Supplement to Foveated 3D Graphics: User Study Details

Brian Guenter Mark Finch Steven Drucker Desney Tan John Snyder
Microsoft Research

1 Overview

We conducted three different experiments with the same 15 subjects to test the conditions for which foveated rendering produces both acceptable quality and quality equivalent to non-foveated rendering: a pair test, a ramp test, and a slider test. All tests were based on a 1D space of foveation quality described in Section 2.

The pair test presented each user with pairs of short animated sequences, each 8 seconds long and separated by a short interval (0.5s) of black. The reference element of the pair used non-foveated rendering; the other used foveated rendering at quality levels from $j = 8$ (low quality) to $j = 22$ (high quality). Pairs at all quality levels in this range were presented twice, in both orders (non-foveated then foveated, and foveated then non-foveated). After seeing each pair, users reported whether the first rendering was better, the second was better, or the two were the same quality. The experiment was designed to interrogate what foveation quality level was comparable to non-foveated rendering.

The ramp test presented each user with a set of short sequences, in which the foveation quality incrementally ramped either up to or down from a reference quality of $j_0 = 22$ to a varying foveation quality in the set $j_1 \in \{4, 5, 7, 10, 12, 15, 17, 22\}$. Users were then asked whether the quality had increased, decreased, or remained the same over each sequence. Each ramp was presented in both directions ($j_0 \rightarrow j_1$ and $j_1 \rightarrow j_0$), sampled using 5 discrete steps, each 5 seconds long and separated by a short interval of black. The study was designed to find the lowest foveation quality perceived to be equivalent to a high quality setting in the absence of abrupt transitions.

Finally, the slider test let users navigate the foveation quality space themselves. Users were first presented with a non-foveated animation as a reference. Then starting at a low level of foveation quality ($j = 4$), users could ask the study administrator[1] to increase the level, show the non-foveated reference again, or decrease the level, with the stated task of finding a quality level equivalent to the non-foveated reference. We then recorded the first quality level index at which users stopped increasing the level and instead compared it to the reference. This test also explored the effect of animation speed on the demand for foveation quality, by running the slider test across six different speeds of the moving camera, a stopped but panning camera, and a completely static camera, yielding a total of 8 different camera motions. Each camera motion was presented to each subject twice, for a total of 16 separate slider tests.

We used the LG W2363D 120Hz LCD monitor of resolution 1920×1080 for our user study. Display configuration parameters were $V^* = 59\text{cm}$, $W^* = 51\text{cm}$, $D^* = 1920$, and $\alpha^* = 9/16 = 0.5625$. This yields an angular display radius of $e^* = 23.37°$, and an angular display sharpness of $\omega^* = 0.0516°$, which represents only a fraction of human foveal acuity, $\omega_0/\omega^* \approx 40\%$. Refer to the main paper for an explanation of these quantities.

[1]Because looking down at the keyboard can cause the eye tracker to momentarily lose tracking, users were not given keyboard access directly. Instead they communicated with the study administrator who controlled the UI.

Figure 1: Scene from our formal user study.

Graphical content for the study involved a moving camera through a static 3D scene, composed of a terrain, a 20×20 grid of various objects positioned above it, and an environment map showing mountains and clouds; see Figure 1. The objects range from diffuse to glossy and were rendered with various types of procedural shaders, including texture and environment mapping. The non-foveated reference renders at about 40Hz, as high as our system supports.

Synchronization to vertical display refresh (v-sync) was disabled for both the foveated and non-foveated methods.[2]

Representative animations are contained in the accompanying video.

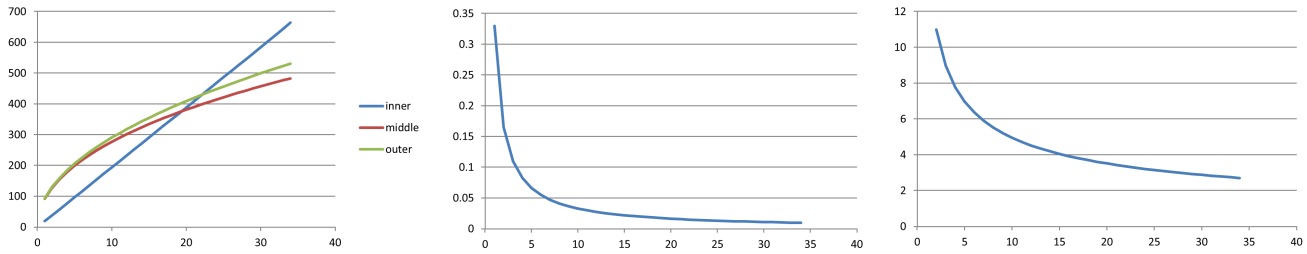
2 Design of a 1D Space of Foveation Quality

