Perceptual Comparison of Four Upscaling Algorithms for Low-Resolution Rendering for Head-mounted VR Displays

Mario König*, Martin Mišiak*†, Arnulph Fuhrmann*

* TH Köln
Computer Graphics Group
mario.koenig@smail.th-koeln.de
arnulph.fuhrmann@th-koeln.de

† Universität Würzburg
HCI Group
martin.misiak@th-koeln.de

Abstract: When rendering images in real-time, shading pixels is a comparatively expensive operation. Especially for head-mounted displays, where separate images are rendered for each eye and high frame rates need to be achieved. Upscaling algorithms are one possibility of reducing the pixel shading costs. Four basic upscaling algorithms are implemented in a VR rendering system, with a subsequent user study on subjective image quality. We find that users preferred methods with a better contrast preservation.

Keywords: real-time rendering, virtual reality, upscaling, upsampling

1 Introduction

In recent years, the field of virtual reality has seen increased activity in research as well as general adoption of the technology. In response to user demands for more clarity and less "screen-door effect", hardware manufacturers have been steadily increasing the display resolutions. This increase however puts even more pressure on the already heavily utilized pixel shading stage as more pixels need to be processed. Although a high resolution is desirable, a fast frame rate has to be maintained, as it strongly influences user performance [CC09] and is a prerequisite for immersive systems [Abr14, CB16]. Therefore techniques are required, that significantly decrease the pixel shading costs. Foveated rendering techniques look very promising in this regard, however they are very dependent on the quality of the used eyetracking solution, and are not widely available.

Another class of techniques are upscaling algorithms. Their goal is to reconstruct an image from a lower resolution one by interpolating between existing information. In our work, a perceptual comparison of four upscaling algorithms in the context of VR rendering is done. While a wide range of algorithms is available, we constrain our comparison to the methods: *nearest neighbour*, *bilinear*, *bicubic* and *bilateral upscaling*. We chose these basic methods as a first step towards more sophisticated methods in the future, as to our knowledge, a perceptual comparison of upscaling algorithms for VR has not been done before. Additionally, with the advent of standalone VR devices such as the *Oculus Quest*, the low

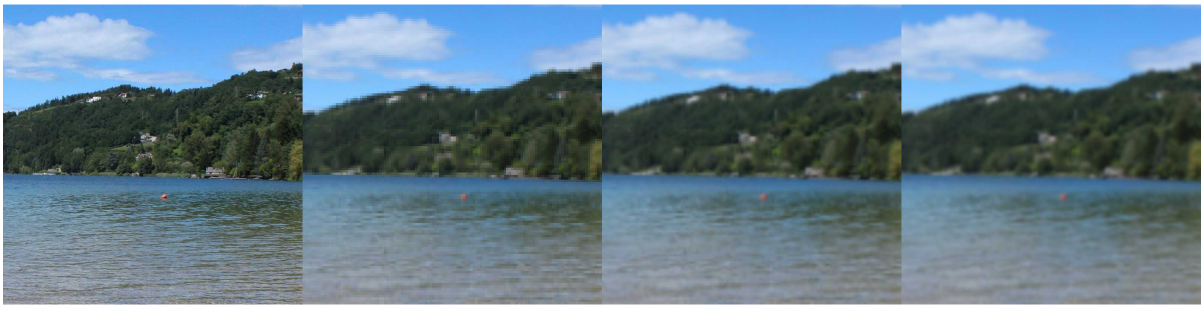


Figure 1: Comparison between different upscaling techniques. The original is rendered in 1000x1000 pixels, the other images are upscaled from 125x125. From left to right: Full render, Nearest Neighbor, Bilinear Interpolation and Bicubic Interpolation.

computational cost of these techniques becomes very appealing.

Related work from the field of image processing focuses on algorithms which are computationally too expensive to be used in a real-time context [NM14]. As are current machine learning approaches, e.g. Shi et al. [SCH$^+$16], especially considering rendering for VR. Nevertheless vendors are supporting these techniques by adding hardware acceleration like Nvidia's Tensor Cores and machine learning algorithms are already in use in real-time in related contexts, e.g. Deep Learning Super Sampling for anti-aliasing [KKN$^+$18].

Evaluation of upscaling algorithms is often done via perceptually based automatic image quality assessment metrics, which can only be directly applied to still images [YXPW10].

