



Assessing Depth Perception in Virtual Environments: A Comprehensive Framework


Sahra K. G. Silva   [Universidade Federal do Piauí | sahrask@gmail.com]

Cléber G. Corrêa  [Universidade Tecnológica Federal do Paraná | clebergimenez@utfpr.edu.br]

Silvio R. R. Sanches  [Universidade Tecnológica Federal do Paraná | silviosanches@utfpr.edu.br]

Marcelo S. Lauretto  [Universidade de São Paulo | marcelolauretto@usp.br]

Fátima L. S. Nunes  [Universidade de São Paulo | fatima.nunes@usp.br]

 Federal University of Piauí, High-Performance Computing Center, Minister Petrônio Portella University Campus, Ininga district, Teresina, PI, ZIP Code: 64049-550, Brazil.

Received: 06 July 2023 • **Accepted:** 18 October 2023 • **Published:** 01 January 2024

Abstract Understanding humans' perception of depth and how they interact with virtual environments is a challenging task. This context involves investigating how features of these environments affect depth perception, which is crucial for tasks like object manipulation and navigation that require interpreting spatial information. This article presents a comprehensive (general, extensible and flexible) framework to assess depth perception in different virtual environments to support the development of more effective and immersive virtual experiences. This approach can assist developers in decision-making regarding different approaches for assessing depth perception in virtual environments, considering stereoscopic and monoscopic techniques for visualization. The framework considers parameters such as the distance between the user and virtual objects and the sizes of virtual objects. Metrics such as hit rate, response time, and presence questionnaire responses were utilized to assess depth perception. The previous experiments are presented (anaglyph and shutter glasses), as well as the new experiments, considering cave environments with and without anaglyph glasses.

Keywords: Framework, Depth Perception Evaluation, Visualization Techniques, Stereoscopy

1 Introduction

An Immersive Virtual Environment (IVE) is a Virtual Reality (VR) system composed of a three-dimensional (3D) environment developed to provide interaction between a human participant and a world simulated by a computer [Slater and Usoh, 1993]. An IVE should offer participants or users the feeling of being in a world different than the one where their real bodies are physically located. This goal may be accomplished by using visual, auditory and haptic devices, which provide sensory inputs to users.

Understanding how people explore IVEs is crucial for many applications, such as designing VR content, developing new image compression algorithms, or learning computational models of saliency or visual attention [Sitzmann *et al.*, 2017]. Stereoscopic techniques, such as color filtering, light shutter and polarized light, are visualization resources that can offer the feeling of immersion in these environments.

Stereoscopy is the acquisition and projection of images of a scene to both the left and the right eyes at the same time for the conversion in a single image by brain [Yanoff and Duker, 2018]. This mechanism occurs because humans have binocular vision, so the two eyes capture two different images of a scene, and the brain interprets these images to provide the perception of the depth of the observed scene.

In a previous systematic reviewer [Silva *et al.*, 2016], we have found some studies concerning the assessment of some effects provided by stereoscopic techniques. Certain effects have been evaluated: immersion [McMahan *et al.*, 2006;

Slater *et al.*, 2010] and depth [Vinnikov and Allison, 2014; Livatino *et al.*, 2015]. The objective of evaluating such effects usually refers to checking the realism and usefulness of IVEs. The effects are studied according to parameters, called effects in some studies, such as distance, which refers to the interval perceived from the viewer to the target (egocentric) or from one target to another (exocentric) [Geuss *et al.*, 2012]. Immersion refers to how much technology can provide an inclusive, extensive, surrounding, and vivid illusion of reality to the senses of an observer [Slater and Usoh, 1993]. Finally, depth refers to the 3D visual perception of a scene [Armbrüster *et al.*, 2008].

The relationship between movement and vision in IVEs has been fairly explored under several evaluation approaches. The literature encompasses studies aimed at validating specific IVEs, investigating depth, distance, immersion and some variations [Cecotti, 2022; Hattori *et al.*, 2022; Leopardi *et al.*, 2021; Ochs *et al.*, 2019; Thalmann *et al.*, 2016; dos Santos *et al.*, 2017], as well as works aimed at investigating such effects from IVEs built exclusively to execute such evaluations [Ng *et al.*, 2016; Lin *et al.*, 2019; e Silva and Nunes, 2015]. However, we have not found studies proposing more general methods that could be replicated in different experiments.

In the literature, there is neither consensus on the use of stereoscopy in IVEs for the performance of some tasks, nor on the degree of adequacy of different stereoscopic techniques for different contexts. As stereoscopic techniques become more diverse, it is necessary to establish methods ca-

pable of measuring and comparing the depth perception provided by different techniques within different contexts, also comparing with monoscopic technique.

This article presents a framework to evaluate depth perception in different IVEs, comparing multiple visualization techniques, especially stereoscopic techniques. A developer can select objective and subjective metrics, as well as the features and a low number of users to test. Although some understanding of variables used in an IVE is desirable, the developer does not need previous experience in research, since a guide with examples and detailed explanation is provided together with the framework itself.

The paper is organized as follows: Section 2 discusses related work on the evaluation of stereoscopic techniques; Section 3 presents the framework; Section 4 presents two experiments conducted to illustrate the framework's application; the results are shown in Section 5; Section 6 indicates the framework benefits and constraints; and Section 7 presents some final remarks.

2 Related Work

Current literature presents evaluation studies to compare stereoscopic technologies in IVEs, as well as to validate hypotheses regarding the influence that depth perception has on the users' performance. Considering a first scope, some studies can be characterized by *ad-hoc* experiments designed and conducted within the context of each IVE [Leopardi et al., 2021; Ochs et al., 2019; Thalmann et al., 2016; dos Santos et al., 2017]. In contrast, considering a second scope, few studies make efforts to propose evaluation methods applied to a particular context, where IVEs are developed strictly to investigate specific effects and depth perception [Cecotti, 2022; Hattori et al., 2022; Lin et al., 2019; Zhao et al., 2020; Vienne et al., 2020].

The studies conducted by Cecotti [2022] and Hattori et al. [2022] are examples of the first-mentioned scope. According to Cecotti [2022], VR has a key impact on users' immersion in learning activities. In his work, a serious game in fully immersive VR related to astronomy education was proposed, which was assessed with undergraduate students. Hattori et al. [2022] evaluated users' performances in dental training simulators. They found that unique characteristics of VR, such as the simulated cutting sensation and the simulated 3D images created by stereo viewers, affect performance.

Considering the same scope, some studies adapt presence questionnaires available in previous literature [Slater et al., 1994; Witmer and Singer, 1998; Lessiter et al., 2001; Schubert et al., 2001] and use them to analyze subjective data, considering Likert scale. These studies usually evaluate IVEs like Virtual Medical training room [Ochs et al., 2019], Virtual Museum system [Leopardi et al., 2021], Virtual Volleyball game [Thalmann et al., 2016], comparing Head Mounted Display (HMD), desktop, Oculus Rift, Cave Automatic Virtual Environment (CAVE) and auto-stereoscopic display. There are few studies that analyze objective data in addition to subjective data, like Dos Santos et al.'s work [dos Santos et al., 2017] and their IVE to teach robotics. They used automatic reports, time spent and movement precision as quanti-

tative data, as well as a questionnaire for qualitative analysis to compare different stereoscopic technologies. These comparisons aim to investigate which one, for example, increases precision and decreases time in tasks.

In a study conducted by Lin et al. [2019], virtual targets using HMD and Stereoscopic Widescreen Display (SWD) was presented to participants, who had to estimate distances by direct reaching, computing accuracy and task completion time. Zhao et al. [2020] evaluate distance stereoacuity, where participants execute a searching task aiming to analyze the distance between separate images, based on red-green anaglyphs, polarized light technology, active shutter and autostereoscopic. In Vienne et al. [2020], participants execute manipulation tasks to judge and adjust angles of a virtual dihedral in a L-shaped VR system, which considered HMD and a CAVE to a depth perception evaluation. These studies are examples of the second mentioned scope.

As observed, studies evaluate stereo effects considering different stereoscopic techniques and the evaluation is generally specific to one system. To our best knowledge, there is no systematic, flexible and extensible method that considers both objective and subjective data to evaluate different IVEs considering different techniques. The previously cited studies contributed to build our approach, since they indicated parameters to be evaluated, the type of environments and tasks that the framework should consider, as well as tools to gather data for the subjective score.

3 Framework Description

The framework offers a way to evaluate depth perception by comparing different visualization techniques, especially stereoscopic techniques, in the same IVE. It shall be suitable for IVEs developers, who are responsible for choosing and implementing stereoscopic techniques.

The following concepts are considered in this work. **Effect** is the product of the stereoscopic technique (depth perception). **Parameters** are factors that can influence an effect; in this work, the parameters considered are distance and size of virtual objects (Section 3.1), since we identified these factors as the most evaluated in literature [Silva et al., 2016]. **Metrics** are data collected during the execution of a task to indicate qualitative and quantitative results; here, the metrics considered are hit rate, error rate, time and presence questionnaire responses.