Navigating a high-dimensional space of eccentricity layer rendering parameters complicates user study, so we collapsed it into a 1D space. Each point in this space, Q_j, represents a set of eccentricity layer sizes and their corresponding sampling factors. A small index j denotes an aggressively foveated rendering which assumes a rapid falloff of peripheral acuity and provides more savings over non-foveated rendering, but may yield noticeable blurring or twinkling artifacts in the periphery. Larger indices denote a less aggressive peripheral falloff which saves less. Our 1D space involves a straightforward application of layer optimization, driven by a linearly increasing angular size for the inner layer (Eq. 1, below) from which the MAR slope can be derived (Eq. 2, below).

Our space uses three eccentricity layers. The angular radius of the inner layer, e_1, is given by

$$Q_j.e_1 = j \triangle e. \tag{1}$$

[2]Disabling v sync can cause tearing, and in hindsight we might better have enabled it. Disabling it does not affect worst-case system latency assuming the frame render time is less than but nearly equal to the vertical refresh interval (8ms).



(a) $\sqrt{P_i}(j)$ **(b)** $m(j)$ **(c)** $s_3(j)$

Figure 2: 1D space of foveation quality for user studies, Q_j. Our space fixes sampling factors for the inner and middle layers at $s_1 = 1$ and $s_2 = 2$, uses an inner layer angular size, e_1, that increases linearly with j, and then searches for an optimal middle layer angular size e_2 and an outer layer sampling factor s_3. (a) plots the square root of the number of pixels in each layer as a function of j. For comparison, the square root of total (non-foveated) pixels on our display is $\sqrt{P^*} = 1440$. (b) plots MAR slope m as a function of j. (c) plots the optimized sampling factor of the outer layer, s_3, as a function of j.

The sampling factor of the intermediate layer is restricted to $s_2 = 2$. With these assumptions, the MAR slope at the j-th foveation quality point is determined by Eqs. 1 and 2 in the main paper, via

$$Q_j.m = \frac{\omega_1 - \omega_0}{Q_j.e_1} = \frac{s_2\,\omega^* - \omega_0}{j \triangle e}. \tag{2}$$

Using this slope, we then derive the angular size of the middle layer and the sampling factor of the outer layer as explained in Section 4.4 in the main paper. The required optimization searches over the single parameter e_2, since e_1 is specified by Eq. 1. An unrestricted three-layer optimization (where s_2 can also vary) provides only slightly more savings under our study's display configuration settings.

All experiments used the same 1D foveation space with $\triangle e = 1/4°$ and angular display radius $e^* = 23.37°$, yielding a total of 92 discrete points Q_j. Figure 2 shows how the sizes of the three eccentricity layers, the MAR slope m, and the outer sampling factor s_3 vary with increasing foveation quality j. At very large quality levels, $j > 34$, a two-layer decomposition becomes more efficient than a restricted three-layer one. Users perceive quality comparable with non-foveated rendering significantly before reaching this limit.

Reprojection blur parameters The space must also specify the reprojection blur parameters. The parameter b represents how much to blend in the previous, reprojected frame. A blur factor is associated with each layer, denoted by a subscript ($b_1$for the inner layer's blur factor, $b_2$for the middle layer, and $b_3$for the outer layer). We chose a fixed blur factor for the inner layer at $b_1 = 0.3$ and for the intermediate layer at $b_2 = 0.47$. These settings provide good antialiasing of our graphical content at the corresponding fixed sampling factors of $s_1 = 1$ and $s_2 = 2$. The sampling factor of the outer layer, s_3, varies with j based on the optimization above. We assume that good antialiasing at a spatial sampling factor of s requires the temporal averaging of $\beta\, s^2$ jittered frames, for some constant β.