2 Upscaling Algorithms

For our study four different interpolation techniques are implemented and tested. *Nearest neighbor* is the most basic texture filtering technique. When sampling an image or a texture, the color of the pixel that has its pixel center closest to the sampling point is reused. No interpolation is done. This often leads to a blocky looking result. *Bilinear interpolation* can be considered as the standard filtering technique in computer graphics. For a sampled point, the resulting color will be interpolated between the colors of the four closest surrounding pixels. First, two linear interpolations are done for the two pixels in each row, i.e. in x-direction. The two results are then filtered linearly again in y-direction. The interpolation weights are determined by the distances of the pixel centers to the sample point. *Bicubic interpolation* works very similar to linear interpolation, but instead of the linear function a third-degree polynomial function is used. The sampling area used is 4×4 pixels, meaning that the final color is interpolated between 16 different pixels. While the aforementioned techniques operate only on a pixels color, *bilateral filtering* uses additional scene information, like depth and normals, which are usually only at disposal in virtual 3D scenes. The interpolation itself is done on four pixels as in *bilinear interpolation*, however the individual contributions are additionally weighted. The weight depends on the difference in depth: the more similar the depth, the higher the weight; and on the angle between the surface normals: the smaller the angle, the higher the weight. This technique provides better quality by preserving edges.

2.1 Implementation

The implementation uses a Java-based *OpenGL* rendering framework using *LWJGL 3.15* with support for Valve's *OpenVR* API. All four techniques work by rendering the low resolution image to a texture, then doing a full screen post-process render pass to upscale the image.

For the *nearest neighbor* and *bilinear upscaling* techniques the implementation is as easy as setting the low resolution textures magnification filters to GL_NEAREST, respectively to GL_LINEAR, then sampling the texture. These techniques are hardware accelerated, while the other algorithms need to be implemented in software. The implementation of the *bicubic interpolation* in this paper follows the method proposed by Pharr and Fernando [PF05]. In order to reduce the amount of texture lookups from 16 to 4, they break down the sampling grid to a 2×2 grid with additional precomputed weights, which can be sampled using bilinear filtering. There are different takes on the implementation of *bilateral filtering.* Our paper follows the presentation of Shopf [Sho09]. At first, normals and depth are written to a second texture during the low resolution render pass making use of multiple render targets. Weights are then calculated independently for depth and normals using the formulas proposed by Shopf [Sho09] and multiplied with the weights calculated for a standard bilinear interpolation.

3 User study

The aim of the user study was a perceptual comparison of the four upscaling algorithms and determine the order of subjective preference. These four algorithms were only compared against each other, as a comparison to a high resolution ground truth rendering would not yield any relevant information regarding our specific research question. To assess if the magnitude of the scaling influences the perceived quality, the techniques were evaluated at 50% and 25% resolution resulting in 8 different rendering conditions.

As rating and ranking all 8 different rendering techniques within a 3D scene at once would be an unnatural and very strenuous task for the participants, *pairwise comparisons* were conducted instead [SCJ88]. The setup and the evaluation of our study is based on the experimental design of Ledda et al. [LCTS05].

3.1 Experimental Design and Participants

The study was conducted with 10 participants, five male and five female, in a within-group design. Seven of the participants were in their twenties, three in their fifties. While three of them had never experienced virtual reality, seven of them had already participated in at least two VR experiments.

The study was conducted using an Oculus Rift consumer version and a computer capable of always providing a frame rate of at least 90 FPS, the displays refresh rate. For each participant a full calibration of the HMD was done to prevent any vision problems that did

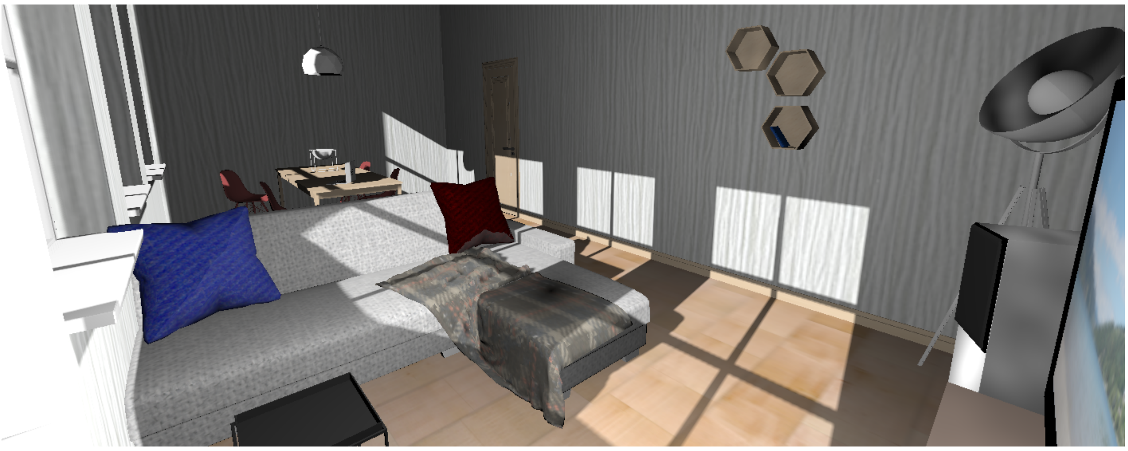


Figure 2: Overview of the used test scene. During the study participants were positioned directly in front of the sofa.

