Is Foveated Rendering Perceivable in Virtual Reality? Exploring the Efficiency and Consistency of Quality Assessment Methods

Chih-Fan Hsu[12], Anthony Chen[12], Cheng-Hsin Hsu[3], Chun-Ying Huang[4], Chin-Laung Lei[2], and Kuan-Ta Chen[1]

[1]Institute of Information Science, Academia Sinica
[2]Department of Electrical Engineering, National Taiwan University
[3]Department of Computer Science, National Tsing Hua University
[4]Department of Computer Science, National Chiao Tung University

ABSTRACT

Foveated rendering leverages human visual system to increase video quality under limited computing resources for Virtual Reality (VR). More specifically, it increases the frame rate and the video quality of the foveal vision via lowering the resolution of the peripheral vision. Optimizing foveated rendering systems is, however, not an easy task, because there are numerous parameters that need to be carefully chosen, such as the number of layers, the eccentricity degrees, and the resolution of the peripheral region. Furthermore, there is no standard and efficient way to evaluate the Quality of Experiment (QoE) of foveated rendering systems. In this paper, we propose a framework to compare the performance of different subjective assessment methods on foveated rendering systems. We consider two performance metrics: *efficiency* and *consistency*, using the *perceptual ratio*, which is the probability of the foveated rendering is perceivable by users. A regression model is proposed to model the relationship between the human perceived quality and foveated rendering parameters. Our comprehensive study and analysis reveal several insights: 1) there is no absolute superior subjective assessment method, 2) subjects need to make more observations to confirm the foveated rendering is imperceptible than perceptible, 3) subjects barely notice the foveated rendering with an eccentricity degree of 7.5°+ and peripheral region of a resolution of 540p+, and 4) QoE levels are highly dependent on the individuals and scenes. Our findings are crucial for optimizing the foveated rendering systems for future VR applications.

CCS CONCEPTS

•Human-centered computing →User studies; Virtual reality; •Computing methodologies →Perception;

KEYWORDS

Virtual Reality, foveated rendering, human perception, Quality of Experience

MM '17, Mountain View, CA, USA

978-1-4503-4906-2/17/10…$15.00
DOI: 10.1145/3123266.3123434

1 INTRODUCTION

Virtual Reality (VR) not only becomes popular for gaming but also provides opportunities to other applications, such as rehabilitation [1], manufacturing [2], education [3], and tourism [4]. Making users immerse in the content is the ultimate goal for content providers and hardware vendors of VR. Providing high-resolution videos is one of the solutions to raise the immersion level to users [5]. However, increasing the resolution solely is insufficient, and the frame rate (frame per second, FPS) must be maintained above the minimal threshold. In recent years, 1080p video resolution at 90Hz become the mainstream requirement of VR [6]. Although the improvement of graphics hardware is rather rapid, it is still too costly for the general publics to purchase the hardware that is capable of delivering such high-quality videos. Minimizing the hardware requirement while achieving high perceived quality becomes the holy grail challenge of VR.

Foveated rendering delivers higher video quality with lower computing power by allocating more computing resources on more crucial video regions. In human visual system, the retina contains two types of photoreceptors, cones and rods [7]. Cones are dense in the central of the fovea (called fovea centralis); their density rapidly decreases toward the periphery of fovea centralis. However, there are no rods in the central of fovea. Cones are responsible for sensing colors, brightness, fine details, and sudden changes. Rods are sensitive to less intensive lights, and they are less sensitive to details. Foveated rendering systems keep the highest-quality video at the *foveal region* and gradually degrade the video quality toward the *peripheral region*. The goal is to reduce the overall workload without negatively affecting the perceived quality of users.

However, there is a trade-off between the foveal parameters and limited hardware resources. To maximize the human satisfaction, quantifying the human perceived quality under various foveal parameters is necessary. Quality of Experience (QoE) [8] is used to evaluate the user satisfaction, expectations, and perceptions with respect to a service or a product. However, measuring QoE on foveated rendering systems is challenging, since there are numerous parameters that need to be considered, such as user profiles, image contents, eccentricity degrees (the radiation angle of the gaze), and the resolution of the peripheral region [9–11]. Moreover, there are no general nor fast subjective assessment methods to evaluate the QoE for foveated rendering systems. Therefore, conducting a comprehensive and detailed analysis to quantify the

This work was supported in part by the Ministry of Science and Technology of Taiwan under the grant 106-2221-E-001-013-MY2.

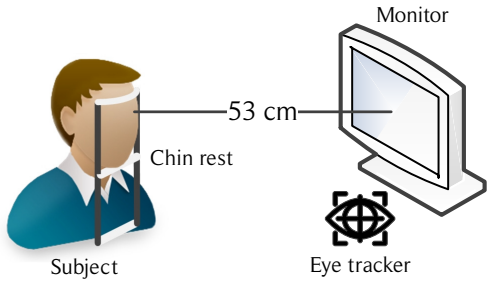


Figure 1: Testbed setup.

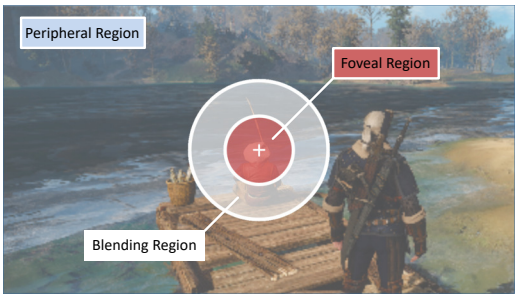


Figure 2: A three-layer foveated image. The cross is the gaze point.

