Research on Computer Vision 3D Reconstruction and Interactive Perception Technology for the Metaverse

Hongyi Li
Sino-French Institute of Nuclear Engineering & Technology, Sun Yat-sen University, Zhuhai, China

Abstract: As an emerging field where virtual and reality are deeply integrated, the metaverse has put forward higher requirements for computer vision 3D reconstruction and interactive perception technologies. This paper delves deeply into the computer vision 3D reconstruction technology for the metaverse, including the principles and processes of 3D reconstruction based on multi-view geometry, deep learning and other methods, and analyzes its application advantages and challenges in the construction of metaverse scenes. Meanwhile, the interactive perception technology was studied, covering key technologies such as gesture recognition, pose estimation, and eye-tracking, as well as their significant roles in achieving natural interaction in the metaverse. In addition, the integrated development of 3D reconstruction and interactive perception technologies, as well as their future development trends and application prospects in the metaverse, was also discussed, aiming to provide theoretical support and technical references for the further development of metaverse-related technologies.

Keywords: Metaverse; Computer Vision; Three-Dimensional Reconstruction; Interactive Perception.

1. Introduction

The metaverse, as a comprehensive concept integrating multiple technologies such as virtual reality, augmented reality, blockchain, and artificial intelligence, is gradually becoming a popular research direction in the field of science and technology [1]. It aims to build a highly interactive and immersive virtual space with the real world, providing users with brand-new experiences and interaction methods. In the metaverse, the construction of 3D scenes and the realization of natural interaction are key elements, and computer vision 3D reconstruction and interactive perception technologies are important supports for achieving these goals.

Computer vision 3D reconstruction technology can restore the three-dimensional shape and structure of objects from two-dimensional images or videos, providing realistic virtual scenes for the metaverse[2]. Interactive perception technology enables users to interact with virtual environments in a natural way in the metaverse, enhancing their sense of immersion and participation[3]. Therefore, in-depth research on computer vision 3D reconstruction and interactive perception technology oriented to the metaverse has significant theoretical significance and practical application value.

2. Computer Vision 3D Reconstruction Technology for the Metaverse

2.1 3D Reconstruction Based on Multi-view Geometry

Multi-view geometry is an important theory in computer vision used to restore three-dimensional structures from images with multiple perspectives [4]. The basic principle is to utilize the corresponding relationship between images from different perspectives and calculate the three-dimensional coordinates of the object through methods such as triangulation.

2.1.1 Feature extraction and matching

Before conducting 3D reconstruction, it is necessary to extract feature points from images from different perspectives and find the corresponding relationships of these feature points among different images. Common feature extraction algorithms include SIFT (Scale-Invariant Feature Transform), SURF (Accelerated Robust Feature), ORB (Oriented FAST and Rotated BRIEF), etc. [5] These algorithms can extract feature points with scale invariance, rotation invariance and illumination invariance, thereby improving the accuracy of feature matching.

2.1.2 Camera calibration

Camera calibration is the process of determining

the internal and external parameters of the camera [6]. Internal parameters include focal length, principal point coordinates, etc., while external parameters include the position and posture of the camera. Through camera calibration, the image coordinate systems from different perspectives can be unified into the world coordinate system, providing accurate geometric information for subsequent 3D reconstruction.

2.1.3 3D Reconstruction algorithm

The 3D reconstruction algorithms based on multi-view geometry mainly include SfM (Motion Recovery Structure) and MVS (Multi-view Stereo), etc. [7] The SfM algorithm estimates the motion trajectory of the camera and the three-dimensional structure of the scene by analyzing the motion information of feature points in the image sequence. The MVS algorithm, based on SfM, utilizes the depth information of multiple images to generate a dense three-dimensional point cloud model.