Considering that immersion and presence are studied in the context of virtual environments, immersion can be defined as a medium's technological capacity to provide realistic experiences that can put users in another reality, removing them from their physical reality. Certain features, such as audio and visual quality, frame rate, field of view and **stereoscopy** can influence the immersion offered by a system. Presence is the subjective experience of these users in the mediated virtual environment [Oh et al., 2018]. It is the feeling of being in another place, a virtual place different from the physical one where it actually is – a sensation of being in the virtual environment as opposed to the real one [Meehan et al., 2002]. Presence is traditionally considered as the psychological perception of “being” in the virtual envi-

ronment in which one is immersed (Heeter [1992]; Sheridan et al. [1992]; Steuer [1992]; Witmer and Kline [1998]), and it can be measured using questionnaires, which can show how connected and engaged a user feels within the virtual space [Grassini and Laumann, 2020]. Presence is about how users perceive and experience a virtual environment on a cognitive and emotional level. It is not solely dependent on technological factors but it is also influenced by individual perception and engagement.

Figure 1 presents the main steps of the framework, which are further detailed in the following subsections.

3.1 Environment Preparation

This first step is the preparation of the IVE that will be used in the experiment, in order to gather the objective and subjective metric data and it must comply with the following requirements: (i) to have versions that use different visualization techniques (stereoscopic and monoscopic techniques) – one version for each technique to be evaluated; (ii) to build tasks which allow gathering performance data related to the tasks to be completed (e.g., error/hit rates and time spent); and (iii) to allow the creation of at least two different scenarios through variations in the parameters of objects with respect to the observer’s point of view.

The different scenarios aim to avoid potential bias in the evaluation of a technique due to the characteristics of the IVE. Thus, when we vary parameters, such as the size and distance of the virtual objects to the virtual camera, we can favor a more impartial evaluation. Besides the requirements above, the experimental design must be planned and implemented carefully in order to avoid the interference of *confounding factors*, i.e., overlooked experimental conditions whose effects cannot be distinguished from those of the techniques to be compared [Oehlert, 2010]. For example, the order of techniques to which users are submitted may influence the results, insofar as successive interactions of users with the IVE may lead to effects of fatigue or adaptation along the tasks. To mitigate this risk, an experimental block design must be applied, where each block (group of users) is characterized by a sequence of techniques to be used by the participants [Oehlert, 2010]. The significance test procedure proposed in Section 3.4.1 includes extensions for block designs. In Section 4.3, we show examples of varying such parameters, in which four different scenarios were generated by changing objects’ distances and sizes.

3.2 Objective Data Acquisition

Objective data acquisition step consists of gathering data from users during their interactions with the IVE.

The objective evaluation consists of capturing performance metrics related to the users’ interactions with the system. In the experiments conducted (Section 4), the metrics considered were error rate (positioning objects in the wrong place), hit rate (collisions with suitable objects) and time to complete the task. Nonetheless, the framework is extensible to other metrics, and allows manual data gathering by one external observer during the user’s interactions; however, the source code can be changed to collect the data.

3.3 Subjective Data Acquisition

The subjective data acquisition step consists of gathering data from users after their IVE interactions. When users have completed all tasks with the same visualization technique, they answer a questionnaire with their opinions about the perceived depth.

Based on the literature, we defined a questionnaire (Tables 5 and 6) to our experiments, which considers ten levels of possible responses, adapted from Witmer and Singer’s Presence Questionnaire (PQ) [Witmer and Singer, 1998]. The statistical analysis routines included in the comprehensive framework are able to deal with any questionnaire composed of ordered single-answer questions, in any order scale (i.e., lower levels representing more negative answers and upper levels the more positive answers, or vice-versa).

3.4 Statistical Model

Once the experimental phase has finished and usage data has been gathered, the statistical analysis is conducted in order to assign objective and subjective scores for each technique. Essentially, the scores are computed through pairwise comparisons between techniques, in which each one earns or loses points if it is significantly better or worse than the other, at a prescribed significance level. The net balance of each technique (wins – losses) is then converted into a more intuitive scale.

3.4.1 Significance Test Procedure

The significance test procedure for comparison between techniques is based on *randomization tests*, a subclass of statistical tests called *permutation tests* [Edgington and Onghena, 2007]. The p-value is given as the proportion of data permutations, providing a test statistic (e.g., the difference between sample means) as large as (or as small as) that obtained in the experimental results. Randomization tests share similar principles of permutation tests, except that the p-value is not computed over all data permutations (which is usually unfeasible even for moderate sample sizes), but instead on a subset of randomly generated permutations.

In contrast to traditional parametric tests (e.g., *t*-tests and ANOVA (Analysis of Variance)), randomization tests have several theoretical advantages: (i) it is possible to draw valid statistical inferences about experimental treatment effects on non-probabilistic samples (usually called also “convenience samples”), which are typical in experiments conducted in the Computer Science fields; (ii) they not require any assumptions about the distribution of the variables being tested; (iii) they are less sensitive to skewed distributions and outliers, which are frequent for some metrics in our context (e.g., time to complete the task); and (iv) they do not depend on asymptotic approximations valid only for large sample sizes [Edgington and Onghena, 2007].

Algorithm 1 presents a simplified version of the randomization test procedure adopted in our framework (“RANDTEST”) for the pairwise comparison between measurements provided by two visualization techniques, regarding a metric. \mathbf{x}^1 and \mathbf{x}^2 denote vectors of size N (number

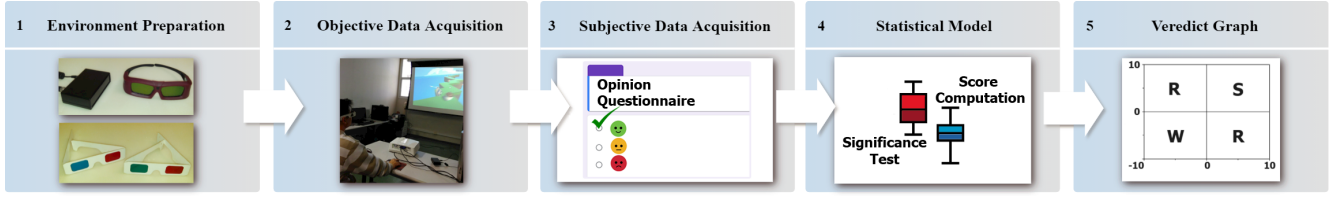


Figure 1. Framework steps

of users), where x_i^j is the metric value for user i under the technique j , for $i = 1, 2, \dots, N$ and $j = 1, 2$.

Briefly, the procedure starts computing test statistics S_x from original observations, detailed in the next paragraph. Next, it randomly swaps metric values between techniques, i.e., some users are randomly draw and their observations are swapped between techniques 1 and 2 (lines 5–8); the corresponding value of the statistic test for the permuted data S_y , is then computed (line 9). This step is repeated B times and the p-value is computed as the proportion of permuted data such that $|S_y| \geq |S_x|$ (lines 10–11).

Computation of statistics S_x is described in the second procedure (“STATISTICS”) of Algorithm 1. For quantitative (discrete/continuous) variables, namely, users’ performance metrics, S_x is the average of differences between techniques outcomes (line 3). For ordinal variables, such as Likert-Type variables, S_x is the standardized difference between $N_{pos} - N_{neg}$, where N_{pos} and N_{neg} denote, respectively, the number of times $x_i^1 > x_i^2$ and $x_i^1 < x_i^2$ (lines 5–7); ties between x_i^1 and x_i^2 are ignored. This statistic is based on Putter’s sign test [Putter, 1955], a robust procedure that consistently holds its significance level and provides a good comparative test power, even under a moderate prevalence of ties [Coakley and Heise, 1996].

Algorithm 1

Randomization test procedure for pairwise comparison between techniques (repeated measures, two-tailed test)

```

1: procedure RandTest( $x^1, x^2, VType, B$ )
2:    $S_x \leftarrow \text{STATISTICS}(x^1, x^2, VType)$    ▷ See “STATISTICS”
   procedure below
3:    $C \leftarrow 1$ 
4:   for  $b \leftarrow 1$  to  $B - 1$  do
5:      $y^1 \leftarrow x^1; y^2 \leftarrow x^2$ 
6:     for  $i \leftarrow 1$  to  $N$  do
7:        $u \leftarrow \text{Uniform}(0,1)$ 
8:       if  $u \geq 0.5$  then swap  $y_i^1$  with  $y_i^2$ 
9:      $S_y \leftarrow \text{STATISTICS}(y^1, y^2, VType)$ 
10:    if  $|S_y| \geq |S_x|$  then  $C + +$ 
11:  p.value  $\leftarrow C/B$ 
12:  return p.value

1: procedure Statistics( $z^1, z^2, VType$ )
2:  if  $VType = \text{“quantitative”}$  then   ▷ Quantitative variable
3:     $S \leftarrow \sum_{i=1}^N (z_i^1 - z_i^2)/N$ 
4:  else                               ▷ Ordinal variable
5:     $N_{pos} \leftarrow \sum_{i=1}^N \mathbf{1}(z_i^1 > z_i^2)$    ▷  $\mathbf{1}$ : indicator function
6:     $N_{neg} \leftarrow \sum_{i=1}^N \mathbf{1}(z_i^1 < z_i^2)$ ;
7:     $S \leftarrow (N_{pos} - N_{neg})/\sqrt{N_{pos} + N_{neg}}$ 
8:  return  $S$ 

```

To determine B , Jockel [1986] established a criterion based on the test power, defined as the probability of a significance test procedure to reject the null hypothesis when it is false. The author derived an upper bound for the decrease

in the power of a randomization test compared with its analogous complete permutation test, as a function of B and the significance level for rejection of the null hypothesis, α . In this work, we considered $B = 10,000$ and a significance level $\alpha = 0.1$, which led to a decrease in performance of less than 2%.