Since the previous frame itself contains a blend of previous frames, reprojection sums up progressively older frames via a simple IIR (infinite impulse response) filter. A frame of age k contributes to the current frame with an exponentially diminishing weight b^k where $b < 1$ is the blur factor. Smaller blur factors thus lead to faster attenuation of an older frame's contribution. The steady-state ratio of the contribution of the last M frames to the total contribution summed over all frames is given by

$$R(b;\, M) = \frac{\sum_{k=0}^{M-1} b^k}{\sum_{k=0}^{\infty} b^k} = 1 - b^M \tag{3}$$

Ensuring this ratio exceeds some desired threshold, $R(b;\, M) \geq \tau$, requires that $b \leq (1-\tau)^{1/M}$. Finally, letting $M = \beta\, s_3^2$, we obtain the following formula for the blur factor of the outer layer given its sampling factor s_3:

$$b_3 = (1-\tau)^{\frac{1}{\beta\,(s_3)^2}} = (\gamma)^{\frac{1}{(s_3)^2}}. \tag{4}$$

Assuming a reasonable threshold $\tau = 0.75$, we chose $\beta = 0.2$ ($\gamma = 0.000977$) based on informal study of a few users.

3 Study Results

All tests were designed as a within-subjects experiment: each subject observed all conditions. Stimuli were presented in four pseudo-random sequences that varied presentation order of the foveated/non-foveated rendering for the pair test, or reference-quality/variable-quality foveated rendering for the ramp test. Raw data for the pair test is shown in Figure 3, the ramp test in Figure 4, and the slider test in Figure 5.

The 15 subjects comprised 8 males and 7 females ranging in ages from 18 to 48, with a median age of 29 and a mean age of 32. Subjects were recruited from the broader community and worked in a wide range of professions. Six of the subjects had corrected vision; five wore contact lenses (1, 5, 7, 8, 9, marked in light green in the result plots) and one wore glasses (2, marked in dark green). Subjects 4, 6, and 15 (marked in yellow) required multiple calibration sessions with the eye tracker, indicating that eye tracking may not have been as accurate for them as is typical.

For the pair test, we identified a foveation quality threshold for each subject as the lowest variable index he or she reported as equal to or better in quality than the non-foveated reference. For the ramp test, we identified this threshold as the lowest quality for which each subject incorrectly labeled the ramp direction or reported that quality did not change over the ramp. In cases where that choice seemed a possible mistake, we conservatively picked a higher threshold (subjects 5, 6, 13, 14, and 15 for the pair test, and subjects 3, 6, and 13 for the ramp test). The resulting extracted thresholds for the pair test were 15, 14, 10, 22, 13, 17, 12, 11, 14, 16, 18, 14, 17, 17, 13 with mean 14.9 and standard deviation 3.07. For the ramp test, the thresholds were 15, 8, 10, 10, 10, 10, 12, 10, 17, 15, 12, 15, 15, 15, 4 with mean 11.9 and standard deviation 3.48. Histograms for these thresholds are shown in Figure 6.

For the slider test, the mean threshold was 14.5 with standard deviation 4.1, across all subjects and speeds.

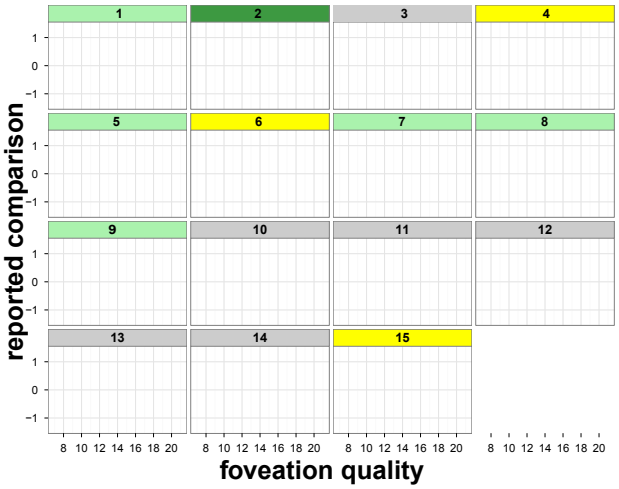


Figure 3: Pair test experimental results. Each tile shows the results for one subject, identified at the top. Foveation quality index j is plotted on the x-axis; the user's reported comparison on the y-axis (+1 = non-foveated reference better, -1 = foveated rendering better, 0 = two of equal quality). The same comparison was presented twice for each j, in two different orders. Subjects marked in light green wore contact lenses. The subject marked in dark green wore glasses. Subjects marked in yellow required multiple calibration sessions, indicating less-than-ideal tracking accuracy. These markings are informational only; no subjects were excluded in the subsequent analysis.

