Objective Metrics and Theoretical Model for Evaluating the Spatial Reality Reproduction Performance of Head-Mounted Display

Liang Gu, LongYun Xiao*, WangZan Jin*, QiNing Wang**, CongShan Rui**, ShiSen Yan*, Tin Pan*, HongJun Fang*, Lei Zhao**

*GravityXR Electronics and Technology Co., Ltd., Ningbo, Zhejiang, China

**Yongjiang Laboratory, Ningbo, Zhejiang, China

Abstract

This paper explores the issue of spatial reality reproduction (SRR) that arises from the inherent differences between the human eye and the camera viewpoint in video see-through (VST) of the head-mounted display. We present a set of indicators and methods to objectively evaluate the performance of SRR. These indicators include scaling, positioning, distortion, motion distortion, and stereo depth. Additionally, we delineate the corresponding measurement equipment and methods. A comprehensive evaluation of these indicators can help us establish a theoretical model, which may incorporate an appropriate weighting for each indicator. This model could facilitate the calculation of the final score through a weighted average, enabling us to assess the strengths and weaknesses of the headset's SRR performance.

Author Keywords

Video See-Through; Spatial Reality Reproduction(SRR); Mixed Reality; Objective Metrics and Theoretical Model of SRR

1. Introduction

The rapid advancements in information technology have led to a significant increase in the use of augmented reality (AR) and virtual reality (VR) across various disciplines. Typically, AR utilizes optical see-through (OST) technology, allowing users to perceive both virtual and real images simultaneously. In contrast, VR integrates virtual and tangible elements through the use of VST.

In an ideal MR system, there should be no perceptible difference between the user's natural world view and the enhanced view of the video perspective. However, in VST technology, the user's visual perception of the real world is primarily derived from the acquisition camera and the near-eye display system, resulting in a discrepancy between the direct visual experience and that of the enhanced view. As shown in Figure 1, the target perceived by the VST at close range may appear slightly larger than the actual target, and its position may be slightly offset. The inherent discrepancy between the human eye and camera viewpoint is likely to give rise to user perception errors, hand-eye incoordination, and other discomforting issues.

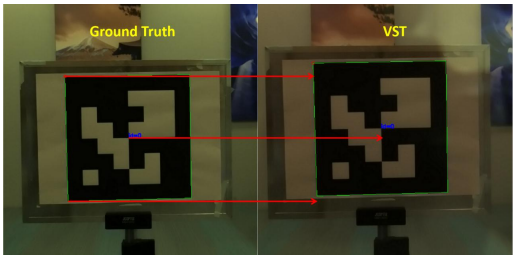


Figure 1. Schematic illustrating the comparison between Ground Truth and VST image.

Extended immersion in mixed reality (MR) can lead to users adapting to an inaccurate environment. When they return to the real world, they may experience a reverse distortion. This adaptation phenomenon is commonly observed in magnifying or reducing lenses, as well as in head-mounted displays when there is a discrepancy between the rendered field of view and the physical field of view[1]. Before users readjust to reality, they may experience aftereffects such as drowsiness, disturbances in motor and postural control, and reduced hand-eye coordination[2].The reproduction accuracy of the rendered field of view in MR systems is crucial, and we refer to this process as spatial reality reproduction (SRR)[3].

SRR can be corrected to compensate for this by detecting the depth of objects in the scene and reprojecting the camera view to the eye position[4,5,6]. However, depth estimation errors and a lack of occlusion boundary information lead to distortions in perspective images, and multi-depth target scenes can distort object edges where they overlap. Many solutions have been proposed in the industry to address these problems. Chaurasia et al.[7] proposed an end-to-end perspective algorithm in which the depth is estimated by a pair of stereo cameras. However, the binocular stereo depth has a large error at medium and long distances. Xiao et al.[8] proposed a neural network perspective approach using modern machine learning techniques to improve the depth estimation and fill in the missing information of occluded regions, but this algorithm is not useful for mobile applications. Meta Reality Labs Research[9] even proposes a light field perspective scheme of compound eye to correct the viewpoint difference between the human eye and the camera, but it still causes problems such as VST clarity degradation and FOV reduction.