The randomization test version presented in Algorithm 1 has some simplifications. It only considers a two-tailed test, since the comparisons between S_y and S_x are given in modules and therefore ignore the signal of differences. The extension to one-tailed tests is straightforward, by properly adapting the condition in line 10. Vectors x^1 and x^2 are assumed to have only one observation for each user. As mentioned earlier, it is advisable that the objective metrics are gathered under different scenarios (e.g., permutations of *objects’ sizes* and *distances* parameters), which implies that each vector has multiple observations for each user – more precisely, K observations per user, where K denote the number of scenarios. The extension for this case is also straightforward, by swapping all data from the same user between x^1 and x^2 , once the user has been randomly selected for that. This guarantees that the data for each user are treated in blocks, in such a way that all configurations of object size and distance are equally distributed in both techniques, thus avoiding that performance differences due to objects sizes and/or distances be confounded with differences due to techniques. These extensions were considered and implemented in our framework.

3.4.2 Score Computation

Two scores, computed from objective and subjective data, are assigned to each visualization technique. Algorithm 2 briefly presents the score computation procedure. Each metric is identified by an index $m \in \{1, 2, \dots, M\}$, where M denotes the number of metrics, and each technique is identified by an index $t \in \{1, 2, \dots, T\}$, where T denotes the number of techniques. \mathbf{X} is a matrix of N rows (or $N * K$, when K scenarios are considered for each user) and $T * M$ columns, where $\mathbf{X}^{m,t}$ denotes the column of \mathbf{X} containing user records of metric m under the technique t . $VType$ is the type of metrics under analysis. \mathbf{Netb} is a vector of size T , where \mathbf{Netb}_t denotes the net balance (wins – losses) earned by technique t . \mathbf{Score} is the vector of final scores for all techniques. The procedure performs all possible pairwise comparisons between techniques, under all metrics. For each metric m ; and each pair of techniques (t_1, t_2) and $t_1 < t_2$, the randomization test assesses the significance of S_x (lines 6–8). If no significant difference is found ($p.value > \alpha$), neither technique earns or loses any point. Otherwise, each technique earns (loses) one point if it wins (loses) the comparison (lines 10–11).

Algorithm 2

Score computation procedure for pairwise comparison between techniques (repeated measures, two-tailed test)

```

1: procedure ScoreComp( $\mathbf{X}, T, M, \text{VType}, \alpha, B$ )
2:    $\text{Netb}_t \leftarrow 0, t = 1, \dots, T$ 
3:   for  $m \leftarrow 1$  to  $M$  do ▷ Iterations over metrics
4:     for each  $(t_1, t_2) \in \{1, \dots, T\}^2, t_1 < t_2$  do
5:       ▷ Iterations over pairs of techniques
6:        $\mathbf{x}^1 \leftarrow \mathbf{X}^{m,t_1}; \mathbf{x}^2 \leftarrow \mathbf{X}^{m,t_2}$ 
7:        $S_x \leftarrow \text{STATISTICS}(\mathbf{x}^1, \mathbf{x}^2, \text{VType})$ 
8:        $\text{p.value} \leftarrow \text{RANDTEST}(\mathbf{x}^1, \mathbf{x}^2, \text{VType}, B)$ 
9:       if  $\text{p.value} \leq \alpha$  then
10:        if  $S_x < 0$  then  $\text{Netb}_{t_1} - -; \text{Netb}_{t_2} + +$ 
11:        if  $S_x > 0$  then  $\text{Netb}_{t_1} + +; \text{Netb}_{t_2} - -$ 
12:    $c \leftarrow 10 / [(T - 1) \cdot M]$ 
13:    $\text{Score}_t \leftarrow c \cdot \text{Netb}_t, t = 1, \dots, T$ 
14:   return  $\text{Score}_t$ 

```

After all iterations, the vector Netb is multiplied by a normalization constant c , in such a way that the final scores (vector Score) range from -10 to 10 (line 13). In the experiments presented in this work, c is given as follows: for the objective scores, we have $T = 3, M = 2$, resulting in $c = 2.5$; for the subjective scores, we have $T = 3, M = 12$, resulting in $c \approx 0.417$. Notice that extreme scores occur when one technique loses or wins all comparisons.

The procedure shown in Algorithm 2 assumes that all metrics are positively ordered, i.e., the higher the value, the better the technique; dealing with negative ordering is straightforward, by multiplying \mathbf{X}^{m,t_1} and \mathbf{X}^{m,t_2} by (-1) . This extension is also implemented, the developer can set up the ordering signal for each metric.

Nonetheless its simplicity, this approach is sufficiently flexible to allow other extensions such as handling other metrics or computing metrics with different weights, setting different weights to objective and subjective scores, adapting scale limits according to domain needs or customs, and setting other criteria for score assignment.

3.5 Verdict Graph

This last step consists of plotting the objective and subjective scores for all techniques in a two-dimensional graph, in order to allow a visual analysis of their relative performances with respect to the provision of depth perception. Each technique t is represented by a coordinate $(\text{Score}_t^o, \text{Score}_t^s)$, where Score_t^o is the objective score and Score_t^s is the subjective score of technique $t, t = 1, \dots, T$.

Figure 1 (Step 5 - Verdict Graph) also presents the base graphs representing the space of possible coordinates. We consider four quadrants, each one corresponding to a verdict about the techniques' performances:

- **Weak:** techniques in this quadrant are those with negative objective and subjective scores, indicating a poor performance in comparison with other competitors; unless a technique in this quadrant is near the center of the graph (coordinate $(0, 0)$), its use should be considered only for non-critical systems and when budget constraints preclude the use of better (and more costly) techniques;
- **Regular:** this verdict includes the two quadrants in which objective and subjective scores are negatively

correlated, and represent techniques with good relative performance under one criterion but weak performance under the other. Their use should also be considered with caution. Nonetheless, these quadrants should have a lower probability density, insofar as techniques providing a better depth perception (consequently, with higher ratings in the questionnaire) should yield a better performance during the tasks (and therefore higher ratings in the objective metrics);

- **Strong:** techniques in this quadrant have, on average, performed better than their competitors, and should be the preferable choice. The closer their positions to the upper right corner (coordinate $(10, 10)$), the more evident their superiority.

4 Framework Validation

As an illustration of our framework's application, we present four experiments conducted in different moments to compare visualization techniques (three techniques in the first two experiments and two techniques in the last), within two different IVEs: a simulator in the health area and an endless racing game (called "3D Running Squirrel").

Dental training simulations have drawn attention as an educational strategy in Covid-19 pandemic [Hattori *et al.*, 2022]. The simulator used in our experiment is an immersive tool, based on 3D interaction, to train dental anesthesia (Figure 5). The user's goal is to manipulate a virtual syringe and insert its needle in a specific region, to inject the anesthetic to block the nerve's electrical signals. Better performance is achieved when the user completes this task with fewer errors (inserting in other regions) and in a shorter time. Additionally, the simulator allows users to navigate the environment using the keyboard changing their viewpoints during interaction to explore virtual objects. In this case, better performance is to find objects and details in a shorter time.

The game "3D Running Squirrel" is an immersive infinite racing game available for desktops and mobile devices (Figure 6). The player can observe the performance achieved from the number of hits (walnuts capture, without falling out the path) and elapsed time. The greater the number of hits and the greater the time spent, the better is the player's performance. From now, we will refer to this IVE as "game" in this work.

4.1 Participants

Two groups of participants were recruited for the experiments.

The participants of the first two experiments (Group 1), comprised 20 students and teachers in the Computer Science area, with sixteen male participants and four female participants (twenty-nine years average age). The gender distribution is consistent with that found among Brazilian students in higher education courses within the computing field [Maciel *et al.*, 2018]. More than half (65%) of the participants mentioned having some type of eye problem; therefore, they used visual correction with their respective eyeglasses during the

experiments. All participants of Group 1 said that they had some experience with 3D virtual environments.

