A Subjective Quality Assessment of Temporally Reprojected Specular Reflections in Virtual Reality

Martin Mišiak *
TH Köln, Computer Graphics Group
University Würzburg, HCI Group

Arnulph Fuhrmann†
TH Köln, Computer Graphics Group

Marc Erich Latoschik‡
University Würzburg, HCI Group

Abstract

Temporal reprojection is a popular method for mitigating sampling artifacts from a variety of sources. This work investigates it's impact on the subjective quality of specular reflections in Virtual Reality(VR). Our results show that temporal reprojection is highly effective at improving the visual comfort of specular materials, especially at low sample counts. A slightly diminished effect could also be observed in improving the subjective accuracy of the resulting reflection.

Index Terms: Computing methodologies—Computer graphics—Rendering; Human-centered computing—Human computer interaction (HCI)—Empirical studies in HCI

1 Introduction

Specular reflections have a vital role in the perception of material properties and can greatly improve the visual realism of a scene [1]. Current real-time applications are increasingly relying on raytraced reflections to elevate the visual fidelity of virtual environments. With increasing GPU performance, this trend is slowly starting to appear in mainstream Virtual Reality (VR) applications as well.

Raytraced reflections, however, are prone to noise artifacts, as reflection rays are generated using a spatio-temporal varying random number distribution. Due to the large pixel magnification of current head-mounted displays, such artifacts are easily noticeable by the users. One popular approach of mitigating this problem is Temporal Anti-Aliasing (TAA) [7], which reprojects color values from previous frames, hence implicitly increasing the effective number of rays. TAA is a fundamental technique in the modern rendering stack as it can be used to reduce a wide range of aliasing sources. Mäkitalo et al. [2] have shown that spatio-temporal reprojection can improve the effective sample count of stereoscopic path-traced images by a factor of up to 25 when considering a low number of input rays per pixel. Despite these findings, the resulting perceptual quality of such images remains largely unexplored in a VR context. In this work, we investigate the impact of temporal reprojection on the subjective quality of specular reflections presented on a VR headset.

2 Stimuli

In our experiment, we compare a reference reflection method against a method with real-time parametrization and temporal sample accumulation. The reference method uses a very high sample count of 256 rays per pixel. As this sample number is not achievable in real-time when using raytraced reflections on current consumer hardware, we employ a parallax-corrected cubemap technique [6]. The test stimuli are rendered with the same method, however, with a significantly lower sample count of $r = [8, 16, 32]$. In addition, the test stimuli are temporally accumulated with an exponential-weight of $\alpha = [1, 0.5, 0.2, 0.05]$. Our TAA-implementation is only applied to the reflective pixels and uses a variance based nearest neighbourhood clipping [4] (with $\gamma = 1$). To preserve the sharpness of specular reflections a bicubic Catmull-Rom filter is used when sampling the history buffer.

All comparisons are made for three different material smoothness values of $m = [0.95, 0.9, 0.8]$, which results in a total of 3x4x3 = 36 comparisons.

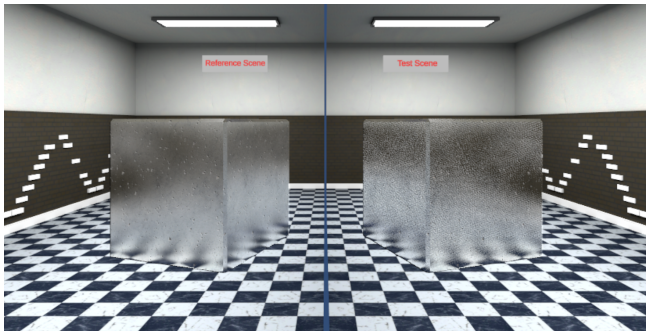

Figure 1: Example trial: Left side shows the reference reflection method with 256spp. Right side shows the test condition method with 8spp and a α value of 1 (TAA off). Both cubes use the same material smoothness level of 0.8.

