Metaverse (2023) Volume 4 Issue 2
doi: 10.54517/m.v4i2.2336

Asia Pacific Academy of Science Pte. Ltd.

ORIGINAL RESEARCH ARTICLE

From motion to magic: Real-time virtual-real stage effects via 3D motion capture

Xiongbin Lin, Xun Wang, Wenwu Yang*

School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China

*** Corresponding author**: Wenwu Yang, wwyang@zjgsu.edu.cn

ABSTRACT

Immersive cultural performances with virtual-real fusion effects are the future development trend in the exhibition and stage industry. However, current virtual-real stage performances heavily rely on traditional sequential design and arrangements. During the performance, actors must move to specific positions based on the musical beat and execute predetermined actions with a pre-designed amplitude and frequency to synchronize with the fixedly played stage visual effects; otherwise, major performance accidents such as plot inconsistencies or continuity errors may occur. To address the problem, this paper introduces a real-time generation system for stage visual effects based on multi-view multi-person 3D motion capture. The system utilizes multi-view 3D motion capture technique to achieve non-intrusive real-time interaction perception of target actors in the stage space. By perceiving the spatial position and performance actions of the target actors, corresponding stage visual effects are generated in real-time. This is followed by the seamless integration of sound effects and immersive high-definition display, ultimately realizing multidimensional real-time interaction between real actors and virtual visual effects in the stage space. We conducted an experimental virtual-real stage performance, lasting approximately two minutes, in a physical theater to validate the effectiveness of our proposed method. The experiment not only produced a unique innovative effect of blending stage and technology but also effectively enhanced the sense of presence and interactivity of the stage performance, providing actors with more freedom and control in their performances.

Keywords: immersive stage performance; interactive generation of stage effects; non-intrusive multi-person motion capture; action recognition

1. Introduction

In recent years, with the integrated innovation and development of the cultural tourism industry, cultural performances have rapidly emerged as an emerging consumer experience format, gradually becoming the core experience of various cultural tourism destinations. A number of well-produced cultural performance programs have emerged, such as “Impression: West Lake” created by renowned Chinese director Yimou Zhang and “Romance of Song Dynasty” in the Songcheng Scenic Area in Hangzhou, China, as shown in **Figure 1**.

These performance programs effectively promote an excellent traditional culture with Chinese characteristics and serve as the core experiential attractions of scenic areas, igniting the industry and attracting

ARTICLE INFO
Received: 16 October 2023 | Accepted: 6 November 2023 | Available online: 15 November 2023

CITATION
Lin X, Wang X, Yang W. From motion to magic: Real-time virtual-real stage effects via 3D motion capture. *Metaverse* 2023; 4(2): 2336. doi: 10.54517/m.v4i2.2336

Figure 1. Representative cultural live performances: "Impression: West Lake" and "Romance of Song Dynasty".

crowds. Traditional cultural performance forms lack innovative and real-time control methods, failing to meet the market and audience's demand for the deep integration of advanced technology and modern cultural performances. Virtual-real performance techniques are considered an important approach to drive innovation in cultural performances, enabling more flexible, free, and creative forms of expression. However, current research on virtual-real fusion and rehearsal editing techniques, both domestically and internationally, mainly focuses on program simulation and lacks relevant research outcomes regarding the introduction of interactive events and the collection of multidimensional interactive sensing signals in live performances. Specifically, current domestic and international performances that combine virtual and real elements still heavily rely on traditional temporal design and arrangement. During the performance, actors need to position themselves at designated spots and perform predetermined actions based on the musical beats and pre-designed amplitude and frequency to synchronize with fixed playback of stage effects. Failure to do so may lead to plot mistakes and other significant performance accidents. This temporal requirement severely affects the presentation effect of stage art and restricts the development of innovative virtual-real fusion performance forms.

To address the aforementioned problem, we have developed a real-time interactive generation system for stage effects based on multi-view 3D motion capture. It utilizes tracked actor positions and performance actions to control the coordinated triggering and real-time generation of stage effects. Firstly, camera parameters and layout schemes are analyzed and determined based on the characteristics of the stage environment to achieve high-quality video capture of performances. Secondly, we have developed a real-time generation system for 3D stage effects using a lightweight real-time multi-person 3D motion capture system and the Unity3D engine. In Unity3D, we have established a mapping relationship between the camera's world coordinate system, where the captured objects are located, and the spatial coordinate system of the virtual stage effects content. Finally, a relatively simple and controllable performance environment has been designed and set up in the "*Bianliang Yimeng*" theater at Zhejiang Hengdian World Studios scenic area. A 2-minute experimental stage performance has been choreographed, where the actor positions and actions, which are captured in real-time, are employed to control the real-time generation and coordinated triggering of stage effects. Experimental results have demonstrated that our developed real-time stage effects system achieves a unique fusion of stage and technology, enhancing the sense of presence and interactivity of stage performances while providing actors with more freedom and control over their performances.