The participants of the last two experiments (Group 2) comprised nine students in the Computer Science area, with seven male participants and two female participants (twenty-one years average age). Forty percent (40%) of the participants mentioned having some type of eye problem; and they also used visual correction with their respective eyeglasses during the experiments. All participants of Group 2 said that they had some experience with 3D virtual environments.

4.2 Visualization Techniques and Devices

Five visualization techniques, one monoscopic technique and four stereoscopic techniques were evaluated in our experiments: color filtering technique by means of true anaglyph glasses, here named “True Anaglyph Technique” (TAT); color filtering technique by means of color anaglyph glasses, named “Color Anaglyph Technique” (CAT); light shutter by using shutter glasses, here named “Shutter Glasses Technique” (SGT); CAVE without glasses (CT); and CAVE with glasses (CGT). The glasses used in CGT were specifically the color anaglyph glasses.

The glasses for TAT are made of lenses with red and blue filters. Similarly, glasses for the CAT use red and cyan lenses. On the other hand, SGT is made with lenses that alternate the scene for each eye in a frequency synchronized with the monitor or projector refresh rate. Figure 2 illustrates an example of the evaluation scenario utilized for conducting experiments with Group 1.



Figure 2. A participant from Group 1 actively engaged in the experiment by playing the game.

CGT is a cave and the participants used anaglyph glasses for 3D visualization. CT did not have the glasses and it was considered a monoscopic environment, although there are several viewpoints of the environment. CT and CGT were specifically designed to create an engaging environment within a classroom setting. Four multimedia projectors were strategically positioned in the classroom to provide a seamless visual experience for the participants. The projectors were carefully adjusted and calibrated to project images onto the targeted walls, ensuring a cohesive and synchronized display. This configuration allowed for a substantial

portion of the classroom walls to be utilized as a canvas for the virtual environment. While the other two walls of the classroom were not directly covered by the projectors, they still contributed to the overall immersive experience. The ambient lighting in the room was adjusted to minimize distractions and enhance the perception of being enveloped in the virtual environment to perform navigation tasks. Figure 3 illustrates an example of the evaluation scenario utilized for conducting experiments with Group 2.



Figure 3. A participant from Group 2 involved in the experiment utilizing the simulator in the navigation task. He was passing through the modeled ceiling in the virtual dental office environment at that moment.

These techniques were chosen due to their differences in cost versus benefit ratio, especially differences between TAT/CAT and SGT/CT/CGT. This difference is evidenced by the fact that shutter glasses and the projectors required for building the CAVE are high-cost equipment, whereas anaglyph glasses can be made from inexpensive materials. Since they present different cost levels, our investigation intends to verify if a technique is suitable for a determined system even if it presents a low cost. However, other devices, such as modern head-mounted displays (HMD), may be used.

To perform the procedures offered by the simulator, it was used a Leap Motion device that captures movements that are transferred to a syringe in the virtual space. To physically represent the syringe, a common straw was used (Figure 4). In the simulator, it is possible to navigate in the virtual environment using the keyboard, with keys to move the viewpoint or the virtual camera (translation and rotation). A projection equipment was used to enable the correct operation of SGT, which requires the visualization device to operate at a refresh or frequency rate of 120Hz. Concerning the game, we used a standard mouse to control the squirrel’s actions in the virtual space.

4.3 Design

Both experiments considered **distance** and **size** as parameters. To assess users’ performance – collecting hits (game), errors (simulator), and time (game and simulator) – under different configurations of parameters, four different scenarios were built in each IVE.

Table 1. Configurations of parameters for the simulator

Scenario	Description
Longer distance/ Larger size	The virtual camera is far from the patient and the target (yellow sphere) is bigger
Longer distance/ Smaller size	The virtual camera is far from the patient and the target (yellow sphere) is smaller
Shorter distance/ Larger size	The virtual camera is closer to the patient and the target (yellow sphere) is bigger
Shorter distance/ Smaller Size	The virtual camera is closer to the patient and the target (yellow sphere) is smaller

Table 2. Configurations of parameters for the game

Scenario	Description
Longer distance/ Larger size	The virtual camera is far from the squirrel and the nuts; and the path has a wide width
Longer distance/ Smaller size	The virtual camera is far from the squirrel and the nuts; the path has a narrow width
Shorter distance/ Larger size	The virtual camera is closer to the squirrel and the nuts; and the path has a wide width
Shorter distance/ Smaller size	The virtual camera is closer to the squirrel and the nuts; and the path has a narrow width



Figure 4. Environment with the use of a straw (red circle) that represents a virtual syringe and Leap Motion (blue circle) for virtual tool tracking. In order to prevent interference caused by the infrared rays emitted by the Leap Motion device on the shutter glasses, a specially designed apparatus made of cardboard and ethyl vinyl acetate (EVA) sheets was used.

In the simulator, each scenario was obtained with the variation of: (i) distance between the virtual camera (representing the user's viewpoint) and the virtual patient, (ii) target's size or region to be reached. The target region is a yellow sphere placed at the virtual patient's inner mouth surface that shows the nerve direction location to be reached. Table 1 and Figure 5 show these configurations.

In the game, each scenario was obtained with the variation of: (i) distance between the virtual camera and virtual objects, (ii) virtual objects sizes (squirrel, walnuts and the path), as shown in Table 2 and Figure 6.

Both IVEs were rendered with the techniques mentioned in Subsection 4.2, as shown in Figures 7, 8, 9 and 10.

Each participant of Group 1 performed twelve tasks within each IVE, each corresponding to a stereoscopy technique and a scenario. The experiments were conducted between 1 p.m. and 6 p.m., maintaining similar conditions by using artificial lighting for all users, in order to avoid lighting influence on the user perception during the experiments.

Each participant of Group 2 performed eight tasks, each one corresponding to a visualization technique (monoscopic and stereoscopic) and a scenario. The experiments were conducted between 3 p.m. and 9 p.m., also maintaining similar conditions by using artificial lighting for all users.

4.4 Tasks and Metrics

In the simulator, in the first two experiments the task consisted of interacting with the system by manipulating a syringe in the virtual space to reach a small target and its respective bigger **size** target, from two different camera's **distance** perspectives. During the interaction the system collected the metrics: error rate (frequency with which the needle missed the target) and the time spent to complete the task.

In one of the last two experiments, related to the simulator, the task consisted of navigating in the virtual environment to find an object, a yellow sphere, placed at the virtual patient's inner mouth surface.

The task related to the game in all experiments consisted of controlling the squirrel that runs in a virtual wall to capture the largest amount of walnuts, without falling out a narrow path and a wider one, from two different camera's **distance** perspectives. During the interaction the system collected the metrics: hit rate (frequency with which the squirrel correctly turned left or right and captured walnuts) and time elapsed until falling out of the path.

4.5 Procedure

This experiment was carried out with the approval of the Committee of Ethics in Research with Human Beings of the Faculty of Medicine of the University of São Paulo (CAAE 54691916.3.0000.0065). Participants signed a Consent Form and they were previously informed about the tasks to be performed and about the devices to be used. The participants of Group 1 first tested the simulator and then the game. In order to minimize any bias and account for factors such as participants' fatigue, the testing order of each visualization technique was alternated for each volunteer during each test session for each one of the IVEs. The last two experiments conducted with participants from the second group were carried out alternating the systems.

After previous explanations by the researcher, for each technique, each participant alternated between performing a task and answering a questionnaire. During all the sessions, the researcher gave instructions to the participant whenever necessary up to the end of the session. Each session lasted on average 40 minutes.

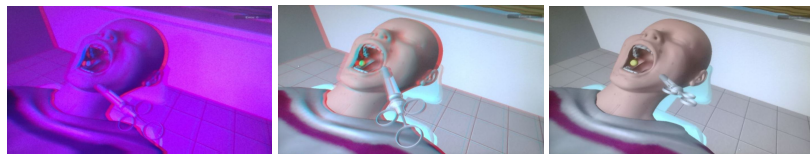
The participants of Group 1 (the first two experiments), were invited to sit facing the video monitor and record their personal data. Next, a pair of stereoscopic visualization glasses from CAT, TAT or SGT were provided, the partici-



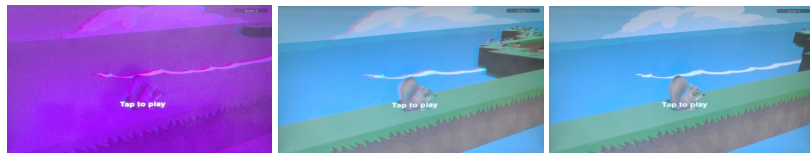
(a) Longer distance, Larger size (b) Longer distance, Smaller size (c) Smaller distance, Larger size (d) Smaller distance, Smaller size
Figure 5. Scenes under different parameter setups in the simulator