not arise from the evaluated algorithms, like blur from a poorly fitting headset. In addition, each participant was provided the chance to experience three to four minutes of Oculus standard scenes to get accustomed to virtual reality. The test scene used during the study is depicted in Figure 2.

During the experiment every participant had to evaluate every condition against each other, resulting in $\binom{8}{2} = 28$ comparisons. For each comparison the participant had to decide, which technique yielded the subjectively higher quality rendering. To avoid ordering effects, the order of comparisons was completely random.

The results can be presented in a 8×8 preference matrix. Figure 3 shows this matrix for one participant on the left and for all participants with summed up values on the right. One can see for example that the single observer considered *nearest neighbor* at 50% resolution (NN50) better than *bicubic interpolation*, but worse than *bilateral interpolation* at the same resolution. The sums of the rows show how often the respective algorithm was considered better than the alternative, e.g. *bilinear interpolation* at 50% (BiLin50) was considered better in six out of seven comparisons by this participant.

3.2 Evaluation

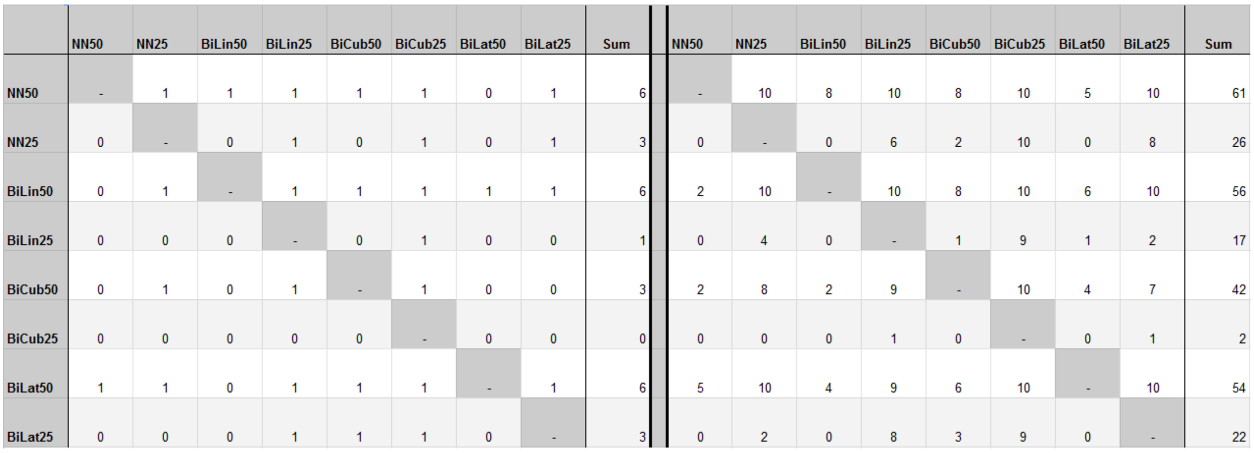
To correctly classify the results the *coefficient of agreement* as proposed by Kendall and Babington-Smith [KS40] is calculated. It is defined as

$$u = \frac{2\Sigma}{\binom{s}{2}\binom{t}{2}} - 1, \tag{1}$$

where

$$\Sigma = \sum_{i \neq j} \binom{p_{ij}}{2}, \tag{2}$$

where s is the number of subjects, t is the number of techniques (8 in our case) and p_{ij} is the number of times that algorithm i is preferred over algorithm j. The smaller the agreement



	NN50	NN25	BiLin50	BiLin25	BiCub50	BiCub25	BiLat50	BiLat25	Sum	NN50	NN25	BiLin50	BiLin25	BiCub50	BiCub25	BiLat50	BiLat25	Sum
NN50	-	1	1	1	1	1	0	1	6	-	10	8	10	8	10	5	10	61
NN25	0	-	0	1	0	1	0	1	3	0	-	0	6	2	10	0	8	26
BiLin50	0	1	-	1	1	1	1	1	6	2	10	-	10	8	10	6	10	56
BiLin25	0	0	0	-	0	1	0	0	1	0	4	0	-	1	9	1	2	17
BiCub50	0	1	0	1	-	1	0	0	3	2	8	2	9	-	10	4	7	42
BiCub25	0	0	0	0	0	-	0	0	0	0	0	0	1	0	-	0	1	2
BiLat50	1	1	0	1	1	1	-	1	6	5	10	4	9	6	10	-	10	54
BiLat25	0	0	0	1	1	1	0	-	3	0	2	0	8	3	9	0	-	22