2. Related work

Common motion capture methods often require the use of specific equipment or sensors, including optical sensors, magnetometers, inertial sensors, acoustic sensors, and pressure sensors. Optical motion capture technology[1–4] places reflective optical markers at human body joints and calculates their positions in space based on triangulation principles (i.e., if a point on the human body can be captured by two cameras simultaneously, its three-dimensional spatial coordinates can be determined). This technology is precise and

efficient but expensive. Magnetometer-based motion capture technology[5] uses magnetic field induction to detect the current position and orientation of joints. It has high power consumption, high cost, and sensitivity to metal objects in the environment. Inertial sensor motion capture technology[6,7] measures joint rotation angle by placing gyroscopes or accelerometers on human limbs. Its main disadvantage is that the measurement result significantly drifts with increasing capture time. Acoustic motion capture technology[8-10], on the other hand, attaches ultrasound pulse emission markers to specific parts of the human body and calculates their position based on sound reception time. In addition, some hybrid motion capture technologies combine multiple sensors or cameras to compensate for their respective shortcomings.

In addition, some hybrid motion-capture technologies combine multiple sensors or cameras to make up for their shortcomings. Examples include combining inertial and acoustic sensors to achieve precise motion capture[11], using monocular cameras to correct the drift error of inertial sensors[12], using deep learning techniques to estimate the position of the human body in space to make up for the absence of inertial sensors that can only measure human posture[13,14], and using three depth cameras and two pressure sensors to record human motion data[15].

The aforementioned motion capture techniques, however, frequently call for wearing particular apparatus, which is inappropriate for stage actors who must wear costumes for performances. Therefore, for our application scenario, vision-based motion capture techniques are preferable. With vision-based motion capture, the movements of actors are only captured by capturing RGB image data from cameras, not by having actors wear special equipment. One common strategy is to directly estimate the 3D skeletal structure of the human body from a monocular camera[16]. However, due to occlusions and inherent depth ambiguities in single-view images, these methods often struggle to generate high-quality 3D human pose estimation results. To address this issue, many recent works utilize multiple viewpoints to estimate the 3D skeletal structure of the human body, using information from multi-views[17,18] to compensate for the limitations of single-view approaches.

3. Real-time generation system of virtual-real stage effects

3.1. Stage environment analysis and camera layout

In order to build a real-time interactive generation system for stage visual effects based on a multi-view 3D motion capture system for the physical stage performance environment, we conducted field research at the "*Bianliang Yimeng*" theater in the Hengdian World Studios scenic area in Zhejiang. We summarized and analyzed some basic characteristics of the physical stage environment:

(1) The width, height and depth of typical stage spaces are usually not less than 16 m × 10 m × 16 m.

(2) A physical stage typically has a wooden floor that vibrates a lot while performances are taking place.

(3) There are various equipment stands (such as projection screens) around the stage, with multiple entrances and exits for actors, and various performance props are usually present on the stage.

(4) The lighting on the stage frequently and dramatically changes while performances are taking place, with actors running and overlapping each other.

According to the characteristics of the stage environment mentioned above, we positioned cameras in various locations on the concrete floor in front of and to the sides of the stage to detect motion and track actors' positions. By using this method, camera position changes brought on by stage tremors during performances are avoided, which could ruin the previously calibrated camera parameters. Actor occlusion issues have a negative effect on the motion capture algorithm, but using complementary multi-view information can help mitigate this effect. Moreover, due to dramatic changes in stage lighting during performances, when capturing large areas using a camera, the images can easily become too bright or too dark. However, by using a multi-

camera layout, each camera is responsible for observing only a small fixed area of the stage, thus avoiding overly bright or dark images.

Furthermore, we have condensed the following camera selection criteria by contrasting the shooting outcomes of various cameras in the physical stage performance environment:

(1) Dynamic range: The dynamic range describes how well a digital camera can simultaneously capture details in both dark and light areas. Richer details can be captured in the image with a higher dynamic range. We discovered that stage lighting frequently changes dramatically, and regular cameras may not have enough dynamic range to support these quick changes in lighting, which can result in overexposure problems, as shown in **Figure 2**.