QoE is in need. In this paper, we propose a framework to compare different subjective assessment methods using foveated images which are generated with two key parameters: *eccentricity degrees* and *resolution of peripheral region*. We define the *perceptual ratio* as the probability of the foveated rendering is perceived by users. We propose two performance metrics: *efficiency* and *consistency*, to evaluate each method based on the perceptual ratio. Several heterogeneous methods are evaluated with our framework, and we carefully study the experiment results and the performance of these methods.

Our contributions are listed as follows:

- We propose a framework to compare subjective assessment methods. Moreover, we apply our framework to compare the performance among four common methods. The results show that there is no single absolute superior method.
- We propose a regression model for understanding the relationship between the perceptual ratio and foveal parameters. This is useful for maximizing the video quality of foveated rendering without being perceived by users.
- We confirm that the perceptual ratio on the foveated rendering is highly dependent on individuals and scenes. Therefore, dynamical and personalized foveal setting adaptation methods are essential for future VR.

2 RELATED WORK

2.1 Foveated Rendering

Foveated rendering systems based on eye tracking hardware have been studied for a few years. Guenter et al. [12] rendered three eccentricity layers around the user's gaze point, where the parameters, resolution and radius, of each layer were defined by a Minimum Angle of Resolution (MAR) line of slope m. Swafford et al. [13] proposed four kinds of quality degradations in the peripheral region on their system, including resolution, screen-space ambient occlusion, tessellation, and ray-casting steps. Patney et al. [14] designed a foveated rendering system with a multiresolution- and saccade-aware temporal anti-aliasing algorithm to meet the low latency requirement of VR systems. These studies are complementary to our work that focuses on the systems' aspects of foveated rendering. To verify the effectiveness of foveated rendering systems, a carefully-designed user study is necessary. Lungaro et al. [15] reproduced specific sample videos in different foveated qualities and got the QoE level in 1-5 Mean Opinion Score (MOS) scales for each video from their experiments. Several papers [12–14] conducted different user studies. However, the purpose of these user studies is to find the parameter setting that saves the most computing power or network bandwidth without being perceived by users. Lee et al. [16] developed the Foveal Peak Signal-to-Noise Ratio (FPSNR) and Foveal Weight Signal-to-Noise Ratio (FWSNR) for foveated image quality assessment. These two methods were based on the objective metrics: PSNR and WSNR. To the best of our knowledge, subjective assessment methods have not been rigorously applied to foveated rendering systems.

2.2 Subjective Assessment Methods

Four common subjective assessment methods are investigated in this paper: 1) Single-Stimulus Absolute Category Rating (ACR), 2) Double-Stimulus Quality Comparison (DSQC), 3) the Descending Method (DM) and 4) the Ascending Method (AM). DM and AM are two variants of method of limits. In ACR [17, 18], a subject is asked to observe a test image once, then the subject rates his/her opinion immediately with a grading scale, such as MOS. In DSQC [17, 18], two images with different qualities of the same content are sequentially shown in a random order, and the subject is asked to select the image with better quality. Method of limits [19, 20] is one of the classical research methods in Psychophysics, which is used to measure a subject's perception of stimuli. In this method, a subject is asked to observe a sequence of images with decreasing (or increasing) quality, which is called the DM (or AM). The subject is asked to determine at what level of the impairment in a stimulus is perceptible (or imperceptible).

3 EXPERIMENT MATERIALS

We conduct laboratory experiments with four subjective assessment methods to quantify human perceptiveness on foveated images. Our setup and test materials are presented in the following.

3.1 Testbed

Our testbed contains the following components: 1) a Dell S2716DG monitor with 2560x1440 native resolution at 144Hz, 59.6 cm × 33.5 cm display area, and 108.8 PPI pixel density, 2) a desktop computer with an Intel® Core™ i7-6700 CPU, 32 GB DDR4 RAM, and an NVIDIA GeForce GTX 980 Ti GPU, 3) a Tobii eye tracker with a 90Hz sampling rate, and 4) a chin rest that supports the subject's head and fixes the distance between the eyes and the monitor at

(a) banquet (b) pub

(c) market (d) forest

(e) plain (f) riverside

Figure 3: Six representative scenes in *The Witcher 3: Wild Hunt*.

53 cm for maintaining stable gaze detection results from the eye tracker. The testbed in our experiment is illustrated in Figure 1.

3.2 Foveated Images

We generate foveated images with two key parameters in real time before each round of the experiments. Figure 2 illustrates a three-layer foveated image in our experiments. The foveal region is composed of the image with the highest quality; the peripheral region is composed of the image with low quality; the blending region is blended with the image of highest quality and the image of low quality. The procedure of producing a foveated image is as follows. Firstly, we load two images with different resolutions of the same scene. Among these two images, one is the reference image which has the native resolution (for the foveal region), and the other has a lower resolution (for the peripheral region). Next, the image at the lower resolution is upscaled to the native resolution using bilinear interpolation. Lastly, we blend these two images into a three-layer foveated image.