2.2 3D Reconstruction based on Deep Learning

In recent years, deep learning has achieved great success in the field of computer vision and has also been widely applied in 3D reconstruction tasks [8]. 3D reconstruction methods based on deep learning can be classified into direct regression method, voxel method, point cloud method and mesh method, etc. The direct regression method directly regents the three-dimensional coordinates or depth information of an object from the input image through a neural network [9]. This method is simple and straightforward, but it requires a large amount of labeled data for training and has limited reconstruction effects on complex scenes. The voxel method divides the three-dimensional space into regular voxel grids, and then predicts through neural networks whether each voxel is occupied by an object [10]. This method can generate regular three-dimensional models, but it involves a large amount of calculation and the resolution is limited by the size of the voxel mesh. The point cloud method directly generates a three-dimensional point cloud model, and each point contains three-dimensional coordinates and possible color information. The point cloud method has the advantages of high flexibility and concise representation, but point cloud data is usually disordered and relatively complex to process. The grid method generates a three-dimensional grid model of an object through neural networks. The grid model is composed of vertices and patches and can represent the surface shape of the object more accurately. The difficulty of the grid method lies in how to generate high-quality and non-self-intersecting grid models.

2.3 Application of 3D Reconstruction Technology in the Construction of Metaverse Scenes

In the metaverse, 3D reconstruction technology can be used to construct various virtual scenes, such as urban buildings, natural landscapes, indoor environments, etc. By 3D reconstructing scenes from the real world and importing them into the metaverse, a more realistic and immersive experience can be provided for users. Meanwhile, 3D reconstruction technology can also be used to create virtual characters and objects, enriching the content of the metaverse.

However, 3D reconstruction technology for the metaverse also faces some challenges. For example, how to improve the accuracy and efficiency of reconstruction, how to handle the 3D reconstruction of large-scale scenes, and how to achieve the 3D reconstruction of dynamic scenes, etc.

3. Interactive Perception Technology for the Metaverse

In the metaverse, a digital world where the virtual and the real are deeply integrated, interactive perception technology acts as a bridge connecting users with the virtual environment, and it is a core element for achieving natural and immersive interactive experiences. It can precisely capture various behaviors and intentions of users and transform them into specific operations and feedback in the metaverse, making users feel as if they were in a real and fantastical virtual space. The following will delve into three key interactive perception technologies: gesture recognition, pose estimation, and eye movement tracking.

3.1 Gesture Recognition Technology

Gesture recognition technology provides users with an intuitive and natural way of interaction. In the metaverse, users do not need to operate complex devices. They can easily control the movement, rotation and zooming of virtual objects through simple gesture actions, and also quickly select and operate menus, greatly

enhancing the convenience and fun of interaction.

Vision-based gesture recognition mainly relies on cameras to collect gesture images, and then uses image processing and machine learning algorithms to identify the types and postures of gestures. In the early days, gesture segmentation based on skin color models was more commonly used. By establishing a skin color model, it segmented the regions in the image that conformed to the skin color features and initially determined the gesture positions. However, this method is vulnerable to the influence of lighting conditions and background complexity, and the segmentation effect is unstable. Later, gesture classification methods based on feature extraction emerged. They extract key features such as contours, edges, and textures of gesture images and classify them using machine learning algorithms, which improves the recognition accuracy. In recent years, deep learning technology has brought new breakthroughs to this field. Deep learning models can automatically learn the complex features of gesture images. For example, after being trained with a large amount of gesture image data, convolutional neural networks (CNNS) can quickly and accurately recognize different gestures, and have stronger robustness.

In addition to vision-based methods, using sensors to obtain gesture information is also very effective. Data gloves are typical devices that can precisely measure the bending angle and position information of fingers. By installing multiple sensors on the gloves, the real-time movement status of the fingers can be obtained, and the data can be transmitted to the computer for processing and analysis, achieving high-precision gesture recognition. This is suitable for scenarios with high requirements for interaction accuracy, such as virtual surgery and virtual assembly. Sensors such as inertial measurement units (IMUs) can also be used for gesture recognition. They can measure motion parameters of objects such as acceleration and angular velocity, and infer the motion trajectory and posture of gestures by analyzing the changes in these parameters. IMU has the advantages of small size, low cost and low power consumption. It is often integrated into wearable devices to provide users with a convenient gesture interaction experience.

