. For example, in *Stillness · Reading the Air*, the choreography incorporates floating arm movements and expansive bodily gestures. These sequences of touching and swaying movements interact with the notion of "air" in different ways, with the forms of interaction continuously changing, intensifying, and becoming more elaborate as the music progresses. In *Lunar Tune*, concentrated percussion performance and pose-based interaction are assigned to two different sections. In the section employing pose-based interaction, the performer stands at the center of the stage, with the digital percussion MIDI controller positioned in front of her like the central control console of a spacecraft. As she performs with drumsticks, her movements resemble the control of a spacecraft's trajectory; when she raises and sweeps her arms through the air, her movements evoke the observation and manipulation of celestial motion. Combined with visual imagery resembling drifting sand or granular matter, these bodily movements contribute to an impression of the vast and tranquil environment beyond Earth.

At the technical level, MediaPipe provides an important foundation for the formation of "the body as image". Its various solutions, including Pose, Hands, and Face Mesh, can acquire extensive coordinate data from human skeletal landmarks, hand landmarks, and facial features, providing rich and stable parameter sources for incorporating bodily movement into electronic music performance. These solutions also include functions for recognizing commonly used gestures. During the creative process, creators can identify specific movements and gestures and further develop analytical methods to examine movement speed, direction, distance variation, spatial position, and continuous movement trajectories, thereby enabling bodily behavior to be mapped more precisely to sound, visual, and spatial feedback.

This rich data structure means that the body is no longer merely a simple "trigger" but can genuinely participate in generating the atmosphere and perceptual imagery of a performance. For example, movement amplitude can influence the range of visual-particle dispersion, movement speed can alter sonic dynamics or rhythmic density, and changes in the body's spatial position can further contribute to the organization of spatial audio, projections, or real-time visual structures. Bodily movement therefore serves not only a control function but also becomes part of the stage narrative and perceptual structure. The sound, visual elements, and spatial transformations perceived by the audience no longer originate from independent media systems but instead form synchronized feedback relationships under the unified influence of bodily behavior. Moreover, compared with earlier interactive-art workflows that relied heavily on customized hardware and complex program development, the community resources, open-source examples, and ready-made component libraries that have emerged around MediaPipe have created a more mature technological ecosystem for real-time interactive performance. Creators can therefore devote more attention to designing relationships among sound, visual elements, and the body rather than continually investing substantial effort in rebuilding and maintaining the underlying motion-capture system. The maturation of this technological environment further expands the possibility of returning the body to the center of the electronic music performance stage.

Overall, "the body as image" represents a further transformation in the role of the body within pose-based interactive music. In this model, the body is no longer merely an external control medium for sound and visual systems but gradually becomes an important component of the work's perceptual structure and stage narrative. The performer's movement trajectories, spatial postures, and physical states not only participate in the real-time generation of sound and visual elements but also directly influence how audiences understand the work's atmosphere, spatial relationships, and emotional expression. The introduction of pose-based interactive technology reinforces the visibility and presence of the body in electronic music performance, which has historically tended toward de-embodiment, and gradually shifts electronic music away from a performance logic centered on operating technological equipment toward a more embodied performance structure emphasizing bodily participation, real-time feedback, and live perception.

5. Conclusion

This study has examined the application of MediaPipe technology in interactive music creation, focusing on its technical characteristics, creative practices, and modes of artistic expression. First, the study summarizes MediaPipe's characteristics in three areas: the provision of diverse solutions and parameters, support for communication protocols and community resources, and compatibility and flexibility across software and hardware platforms. In analyzing creative practices, the study presents case analyses of two representative interactive music works. Finally, based on these creative examples, the study identifies three principal characteristics of MediaPipe-based interactive music from the perspectives of sound, audiovisual relationships, and bodily expression. Overall, this study attempts to examine the methodological characteristics and aesthetic orientations of pose-based interactive music creation from the dual perspectives of creative practice and interactive music theory.

At the technical level, MediaPipe's three principal characteristics—its rich range of solutions and parameters, support for multiple communication protocols and community resources, and high degree of software and hardware compatibility and flexibility—not only broaden the accessibility of interactive music but also transform the technology itself from a "system burden" into a "creative tool", allowing it to serve musical structure, sound organization, and stage expression more directly.

At the level of creative practice and artistic characteristics, the analyses of *Flower in Hand, Shadow in Transformation* and *Lunar Tune* demonstrate that MediaPipe enables the body to function not only as an "instrument" in interactive music but also as a central control element within the performance system, coordinating audiovisual interaction. Bodily posture in these works likewise serves more than the functions of an "interface" or "central control mechanism". The physical form and movement of the performer themselves constitute important components of the stage presentation, working together with sound and visual elements to shape the overall performance.