3.3 Scenes

We capture high-quality gaming screenshots with 2560x1440 resolution from a role-playing action game, *The Witcher 3: Wild Hunt*. This game provides a vivid world with diverse scenes and numerous options in the graphics settings. Moreover, adopting gaming screenshots as the image contents in our user study is closer to the real usage scenarios of the foveated rendering in the future. Figure 3 shows six representative scenes in our experiments.

Figure 4: The saliency map of the market scene. The red (darker) area represents the region that interests subjects the most; while the white areas is the least.

3.4 Fixation Points

In real life scenarios, the foveated images based on the user's gaze point should be generated in real time. However, since our eye tracker has a limited refresh rate at 90Hz, a subject may notice that the quality of the foveated images around his/her gaze point is blurred at first and then become clear. This can be attributed to the blending latency when the user moves his/her eyes rapidly. To get around this issue, we ask subjects to fix their gaze at the assigned fixation point. The appropriate fixations are determined by the visual saliencies from the fixation experiments with 11 participants (2 females and 9 males with an average age of 26), who are asked to freely observe the given images for 10 seconds. According to the gaze points collected from the fixation experiments, we select five fixations for each scene. Figure 4 illustrates a sample of five fixations in the market scene.

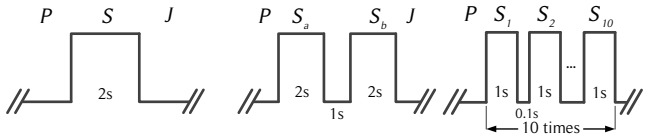
4 EXPERIMENT PROCEDURE

There are 17 subjects (10 males and 7 females), with an average age of 24.7 and a standard deviation age of 5.1 years old, where the range of ages is from 20 to 39. All subjects have corrected 20/20 vision. At the beginning of each test, several instructions are given to each subject about the type of test, the procedure, and the timing. During each round, we use an eye tracker to detect the subject's gaze and limit his/her gaze at certain fixation points. To avoid fatigue, each subject is asked to take a five-minute break between sessions and is allowed to have a short rest at any time.

4.1 Stages

There are three categories of stages, and each round consists of several stages. All the rounds begin with the preparation stage followed by the stimulation stages. In a round of ACR, DSQC, and DM (or AM), there are at most one, two, and 10 stimulation stages, respectively. The ACR and DSQC rounds end up with the judgment stage, while there is no judgment stage in the DM (or AM). These three stage categories are listed in details as follows:

- **Preparation stage.** Several instructions are shown on the screen, and each subject needs to follow these instructions to move forward to the next stage.
- **Stimulation stage.** Each subject is asked to observe the image while maintaining the gaze at the assigned fixation



(a) ACR (b) DSQC (c) DM and AM

Figure 5: The stimulus presentations of a round for all methods. P, S, and J stand for preparation, stimulation, and judgment stages, respectively. There are two stimulation stages, S_a and S_b, in a DSQC round, and 10 stimulation stages, S_1 to S_{10}, in a DM and an AM round. In a DSQC and DM (or AM) round, a short time interval is inserted between two stimulation stages.

point. If the subject's gaze drifts away from the fixation point, the round will be interrupted immediately, and the subject needs to redo the round.

- **Judgment stage.** Each subject is asked to judge the quality of the stimulus shown in the stimulation stage according to a given grade scale.

The stimulus presentations in a single round of all methods are shown in Figure 5.

4.2 Foveal Parameters

Two essential parameters of the foveated rendering are: *eccentricity degrees* (eccentricity in short) and *resolution of peripheral region* (resolution in short). In ACR and DSQC, each parameter comprises three levels. The eccentricity levels are 2.5° (equal to fovea portion of the retina), 5° (slightly larger than paracentral portion), and 7.5° (close to macula portion). These levels correspond to approximately 100, 200, and 300 px radius of the foveal region, while the distance between eyes and the screen is 53 cm; the resolution levels are 360p, 540p, and 720p. In DM and AM, we extend each foveal parameter to six levels. The additional eccentricity levels are 4°, 6°, and 9° which correspond to approximately 160, 240, and 360 px; the additional resolution levels are 900p, 1080p, and 1260p. A blending region with an extra 100 px border outside the foveal region is added to prevent having a noticeable boundary between the foveal and the peripheral regions. The weights of blending region increase linearly from 0 (inner) to 1 (outer), which indicate the proportion of composition between the reference and the low-resolution images.

4.3 Single-Stimulus Absolute Category Rating

The stimulus presentation of a single round in ACR is shown in Figure 5(a). Subjects can freely observe the reference image of the given scene for five seconds before performing a series of judgments. In the preparation stage, the monitor displays a grey screen with a red cross pointing out where the subject should fixate at. Once the subject fixes his/her gaze on the cross, the cross size begins to shrink, indicating the gazed is fixed. After the subject's gaze fixes for one second, the experiment proceeds to the stimulation stage, and the monitor displays the stimulus for two seconds. Each subject is asked to observe the image maintaining the fixed position. In the judgment stage, the subject assesses the overall quality of the stimulus based on a 5-grade scale [17]. The levels from 5 to 1 of this 5-grade scale are: 5) imperceptible, 4) perceptible, but not annoying, 3) slightly annoying, 2) annoying, and 1) very annoying. ACR contains 60 rounds which are equally divided into six scenes, leading to 10 rounds per scene. Of these 10 rounds, there are 9 foveated images with different foveal settings (3 eccentricities × 3 resolution) and one reference image for examining the subject's reliability.