Different SRR algorithm strategies yield different video perspective effects, and an intuitive user experience is insufficient to address the multi-dimensional problems. Therefore, there is an urgent need for a set of theoretical models to evaluate them objectively. In this regard, we have extracted the five indicators that users are most concerned about in the SRR, and we have developed equipment and designed experiments to test them. We use the Apple Vision Pro(AVP) and Quest3 as examples, and the data for the five indicators can be found in the following sections of this paper. We add their respective weights to the five indicators and form a theoretical model after weighting and calculation. This model allow us to objectively evaluate the SRR performance of the headset.

2. System Overview

The disparity between the viewpoint of the human eye and the camera is the direct cause of the SRR problem, as illustrated in Figure 2. Typically, the VST camera is positioned in front of the lower part of the human eye point, so the size and position of its shooting target must be different from that seen by the eye point. As a result, when the spatial reality is reproduced, the lack of to

ISSN 0097-996X/25/5602-0778-$1.00 © 2025

mapping algorithms to correct and compensate for this ultimately results in the difference between the physically real image and the VST image.

The distinction between the physically real image and the VST image can typically be evaluated through subjective means. Human perception is highly sensitive to changes in visual imagery prior to and following headset use. However, the human factors approach can only be evaluated qualitatively. To address this limitation, we have proposed a series of objective evaluation indicators, which form a theoretical model used to quantify the SRR performance of the headset. This model includes scaling, positioning, distortion, motion distortion, and stereo depth. In addition, we have developed an exclusive binocular camera system and designed experiments to test the SRR performance of AVP&Quest3. The following is a summary of these experiments.

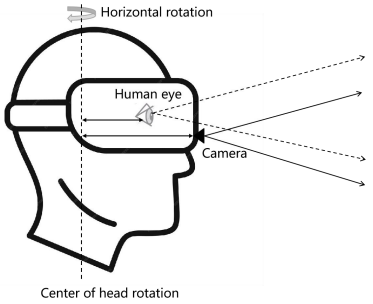


Figure 2. Schematic illustrating the differences between center of head rotation, human eye, and camera position.

(a) Metrics for Objective Evaluation of SRR Performance.

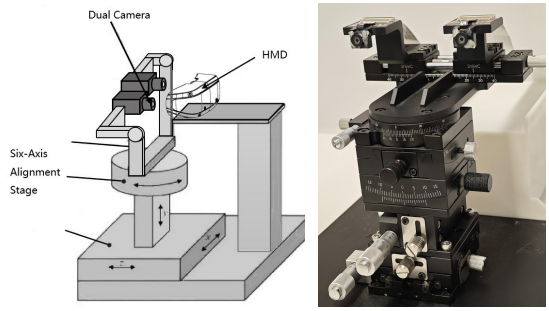
We propose here some objective metrics that can quantify the performance of SRR, which are presented in turn as follows:

1. Scaling: the ratio of virtual/real object sizes at different distances.
2. Positioning: the relative difference in the spatial offset of virtual/real objects at different distances.
3. Distortion: shape variables and their ranges for distortions of the VST with respect to the ground truth reference system for still observation.
4. Motion Distortion: variations in the rotation angle of the VST relative to the ground truth reference system as the simulated human head moves.
5. Stereo Depth: the binocular stereo depth of the VST target is measured at different distances, and this measurement includes axial accuracy and in-plane precision.

(b) Measurement Equipment and Methods.

We have developed a binocular camera system to test the SRR performance of the headset. The two cameras are designed to mimic the human eye, ensuring that their optical axes are aligned in the forward direction. This device can align the headset and the test cameras binocularly, allowing data collection from both the left and right eyes simultaneously, thus significantly improving testing efficiency. We have established strict requirements for the displacement and angular accuracy of the binocular device, demanding precision within 0.01 mm and 0.01° respectively.