(2) Color depth range: When dark light or light contrast is strong, ordinary cameras are prone to noise, automatic exposure failure, and insufficient tolerance, as shown in **Figure 3**.

Figure 2. Testing the camera's dynamic range in a live stage lighting environment.

Figure 3. Testing the camera's color depth range in a live stage lighting environment.

Based on the experimental tests and analysis, we have chosen a cost-effective camera, the Panasonic GH5S mirrorless camera, for data collection in the system. This camera provides clear imaging and strong color depth rendering, offering image quality consistent with high-performance DSLR cameras. It not only avoids overexposure issues under stage lighting but also provides rich color depth, ensuring the system can capture multi-view data of higher quality. This helps to avoid low-quality actor motion capture caused by data quality issues.

We employed the classical planar chessboard calibration method to obtain the camera's intrinsic and extrinsic parameters. Considering that the corners of the calibration board may not be detected properly when the board is too blurry due to its distance from the camera, we used an enlarged calibration board with dimensions of 1.1 × 0.8 m, as shown in **Figure 4**.

Figure 4. Using a black and white chessboard for camera calibration in a scaled stage space.

3.2. Non-intrusive multi-target actor 3D human motion capture

To achieve non-contact multi-target 3D actor motion capture, we designed and implemented a lightweight real-time multi-person 3D motion capture system based on multiple views[19], as shown in **Figure 5**.

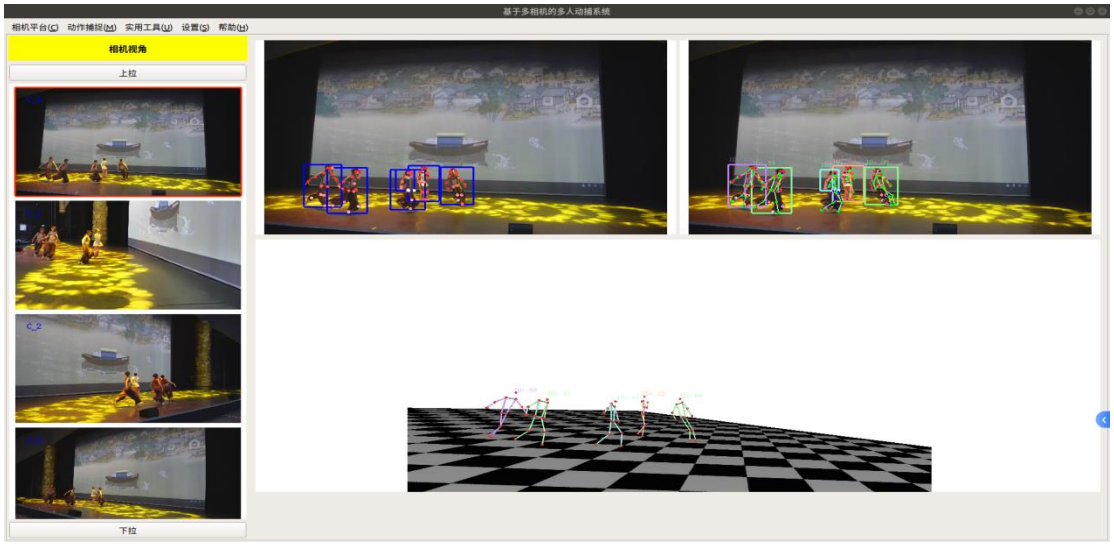

Figure 5. Non-intrusive multi-target actor 3D motion capture.

The system does not have specific requirements for the background and dressing of objects in the scene. Its core algorithm is a robust multi-view multi-person 3D pose estimation method. The algorithm consists of three processing steps: (1) detecting the 2D human poses of people in each view; (2) associating the 2D human poses of individuals in various views; (3) reconstructing the 3D human poses of each individual. This algorithm fully utilizes the complementary nature of multi-view information to lessen the impact of occlusion on motion capture, ensuring the accuracy of 3D human motion capture. As shown in **Figure 6**, based on the algorithm's processing steps, the system mainly consists of four modules: data acquisition module (M_0), camera calibration module (M_1), algorithm processing module (M_2), and main control module (M_3). The main functions of each module are as follows:

(1) Module M_0, the data acquisition module, is responsible for connecting SDI video capture cards and synchronizing data from multiple cameras, ultimately achieving real-time output of multi-view data for each frame.