It should be noted that this study remains primarily focused on analysis at the level of creative practice. More systematic experimental research is still needed to investigate the stability and accuracy of MediaPipe under different environmental conditions, as well as to conduct systematic comparisons with other motion-sensing technologies. These issues warrant further investigation in subsequent research and creative practice.

Overall, through an analysis of the integration of MediaPipe technology with interactive music works, this study seeks to demonstrate that contemporary computer-vision-based bodily interaction technologies are providing new methodological approaches and aesthetic possibilities for interactive music creation. MediaPipe is by no means a definitive solution to the challenges of artistic creation. Nevertheless, the technological orientation and creative potential it embodies provide a practically meaningful reference for further exploration of the relationship between the body and music in future interactive music practices.

References

[1] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.

[2] Kim, J. H., & Seifert, U. (2006). Embodiment: The body in algorithmic sound generation. *Contemporary Music Review, 25*(1–2), 139–149.

[3] Schacher, J. C. (2012). The body in electronic music performance. In *Proceedings of the Sound and Music Computing Conference* (pp. 194–200).

[4] Visi, F., Schramm, R., & Miranda, E. (2014). Gesture in performance with traditional musical instruments and electronics: Use of embodied music cognition and multimodal motion capture to design gestural mapping strategies. In *Proceedings of the 2014 International Workshop on Movement and Computing* (pp. 100–105).

[5] Eckel, G., & Pirro, D. (2009). On artistic research in the context of the project embodied generative music. In *Proceedings of the International Computer Music Conference (ICMC)*.

[6] Caramiaux, B., Susini, P., Bianco, T., Bevilacqua, F., & Schnell, N. (2011). Gestural embodiment of environmental sounds: An experimental study. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 144–148).

[7] Tobita, K., & Mima, K. (2024). Enhanced detection of musical performance timings using MediaPipe and multilayer perceptron classifier. *AI, Computer Science and Robotics Technology*.

[8] Bown, O., Eldridge, A., & McCormack, J. (2009). Understanding interaction in contemporary digital music: From instruments to behavioural objects. *Organised Sound*(2).

[9] Eyes, B. J. (2014). *Multimodal performance approaches in electronic music* [Doctoral dissertation, University of York].

[10] Leman, M., Maes, P. J., Nijs, L., & Van Dyck, E. (2018). What is embodied music cognition? In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 747–760). Springer.

Appendix

The following is a Python script that calls the MediaPipe library in conjunction with the OpenCV library to capture hand-coordinate data and output it via UDP:

```
# Mediapipe pose solution control script
import cv2
import mediapipe as mp
from pythonosc import udp_client
osc_client = udp_client.SimpleUDPClient("127.0.0.1", 9009) # Initialize OSC client (send to localhost:9009)
"""Initialize MediaPipe hand detection module and video input device"""
mp_hands = mp.solutions.hands
hands = mp_hands.Hands(min_detection_confidence=0.7, min_tracking_confidence=0.5)
mp_draw = mp.solutions.drawing_utils
cap = cv2.VideoCapture(0) # Open default camera
while cap.isOpened():
    success, frame = cap.read()
    if not success:
        break
    rgb_frame = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB) # Convert BGR image to RGB
   result = hands.process(rgb_frame)
    if result.multi_hand_landmarks: # If hand landmarks are detected
        for hand_landmarks in result.multi_hand_landmarks:
                mp_draw.draw_landmarks(frame, hand_landmarks, mp_hands.HAND_CONNECTIONS) # Draw hand landmarks (optional, for demonstration)
                    pinky_tip = hand_landmarks.landmark[20] # Predefined index of pinky fingertip in MediaPipe is 20
                x, y, z = pinky_tip.x, pinky_tip.y, pinky_tip.z # Get normalized coordinates (x, y, z), z is depth information, usually only x and y are used
                  osc_client.send_message("/pinky_tip", [x, y, z]) # OSend coordinates to port 9009 via OSC, message address is "/pinky_tip"
                """Draw coordinate text on monitoring image (for debugging)"""
                h, w, _ = frame.shape
                cx, cy = int(x * w), int(y * h)
                cv2.putText(frame, f"Pinky: ({x:.2f}, {y:.2f})", (cx, cy),
                            cv2.FONT_HERSHEY_SIMPLEX, 0.5, (0, 255, 0), 2)
      cv2.imshow("MediaPipe Hand Tracking", frame) # Display output image window
      if cv2.waitKey(1) & 0xFF == ord('q'):
          break
"""Release all resources"""
cap.release()
cv2.destroyAllWindows()
hands.close()
```