When no headset is in place, the binocular camera captures an image of the ground truth reference system. Once the headset is placed and binocular alignment is achieved, it can simulate the user's experience with the headset and capture the VST image as the test value.



(a)Schematic diagram (b)Real equipment

Figure 3. Schematic and physical drawings of equipment.

By utilizing this set of binocular camera equipment, we can simulate the user's shooting experience both before and after wearing the headset to view the screen. This allows us to process the data with algorithms to obtain a comparison between virtual and real object targets. The results are summarized into five key indicators that reflect the performance of the SRR. The specific method does not require excessive elaboration.

(c) Test Results of The AVP&Quest3.

We collected data on scaling, positioning, distortion, motion distortion, and stereo depth for AVP&Quest3. The two performed very differently on these metrics.

1. As shown in Figure 4, the ratio of VST to GT in AVP exceeds 1 at a distance of 0.5 meters. There is a gradual decrease in this ratio as it approaches 1, while the entire range for Quest3 remains close to 1.

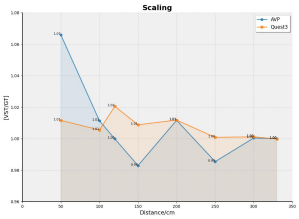
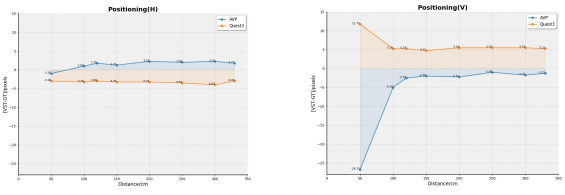


Figure 4. The scaling of AVP&Quest3.

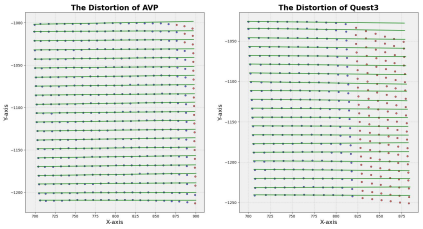
2. In terms of positioning, as illustrated in Figure 5, both of them exhibit slight horizontal deviations in pixel values, while larger vertical deviations occur at close range. This is primarily due to the downward mounting of the VST cameras on the headset. AVP reaches 27 pixels at a distance of 0.5 meters, corresponding to an angle of 1.755°.



(a)Horizontal direction (b)Vertical direction

Figure 5. The positioning of AVP&Quest3.

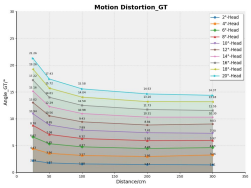
3. As illustrated in Figure 6, distortion occurs in the edge junction area of shooting targets at different depths. In the Quest3, the distortion between near and far targets becomes particularly noticeable when the distance between them exceeds 3 meters, showing nearly three times more distortion compared to the AVP.



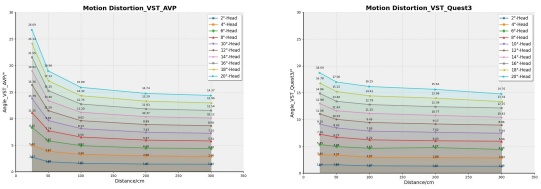
(a)AVP (b)Quest3

Figure 6. The distortion of AVP&Quest3.

4. In the analysis of motion distortion shown in Figure 7, we simulate the horizontal rotation of the human head and observe the rotation angle of a target point within the camera's field of view at various distances. Our findings reveal that the rotation angle displayed on the VST of the AVP significantly differs from the ground truth, particularly at closer distances, where it is noticeably larger. In contrast, the Quest3 shows an angle that is slightly smaller than the ground truth. However, the rotation angles for both the VST and the ground truth are essentially the same.



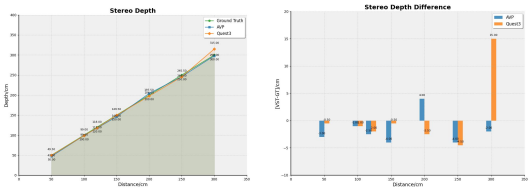
(a)Ground Truth



(b)AVP (c)Quest3

Figure 7. The motion distortion of AVP&Quest3.