3.2 Pose Estimation Technology

Pose estimation technology aims to estimate the pose information of the human body or object, including position, direction, and joint angles, etc. It is a key technology for achieving the synchronization of virtual character movements and user body interaction in the metaverse.

Pose estimation methods based on monocular cameras usually utilize deep learning models to predict the three-dimensional poses of human bodies or objects from two-dimensional images. This method is low-cost and easy to deploy, but its accuracy is relatively low and it is vulnerable to occlusion and changes in viewing angles. For example, when a part of the human body is occluded, the deep learning model may not be able to accurately predict the pose information of the occluded part. To improve the accuracy, researchers have proposed improved methods, such as introducing a multi-task learning framework to simultaneously learn tasks like pose estimation and human body segmentation, and utilize the correlations among tasks to enhance performance. The timing information can also be utilized to analyze continuous video frames and estimate the pose changes more accurately.

To further improve the accuracy of attitude estimation, multi-cameras or depth cameras are widely used. The multi-camera system calculates the three-dimensional coordinates of the human body or object through methods such as triangulation to obtain more accurate posture information. By installing multiple cameras at different positions to capture human bodies or objects from multiple perspectives, and then using computer vision algorithms to process and analyze the images, three-dimensional postures are calculated. Depth cameras can directly obtain the depth information of objects, providing more abundant data for attitude estimation. It acquires depth information by emitting infrared light and measuring the reflection time, and then fuses the depth information with the color image to achieve three-dimensional pose estimation. This method has high accuracy and strong robustness, and is suitable for scenarios with high requirements for pose estimation accuracy, such as virtual fitness and virtual dance.

3.3 Eye-Tracking Technology

Eye-tracking technology can monitor the user's eye movements in real time, obtain information on the direction of the gaze and the point of fixation, and provide a more natural and efficient

way for interaction in the metaverse.

The video-based eye tracking method collects the user's eye images through the camera, uses the image processing algorithm to detect the positions of the pupil and corneal reflection points, and calculates the direction of the line of sight. This method is low in cost and easy to implement, but it is susceptible to lighting conditions and head movements. The pupils constrict in a strong light environment, making detection difficult. Large movements of the head can cause significant changes in the eye image and affect the accurate calculation of the line of sight direction. To improve accuracy and robustness, researchers have proposed improved algorithms, such as adopting an adaptive pupil detection algorithm to automatically adjust the detection parameters according to the lighting conditions; The head pose estimation technology is utilized to compensate for head movements and reduce the impact on eye movement tracking.

Sensor-based eye tracking methods use specialized eye trackers to measure eye movements. The eye tracker can provide precise eye movement data and accurately record the trajectory of eye movement and the information of fixation points in real time. It usually uses infrared light sources and optical sensors to detect eye movements, and calculates the position and direction of eye movement by analyzing the reflection of infrared light on the surface of the eye. However, this method is costly and inconvenient to use. Eye trackers require professional equipment and technical support. They are usually used in specific experimental environments, and users need to wear specialized devices, which may cause discomfort. Therefore, they are mainly applied in professional fields such as scientific research and medical care.

4. The Integrated Development of 3D Reconstruction and Interactive Perception Technologies

4.1 The Necessity of Technological Integration

In the metaverse, 3D reconstruction and interactive perception technologies are interdependent and mutually reinforcing. 3D reconstruction technology provides virtual scenes and objects for interactive perception, and interactive perception technology enables users to interact naturally with the 3D reconstructed virtual environment, thereby further enriching and improving the content and experience of the metaverse. Therefore, it is of great significance to achieve the integrated development of 3D reconstruction and interactive perception technology.

4.2 Ways of Integrated Development

At the data level, the geometric information obtained from 3D reconstruction and the user behavior information acquired through interactive perception can be fused. For example, based on the user's gesture and posture information, operations and deformations are performed on the 3D reconstructed virtual objects; Adjust the display content and perspective of the virtual scene based on the user's eye movement information.