(2) Module M_1, the camera calibration module, uses the images of planar chessboards captured by the cameras through module M_0 to calculate the intrinsic and extrinsic parameters of all cameras using camera calibration algorithms.

(3) Module M_2, the algorithm processing module, sequentially receives multi-view data frames from

module M_0 and processes them to generate real-time 3D human motion poses for all target objects in each frame.

(4) Module M_3, the main control module, is mainly responsible for integrating various modules and organizing related data.

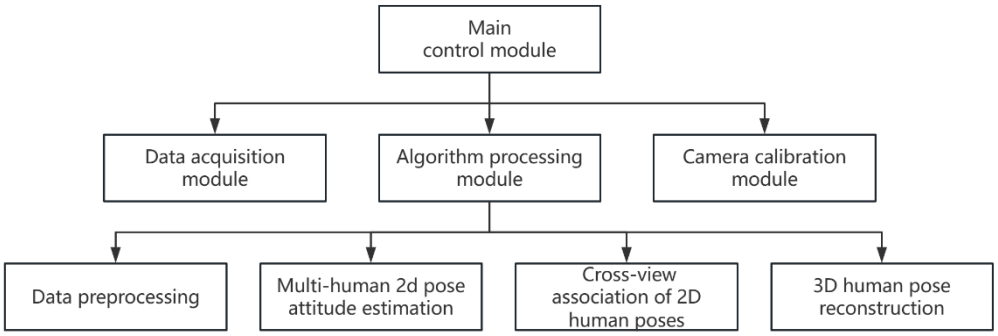

Figure 6. Software modules and their structural relationships of the 3D motion capture system.

The algorithm processing module M_2 primarily entails data preprocessing, including procedures like image undistortion and YUV to RGB conversion. A multi-view, multi-person 3D pose estimation algorithm is also included. This algorithm entails four processing steps: data preprocessing, multi-person 2D pose detection in each view, cross-view association of 2D poses, and reconstruction of 3D human poses. In order to get excellent results, we also employ a top-down multi-person 2D pose detection algorithm for the 2D pose detection step. The algorithm first extracts the 2D bounding boxes of each person in the view, then segments the individual body images based on the 2D bounding boxes and performs single-person 2D pose estimation on each body image separately. Therefore, we further divide Module M_2 into the following five sub-modules: data preprocessing, 2D body bounding box detection, single-person 2D pose estimation, cross-view association of 2D poses, and reconstruction of 3D human poses. Through this modular design, not only can the complexity of system maintenance and expansion be effectively reduced, but it also facilitates efficient optimization of each module in the system.

3.3. Real-time interactive generation of stage visual effects

Based on the captured 3D positions and performance actions of the target actor, the system recognizes the actions and triggers corresponding stage effects. First, we pre-record specific actions of the actor in the system. Specifically, we take a sequence of 15 continuous frames of 3D human body skeleton poses captured by the system as one action sequence, and use the ST-GCN++ action recognition deep network model[20] to output a corresponding action feature vector. During the live stage performance, the system captures the real-time 3D performance actions of the target actor and inputs the obtained action sequence into ST-GCN++ to obtain the corresponding action feature vector. By calculating the real-time action feature vector of the target actor and the pre-recorded specific action feature vectors, we can determine whether the target actor is currently performing a pre-recorded specific action, and then trigger and display the corresponding stage effects.

The position coordinates of the target actor captured by the system must be converted into the 3D stage modeling space in order for the stage effect content to be displayed on the screen in the same position as the target actor. This is necessary because the coordinate system used to calibrate the camera beneath the stage scene differs from the coordinate system used to create the stage space in Unity3D. Assuming that the stage coordinate space is P and the 3D stage modeling space is C, we know the origin position of coordinate space C and the three unit coordinate axes in coordinate space P. We need to be able to transform a point or vector A_C represented in coordinate space C into a point or vector A_P represented in coordinate space P. Conversely,

we also need to transform a point or vector B_P represented in coordinate space P into a point or vector B_C represented in coordinate space C:

$$A_P = M_{C\to P} A_C$$

$$B_C = M_{P\to C} B_P$$

where $M_{C\to P}$ represents the transformation matrix from coordinate space C to coordinate space P, and $M_{P\to C}$ represents its inverse matrix. If we know the representation of the three coordinate axes of the coordinate space C under the coordinate space P, x_C, y_C and z_C (these are three basis vectors, for example, $x_C = (x_{Cx}, x_{Cy}, x_{Cz})$), and their origin position O_C, then the transformation matrix from C to P, $M_{C\to P}$, is given by:

$$M_{C\to P} = \begin{bmatrix} x_{Cx} & y_{Cx} & z_{Cx} & O_{Cx} \\ x_{Cy} & y_{Cy} & z_{Cy} & O_{Cy} \\ x_{Cz} & y_{Cz} & z_{Cz} & O_{Cz} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Next, we explain how to obtain the basis vectors x_C, y_C, and z_C in coordinate space P. As shown in **Figure 7**, we first determine the three-dimensional coordinates of four points in coordinate space C, namely P_{C0}, P_{C1}, P_{C2}, and P_{C3}. Then, these points are projected onto the stage space through the stage curtain. By using the camera parameters, we can solve for their corresponding three-dimensional coordinates in the stage space, denoted as P_{P0}, P_{P1}, P_{P2}, and P_{P3}, respectively. Since P_{P2} only differs from P_{P1} in the x-axis direction, we can obtain the basis vector x_C by subtracting the coordinates of these two points:

$$x_C = P_{P2} - P_{P1}$$

In a similar way:

$$y_C = P_{P3} - P_{P2}$$

By using the average of the two sets of basis vectors as the final basis vectors, we can minimize the error that may have occurred during the manual selection of P_{P0}, P_{P1}, P_{P2}, and P_{P3}. Consequently, we have:

$$x_C = \frac{(P_{P2} - P_{P1}) + (P_{P3} - P_{P0})}{2}$$

$$y_C = \frac{(P_{P0} - P_{P1}) + (P_{P3} - P_{P2})}{2}$$

The basis vector z_C should be perpendicular to the plane formed by x_C and y_C, and it is obtained by taking the cross product of x_C and y_C. Additionally, we will take P_{P1} as the origin O_P in coordinate space P. Finally, we can calculate the transformation matrix $M_{C\to P}$ and its inverse matrix $M_{P\to C}$ accordingly.

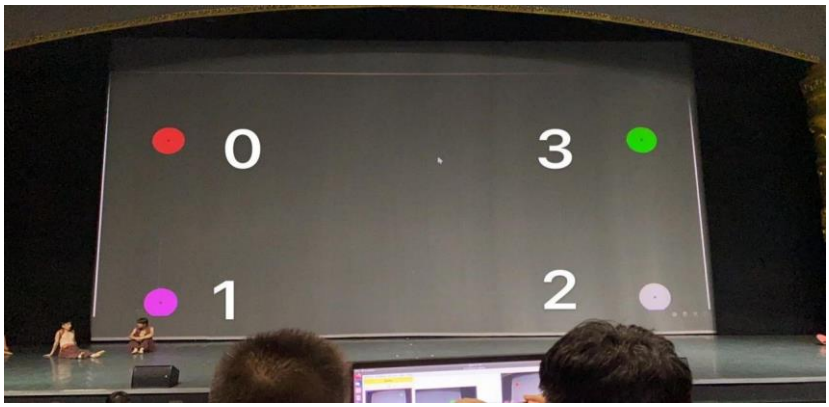

Figure 7. Illustration of projected coordinate points.

4. Experimental results

4.1. Stage design

We collaborated with performance art experts from Zhejiang Hengdian Film and Television City to design an experimental mixed reality stage performance titled "Bianhe—Love of the Boatman". The plot of this performance segment is relatively simple: on the banks of the Bianhe River, the boatman's wife and the boatman are bidding a reluctant farewell. During the dance choreography, the background stage effects depict a scene from the painting "Along the River During the Qingming Festival". The Bianhe River glistens, with a wooden boat floating on it, and fish swimming around the boat. During the performance, there will be five actors: one female actor (the boatman's wife) and four male actors (the boatman and his colleagues). They will elegantly dance by the side of the Bianhe River, while the wooden boat in the river will follow the movements of the female actor. When the female actor performs a side kick, it will trigger an animation of fish leaping out of the river.

4.2. 3D stage modeling and simulation

We use the Unity3D engine for real-time rendering of 3D stage art modeling in accordance with the design of the stage art content. As shown in **Figure 8**, the stage modeling consists of four elements: riverbank houses, river water, wooden boats, and a school of fish. The background houses are selected from the painting "Along the River During the Qingming Festival" with the water surface removed. The river water is rendered in real-time using a water surface shader, including the bubbles generated by the movement of the wooden boat and the school of fish, as well as their reflections on the water's surface. A fish swarm algorithm in the engine simulates the movement of the school of fish. Based on the pre-set action number received, the fish swarm's control algorithm synchronizes the fish leaping effect with the actress's movements. The motion capture system uses the actress's position to determine the position of the wooden boat as it moves, creating an interactive effect where the boat moves in sync with the target actress.