5. Regarding stereo depth, as illustrated in Figure 8, both models demonstrate satisfactory performance, with the Quest3 exhibiting a deviation of 15 centimeters at a distance of 3 meters.



(a)The stereo depth (b)The stereo depth difference

Figure 8. The stereo depth of AVP&Quest3.

3. The Theoretical Model for Evaluating SRR Performance

Based on the likelihood of use in each scenario, distortion emerges as the most significant factor on user experience, followed by positioning, scaling, stereo depth, and motion distortion, in that order. Weighted metrics from previous sections were used to create a theoretical model for comprehensively evaluating the performance of the headset's SRR. Given the varying priorities of these metrics for user experience, we assigned a score to each metric and established the corresponding weight coefficients.

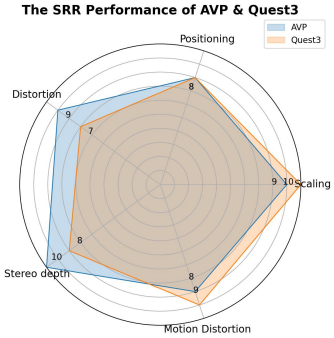


Figure 9. The score for each metric of AVP&Quest3.

First, an analysis of AVP's performance in these five indicators is warranted. As demonstrated in Figure 9, its SRR clearly demonstrates a notable advantage in distortion, indicating a potential optimization of AVP's algorithm strategy in this area. According to the test data on scaling and positioning, it seems that the SRR algorithm utilizes a strategy involving dual spheres or dual planes. This approach ensures that the VST image closely resembles the physical real image at mid-range and long distances. This approach involves sacrificing the accuracy of scaling and positioning at close distances to reduce the impact of the algorithm on distortion.

Secondly, Quest3 employs an alternative SRR algorithm strategy that effectively handles scaling and accurately restores the object's true dimensions. This results in scaling ratio values that closely approximate 1 at various distances. However, this approach can lead to significant distortion of targets with different depths, particularly when the depth disparity between the near and far targets is substantial. The area where the edges of the two targets meet shows significant distortion, as illustrated in Figure 10.

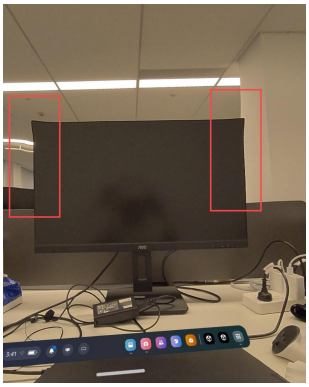


Figure 10. Schematic illustrating the distortion of Quest3's VST image.

The core strategy of the SRR algorithm is to adjust the scaling ratios of feature targets with different depths. Two issues must be addressed: First, acquiring the depth information of the feature target is essential. Second, the design of the curve of the [target depth]-[scaling ratio] is crucial. Various sensor solutions can be used to acquire the depth information, including DTOF, structured light, and binocular camera. The scaling ratio is then set according to the target's depth, and the curve can be a naive SRR, which uses a fixed scaling ratio across the entire depth range, or a biplane or bisphere combination line, which assigns a scaling curve based on the target's depth. Alternatively, a more direct approach involves setting the scaling curve to restore the size of each depth target to match the real object. Figure 11 shows three kinds of scaling curves.

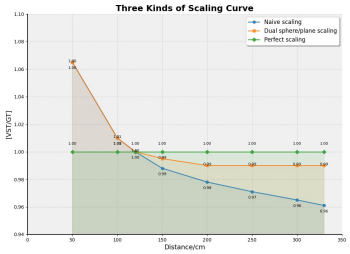


Figure 11. Schematic of the three kinds of scaling curves.