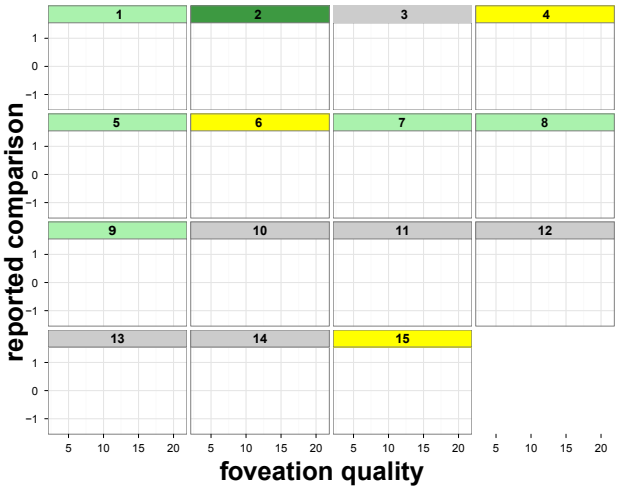


Figure 4: Ramp test experimental results. Each subject's reported comparison (+1 = reference quality [j_0] better, -1 = variable quality [j_1] better, 0 = two of equal quality) is plotted as a function of varying foveation quality j_1. The same comparison was presented twice, once ramping from $j_0 \rightarrow j_1$ and once from $j_1 \rightarrow j_0$.

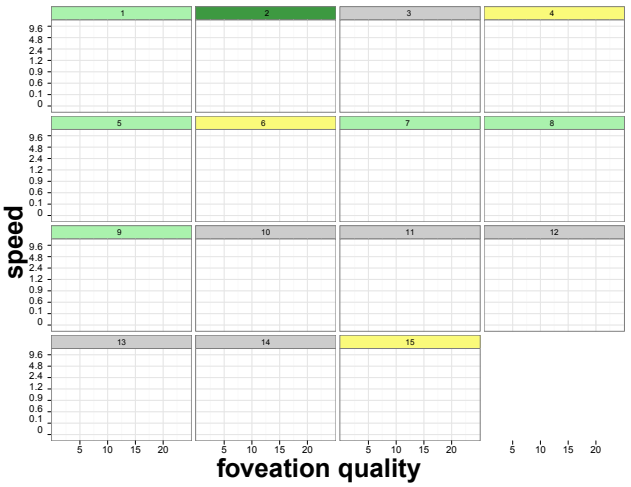


Figure 5: Slider test experimental results. We plot foveation quality where the subject first called for a comparison with the non-foveated reference. The vertical axis graphs animation speed based on 8 settings. A test at each speed was run twice. Subjects 1 and 2 were tested with a reduced set of speeds. Subject 10 asked to see the reference in only 4 out of the 16 runs.

In the pair test, subject 4 never choose foveated rendering at any available quality setting as equivalent in quality to the non-foveated reference. For this subject, we chose a threshold at the highest quality setting available in the test ($j = 22$) to compute the mean.

4 Analysis and Discussion

We analyzed the distribution of thresholds for both conditions (the ramp test and the pair test) to test whether the difference in their means was statistically significant. The means were 11.9 and 14.9 respectively, with corresponding standard deviations 3.48 and 3.07. Using Welch's t-test, we find a significant effect of condition versus the threshold parameter: $t(27.54) = -2.504,\ p = 0.018 < 0.05$. The probability that the two means are identical is estimated to be only 1.8%. We observed no statistically significant dependence of foveation quality demand on animation speed in the slider test.

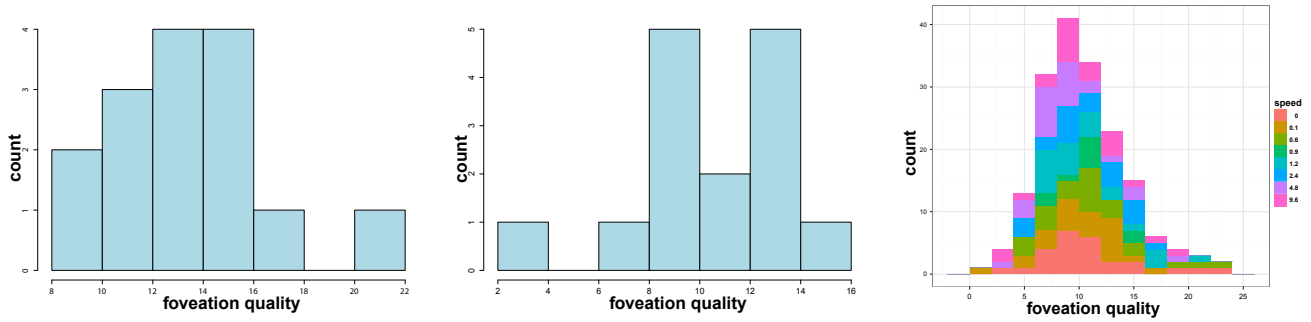
We therefore distinguish two quality targets, A and B. From the ramp test, we obtain the mean threshold $j_A \approx 12$, where quality is neither perceptibly degraded or enhanced over a short progression compared to a high-quality foveated reference ($j_0 = 22$). This represents the more aggressive of the two targets. From the pair test, we obtain the mean threshold $j_B \approx 15$, where foveation quality is equivalent to a non-foveated reference. This represents the more conservative of the two targets. Slider test results generally confirm those from the pair test, probably because we showed users the non-foveated rendering at the start of each trial and identified it as a quality reference.