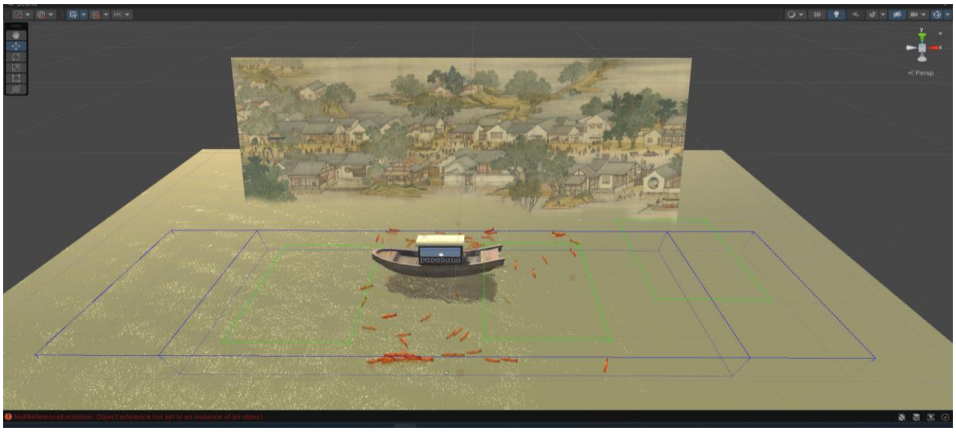

Figure 8. Illustration of stage design modeling.

4.3. Real-time generation and collaborative triggering of stage visual special effects

During the live stage performance, the 3D motion capture system captures the three-dimensional performance actions and spatial positions of the target actors in real-time. The captured information is then transmitted to the stage effects system developed in Unity3D engine via the TCP network protocol. The stage effects system continuously receives the target actor's information from the motion capture system and updates the position of the wooden boat in real-time based on the target actor's position information. It also triggers specific stage effects based on the target actor's action information. For example, when the target actress

performs a side kick, the school of fish leaps. Some experimental results are shown in **Figure 9** and **Figure 10**. These results show that the suggested technique is capable of achieving real-time and accurate perception of the three-dimensional performance actions and spatial positions of the target actors in the stage space. It allows for the real-time generation and coordinated triggering of stage effects based on how the performance of the target actor interacts with the stage effects. The conventional sequential stage performance mode is broken by this novel approach. The stage and technology are uniquely combined, the sense of presence and interaction on stage is improved, and actors have more creative freedom and control over their performances.

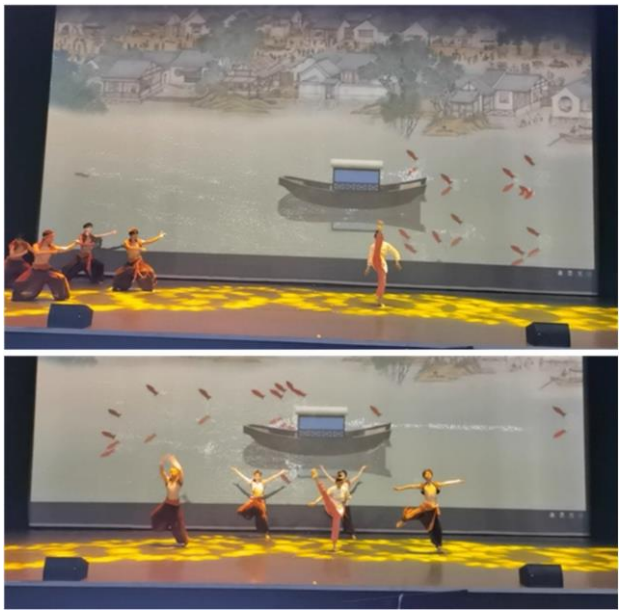

Figure 9. The target actor's actions trigger stage effects. When the motion capture system captures the actress lifting her leg, the stage effects system generates real-time animation content of a school of fish leaping and projects it. The two rows correspond to different times.