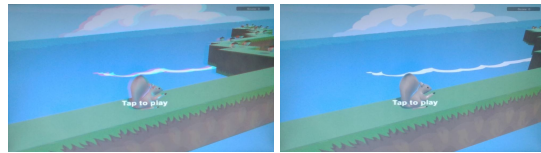
(a) Longer distance, Larger size (b) Longer distance, Smaller size (c) Smaller distance, Larger size (d) Smaller distance, Smaller size
Figure 6. Scenes under different parameter setups in the game



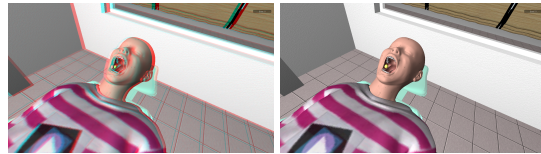
(a) Simulator with True Anaglyph Technique (TAT) (b) Simulator with Color Anaglyph Technique (CAT) (c) Simulator with Shutter Glasses Technique (SGT)
Figure 7. Evaluation environments for the conducted experiment with the simulator for Group 1 experiment



(a) Game with True Anaglyph Technique (TAT) (b) Game with Color Anaglyph Technique (CAT) (c) Game with Shutter Glasses Technique (SGT)
Figure 8. Evaluation environments for the conducted experiment with the game for Group 1 experiment



(a) Game in CAVE with Color Anaglyph glasses technique (CGT) (b) Game in CAVE without glasses technique (CT)
Figure 9. Evaluation environments for the conducted experiment with the game for Group 2 experiment



(a) Simulator in CAVE with Color Anaglyph glasses technique (CGT) (b) Simulator in CAVE without glasses technique (CT)
Figure 10. Evaluation environments for the conducted experiment with the simulator for Group 2 experiment

part took place close to a table (in front of the projection) and the tests started. After completing all tasks with the same stereoscopy technique, participants answered the evaluation questionnaire.

The participants of Group 2 (the last two experiments), were invited to sit facing the video monitor and record their personal data. Next, the participants should enter a virtual room. A pair of stereoscopic visualization glasses was provided to participants to interact with the simulator and game when the experiment refers to the CGT. For CT the visual-

ization technique was monoscopic. Finally, the participants completed the tasks and answered the evaluation questionnaire.

5 Results

With respect to the objective metrics, Table 3 and 4 present the results of pairwise comparisons between CAT, TAT and SGT (Group 1 and the first two experiments), as well as

the results of pairwise comparisons between CGT and CT (Group 2 and the last two experiments), concerning the objective metrics. Each entry contains the average difference of ratings in each metric and its respective p-value. Significant differences (p-value < 0.1) are highlighted in bold.

In both IVEs (Group 1 and the first two experiments), the relative performance between techniques showed similar patterns. Nonetheless, CAT has yielded a slight superiority over TAT (lower error rate and time averages within the simulator; higher hit rate and time average within the game), no significant difference was found in either metric. On the other hand, the significant differences found in most pairwise comparisons CAT \times SGT and TAT \times SGT reveal that SGT stands out with the best performance.

In both IVEs (Group 2 and the last two experiments), CT obtained a superior performance compared with CGT. It is important to mention that the tasks and systems were different and only time was used as a metric for the simulator. Significant differences were found in two experiments (CT and CGT for both systems).

Concerning the subjective metrics, Tables 5 and 6 present the results of pairwise comparisons between CAT, TAT and SGT (the first two experiments - Group 1) and pairwise comparisons between CGT and CT (the last two experiments - Group 2).

The relative performance of techniques on subjective metrics reproduced approximately the same pattern found on the objective metrics. In the case of Group 1 and the first two experiments, CAT and TAT obtained very similar ratings, and significant differences (in favor of CAT) were found in only one of the 12 questions, for both IVEs. Similarly, SGT achieved a strong superiority over CAT and TAT, where most of pairwise comparisons showed significant differences: within the simulator, 10/12 for CAT \times SGT and 12/12 for TAT \times SGT; within the game, 9/12 for CAT \times SGT and 9/12 for TAT \times SGT. In the case of Group 2 and the first last experiments, CT obtained strong superiority over CGT, where most of pairwise comparisons showed significant differences within the simulator, 9/12 for CGT \times CT. Within the game, 3/12 pairwise comparisons showed significant differences for CGT \times CT.

The computation of the objective and subjective scores is straightforward, as described in Section 3.4.2. For example, to compute the objective score of the CAT in the simulator, we notice that it obtained 2 ties and 2 losses (Table 3), resulting in a net balance -2 . Multiplying the net balance by the normalization constant $c = 2.5$ (Section 3.4.2) yields its objective score $\text{Score}_{CAT}^o = -5.0$. As for the subjective score, we notice that CAT obtained 13 ties, 1 win and 10 losses (Table 5), resulting in a net balance -9 . Multiplying the net balance by the normalization constant $c \approx 0.417$ yields the subjective score $\text{Score}_{CAT}^s \approx -3.8$.

Figure 11 presents the scores for CAT (Color Anaglyph), TAT (True Anaglyph) and SGT (Shutter Glasses), each one with their subjective and objective scores combined in a triangle, as well as their respective quadrant-based verdicts, considering Group 1 and the first two experiments. As noticed in the pairwise comparisons, SGT has shown a remarkable superiority over CAT and TAT, being therefore considered a **strong** stereoscopy technique. CAT and TAT, in their turn,

have shown very similar performances, quite below SGT. Insofar as neither of them has shown a convincing superiority over the other, both are considered **weak** techniques in our experiments. In both IVEs, CAT presents a slight superiority over TAT, evidenced by its lower distance from the center of the graph. Nonetheless, its use in the simulator should be considered with caution, due to the system's criticality. On the other hand, for non-critical systems such as the game, CAT is preferable over TAT in case cost constraints preclude the acquisition of more expensive devices.

Figure 12 presents the scores for CGT (CAVE Glasses Technique) and CT (CAVE Technique), considering Group 2 and the last two experiments. CT was superior to CGT for the simulator and game, although the tasks differed for each system. We can observe that a technique (CGT) was considered **weak** and another technique (CT) was considered **strong**.

It is worthy mentioning that, as described in Section 4.2, although participants in Groups 1 and 2 were submitted to the same IVEs (dental training simulator and 3D running squirrel game), the visualization techniques used by each group were different, which precludes a joint or comparative analysis of their results.

6 Discussion

The existing evaluation methods from literature are executed for particular scenarios and, in general, inserted in a single IVE and without the possibility of being reused. The framework can be applied to evaluate IVEs that consider the manipulation of objects and the navigation in virtual environments, and its application depends neither on the scope of the IVE nor on the visualization techniques to be evaluated. Besides, the framework is able to incorporate different objective metrics, parameters and questionnaires (as long as they are composed exclusively of ordered single-answer questionnaires). It is important to mention that the framework allows comparing stereoscopic and monoscopic techniques, which lack the presentation of distinct scenarios for each eye.

Furthermore, the literature indicated objective and subjective evaluations, generally disconnectedly. The two-way evaluation is based on the consideration that, taking into account only the objective or only the subjective dimension, may result in biased or incomplete assessments. Thus, our framework integrates both dimensions to form a verdict about the depth perceived by an individual when using different techniques. Such a verdict (Figure 1) is an initial proposal that can be adapted to the developer's needs.

Our framework requires a questionnaire to assess the users' point of view concerning their sensations perceived during the performance of tasks. Among the several questionnaires available in the literature, we adapted the one proposed by Witmer and Singer's [Witmer and Singer, 1998]. The authors suggest that "involvement" is an important determinant of presence in virtual environments, so we adapted some questions from their questionnaire about this factor, resulting in questions 3 to 12 (Tables 5 and 6). Besides, two new questions (1 and 2) were included to assess the visual strain and the users' comfort when using visualization devices. The framework can incorporate other questionnaires,

Table 3. Average differences with respect to (w.r.t.) objective metrics for pairwise comparisons between stereoscopic techniques within the Simulator, CAT × TAT, CAT × SGT and TAT × SGT, in the first two experiments (Group 1); CGT × CT in the last two experiments (Group 2).

Metric	CAT × TAT (*)	CAT × SGT (*)	TAT × SGT (*)	CGT × CT (**)
Error rate	-3.71 (p=0.383)	7.26 (p<0.001)	10.98 (p=0.002)	-
Time	-5.38 (p=0.257)	8.95 (p=0.003)	14.33 (p<0.001)	57.63 (p=0.0036)

*CAT = Color Anaglyph Technique; TAT = True Anaglyph Technique; SGT = Shutter Glasses Technique

**CGT = CAVE Glasses Technique; CT = CAVE Technique

Table 4. Average differences w.r.t. objective metrics for pairwise comparisons between stereoscopic techniques within the Game, CAT × TAT, CAT × SGT and TAT × SGT, in the first two experiments (Group 1); CGT × CT in the last two experiments (Group 2).