We then obtain the following estimates for model slope m using Eq. 2:

$$\begin{aligned} j_A = 12 \ &\Rightarrow \ m_A = 0.0275 = 1.65' \text{ per eccentricity } ^\circ \\ j_B = 15 \ &\Rightarrow \ m_B = 0.0220 = 1.32' \text{ per eccentricity } ^\circ \end{aligned}$$

Note that taking means in the space of foveation quality j is more conservative than doing so in the space of slopes m. Foveated rendering cost is roughly proportional to $j^2 \propto 1/m^2$.

Our use of means across multiple subjects to determine rendering parameters may appear problematic. It is certainly possible to calibrate foveated rendering parameters to an individual user. We think



(a) pair test **(b)** ramp test **(c)** slider test

Figure 6: Foveation quality histograms for the three tests in our formal user study.

most CG applications will target a generic user to better manage computational resources and graphical content. The informal study documented in the next section provides evidence that a single set of parameters works well for nearly everyone.

5 Informal Study

We also conducted an informal user study in addition to the formal user study described previously. It involved a fixed set of parameters for the eccentricity layers, derived *ad hoc* before we developed the acuity falloff model, rather than with the procedure in Section 2. We intentionally did not show users the non-foveated rendering as a reference in these demonstrations.

Fixed layer diameters (in pixels) were used: $D_1 = 188$, $D_2 = 383$, and $D_3 = 333$. Sampling factors were $s_1 = 1$, $s_2 = 2$, and $s_3 = 5.77$. From these, we can derive eccentricity layer angular radii from Eq. 3 in the main paper to yield $e_1 = 2.42°$ and $e_2 = 9.78°$. We can then estimate the MAR slope (as in Eq. 2) via $m_1 = (s_2\omega^* - \omega_0)/e_1 = .0340$ and $m_2 = (s_3\omega^* - \omega_0)/e_2 = .0283$, giving us a more aggressive slope range of .028-.034 than we found in the formal user study. Reprojection blur factors were fixed at $b_1 = .15$, $b_2 = .4$, and $b_3 = .7$.

We showed our system to 90 subjects and then surveyed their opinion of its quality on a scale of 1 to 5. Eight of these subjects tracked poorly with our eye tracker and were excluded. Six were shown a 2D version of the demo but not the 3D version, and are also excluded here. Of the 76 remaining subjects, 61 gave the system a "5" quality rating, 10 gave it a "4", 3 gave it a "3", and 2 gave it a "2". 62 of these subjects agreed that the rendering "looked all high-resolution". Seventeen subjects said they could see a "pop with fast eye movement".[3] Two noted that they could see a "pop on blink". Seven said the "periphery looked blurry". Many made comments (e.g., "Is it on?", "I feel cheated; it's not doing anything.", "The demo totally works.", "[Foveation was] completely unnoticeable.", "I can't see any artifacts."), that indicated unawareness of peripheral resolution manipulation. One subject even reported the system's quality as "higher than if [rendered] all sharp."

In summary, most users reported being satisfied by the experience. We thus believe there is a "comfortable" or "effective" level of foveated rendering which provides further savings over the quality levels identified in the formal user study. This is a level at which there is no peripheral degradation or other consciously noticeable artifact from foveated rendering, without reference to any comparison "ground truth". Confirming this more aggressive quality level in a formal user study remains for future work.

[3] This implies that our informal settings were too aggressive for some users. We suspect that eye tracker calibration and performance was less than ideal in many of these cases.