3 Experiment

Our experiment is based on the Double Stimulus Impairment Scale (DSIS) methodology as proposed by Nehmé et al. [3], where participants simultaneously compare a test stimulus against a known reference. The participants are seated in front of two reflective cubes. The left cube is rendered using the reference method, and the right cube is rendered using a random test parametrization (Figure 1). In each trial, the participants are tasked with comparing the reflections in both cubes. After 20 seconds, the participants are asked if they perceived a difference between both stimuli, as well as how annoying the difference is. The rating options are on a 5-point scale with *(1) very strong difference / very strongly annoying* to *(5) no perceivable difference / not annoying*. For annoyance, the participants received the following definition: " If the reflections of an object in a VR experience would behave like in the current test condition. How annoying would this be to you ?" Participants remain seated for the duration of the experiment, however, they are strongly encouraged to lean in all directions and view the cubes from varying positions. To detect any form of "wild guessing" from the participants, we included three (one for each material smoothness) reference-reference comparisons into the existing trials.

Thirteen individuals (7M, 6F) aged 24–57 were recruited for the experiment. All participants had normal or corrected to normal vision and were naive to the goals of the study. The prototype is implemented in Unity and presented on a HP Reverb G2, with a

*e-mail: martin.misiak@th-koeln.de
†e-mail: arnulph.fuhrmann@th-koeln.de
‡e-mail: marc.latoschik@uni-wuerzburg.de

rendering resolution of 2236x2184 pixels per eye and a refresh rate of 90 Hz.

4 Results

One participant failed to identify the reference-reference conditions and was excluded from further analysis. For the remaining 12 participants, the collected difference and annoyance ratings (D,A) are averaged into a Mean Opinion Score (MOS) for each tested condition (r,m,α):

$$MOS_D(r,m,\alpha) = \frac{1}{N}\sum_i^N D_i^{rm\alpha} \qquad MOS_A(r,m,\alpha) = \frac{1}{N}\sum_i^N A_i^{rm\alpha} \tag{1}$$

The obtained scores can be seen in Figure 2.

Spearman's rank correlation was computed to assess the relationship between overall annoyance and difference ratings. There was a very strong positive correlation between the two output variables ($r(466) = 0.822, p < 0.001$). In general, participants rated visual annoyance higher (32.6%) or equal (64.2%) to visual difference, while the latter received higher ratings only in a very small portion (3.2%) of the gathered data. The highest difference between MOS_A and MOS_D scores can be seen for conditions with a low sample count ($r = 8$) and active TAA ($\alpha \leq 0.2$), where sampling noise is heavily suppressed, yet the sample count is insufficient to reconstruct the reflection correctly. An example for this is the ($m = 0.8$, $r = 8$, $\alpha = 0.05$) condition, where a Wilcoxon Signed-Rank test indicated a significant difference between the annoyance and difference ratings (W = 3, p < .004).

Because of the strong correlation between annoyance and difference scores, going forward we will consider only the annoyance ratings in our analysis. To determine the influence of material smoothness and sample count on the resulting MOS_A, we consider data points where TAA is turned off ($\alpha = 1.0$). As expected, the MOS_A decreases rapidly with material smoothness ($MOS_A(*,0.95,1.0) = 4.58$ | $MOS_A(*,0.9,1.0) = 3.61$ | $MOS_A(*,0.8,1.0) = 2.92$). Given a fixed number of reflection rays, the risk of undersampling the material appearance increases with the size of its specular lobe. In contrast, the number of reflection samples has a positive influence on the MOS_A scores. As more samples are available, the specular lobe can be sampled more densely ($MOS_A(8,*,1.0) = 2.83$ | $MOS_A(16,*,1.0) = 3.94$ | $MOS_A(32,*,1.0) = 4.33$), resulting in less visible sampling noise.

An important parameter of interest is the temporal accumulation weight α. A Friedman's test showed that there was a significant difference between the annoyance ratings obtained for the 4 tested α-values ($\chi^2(3,N = 12) = 55.64, p < .001$). Post-Hoc comparisons using the Wilcoxon Signed-Rank test with Bonferroni correction showed significant differences between all pairs of α-values, except between $\alpha = 0.20$ and $\alpha = 0.05$. As α is lowered, the previous frames are weighted more during image composition. This increases the effective number of reflection samples, as samples from previous frames are reused in the current frame. This is reflected in an increasing MOS_A when grouped by the α parameter ($MOS_A(*,*,1.0) = 3.70$ | $MOS_A(*,*,0.5) = 4.18$ | $MOS_A(*,*,0.2) = 4.69$ | $MOS_A(*,*,0.05) = 4.66$).