Metric	CAT × TAT (*)	CAT × SGT (*)	TAT × SGT (*)	CGT × CT (**)
Hit rate	6.48 (p=0.137)	-9.94 (p=0.058)	-16.41 (p=0.013)	-21.11 (p=0.0035)
Time	2.29 (p=0.134)	-2.24 (p=0.209)	-4.53 (p=0.020)	-6.27.29 (p=0.0032)

*CAT = Color Anaglyph Technique; TAT = True Anaglyph Technique; SGT = Shutter Glasses Technique

**CGT = CAVE Glasses Technique; CT = CAVE Technique

Table 5. Differences w.r.t. Subjective Metrics for Pairwise Comparisons Between stereoscopic techniques within the Simulator, CAT × TAT, CAT × SGT and TAT × SGT, in the first two experiments (Group 1); CGT × CT in the last two experiments (Group 2).

#	Question	CAT × TAT (*)	CAT × SGT (*)	TAT × SGT (*)	CGT × CT (**)
1	Were you comfortable while using the stereoscopy device?	1.15 (p=0.394)	-3.58 (p<0.001)	-4.02 (p<0.001)	-2.77 (p=0.045)
2	Did you feel eye strain during and/or after the experience with the stereoscopy device?	1.60 (p=0.181)	-2.98 (p=0.004)	-3.30 (p<0.001)	-2.33 (p=0.035)
3	How quickly did you adapt yourself to the virtual environment experience with the stereoscopy device?	2.32 (p=0.033)	-0.77 (p=0.609)	-3.74 (p<0.001)	-1.32 (p=0.097)
4	How natural were your interactions with the environment?	0.00 (p=1.000)	-3.50 (p<0.001)	-3.87 (p<0.001)	-1.11 (p=0.065)
5	How much did the visual aspects of the environment involve you?	0.50 (p=0.803)	-3.50 (p<0.001)	-3.50 (p<0.001)	-0.66 (p=0.018)
6	How compelling was your sense of objects motion through space?	-0.28 (p=1.000)	-3.21 (p=0.002)	-2.14 (p=0.057)	-0.77 (p=0.062)
7	Were you able to anticipate what would happen next in response to the actions that you performed?	1.00 (p=0.452)	-1.41 (p=0.239)	-1.89 (p=0.090)	-0.66 (p=0.299)
8	How much did the visualization quality provided by the stereoscopy device interfere or distract you from performing assigned tasks or required activities?	1.70 (p=0.148)	-2.67 (p=0.012)	-3.44 (p<0.001)	-1.33 (p=0.158)
9	How well could you move or manipulate objects in the virtual environment?	1.39 (p=0.272)	-2.00 (p=0.077)	-3.00 (p=0.004)	-1.44 (p=0.030)
10	How much did the manipulation of objects in the virtual environment seem consistent with the manipulation of objects in the real world?	0.83 (p=0.588)	-2.50 (p=0.022)	-2.50 (p=0.023)	-0.66 (p=0.370)
11	How engaged were you in the performance of the tasks or activities required in the virtual environment?	0.28 (p=1.000)	-3.05 (p=0.004)	-2.31 (p=0.037)	-1.11 (p=0.095)
12	How capable were you to perceive the 3D effect in the virtual environment?	0.53 (p=0.788)	-3.64 (p<0.001)	-4.12 (p<0.001)	-2.11 (p=0.033)

*CAT = Color Anaglyph Technique; TAT = True Anaglyph Technique; SGT = Shutter Glasses Technique

**CGT = CAVE Glasses Technique; CT = CAVE Technique

such as Immersive Tendencies Questionnaire (ITQ) [Witmer and Singer, 1998], ITC-Sense of Presence Inventory (ITC-SOPI) [Lessiter et al., 2001], and Igroup Presence Questionnaire (IPQ) [Schubert et al., 2001].

Although we intend to continue acquiring data considering other techniques, new IVEs and a greater number of participants, the number of participants considered in our experiments is according to the literature. Several studies carried out experiments with different numbers of participants using questionnaires and physiological measures, evaluating presence in virtual environments. There are studies with 10 [Clemente et al., 2013b], 14 [Clemente et al., 2013a], 18 [Anderson et al., 2017], 19 [Poels et al., 2012], and 20 participants [Burns and Fairclough, 2015].

Thus, we believe the results obtained with the first two experiments presented here are according to the literature; nevertheless, considering the last two experiments, CAVE without glasses (monoscopic) showed a superior result when compared with CAVE with glasses (stereoscopic), in both

tasks (manipulation and navigation). CAVE without glasses used a monoscopic technique; however, the participant had several viewpoints of the environment. This fact can have contributed to performing tasks and the pair of glasses was not necessary, decreasing the cognitive load and facilitating the actions of the participants when compared with CAVE with glasses. That way, it is possible to observe that various aspects can influence depth perception.

The results suggest that, in the context of our experiments, the use of stereoscopic glasses can impose an additional cognitive burden on participants as they need to adapt to the stereoscopic visualization. The absence of glasses in the monoscopic environment can alleviate this cognitive burden and facilitate the execution of tasks, reaching better performance. Moreover, individual preferences of participants can also influence depth perception. Some participants may adapt more effectively to stereoscopic viewing, while others may prefer monoscopic viewing.

Besides, the permutation test procedure presented prop-

Table 6. Differences w.r.t. Subjective Metrics for Pairwise Comparisons Between stereoscopic techniques Within the Game, CAT × TAT, CAT × SGT and TAT × SGT, in the first two experiments (Group 1); CGT × CT in the last two experiments (Group 2).

#	Question	CAT × TAT (*)	CAT × SGT (*)	TAT × SGT (*)	CGT × CT (**)
1	Were you comfortable while using the stereoscopy device?	0.00 (p=1.000)	-3.44 (p<0.001)	-3.77 (p<0.001)	-1.22 (p=0.125)
2	Did you feel eye strain during and/or after the experience with the stereoscopy device?	0.28 (p=1.000)	-3.15 (p=0.004)	-4.24 (p<0.001)	-1.66 (p=0.045)
3	How quickly did adapt yourself to the virtual environment experience with the stereoscopy device?	1.60 (p=0.175)	-3.32 (p=0.001)	-3.74 (p<0.001)	-1.55 (p=0.108)
4	How natural were your interactions with the environment?	0.26 (p=1.000)	-1.94 (p=0.094)	-1.70 (p=0.149)	-1.0 (p=0.320)
5	How much did the visual aspects of the environment involve you?	0.00 (p=1.000)	-1.94 (p=0.092)	-2.14 (p=0.060)	-0.66 (p=0.306)
6	How compelling was your sense of objects motion through space?	-0.53 (p=0.787)	-2.14 (p=0.056)	-2.71 (p=0.014)	-0.55 (p=0.411)
7	Were you able to anticipate what would happen next in response to the actions that you performed?	0.53 (p=0.790)	0.00 (p=1.000)	-0.58 (p=0.779)	-0.55 (p=0.429)
8	How much did the visualization quality provided by the stereoscopy device interfere or distract you from performing assigned tasks or required activities?	1.07 (p=0.423)	-3.64 (p<0.001)	-3.77 (p<0.001)	-2.11 (p=0.310)
9	How well could you move or manipulate objects in the virtual environment?	0.90 (p=0.554)	-2.11 (p=0.065)	-2.31 (p=0.036)	-1.33 (p=0.027)
10	How much did the manipulation of objects in the virtual environment seem consistent with the manipulation of objects in the real world?	2.11 (p=0.066)	-1.73 (p=0.145)	-3.21 (p=0.001)	0.33 (p=0.654)
11	How engaged were you in the performance of the tasks or activities required in the virtual environment?	-0.30 (p=1.000)	-1.26 (p=0.341)	-1.26 (p=0.345)	0.33 (p=0.752)
12	How capable were you to perceive the 3D effect in the virtual environment?	0.58 (p=0.778)	-3.64 (p<0.001)	-3.50 (p<0.001)	-1.55 (p=0.027)

*CAT = Color Anaglyph Technique; TAT = True Anaglyph Technique; SGT = Shutter Glasses Technique

**CGT = CAVE Glasses Technique; CT = CAVE Technique

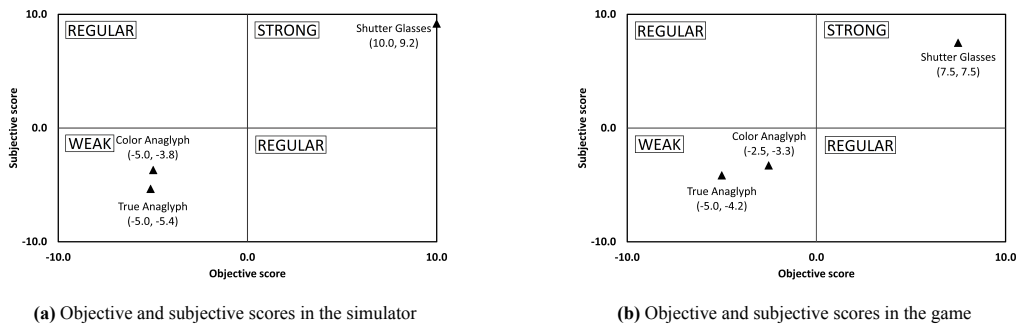


Figure 11. Result for verdict's graph to each stereoscopic technique classification - first two experiments (Group 1)

erly handles quantitative and ordinal metrics and does not rely on asymptotic convergence. Additionally, the method is easily extensible for experimental block designs.