4.4 Double-Stimulus Quality Comparison

DSQC requires subjects to assess two versions of each given scene. Its stimulus presentation for a single round is shown in Figure 5(b). We randomly display one image of each image pair in the first stimulation stage, and the other is presented in the second stimulation stage. After two stimulation stages, a subject chooses between: 1) the first image is better, 2) the second image is better, and 3) both images have equal quality. This method contains 60 rounds which are equally divided into six scenes, leading to 10 rounds per scene, similar to ACR.

4.5 Descending Method and Ascending Method

The stimulus presentation of a single round of DM and AM is shown in Figure 5(c). These tests show an image sequence that is composed of 10 images in a single round. Each image appears for one second, followed by the grey screen for 0.1 seconds. Each subject is asked to carefully observe the image sequence. There is an issue of such method of limits: subjects may *predict* when the stimulus is close to being perceptible (or imperceptible). To mitigate this issue, we duplicate the first level of stimulus for random times. In DM, the quality of the image sequence gradually ramps down from the reference image (highest quality) to the foveated image with the worst setting. In AM, we add an extra stage at the beginning of each round, in which subjects can freely observe the reference image of the given scene for three seconds. Next, the quality of the image sequence gradually ramps up from the foveated image with the worst setting to the reference image. When a subject perceives that the quality of an image sequence is getting worse in DM or becomes as good as the reference one in AM, he/she terminates the round immediately, and the setting of the perceived stimulus is recorded.

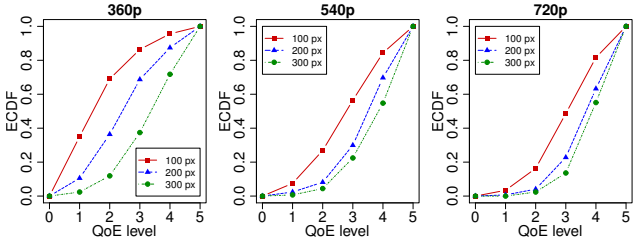
Both DM and AM contain 36 rounds which are equally divided into six scenes, leading to six rounds per scene. Of these six rounds, three image sequences consist of the reference image and the foveated images of which resolutions descend (or ascend) from 1260p to 360p (or 360p to 1260p) with constant eccentricity in 100, 200, and 300 px, respectively. Similarly, the other three image sequences contain the reference image and the foveated images of which eccentricities descend (or ascend) from 360 to 100 px (or 100 to 360 px) with constant resolution in 720p, 540p, and 360p respectively. That is to say, the quality of each image sequence is made up of six test images with different quality levels and a reference image.

5 EXPERIMENT RESULTS

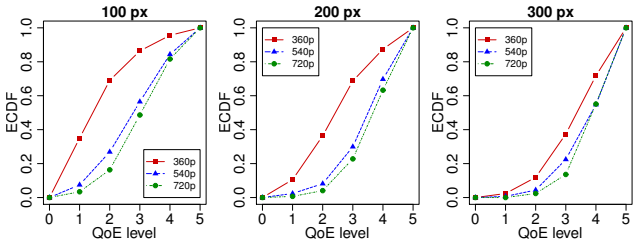
The statistics of four subjective assessments are listed in Table 1. Each subject performs all methods at least once, and the subject is asked to do an additional DSQC test if time permits. We report

Table 1: The statistics for all subjective assessments

	# of sessions	# of rounds per session	Average round time	Average session time
ACR	49	60	6 s	6.1 min
DSQC	58	60	10 s	10.3 min
DM	48	36	14 s	8.5 min
AM	47	36	17 s	10.4 min



(a) Impact of eccentricity on QoE scores



(b) Impact of resolution on QoE scores

Figure 6: The QoE score distributions of the ACR test.

the experiment results of the four subjective assessment methods below. For brevity, we write foveal settings as pairs, e.g., 100 px eccentricity in 360p resolution is written as (100, 360).

5.1 Single-Stimulus Absolute Category Rating

QoE scores of ACR at all foveal settings are shown in Figures 6 and 7. We can see the impacts of eccentricity and resolution on the QoE scores in Figure 6. At 360p resolution in Figure 6(a), the differences among three eccentricities are clear. However, the distributions of 200 px and of 300 px eccentricity become similar to each other at 540p and 720p resolutions. This indicates that the quality difference between the 200 px and 300 px resolutions becomes imperceptive to subjects. In Figure 6(b), the distributions of 540p and of 720p are all relatively close, indicating the foveated rendering at 360p resolution is much more noticeable. However, when it comes to 300 px eccentricity, the foveated rendering can barely be perceived no matter what the resolution is. It means that making the eccentricity larger increases the QoE score efficiently. Figure 7 summarizes the average QoE scores with 95% confidence interval. According to these two figures, we conclude that both parameters, resolution and eccentricity, are positively correlated with the QoE scores in ACR.