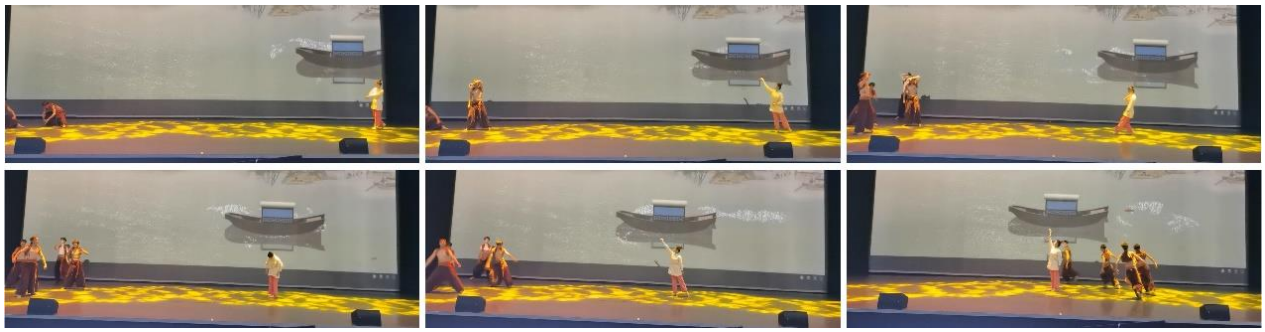

Figure 10. The stage content is updated in real-time based on the target actress's position. As the actress changes her position, the small wooden boat will move in real-time accordingly.

4.4. Discussion

With the support of this system, the actor does not need to appear in a fixed position during the performance, and the special effects screen will change with the position and action of the actor. We call this kind of virtual-real interaction method "high redundancy event-driven virtual-real interaction". Different from the traditional way of virtual and real interaction, the effect of stage art is not a pre-made video but a real-time rendering picture, and it is the stage art with the actor rather than the actor with the stage art. This kind of virtual and real interaction not only frees the actors from repetitive rehearsal tasks but also increases the redundancy of the performance and allows the actors to have a certain freedom to play. As shown in **Figure 11**, the two pictures correspond to two different performances of the same piece of act, and in the pre-designed

performance arrangement, the actress starts from the right end and walks to the center of the stage, and the boat follows the actress all the way. The actress does not stand in the same position in the two performances, but the boat is still in the correct position with her. If the pre-rendered video interaction is used, the change of the actress's position will lead to continuity errors in the scene.

Figure 11. High redundancy event-driven virtual-real interaction.

After the application of the system in the production of "*Bianliang Yimeng*" in Hengdian World Studios, it saves about one-third of the rehearsal time and more than one-quarter of the capital cost. The general manager of Hengdian Studio Art Troupe also highly recognized the system effect.

5. Conclusion

In view of the problem that the current cultural performances combining virtual reality still rely heavily on the sequential design and arrangement in the early stage of creation, and the actors need to rehearse repeatedly according to the rhythm of the music to match the stage content and special effects, we propose and realize the real-time generation system of virtual-real stage visual special effects based on multi-view 3D motion capture technology. It can control the collaborative triggering and real-time generation of the stage visual special effects by tracking the target actor's positions and performance motions. Experimental results demonstrate that our proposed system enhances the sense of presence and interactivity in stage performances, providing actors with more freedom and control over their performances. It achieves a dynamic integration of stage content driven by actor performances.

Author contributions

Conceptualization, WY and XL; methodology, WY and XL; software, WY and XL; validation, XL; formal analysis, WY; investigation, XL; resources, XL; data curation, XL; writing—original draft preparation, XL; writing—review and editing, XW and WY; visualization, XL; supervision, WY and XW; project administration, WY; funding acquisition, WY and XW. All authors have read and agreed to the published version of the manuscript.

Data availability

Data available upon request.

Funding statement

This material is based upon work supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LY21F020010, Key R&D Program of China (No. 2018YFB1404102).

Conflict of interest

The authors declare no conflict of interest.