Figure 3: Preference matrices - left: for one participant; right: summed up for all 10 subjects. The number in a cell indicates the number of times the algorithm in the row was considered better than the one in the corresponding column. NN - *nearest neighbor*, BiLin - *bilinear filtering*, BiCub - *bicubic filteing* and BiLat - *Bilateral Filtering*, each at 50 and 25% resolution.

between participants is the smaller will be u. If all participants make the same choice u will take its maximal value of 1.

The so called *consistency* or *transitivity* will also be considered. If an observer makes a choice for three conditions, **A**, **B** and **C** and votes as **A** → **B**, **B** → **C** and **A** → **C**, the choices are considered consistent (the arrow indicating that the first option was chosen over the second one). If he instead votes **A** → **B**, **B** → **C** and **C** → **A** his choices are circular, meaning inconsistent. Inconsistency can regularly happen when using *pairwise comparison*. If most of the participants are inconsistent, we can conclude that the compared techniques are very similar and it is difficult to make consistent decisions. The *coefficient of inconsistency* is defined as [KS40]:

$$\zeta = 1 - \frac{24c}{t^3 - 4t} \tag{3}$$

where c is the number of circular triads per subject, that can be determined as

$$c = \frac{t}{24}(t^2 - 1) - \frac{1}{2}T \tag{4}$$

with $T = \sum(p_i - (t-1)/2)^2$, where p_i is the number of preferences scored by technique i [Dav63]. $\zeta = 1$ means no circular triads, which in turn leads to the conclusion that the evaluated techniques can be ranked.

3.3 Results

The results of the experiment are shown in the preference matrix in Figure 3. The numbers in each cell present the number of times that the technique in the row was considered better than the one in the corresponding column. In total the algorithms that won the most comparisons are *nearest neighbor* (61, 87%), *bilinear interpolation* (56, 80%) and *bilateral interpolation* (54, 77%), all at 50% resolution.

To test if the results bear any meaning at all the significance of the *coefficient of agreement* is calculated first. This was done analogous to [LCTS05] by using the null hypothesis H0 that there is no agreement between the participants and the alternative hypothesis H1 that the degree of agreement is greater than if results had been acquired randomly. By using the chi-squared test statistics $\chi^2 = \frac{t(1-t)(1+u(s-1))}{2}$ [SCJ88] our experiment proved to be statistically significant with a coefficient of agreement of $u = 0.5317$.

Furthermore a significance test of the score differences is performed in order to see whether the perceptual qualities of any two algorithms are statistically distinguishable. Following [LCTS05] and [Dav63] the formula $R' = \frac{1}{2}W_{t,\alpha}\sqrt{st} + \frac{1}{4}$, with $W_{t,\alpha}$ obtained from a statistical table [PH66], yields the minimal difference in score values R^+ (the smallest integer greater than R') to be statistically significant. In our case R^+ evaluates to 8. Relating this to the results shown in Figure 3, there is no significant difference between the three most preferred algorithms, but the fourth best method, bicubic interpolation at 50%, is significantly rated worse compared to the winners.

Going by this criteria one cannot create a definitive ranking of the algorithms, but certain important observations can be made. All of the renderings at 50% resolution are overall significantly better rated than any rendering at 25% resolution. The only outstanding observation in comparisons between 25% and 50% is that *bilateral upscaling* at 25% was considered better than *bicubic interpolation* at 50% in 3 out of 10 comparisons. The average *coefficient of consistency* ζ is very high at 0.89, with the lowest scoring at 0.8 and two participants being perfectly consistent (i.e. 1.0). This is probably an effect of 50% resolution renderings being considered better than 25% resolution renderings in nearly every case.

4 Conclusion

The study shows that the advantages of higher resolution cannot be caught up to by the tested algorithms. A quick assessment with three participants checking against a full resolution render that was not part of the study points in the same direction - in every single case the full resolution render was perceived as of better quality.