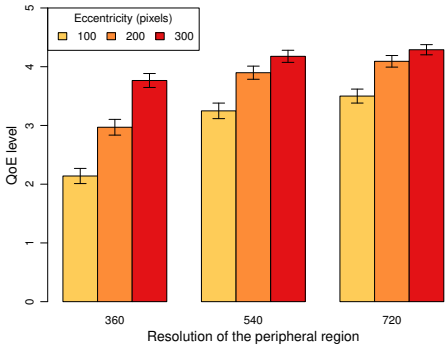
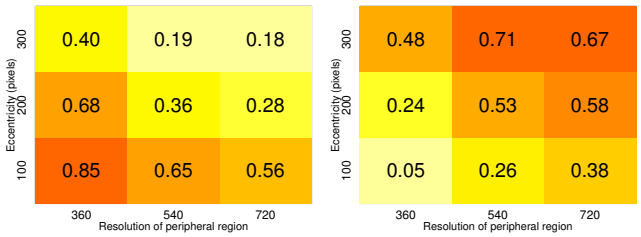


Figure 7: Average ACR QoE scores for all foveal settings. The intervals indicate the 95% confidence intervals.



(a) Probabilities of the reference image is selected correctly

(b) Probabilities of the option *equal quality* is selected

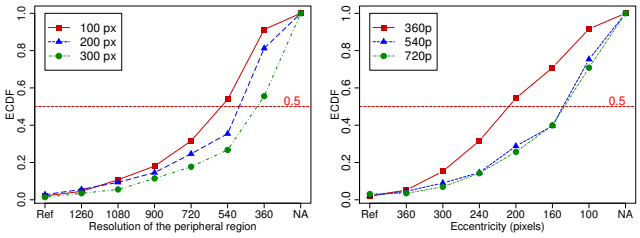
Figure 8: The probabilities in the DSQC test at all foveal settings.

5.2 Double-Stimulus Quality Comparison

Figure 8(a) reports the probabilities of subjects select the correct reference images. If the value in the box is greater than 0.5, the foveated rendering is more likely to be noticed by subjects. Figure 8(b) shows the ratios of the option *equal quality* is selected. The foveated rendering is noticeable at 100 px eccentricity, and subjects can select correct reference images in most of the rounds at (100, 360). When the eccentricity becomes bigger, subjects barely notice the difference between the test and the reference images at 300 px eccentricity. Moreover, the foveated rendering is almost imperceptible while the resolution is greater than 540p at 300 px eccentricity.

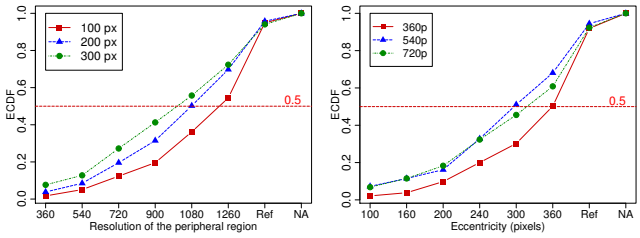
5.3 Descending Method

Figures 9(a) and 9(b) report the distributions of the perceived settings under different eccentricity and resolution, respectively. The horizontal dashed lines represent 50%. For samples above the dashed lines, the foveated rendering is more likely to be perceived. The perceived setting is just above the dashed line is the threshold setting. This threshold setting is called Just Noticeable Degradation (JNDG). In Figure 9(a), the foveated rendering is noticeable to subjects while the foveal setting is less than or equal to the JNDG: (100, 540), (200, 360), and (300, 360); in Figure 9(b), the foveated rendering is noticeable to subjects while the setting is less than or equal to the JNDG: (200, 360), (100, 540), and (100, 720). In general,



(a) Eccentricity (b) Resolution

Figure 9: The effect of two foveal parameters on the distributions of the perceived settings in the DM test. NA indicates that subjects have no action during the round.



(a) Eccentricity (b) Resolution

Figure 10: The effect of two foveal parameters on the distributions of the perceived settings in the AM test. NA indicates that subjects have no action during the round.

the JNDG is negatively correlated with the parameters. However, the JNDG and distributions at 540p and 720p resolutions are similar to each other.

5.4 Ascending Method

Figure 10 reports the distributions of the perceived settings, which are affected by the two foveal parameters. The samples above the horizontal dashed lines are perceived settings where the reference and the foveated images qualities are considered the same by the majority of the subjects. We define the threshold setting that is just above the 50% threshold as Minimal Satisfied Level (MSL). Concretely, we obtain that the perceived settings which are greater than or equal to the MSL: (100, 1080), (200, 1080), and (300, 1260) in Figure 10(a), and obtain the perceived settings which are greater than or equal to the MSL: (360, 360), (300, 540), and (360, 720) in Figure 10(b), where the foveated image quality is good enough to the subjects. The MSL are positively correlated with the fixed parameters except for 540p and 720p resolutions in Figure 10(b).

6 COMPARATIVE ANALYSIS

In this section, we present our proposed comparison framework in detail. We then carefully analyze the performance of the subjective assessment methods.

6.1 Perceptual Ratio: A Unified QoE Metric

QoE scores of all methods are reported in Section 5. However, there is no direct way to compare among these heterogeneous QoE score metrics. The QoE metric of ACR is absolute value from 1 to 5; the QoE metric of DSQC is a set of probabilities; the QoE metric of DM (or AM) test is the threshold foveal settings. We propose a QoE metric, perceptual ratio, which is defined as the probability whether the foveated rendering is perceived. Perceptual ratio 1 means that the foveated rendering will definitely be perceived. We then convert all QoE metrics to the perceptual ratio at 9 foveal settings (3 eccentricities × 3 resolutions).