References

1. Bishop G, Hill C. *Self-Tracker: A Smart Optical Sensor on Silicon* [PhD thesis]. The University of North Carolina at Chapel Hill; 1984.
2. Chen K, Wang Y, Zhang S-H, et al. MoCap-solver: A neural solver for optical motion capture data. *ACM Transactions on Graphics* 2021; 40(4): 1–11. doi: 10.1145/3450626.3459681
3. Woltring HJ. New possibilities for human motion studies by real-time light spot position measurement. *Biotelemetry* 1974; 1(3): 132–146.
4. Yokokohji Y, Kitaoka Y, Yoshikawa T. Motion capture from demonstrator's viewpoint and its application to robot teaching. In: Proceedings of the 2002 IEEE International Conference on Robotics and Automation; 11–15 May 2002; Washington, DC, USA. pp. 1551–1558. doi: 10.1109/ROBOT.2002.1014764
5. Anisfield N. Ascension technology puts spotlight on DC field magnetic motion tracking. *HP Chronicle* 2000; 17(9).
6. Miller N, Jenkins OC, Kallmann M, Mataric MJ. Motion capture from inertial sensing for untethered humanoid teleoperation. In: Proceedings of the 4th IEEE/RAS International Conference on Humanoid Robots; 10–12 November 2004; Santa Monica, CA, USA. pp. 547–565. doi: 10.1109/ICHR.2004.1442670
7. Yi X, Zhou Y, Habermann M, et al. Physical inertial poser (PIP): Physics-aware real-time human motion tracking from sparse inertial sensors. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 18–24 June 2022; New Orleans, LA, USA. pp. 13157–13168. doi: 10.1109/CVPR52688.2022.01282
8. Hazas M, Ward A. A novel broadband ultrasonic location system. In: Borriello G, Holmquist LE (editors). *UbiComp 2002: Ubiquitous Computing*, Proceedings of the 4th International Conference on Ubiquitous Computing; 29 September–1 October 2002; Göteborg, Sweden. Springer-Verlag Berlin; 2002. Volume 2498, pp. 264–280. doi: 10.1007/3-540-45809-3_21
9. Lai J, Luo C. AcousticPose: Acoustic-based human pose estimation. In: Cui L, Xie X (editors). *Wireless Sensor Networks*, Proceedings of the 15th China Conference on Wireless Sensor Networks; 22–25 October 2021; Guilin, China. Springer Singapore; 2021. Volume 1509, pp. 57–69. doi: 10.1007/978-981-16-8174-5_5.
10. Laurijssen D, Truijen S, Saeys W, et al. An ultrasonic six degrees-of-freedom pose estimation sensor. *IEEE Sensors Journal* 2017; 17(1): 151–159. doi: 10.1109/JSEN.2016.2618399
11. Foxlin E, Harrington M, Pfeifer G. Constellation™: A wide-range wireless motion-tracking system for augmented reality and virtual set applications. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques; 19–24 July 1998; Orlando, Florida, USA. pp. 371–378. doi: 10.1145/280814.280937
12. von Marcard T, Henschel R, Black MJ, et al. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (editors). *Computer Vision—ECCV 2018*, Proceedings of the 15th European Conference on Computer Vision; 8–14 September 2018; Munich, Germany. Springer Cham; 2018. Volume 11214, pp. 614–631. doi: 10.1007/978-3-030-01249-6_37
13. Kanazawa A, Zhang JY, Felsen P, Malik J. Learning 3d human dynamics from video. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 15–20 June 2019; Long Beach, CA, USA. pp. 5614–5623. doi: 10.1109/CVPR.2019.00576
14. Schreiner P, Perepichka M, Lewis H, et al. Global position prediction for interactive motion capture. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2021; 4(3): 1–16. doi: 10.1145/3479985
15. Zhang Z, Siu K, Zhang J, et al. Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. *ACM Transactions on Graphics* 2014; 33(6): 1–14. doi: 10.1145/2661229.2661286
16. Cheng Y, Wang B, Yang B, Tan RT. Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 20–25 June 2021; Nashville, TN, USA. pp. 7645–7655. doi: 10.1109/CVPR46437.2021.00756
17. Zhang Y, An L, Yu T, et al. 4D association graph for realtime multi-person motion capture using multiple video cameras. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 13–19 June 2020; Seattle, WA, USA. pp. 1321–1330. doi: 10.1109/CVPR42600.2020.00140
18. Zhou Z, Shuai Q, Wang Y, et al. QuickPose: Real-time multi-view multi-person pose estimation in crowded scenes. In: Proceedings of SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference; August 7–11 2022; Vancouver, BC, Canada. pp. 1–9. doi: 10.1145/3528233.3530746
19. Yang W, Li Y, Xing S, et al. Lightweight multi-person motion capture system in the wild (Chinese). *SCIENTIA SINICA Informationis* 2023. doi: 10.1360/SSI-2022-0397
20. Duan H, Wang J, Chen K, Lin D. PYSKL: Towards good practices for skeleton action recognition. In:

Proceedings of the 30th ACM International Conference on Multimedia; 10–14 October 2022; Lisboa, Portugal. pp. 7351–7354. doi: 10.1145/3503161.3548546