Display and Imaging System Sharpness Modeling and Requirement in High-Resolution VR and AR

Jiawei Lu[1], Trisha Lian[2], Jerry Jia[1*]
[1]Meta Reality Labs; Sunnyvale CA; [2]Meta Reality Labs; Redmond WA
****jerry.jia@meta.com***

Abstract

Pixels per degree (PPD) alone is not a reliable predictor for high-resolution experience in VR and AR. This is because "high-resolution experience" depends not only on PPD but also display fill factor, pixel arrangement, optical blur, and other factors. This often complicates system architecture decisions and design comparisons. Is there a simple way to capture all the contributors and quantitatively match user experience? In this paper, we present a system-level model and metric, system modulation transfer function (system MTF), to predict perceptual quality considering all the key parameter dimensions: pixel per degree (display), pixel shape (display), fill factor (display), optical blur (Optics), and image processing (graphics pipeline). The metric can be defined in much the same way of traditional MTF for imaging systems by examining image formation of a point source and then performing Fourier transform over the response function, but with special mathematical treatments. One application of the model is described on perceived text quality, where two weight functions depending on text orientation and spatial frequency are incorporated into the above model. A perceptual study on text quality across different resolutions was performed to validate the system MTF model.

Introduction

Resolution in virtual reality headsets is measured in pixel per degree (PPD) and it is typically lower than that of retina display in current smart phones. For example, typical phones today held at 25 cm away have 55-80 PPD. Computer monitors viewed at 50cm distance have 30-75 PPD. Today's VR systems are at about 20-30 PPD and people can usually see individual pixels ("screen door effect"). The reasons are not obvious: 1) VR system has a much wider field of view (FOV) than phones. Typical phones only have 20-30 deg total FOV (+/-10 to +/-15 deg) while VR systems requires about 100 deg FOV. This means we need to have significantly more pixels in a VR system than a phone display. Each eye in VR requires approximately 10x more pixels than a typical phone. 2) There needs to be an optical lens in a VR system between the display and human eye. This lens projects the virtual image of the display (in typical VR architectures) at a distance within eyes' accommodation range (e.g., 1 meter to 3 meters). This magnifies the pixels in angle space effectively diminishing resolution. As shown in Figure 1a, assuming pixel to pixel distance is 50um (typical for today's phone displays), with the phone held at 25 cm distance, the resolution perceived by the eye is very high at 87 PPD. In Figure 1b, once we add the lens and the eye observes the virtual image of the pixels at about 1.5 meters, the resolution is drastically reduced to 10 PPD. This means we need to have significantly denser pixels in a VR system than a phone display. Combining reason 1) and 2), we need about 10x more pixels on 1/4 size of a phone display to achieve "retina display" experience in a VR system, which is greatly challenging.

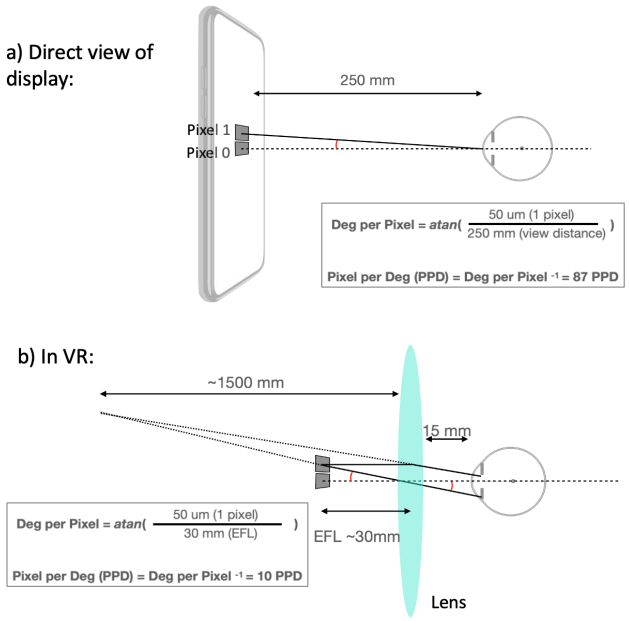

Figure 1. Resolution comparison between a phone display and a VR system with the same display. A VR system has a much larger FOV than a phone and needs significantly more and denser pixels for each eye's display.

There are more hardware factors in the play than just display resolution. Shown in Figure 2b, the pixel shape, arrangement, and fill factor (ratio of lit area of a pixel) also strongly impact the user experience and perception of the artifacts such as "screen door effect." Sub-pixel arrangement (the relationship and layout of R/G/B sub-pixels within a pixel) and sub-pixel rendering (SPR) algorithm optimized for a specific sub-pixel arrangement also contributes to the perceived text quality, as shown in Figure 2c. SPR is particularly important to text content and is not as effective in photographic or natural content. Optical blur introduced by the VR optical lenses, AR optical path, and eye blur (eye performance in the fovea direction which varies with spatial frequency) is another additional factor (Figure 2d).

Additionally, the content determines the spatial frequency that the system should operate well at. Text quality is one of the most challenging use cases. Making text legible is the first step but how good the text appears or text quality (i.e., if the text appears "sharp" and has a "clean edge") requires components of greatly higher spatial frequencies. Figure 3a shows the design heuristics we apply

for photographic content (3-15 cycles per degree, or 6-30 PPDs) and for text content (10-30 cycles per degree, or 20-60 PPDs). Figure 3b is an example of images of eye chart (2nd and 3rd rows representing 20/100 and 20/80 vision respectively) taken by a recent mobile phone with different resolutions. While all the 3 images (first one is the original content as a reference) are legible, the text quality is clearly preferable in the 56 PPD image. Because the frequency of interest is mainly determined by the content, we choose these specific frequencies in our formulation for system MTF.