At the algorithmic level, the 3D reconstruction algorithm and the interactive perception algorithm can be integrated and optimized. For example, the user feedback information obtained by using interactive sensing technology is utilized to guide the 3D reconstruction process and improve the accuracy and efficiency of the reconstruction; or the result of 3D reconstruction can be used as the input of the interactive sensing algorithm to enhance the accuracy and robustness of interactive sensing.

4.3 Challenges Faced in Integrated Development

The integrated development of 3D reconstruction and interactive perception technology also faces some challenges. For example, how to achieve data sharing and interaction among different technologies, how to handle the efficient processing and storage of large-scale data, and how to ensure the real-time performance and stability of the integrated system, etc.

5. Development Trends of 3D Reconstruction and Interactive Perception Technologies for the Metaverse

5.1 High Precision and High Efficiency

In the future, 3D reconstruction and interactive perception technologies will unswervingly move towards the goals of high precision and high efficiency. In terms of 3D reconstruction, the optimization of the algorithm will become a key driving force. Traditional 3D reconstruction

algorithms often have problems such as insufficient accuracy or excessively long computing time when dealing with complex scenes. By introducing advanced technologies such as deep learning, the algorithm can automatically learn the features and patterns in the scene, thereby capturing the shape, texture and detailed information of objects more accurately. For example, using convolutional neural networks (CNNS) to train a large amount of 3D model data can make the reconstruction results more realistic and significantly reduce the error. Meanwhile, the emergence of new hardware devices such as high-performance graphics processing units (Gpus) and dedicated acceleration chips will further enhance computing speed, reduce computational complexity, and enable real-time or near-real-time 3D reconstruction, providing users with a smoother experience.

In the field of interactive perception, high precision means being able to capture users' actions, expressions and intentions more accurately. By optimizing sensor technology and algorithms to enhance the accuracy of data collection and the efficiency of processing, the system can respond to users' operations in real time and reduce delays. For example, in virtual reality (VR) games, high-precision gesture recognition and pose estimation technologies can enable players to interact with the virtual environment more naturally, as if they were in the real game world.

5.2 Multimodal Fusion

Multimodal integration will become an important trend in future development, bringing more realistic and immersive experiences to the metaverse. At present, the interaction in the metaverse mainly relies on visual information, but the future development direction is to integrate multiple sensory information such as touch, hearing and smell. The development of haptic feedback devices will enable users to feel the texture and weight of virtual objects, enhancing the sense of reality in interaction. For instance, when a user touches a virtual apple in the metaverse, the haptic feedback device can simulate the surface texture and hardness of the apple, giving the user a sense of being right there.

Sound positioning technology will also bring more realistic spatial sound effects to the metaverse. By precisely calculating the direction and distance of sound propagation, users can clearly identify the source of the sound and enhance their spatial perception ability. In addition, the introduction of olfactory simulation technology will further enrich the sensory experience of the metaverse, enabling users to smell odors in virtual environments, such as floral scents and the aroma of food, making the metaverse closer to the real world.

5.3 Intelligence and Adaptability

With the continuous development of artificial intelligence technology, 3D reconstruction and interactive perception technology will become more intelligent and adaptive. The system can automatically learn the behavioral habits and preferences of users, and automatically adjust parameters and strategies according to different scenarios and tasks. For example, in a virtual meeting scenario, the system can automatically adjust the atmosphere and pace of the meeting based on the expressions and voice information of the participants, providing more personalized services. In terms of interaction perception, the system can dynamically adjust the difficulty and mode of interaction based on the user's operation habits and skill levels, enabling users to use the metaverse more easily.

5.4 Cross-platform and Standardization

To achieve the interconnection and wide application of the metaverse, 3D reconstruction and interactive perception technologies need to be cross-platform and standardized. Different devices and systems can be seamlessly integrated and interacted with each other, and the data format and interface standards are unified. This will promote the healthy development of the metaverse industry, reduce development costs and improve development efficiency. For example, developers can use a unified standard interface to integrate 3D reconstruction and interactive perception technologies into different platforms and devices without having to develop adaptive code separately for each platform. At the same time, cross-platform compatibility will also enable users to access the metaverse anytime and anywhere on different devices, enjoying consistent services and experiences.