Regarding the tasks, the manipulation tasks were addressed in the first two experiments (Group 1). Specifically, in the first experiment, participants were tasked with **manipulating** a virtual syringe in the simulator using the Leap Motion controller. The second experiment involved **manipulating** a virtual squirrel in the game using a mouse. On the other hand, in the last two experiments, manipulation and navigation tasks were addressed (Group 2). More specifically, in the last two experiments, participants were required to **navigate** through the virtual dental office using the keyboard, and they were tasked with **manipulating** the virtual squirrel in the game using a mouse. Comparing the results (task completion time metric) between tasks in the simulator (Group 1 and Group 2), the times of the navigation task were larger when compared with those of the manipulation task. Thus, the task can influence the results and this issue must be considered. This suggests that the type of task can influence the results obtained when evaluating techniques. Therefore, when comparing different techniques, it is essential to ensure that the same tasks are applied in order to obtain fairer and more meaningful results, as conducted in the mentioned ex-

periments.

The variation of parameters of the virtual environments presented in the experiments is important to analyze depth perception. By manipulating parameters such as size and distance of virtual objects in relation to the virtual camera, researchers can assess how these changes impact participants' perception of depth. This allows a more comprehensive evaluation of the effectiveness of different visualization techniques, such as stereoscopic and monoscopic, in creating a sense of depth and immersion. Additionally, by creating different scenarios with varying parameters, potential biases in the evaluation process can be minimized, ensuring a more objective assessment of the techniques being compared. These variations in the virtual environments help researchers gather valuable data on how different factors influence depth perception, enhancing our understanding of human perception in immersive virtual environments.

Regarding all experiments, (two groups and four experiments), CAVE with glasses, i.e., the results of Group 2 and the last two experiments, were inferior compared to all results. Thus, the techniques based on anaglyphs were classified in the **weak** quadrant in the verdict graph.

Finally, evaluation methods are relevant since other technologies are developed and effects can be created. The study

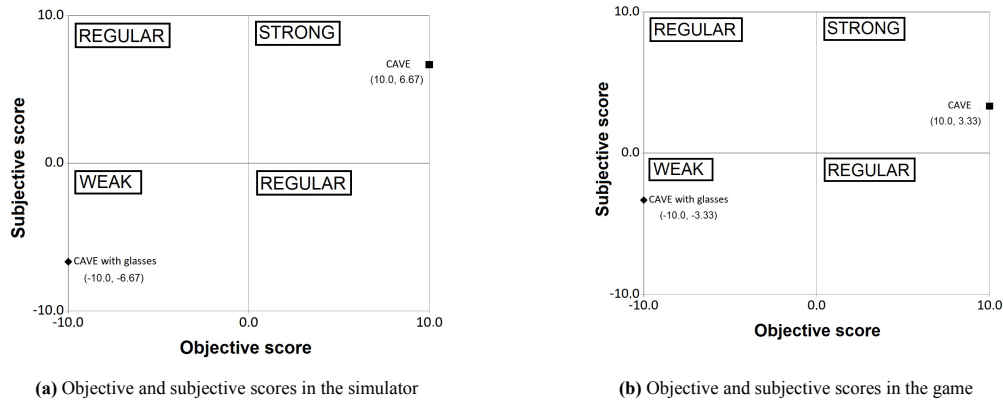


Figure 12. Result for verdict's graph to each technique classification - last two experiments (Group 2)

conducted by Thalmann *et al.* [2016] showed that Oculus Rift was slightly better than CAVE and stereoscopic display. In the following subsections, we present the constraints and benefits of the framework and our experiments.

6.1 Limitations

In the comprehensive framework, the most cited parameters in literature were considered: distance and objects' size. The rendering of IVEs with visualization establishes the construction of different scenarios with variations of these parameters, generating different versions of the same system. Such different scenarios are required to assess, at which level, variations of the parameters facilitate or hinder the manipulation of objects or the navigation in the environment. Considering that each IVE has already been duly studied to define the adequate object's size and distance from the camera for its purpose, such suggested variations could have an impact on errors, hits and times measured during an evaluation. Consequently, they would enable assessing the behavior of the users' perception when the parameters are changed.

Although the experiments conducted were not the main focus of this work, some limitations could be mentioned. Considering the last two experiments executed by Group 2, the results suggest that stereoscopy techniques may not be necessary in some VR systems. However, a CAVE was used and other systems could be assessed. Only distance and objects' size variations were considered as parameters for the scenarios, as well as only the hits rate, errors rate and time metrics were included in the comparisons between techniques. The literature was considered; however, in future applications, other parameters and metrics can be addressed.

The results and verdicts obtained in the experiments are valid only in the context of the experiment. Therefore, their extrapolation to other contexts (even in IVEs with characteristics similar to the IVEs tested here) is not straightforward.

Our results obtained with users with different characteristics are similar to results reported in the literature by Zhao *et al.* [2020] and Livatino *et al.* [2015], where different techniques are compared. However, new experiments with developers of the visualization and related areas could be useful to assess the framework, especially to check the values of the differences between the techniques.

6.2 Benefits

We have not found in the literature an extensible evaluation method, applicable to different systems, to provide comparative performance assessments of visualization techniques, especially stereoscopic techniques, in the same IVE, evaluating depth perception. In this sense, the comprehensive framework is a novelty aiming to address such lack in the VR area.

Besides, the comprehensive framework is flexible in the sense that it allows to: (i) incorporate any ordered objective metrics, in any scale ordering, without the need to change any routines of the source code; this is done using a dictionary of metrics provided by the developer; (ii) incorporate any questionnaire composed by ordered single-answer questions, which is also possible using a dictionary of questions; (iii) incorporate other IVE scenarios based on parameters variations beyond distance and object's size; (iv) make changes or extensions to the scores computation routines, with minor changes in the source code: handling metrics with different weights, setting different weights to objective and subjective scores, adapting scale limits according to domain needs or customs, and setting other criteria for scores assignment.

7 Conclusion

The framework integrates objective and subjective evaluations, gathering data that are analyzed to identify significant differences between stereoscopic and monoscopic techniques.

The comprehensive framework is applicable to different systems within the context of IVEs with object manipulation and navigation, is flexible and allows incorporating any objective and subjective metrics (real or ordinal), and several scenarios based on parameters variations, and changing criteria and parameters in the score computation. It is aimed to provide support for VR systems developers to decide whether a stereoscopy technique must be used in a VR system – and which one. Based on the scores of candidate techniques, developers can choose a given technique based on its cost and effectiveness within a specific domain. Further, the framework can help the developer or another professional to analyze the stereoscopic technique performance on each metric, identifying its strengths and weaknesses.

In future work, we intend to conduct new experiments

considering other parameters found in the literature, such as shade, texture, and lighting. Additionally, other metrics can be included, such as speed and length. We also propose to analyze other IVEs, using the framework in other contexts, as well as other techniques, with other visualization devices. Developers, as engaging experts in the field, could evaluate the framework, although certain results are similar to the results found in the literature when techniques are compared. Such collaboration with developers is planned as part of our future work.

Declarations

Acknowledgements

This article is an extended and revised version of the original work by Silva *et al.* [2022]. The authors would like to thank National Institute of Science and Technology - Medicine Assisted by Scientific Computing (INCT-MACC), second step - 2016-2021; and São Paulo Research Foundation (FAPESP), grants 2014/50889-7 and 2013/07375-0; and the Post-Doctoral National Program of the Brazilian National Council for the Improvement of Higher Education Personnel (PNPD-CAPEs), School of Arts, Sciences and Humanities - Graduate Committee (EACH/CPG number 88/2015).

Authors' Contributions

Sahra K. G. Silva: Conceptualization, Data curation, Investigation, Writing – original draft. Cleber G. Corrêa: Methodology, Investigation, Writing – review and editing. Silvio R. R. Sanches: Writing – review and editing. Marcelo S. Lauretto: Formal analysis, Writing – review and editing. Fátima L. S. Nunes: Supervision, Project administration, Writing – review and editing.

Availability of data and materials

The framework's R programming language code [R Core Team, 2018], along with example files, is available at <https://github.com/lapisusp/statisticalmodel>.