5 Conclusion

Based on these results, we can conclude that temporal reprojection significantly improves the subjective quality of specular reflections in VR. A naive increase in samples does not scale well to rougher materials. In our comparisons, even 32 reflection rays were insufficient to guarantee a pleasant viewing experience ($MOS_A > 4.0$) of the roughest tested material. If we consider that most real-time raytracing applications currently have a budget of less than 8 reflection rays per pixel, the usage of temporal reprojection becomes

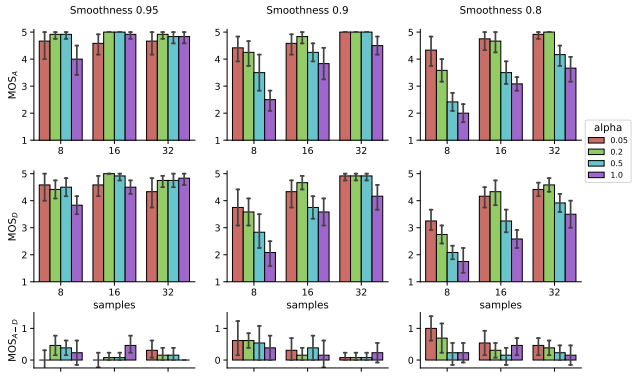

Figure 2: Mean Opinion Scores with 95% CI for visual annoyance (MOS_A) and visual difference (MOS_D) of our experiment. The third row shows the differences between annoyance and difference ratings for each condition. The scores are computed for 3 different numbers of reflection samples (8,16,32) and 4 α values (1 [TAA off], 0.5, 0.2, 0.05). Each figure column represents a different material smoothness value (0.95, 0.9, 0.8).

mandatory if visual comfort is the goal. While the subjective reflection accuracy is also increased when reprojecting samples via a TAA implementation, its influence in this domain is diminished when compared to visual comfort. If an accurate reflection result is needed, employing more samples or using a specialized temporal accumulation method for reflection rays should be considered.

The highest scores were achieved with an accumulation weight $\alpha \leq 0.2$. Interestingly $\alpha = 0.05$ did not provide a major advantage over 0.2, nor did it prove problematic due to an increased likelihood of ghosting artifacts as initially anticipated. Nonetheless, we currently recommend a more defensive parametrization of $\alpha = 0.2$, which is consistent with related work, where TAA was used in a non-VR context [2, 5].

Acknowledgments

This work was funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under grant number 005-2105-0046 as part of the project KoViTReK. .

References

[1] M. Elhelw, M. Nicolaou, A. Chung, G.-Z. Yang, and M. S. Atkins. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Transactions on Applied Perception (TAP)*, 5(1):1–20, 2008.

[2] M. J. Mäkitalo, P. E. Kivi, and P. O. Jääskeläinen. Systematic evaluation of the quality benefits of spatiotemporal sample reprojection in real-time stereoscopic path tracing. *IEEE Access*, 8:133514–133526, 2020.

[3] Y. Nehmé, J.-P. Farrugia, F. Dupont, P. L. Callet, and G. Lavoué. Comparison of subjective methods for quality assessment of 3d graphics in virtual reality. *ACM Transactions on Applied Perception (TAP)*, 18(1):1–23, 2020.

[4] M. Salvi. An excursion in temporal super sampling. In *Game Developers Conference*, vol. 3, p. 12, 2016.

[5] C. Schied, A. Kaplanyan, C. Wyman, A. Patney, C. R. A. Chaitanya, J. Burgess, S. Liu, C. Dachsbacher, A. Lefohn, and M. Salvi. Spatiotemporal variance-guided filtering: real-time reconstruction for path-traced global illumination. In *Proceedings of High Performance Graphics*, pp. 1–12. 2017.

[6] L. Sébastien and A. Zanuttini. Local image-based lighting with parallax-corrected cubemaps. In *ACM SIGGRAPH 2012 Talks*, pp. 1–1. 2012.

[7] L. Yang, S. Liu, and M. Salvi. A survey of temporal antialiasing techniques. In *Computer graphics forum*, vol. 39, pp. 607–621. Wiley Online Library, 2020.