6. Application Prospects

Computer vision 3D reconstruction and interactive perception technologies for the metaverse have demonstrated broad and highly

potential application prospects in multiple fields. In the field of entertainment and gaming, this technology will reshape the gaming experience. Players are no longer confined to traditional operation methods. Instead, they can fully immerse themselves in the virtual game world with the help of natural gestures and postures. Whether it's controlling characters to adventure in fantasy scenes or interacting with various elements in the game, it can significantly enhance the game's fun and immersion, bringing players an unprecedented entertainment experience.

The field of education and training will also undergo changes due to this technology. The metaverse can create a vivid and intuitive teaching environment. With the help of 3D reconstruction technology, realistic virtual scenes and models can be presented. Students and trainees can carry out practical operations and simulation training in such an environment, as if they were in a real working scene, which greatly improves the teaching effect and training quality, making learning more efficient and interesting.

The field of industrial design and manufacturing has also benefited greatly. The metaverse offers a brand-new virtual platform for product design and manufacturing. Designers can conduct 3D modeling and interactive design in a virtual environment, view the appearance and performance of products in real time, and optimize and improve them promptly. In addition, it can also achieve remote collaboration and training, break geographical restrictions, improve production efficiency and quality, and promote the development of industrial design and manufacturing towards intelligence and efficiency.

In the field of medical health, this technology is of great significance. By 3D reconstructing models of human organs and tissues, doctors can conduct surgical operation training in a virtual environment, enhancing surgical skills and safety. Meanwhile, interactive perception technology facilitates remote medical diagnosis and treatment, enabling patients to enjoy convenient and professional medical services at home without having to travel long distances, thus injecting new vitality into the development of the medical and health industry.

7. Conclusion

This paper conducts an in-depth study on the 3D reconstruction and interactive perception technology of computer vision for the metaverse. The 3D reconstruction method based on multi-view geometry and deep learning is introduced, and its application advantages and challenges in the construction of metaverse scenes are analyzed. The interactive perception technologies such as gesture recognition, pose estimation, and eye movement tracking were discussed, as well as their important roles in achieving natural interaction in the metaverse; The integrated development mode and challenges faced by 3D reconstruction and interactive perception technology were studied; The development trends and application prospects of future technologies were prospected.

Computer vision 3D reconstruction and interactive perception technology for the metaverse is a key supporting technology for the development of the metaverse. Although certain progress has been made at present, many challenges and problems are still faced. In the future, it is necessary to further enhance the research and innovation of related technologies, promote the integrated development of 3D reconstruction and interactive perception technologies, and provide a more solid technical foundation for the realization and application of the metaverse. At the same time, it is also necessary to enhance cross-disciplinary cooperation and exchanges to promote the healthy development of the metaverse industry and bring a brighter future to human society.

References

[1] Kozinets, R. V. (2023). Immersive netnography: a novel method for service experience research in virtual reality, augmented reality and metaverse contexts. Journal of Service Management, 34(1), 100-125.

[2] Yanwen, Z., Kai, H., & Pengsheng, W. (2020). Review of 3D reconstruction algorithms. Nanjing Xinxi Gongcheng Daxue Xuebao, 12(5), 591-602.

[3] Smith, J. W. (2015). Immersive virtual environment technology to supplement environmental perception, preference and behavior research: a review with applications. International journal of environmental research and public health, 12(9), 11486-11505.

[4] Hartley, R. (2003). Multiple view geometry

in computer vision (Vol. 665). Cambridge university press.

[5] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60, 91-110.

[6] Zhang, Z. (2002). A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence, 22(11), 1330-1334.

[7] Schonberger, J. L., & Frahm, J. M. (2016). Structure-from-motion revisited. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4104-4113).

[8] Choy, C. B., Xu, D., Gwak, J., Chen, K., & Savarese, S. (2016). 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11-14, 2016, proceedings, part VIII 14 (pp. 628-644). Springer International Publishing.

[9] Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems, 27.

[10] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015). 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1912-1920).