- In ACR, the mean score μ_r of the judgments, where the test image is set to the reference image is calculated, and all judgments J are categorized into two sets: $J_{\geq \lfloor \mu_r \rfloor}$ (the QoE score is greater than or equal to $\lfloor \mu_r \rfloor$) and $J_{< \lfloor \mu_r \rfloor}$ (the QoE score is less than $\lfloor \mu_r \rfloor$). Then we define the perceptual ratio as the ratio of $|J_{< \lfloor \mu_r \rfloor}|$ to $|J|$ for all foveal settings.
- In DSQC, the probability for a subject to select the correct reference images to all selections is the perceptual ratio.
- In DM (or AM), a sequence of the test images with descending (or ascending) quality will be converted to a series of quality comparisons with the reference image. The foveal settings that are less than the JNDG (or MSL) settings at certain fixed foveal parameters are marked with 0 (or 1), and others are marked with 1 (or 0). The value 0 indicates the foveated rendering cannot be perceived by subjects; the value 1 indicates the foveated rendering is perceived by subjects. Then, the perceptual ratio of a certain setting can be calculated with the ratio of the number of quality comparisons which are marked with 1 to the total number of quality comparisons.

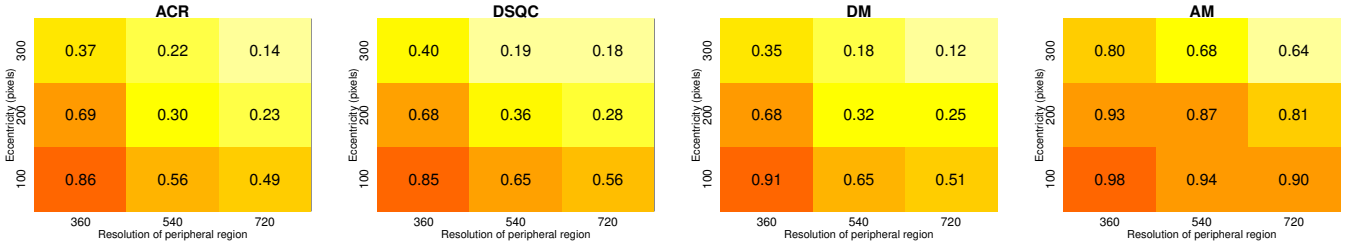
Figure 11 reports the average perceptual ratios of all subjects at 9 foveal settings. The distribution of perceptual ratios of all methods are similar to each other except for AM. This can be attributed to that confirming the changing of image quality is imperceptible is harder than perceptible, therefore, subjects need to make more observations to decide the threshold setting. This fact leads to that the perceptual ratios of the AM test are higher than those of the DM test for all foveal settings. Overall, the foveated rendering with 200 px eccentricity and 540p resolution is close to the boundary between perceptible and imperceptible, and subjects barely notice the foveated rendering if the settings are higher.

6.2 Efficiency and Consistency

We propose two performance metrics, efficiency and consistency, to evaluate the performance of subjective assessment methods based on the perceptual ratio.

Efficiency aims to find *how fast the general consensus of the perceptual ratio will converge.* Naturally, the higher the efficiency is, the lower the cost of performing the experiment can be, such as shorter assessment time or fewer numbers of judgments are required. Assuming that the general consensus of the perceptual ratio μ_p exists in a very large population of user study samples D_p. D_n is the subset of D_p, where p and n denote the cost, time, or the number of judgments. The convergence of general consensus is defined with:

$$lim_{n \to p} d(\mu_n, \mu_p) < \varepsilon, \tag{1}$$



(a) ACR (b) DSQC (c) DM (d) AM

Figure 11: Perceptual ratios of different subjective assessment methods.

where μ_n (or μ_p) is a set of perceptual ratios at 9 foveal settings, which can be calculated from D_n (or D_p), $d(\cdot)$ is a distance function, and ε is a given threshold.

We define D_p *as the whole samples we have collected*, and the distance function $d(\cdot)$ as RMSE value between μ_n and μ_p. To evaluate the efficiency level, we calculate n for each method using:

$$\arg_n min(n), \text{ where } RMSE(\mu_n, \mu_p) < \varepsilon. \tag{2}$$

The obtained cost n indicates the minimal required cost of a method to obtain the perceptual ratios that are close to the general consensus, and we denote the cost of method m as n_m, where $m \in \{ACR, DSQC, DM, AM\}$. Finally, the efficiency level of m is defined as:

$$efficiency_m = 1/(n_m \times RT_m), \tag{3}$$

where RT_m is the average round time of the method m, and the cost n is defined by the number of judgments.