References

- Anderson, A. P., Mayer, M. D., Fellows, A. M., Cowan, D. R., Hegel, M. T., and Buckley, J. C. (2017). Relaxation with immersive natural scenes presented using virtual reality. *Aerospace medicine and human performance*, 88(6):520–526. DOI: 10.3357/amhp.4747.2017.
- Armbrüster, C., Wolter, M., Kuhlen, T., Spijkers, W., and Fimm, B. (2008). Depth perception in virtual reality: distance estimations in peri- and extrapersonal space. *Cyberpsychology & Behavior*, 11(1):9–15.
- Burns, C. G. and Fairclough, S. H. (2015). Use of auditory event-related potentials to measure immersion during a computer game. *International Journal of Human-Computer Studies*, 73:107–114.
- Cecotti, H. (2022). A serious game in fully immersive virtual reality for teaching astronomy based on the messier catalog. In *2022 8th International Conference of the Immersive Learning Research Network (iLRN)*, pages 1–7. DOI: 10.23919/iLRN55037.2022.9815994.
- Clemente, M., Rey, B., Rodríguez-Pujadas, A., Barros-Loscertales, A., Banos, R. M., Botella, C., Alcaniz, M., and Ávila, C. (2013a). An fMRI Study to Analyze Neural Correlates of Presence during Virtual Reality Experiences. *Interacting with Computers*, 26(3):269–284. DOI: 10.1093/iwc/iwt037.
- Clemente, M., Rodríguez, A., Rey, B., and Alcaniz, M. (2013b). Measuring presence during the navigation in a virtual environment using eeg. *Studies in health technology and informatics*, 191:136–140.
- Coakley, C. W. and Heise, M. A. (1996). Versions of the sign test in the presence of ties. *Biometrics*, 52(4):1242–1251. DOI: doi:10.2307/2532840.
- dos Santos, M. C. C., Sangalli, V. A., and Pinho, M. S. (2017). Evaluating the use of virtual reality on professional robotics education. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 448–455. IEEE.
- e Silva, S. K. G. and Nunes, F. L. S. (2015). Depth perception evaluation with different stereoscopic techniques: A case study. In *Virtual and Augmented Reality (SVR), 2015 XVII Symposium on*, pages 52–60.
- Edgington, E. S. and Onghena, P. (2007). *Randomization Tests*. Chapman & Hall/CRC.
- Geuss, M., Stefanucci, J., Creem-Regehr, S., and Thompson, W. (2012). Effect of viewing plane on perceived distances in real and virtual environments. *Journal of Experimental Psychology: Human Perception and Performance*, 38(5):1242–1253. DOI: 10.1037/a0027524.
- Grassini, S. and Laumann, K. (2020). Questionnaire measures and physiological correlates of presence: A systematic review. *Frontiers in Psychology*, 11:349. DOI: 10.3389/fpsyg.2020.00349.
- Hattori, A., Ichi Tonami, K., Tsuruta, J., Hideshima, M., Kimura, Y., Nitta, H., and Araki, K. (2022). Effect of the haptic 3d virtual reality dental training simulator on assessment of tooth preparation. *Journal of Dental Sciences*, 17(1):514–520. DOI: <https://doi.org/10.1016/j.jds.2021.06.022>.
- Heeter, C. (1992). Being there: The subjective experience of presence, telepresence, presence: Teleoperators and virtual environments.
- Jockel, K.-H. (1986). Finite sample properties and asymptotic efficiency of monte carlo tests. *The Annals of Statistics*, 14(1):336–347.
- Leopardi, A., Ceccacci, S., Mengoni, M., Naspetti, S., Gambelli, D., Ozturk, E., and Zanoli, R. (2021). X-reality technologies for museums: a comparative evaluation based on presence and visitors experience through user studies. *Journal of Cultural Heritage*, 47:188–198. DOI: <https://doi.org/10.1016/j.culher.2020.10.005>.
- Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. (2001). A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence: Teleoperators and Virtual Environments*, 10(3):282–297. DOI: 10.1162/105474601300343612.
- Lin, C. J., Abreham, B. T., and Woldegiorgis, B. H.

- (2019). Effects of displays on a direct reaching task: A comparative study of head mounted display and stereoscopic widescreen display. *International Journal of Industrial Ergonomics*, 72:372–379. DOI: <https://doi.org/10.1016/j.ergon.2019.06.013>.
- Livatino, S., De Paolis, L., D'Agostino, M., Zocco, A., Agrimi, A., De Santis, A., Bruno, L., and Lapresa, M. (2015). Stereoscopic visualization and 3-D technologies in medical endoscopic teleoperation. *Industrial Electronics, IEEE Transactions on*, 62:525–535. DOI: 10.1109/TIE.2014.2334675.
- Maciel, C., Bim, S. A., and da Silva Figueiredo, K. (2018). Digital girls program: Disseminating computer science to girls in Brazil. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, GE '18, pages 29–32, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3195570.3195574.
- McMahan, R. P., Gorton, D., Gresock, J., McConnell, W., and Bowman, D. A. (2006). Separating the effects of level of immersion and 3D interaction techniques. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, VRST '06, pages 108–111, New York, NY, USA. ACM.
- Meehan, M., Insko, B., Whitton, M., and Brooks Jr, F. P. (2002). Physiological measures of presence in stressful virtual environments. *Acm transactions on graphics (tog)*, 21(3):645–652.
- Ng, A. K., Chan, L. K., and Lau, H. Y. (2016). Depth perception in virtual environment: The effects of immersive system and freedom of movement. In *International Conference on Virtual, Augmented and Mixed Reality*, pages 173–183. Springer.
- Ochs, M., Mestre, D., De Montcheuil, G., Pergandi, J.-M., Saubesty, J., Lombardo, E., Francon, D., and Blache, P. (2019). Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces*, 13(1):41–51.
- Oehlert, G. W. (2010). *A first course in design and analysis of experiments*. Retrieved from the University of Minnesota Digital Conservancy.
- Oh, C. S., Bailenson, J. N., and Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5:114. DOI: 10.3389/frobt.2018.00114.
- Poels, K., Hoogen, W. v. d., Ijsselsteijn, W., and de Kort, Y. (2012). Pleasure to play, arousal to stay: The effect of player emotions on digital game preferences and playing time. *Cyberpsychology, Behavior, and Social Networking*, 15(1):1–6. PMID: 21875354. DOI: 10.1089/cyber.2010.0040.
- Putter, J. (1955). The treatment of ties in some nonparametric tests. *Annals of Mathematical Statistics*, 26:368–386.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Schubert, T., Friedmann, F., and Regenbrecht, H. (2001). Igroup presence quest.
- Sheridan, T. B. et al. (1992). Musings on telepresence and virtual presence. *Presence Teleoperators Virtual Environ.*, 1(1):120–125.
- Silva, S. K. G., Correa, C. G., and Nunes, F. L. S. (2016). Three-dimensionality perception evaluation in stereoscopic virtual environments: a systematic review. In *Virtual and Augmented Reality (SVR), 2016 XVIII Symposium on*, pages 1–12.
- Silva, S. K. G., Corrêa, C. G., Lauretto, M. S., and Nunes, F. L. S. (2022). A framework for evaluating depth perception in stereoscopic virtual environments. In *Proceedings of the XXIV Symposium on Virtual and Augmented Reality (SVR)*, Natal, RN.
- Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., and Wetzstein, G. (2017). How do people explore virtual environments?
- Slater, M., Spanlang, B., and Corominas, D. (2010). Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *ACM Transactions on Graphics (TOG)*, 29(4):92.
- Slater, M. and Usoh, M. (1993). Representations systems, perceptual position, and presence in immersive virtual environments. *Presence: Teleoperators and virtual environments*, 2(3):221–233.
- Slater, M., Usoh, M., and Steed, A. (1994). Depth of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 3(2):130–144. DOI: 10.1162/pres.1994.3.2.130.
- Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4):73–93. DOI: <https://doi.org/10.1111/j.1460-2466.1992.tb00812.x>.
- Thalmann, D., Lee, J., and Thalmann, N. M. (2016). An evaluation of spatial presence, social presence, and interactions with various 3D displays. In *Proceedings of the 29th International Conference on Computer Animation and Social Agents*, pages 197–204.
- Vienne, C., Masfrand, S., Bourdin, C., and Vercher, J.-L. (2020). Depth perception in virtual reality systems: Effect of screen distance, environment richness and display factors. *IEEE Access*, 8:29099–29110. DOI: 10.1109/ACCESS.2020.2972122.
- Vinnikov, M. and Allison, R. S. (2014). Gaze-contingent depth of field in realistic scenes: The user experience. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14, pages 119–126, New York, NY, USA. ACM. DOI: 10.1145/2578153.2578170.
- Witmer, B. G. and Kline, P. B. (1998). Judging perceived and traversed distance in virtual environments. *Presence*, 7(2):144–167.
- Witmer, B. G. and Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and virtual environments*, 7(3):225–240.
- Yanoff, M. and Duker, J. (2018). *Ophthalmology 5th Edition*. Elsevier.
- Zhao, L., Zhang, Y., Wu, H., and Xiao, J. (2020). The difference of distance stereoacuity measured with different separating methods. *Annals of translational medicine*, 8(7).