Ultimately, the SRR algorithm is unable to perfectly restore the VST image to match the physical real image due to differences in the depth of the target within the field of view.There is a trade-off between scaling and distortion, which necessitates a careful consideration of strategy. We must prioritize the user experience and optimize the indicators that carry higher weights. By aligning with the algorithmic priorities, we assign corresponding weights to each indicator and compute a weighted score to evaluate the overall performance of the SRR in both headsets.

Table 1. Overall rating of AVP&Quest3

	Evaluation metrics	Initial scores for each item/1-10	Weight for each item	Final scores for each item	Overall rating
AVP	Distortion	9	0.3	2.7	8.8
	Positioning	8	0.2	1.6	
	Scaling	9	0.2	1.8	
	Stereo depth	10	0.15	1.5	
	Motion distortion	8	0.15	1.2	
Quest3	Distortion	7	0.3	2.1	8.25
	Positioning	8	0.2	1.6	
	Scaling	10	0.2	2	
	Stereo depth	8	0.15	1.2	
	Motion distortion	9	0.15	1.35	

As shown in Table 1, AVP outperforms Quest3 in terms of overall SRR performance. However, it's important to note that the two algorithms have different focuses. AVP prioritizes the user's experience of distortion, which involves sacrificing scaling and positioning to some extent. In contrast, Quest3 focuses on optimizing scaling and restoring the dimensions of the real object in the VST image. AVP's VST image is characterized by reduced distortion, while Quest3's VST image demonstrates enhanced spatial stability. In terms of overall rating, AVP achieves a high score of 8.8, aligning with our user experience. Its SRR performance is also superior overall.

4. Conclusion

We have proposed a series of objective metrics to quantify the performance of SRR, including scaling, positioning, distortion, motion distortion, and stereo depth. Additionally, we have developed a binocular camera measurement system and designed experiments to test these metrics. Based on the data results of these metrics and their corresponding weight coefficients, we have established a theoretical model to objectively evaluate the performance of SRR in head-mounted displays. These indicators provide a more comprehensive understanding of how the SRR affects the VST. This information is beneficial for optimizing the SRR algorithm on the system side, making the VST image appear more realistic and enhancing the overall user experience by minimizing discomfort. We plan to conduct human factors experiments to optimize the theoretical model by assigning more reasonable weight coefficients to each index. This approach aims to bring subjective and objective evaluation results closer together, ultimately forming a closed loop.

5. References

1. Draper M H. The adaptive effects of virtual interfaces: vestibulo-ocular reflex and simulator sickness[M]. University of Washington, 1998.
2. Keshavarz B, Hecht H, Lawson B D. Visually Induced Motion Sickness: Causes, Characteristics, and Countermeasures[J]. 2014.
3. Aoyama K , Yokoyama K , Nakahata Y Y .48-5: Eye-sensing Light Field Display for Spatial Reality Reproduction[J].SID International Symposium: Digest of Technology Papers, 2021, 52(2):669-672.
4. Ishihara A, Aga H, Ishihara Y, et al. Integrating both parallax and latency compensation into video see-through head-mounted display[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(5): 2826-2836.
5. Guan P, Penner E, Hegland J, et al. Perceptual requirements for world-locked rendering in AR and VR[C]//SIGGRAPH Asia 2023 Conference Papers. 2023: 1-10.
6. Krajancich B, Kellnhofer P, Wetzstein G. Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering[J]. ACM Transactions on Graphics (TOG), 2020, 39(6): 1-10.
7. Chaurasia G, Nieuwoudt A, Ichim A E, et al. Passthrough+ real-time stereoscopic view synthesis for mobile mixed reality[J]. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 2020, 3(1): 1-17.
8. Xiao L, Nouri S, Hegland J, et al. Neuralpassthrough: Learned real-time view synthesis for vr[C]//ACM SIGGRAPH 2022 Conference Proceedings. 2022: 1-9.
9. Kuo G, Penner E, Moczydlowski S, et al. Perspective-Correct VR Passthrough Without Reprojection[C]//ACM SIGGRAPH 2023 Conference Proceedings. 2023: 1-9