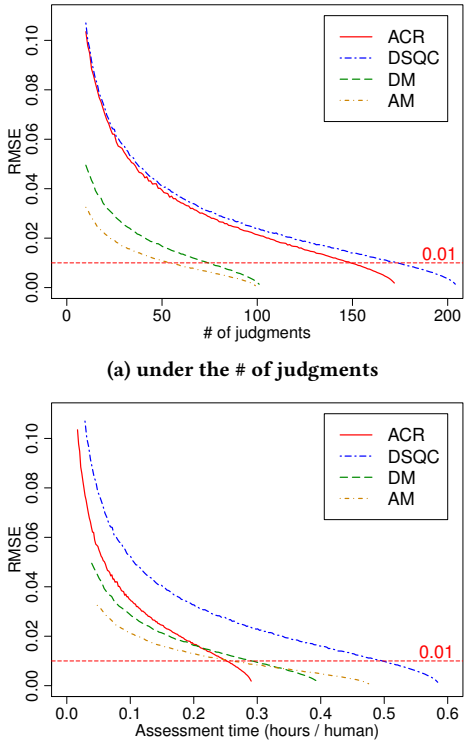
Consistency aims to evaluate *how consistent the individual QoE scores are*. Let μ'_p denote the set of perceptual ratios that are calculated from all the samples from a single subject D'_p, and μ'_n denote the perceptual ratios that are calculated from a subset of D'_p, D'_n, where n is obtained from Eq. (2). We randomly sample D'_n for k times, and then we calculate a set of RMSE values $R = \{r_1, r_2, r_3, ..., r_k\}$, where $r_i = RMSE(\mu'_{n,i}, \mu'_p), i \in \{1, ..., k\}$. If the individual perceptual ratios μ'_n is consistent across all foveal settings, the standard deviation of R is small. Therefore, we define the consistency level as,

$$consistency_m = 1/\sigma_{R_m}, \tag{4}$$

for each method m, where σ_R is the standard deviation of R.

6.3 Performance of Subjective Assessment Methods

Figure 12 shows the convergence of all methods, and each value in this figure is calculated by random sampling D_n and D'_n for 1000 times ($k = 1000$). The horizontal dashed lines in Figure 12 indicate a threshold $\varepsilon = 0.01$. Figure 12(a) reports the distance between μ_n and μ_p under different numbers of judgments. In this figure, we observe that all methods can be separated into two groups: 1) the ACR and DSQC tests and 2) two variations of method of limits. The RMSE values of the first group are higher than the values of the second group under a small number of judgments, because the information which a single round contains in the ACR and DSQC tests is less than that in the DM and AM tests. In the DM and AM tests, each subject at most observes 10 test images in a single round



(a) under the # of judgments

(b) under the assessment time

Figure 12: The convergence of the perceptual ratio under two kinds of costs: number of judgments and assessment time. Y-axis indicates the distance between the general consensus μ_p and μ_n which are calculated from the whole and partial samples respectively. The horizontal dashed red line indicates a threshold $\varepsilon = 0.01$.

before making an assessment. In contrast, in the ACR and DSQC tests, a subject makes a judgment after each foveated image.

Figure 12(b) reports the RMSE between the μ_n and μ_p under different assessment time. In the efficiency level of four methods, the ACR test benefits from the shortest round time, therefore, this method converges earlier; the DSQC test requires more judgments and suffers from longer round time, therefore, this method is less efficient. Between the DM and AM tests, the DM test requires more

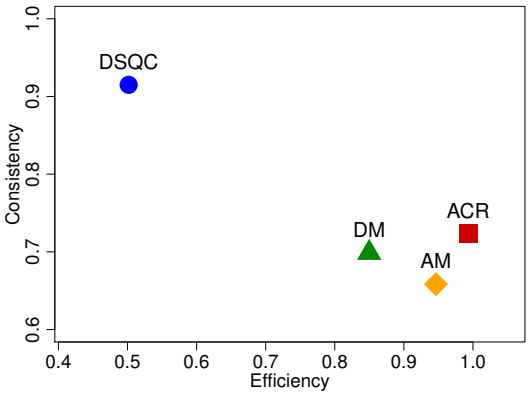


Figure 13: Efficiency vs. consistency.

judgments but takes less round time than the AM test. Hence, the efficiency levels between the DM and AM tests are similar.

Two performance metrics of all methods are shown in Figure 13. The values are scaled to the range from 0 to 1. In the consistency, the DSQC test outperforms other methods. This can be contribute to that each subject only needs to compare between a reference and a foveated images in a single round, and the quality difference between these two images is relatively obvious than the quality difference between two adjacent images in the DM and AM tests. In the ACR test, each subject makes the judgment via the memorized quality of the reference image or the quality of the test image in the previous round, and it is tricky to get consistent scores. In the DM and AM test, the AM test suffers from the uncertainty of the number of the additional observations, therefore, the AM test is less consistent than the DM test.

Overall, the DSQC test achieves highest consistency, but this method requires more judgments and time to converge to the general consensus. The ACR test is the most efficient method. The DM and AM tests require less judgments than the DSQC and ACR tests, however, the average round time of the DM and AM tests are much longer than the DSQC and ACR tests.

6.4 Modeling of Perceptual Ratios

To find the relationship between perceptual ratio and foveal parameters, we make two observations: 1) the relationship between perceptual ratio and resolution is logarithmic and 2) the relationship between perceptual ratio and eccentricity is linear. We then perform regression analysis for all methods with a regression model:

$$perceptual\ ratio = a \times log(resolution) + b \times eccentricity + r. \quad (5)$$

The results of regression analysis are shown in Table 2. The proposed model achieves high R^2 and is significant (p-value <0.05) in all methods. This analysis also reveals that the resolution has more influence on the perceptual ratio. To understand the relationship between the human perception and foveated rendering further, we take individuals and scenes into account. According to the proposed model (Eq. (5)), the χ^2 test is performed to see whether the model is improved with the additional independent variables, individuals or scenes. Table 3 shows that individuals and scenes significantly reduce the residual sum of squares respectively except the model of AM, which impose no significant influence when taking scenes into account. Since the residual sum of squares of the base model of AM has been very small, taking scenes into account doesn't improve the model much.