It seems like sharpness was the number one priority for the participants as the interpolation methods scored lower the more blur they introduce. This observation matches the aforementioned fact, that *bilateral upscaling* at 25% was considered better than *bicubic interpolation* at 50% resolution three times, as the bilateral algorithms preserves edges better. This finding can be explained by the high importance of local contrast, as reported by Patney et al. [PSK$^+$16]. Surprisingly *nearest neighbor* was perceived as well as *bilinear interpolation* and *bilateral filtering.* In no case did a participant complain about blocking artifacts. This is interesting, as *nearest neighbor* is very prone to spatio-temporal artifacts, which limits its use in a pure foveation based approach [HMT18].

In their respective resolution tiers the ordering is the same with *bicubic interpolation* scoring lowest and no significant difference between the remaining three algorithms. There being no definitive ranking for the top three algorithms is confirmed by the coefficients of agreement (0.53) and consistency (0.89): while the participants are consistent in their choice which technique they consider as better themselves, they do not agree as often. Hence it can be resorted to using the fastest of the three algorithms.

4.1 Future work

Since this study focuses on very basic upscaling algorithms, there are still a variety of more sophisticated techniques that should be considered for future evaluations. In particular spatio-temporal techniques [HEMS10], which are considered state-of-the-art in a real-time rendering context and machine learning based approaches [SCH+16]. Additionally, a reference comparison could be made for different coarse pixel shading techniques [VST+14], which are more oriented towards image quality than performance.

Recent studies also indicate that simply adjusting contrast, which is partially lost during interpolation, increases the perceived image quality [PSK+16]. Therefore, it might be worthwhile to evaluate if the basic techniques that result in more blurry images are considered better, if the contrast of the resulting image is adjusted afterwards.

References

[Abr14] Michael Abrash. What VR could, should, and almost certainly will be within two years. *Steam Dev Days, Seattle*, 4, 2014.

[CB16] James J. Cummings and Jeremy N. Bailenson. How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, 19(2):272–309, 2016.

[CC09] Mark Claypool and Kajal Claypool. Perspectives, frame rates and resolutions: it's all in the game. In *Proceedings of the 4th International Conference on Foundations of Digital Games*, pages 42–49. ACM, 2009.

[Dav63] Herbert A. David. *The method of paired comparisons*, volume 12. Charles Griffin and Company, 1963.

[HEMS10] Robert Herzog, Elmar Eisemann, Karol Myszkowski, and H.-P. Seidel. Spatio-temporal upsampling on the GPU. In *Proceedings of the 2010 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 91–98. ACM, 2010.

[HMT18] David Hoffman, Zoe Meraz, and Eric Turner. Limits of peripheral acuity and implications for VR system design. *Journal of the Society for Information Display*, 26(8):483–495, 2018.

[KKN+18] Alexander Keller, Jaroslav Křivánek, Jan Novák, Anton Kaplanyan, and Marco Salvi. Machine learning and rendering. In *ACM SIGGRAPH 2018 Courses*, SIGGRAPH '18, pages 19:1–19:2, New York, NY, USA, 2018. ACM.

[KS40] Maurice G. Kendall and B. Babington Smith. On the method of paired comparisons. *Biometrika*, 31(3/4):324–345, 1940.

[LCTS05] Patrick Ledda, Alan Chalmers, Tom Troscianko, and Helge Seetzen. Evaluation of tone mapping operators using a high dynamic range display. *ACM Transactions on Graphics (TOG)*, 24(3):640–648, 2005.

[NM14] Kamal Nasrollahi and Thomas B. Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014.

[PF05] Matt Pharr and Randima Fernando. *Gpu gems 2: programming techniques for high-performance graphics and general-purpose computation.* Addison-Wesley Professional, 2005.

[PH66] Egon Sharpe Pearson and Herman Otto Hartley. *Biometrika tables for statisticians.* Cambridge University Press, 1966.

[PSK$^+$16] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 2016.

[SCH$^+$16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[SCJ88] Sidney Siegel and John N. Castellan Jr. *Nonparametric statistics for the behavioral sciences.* Mcgraw-Hill Book Company, 1988.

[Sho09] Jeremy Shopf. Mixed resolution rendering. In *Game Developers Conference*, 2009.

[VST$^+$14] Karthik Vaidyanathan, Marco Salvi, Robert Toth, Tim Foley, Tomas Akenine-Möller, Jim Nilsson, Jacob Munkberg, Jon Hasselgren, Masamichi Sugihara, Petrik Clarberg, et al. Coarse pixel shading. In *Proceedings of High Performance Graphics*, pages 9–18. Eurographics Association, 2014.

[YXPW10] Junyong You, Liyuan Xing, Andrew Perkis, and Xu Wang. Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis. In *Proc. Int. Workshop Video Process. Quality Metrics Consum. Electron*, volume 9, pages 1–6, 2010.