Table 2: The performance and significance of proposed regression model for all subjective assessment methods.

Method	Coefficients			*Adj.* R^2	p-value
	a	b	r		
ACR	-0.528	-0.002	4.125	0.91	$<2.75 \times 10^{-4}$
DSQC	-0.443	-0.002	3.664	0.94	$<8.05 \times 10^{-5}$
DM	-0.520	-0.002	4.165	0.94	$<8.47 \times 10^{-5}$
AM	-0.174	-0.001	2.157	0.91	$<2.67 \times 10^{-4}$

Table 3: The model performance of adding additional independent variables.

Method	Add. variable	*Adj.* R^2	p-value of the χ^2 test
ACR	subject ID	0.81	$<1.95 \times 10^{-46}$
	scenes	0.88	$<1.97 \times 10^{-08}$
DSQC	subject ID	0.78	$<9.68 \times 10^{-20}$
	scenes	0.89	$<2.51 \times 10^{-06}$
DM	subject ID	0.85	$<2.54 \times 10^{-56}$
	scenes	0.92	$<4.35 \times 10^{-04}$
AM	subject ID	0.83	$<4.44 \times 10^{-98}$
	scenes	0.87	$<1.20 \times 10^{-01}$

7 CONCLUDING REMARKS

In this paper, we proposed a framework to compare the performance of heterogeneous subjective assessment methods based on the perceptual ratio of various foveal images. Two performance metrics, efficiency and consistency, were evaluated via our framework with different methods. We carefully analyzed all QoE scores and compared the performance among the methods. Additionally, we performed regression analysis to model the perceptual ratio with two foveal parameters, and confirmed that individuals and scenes are highly correlated to the perceptual ratio. Our framework allows researchers and developers to intelligently select the subjective assessment method that meets their demands the most. In addition, the proposed regression model is useful for optimizing the foveated rendering systems. Our comprehensive study reveals several insights which are crucial for developing future foveated rendering systems. Our work can be extended in several dimensions. For example, we plan to study more parameters (such as color distribution, video details, and individual profiles) that may affect the QoE scores for optimizing the foveated rendering systems under various circumstances. Our model can also be integrated with real foveated rendering systems in real-time VR applications, so as to maximize the QoS under diverse and dynamic resource levels.

REFERENCES

[1] Maria Schultheis and Albert Rizzo. The application of Virtual Reality technology in rehabilitation. *Rehabilitation Psychology*, 46(3):296, 2001.

[2] Tariq Mujber, Tamas Szecsi, and Mohammed Hashmi. Virtual Reality applications in manufacturing process simulation. *Journal of materials processing technology*, 155:1834–1838, 2004.

[3] Christine Youngblut. Educational uses of Virtual Reality technology. Technical report, DTIC Document, 1998.

[4] Daniel Guttentag. Virtual Reality: Applications and implications for tourism. *Tourism Management*, 31(5):637–651, 2010.

[5] Doug Bowman and Ryan McMahan. Virtual Reality: How much immersion is enough? *Computer*, 40(7):36–43, 2007.

[6] Asynchronous Timewarp on Oculus Rift. https://developer3.oculus.com/blog/asynchronous-timewarp-on-oculus-rift/.

[7] Brian Wandell. *Foundations of Vision*. Sinauer Associates, 1995.

[8] Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia. A generic quantitative relationship between Quality of Experience and Quality of Service. *IEEE Network*, 24(2):36–41, 2010.

[9] Karlene Ball, Bettina Beard, Daniel Roenker, Richard Miller, and David Griggs. Age and visual search: Expanding the useful field of view. *JOSA A*, 5(12):2210–2219, 1988.

[10] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13–13, 2011.

[11] Eyal Reingold, Lester Loschky, George McConkie, and David Stampe. Gaze-contingent multiresolutional displays: An integrative review. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(2):307–328, 2003.

[12] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. Foveated 3D graphics. *ACM Transactions on Graphics*, 31(6):164:1–164:10, 2012.

[13] Nicholas Swafford, José Iglesias-Guitian, Charalampos Koniaris, Bochang Moon, Darren Cosker, and Kenny Mitchell. User, metric, and computational evaluation of foveated rendering methods. In *Proceedings of the ACM Symposium on Applied Perception*, pages 7–14, 2016.

[14] 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*, 35(6):179:1–179:12, 2016.

[15] Pietro Lungaro and Konrad Tollmar. Eye-gaze based service provision and QoE optimization. In *Proceedings of the Workshop on Perceptual Quality of Systems*, pages 6–10, 2016.

[16] Sanghoon Lee, Marios Pattichis, and Alan Bovik. Foveated video quality assessment. *IEEE Transactions on Multimedia*, 4(1):129–132, 2002.

[17] *Recommendation ITU-R BT.500-13*, 2012.

[18] Rafał Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8):2478–2491, 2012.

[19] George Gescheider. *Psychophysics: The Fundamentals*. Psychology Press, 3rd edition, 1997.

[20] Wanmin Wu, Ahsan Arefin, Gregorij Kurillo, Pooja Agarwal, Klara Nahrstedt, and Ruzena Bajcsy. Color-plus-depth level-of-detail in 3D tele-immersive video: A psychophysical approach. In *Proceedings of the ACM International Conference on Multimedia*, pages 13–22, 2011.