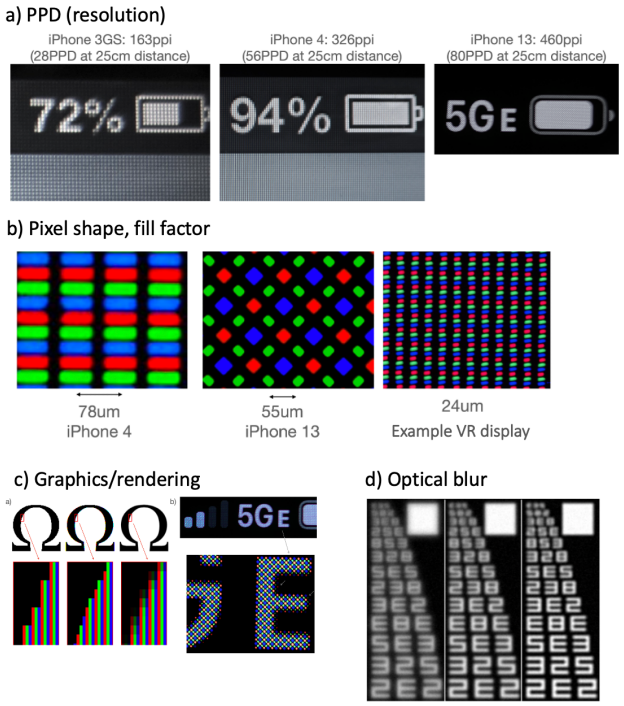

Figure 2. Factors contributing to "high-resolution" experience in VR/AR: a) resolution [3], b) pixel layout/ fill factor, c) graphics pipeline such as sub pixel rendering [4], d) optical blur [7] and eye blur. This is not a complete list of contributing factors.

As stated above, the high-resolution experience depends on many factors – PPD, pixel layout, fill factor, rendering, optics and content itself. Is there a way to capture all the complex contributors, match user experience quantitatively, and guide practical system design? Our goal is to correlate the human perception to engineering metrics in order to design and manufacture quality products.

In this paper, we present a system-level model and metric - system MTF - to predict perceptual quality considering all the key VR dimensions: pixel shape (display), pixel per degree (display), fill factor (display), optical blur (VR/AR optics and eye), and image processing (graphics pipeline). The metric can be defined in much the same way of traditional MTF for imaging systems by examining image formation of a point source and then performing Fourier transform over the response function,[1] [2] with special mathematical treatments.

The system MTF model correlates well the perceptual study results and can be used to predict the perceptual performance of a VR or AR system. The model allows us to optimize the entire system from graphics pipeline to optical system architecture. We demonstrate one application on perceived text quality, where two weight functions depending on text orientation and frequency were incorporated into the above model.

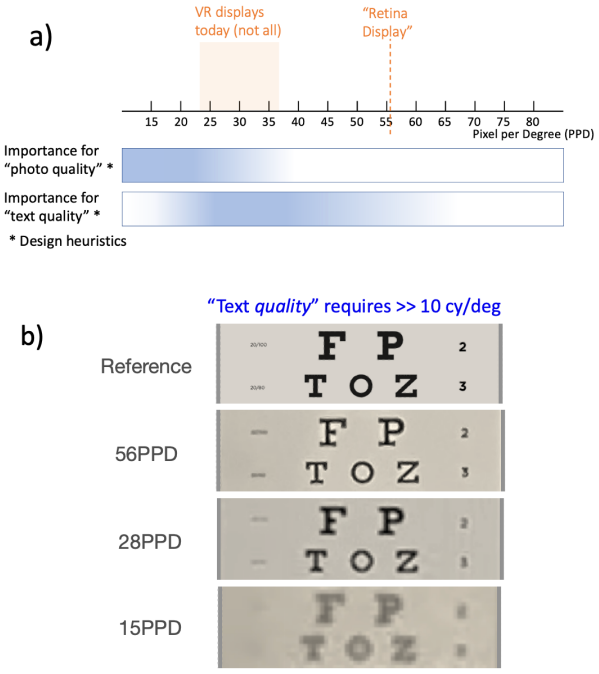

Figure 3. a) A landscape of resolution for typical consumer electronics (displays of mobile phones, laptops, and monitors, and mobile phone cameras). Text requires higher spatial frequencies than graphs. b) example photographs of a Snellen eye chart (2nd and 3rd rows representing 20/100 and 20/80 vision respectively) taken with a recent mobile phone at different resolutions.

System MTF Model and Perception Metric

The modulation transfer function (MTF) is a metric in the spatial frequency domain that measures the performance of an image/ display system. It involves a point spread function (PSF) that fully characterize the system and an image formation process.[2][5] Every pixel in input content results in a shifted PSF. The output image is the overlapped sum of these PSFs. Each point in output image receives contribution from many pixels in input content. This process is also referred to as convolution of content source with a PSF. The MTF is calculated from a slice of the PSF and it can vary with orientation. [2][5]. A key difference in the MTF of a display or a sensing system to that of a smooth optics lens is that the display and sensor are pixelized (i.e., a sampling system). This difference requires additional mathematical treatment to MTF model.

There are unique benefits of using MTF:
1) *MTF is conveniently multipliable*. To predict a system performance, we can simply multiply the several components' MTFs. For example, the MTF of a MR system is the MTF of the display module multiplied by the MTF of the camera module. In addition, display module's MTF is the product of VR lens MTF and display MTF. Similarly, MTF of the MR camera module is the product of lens MTF and sensor MTF.

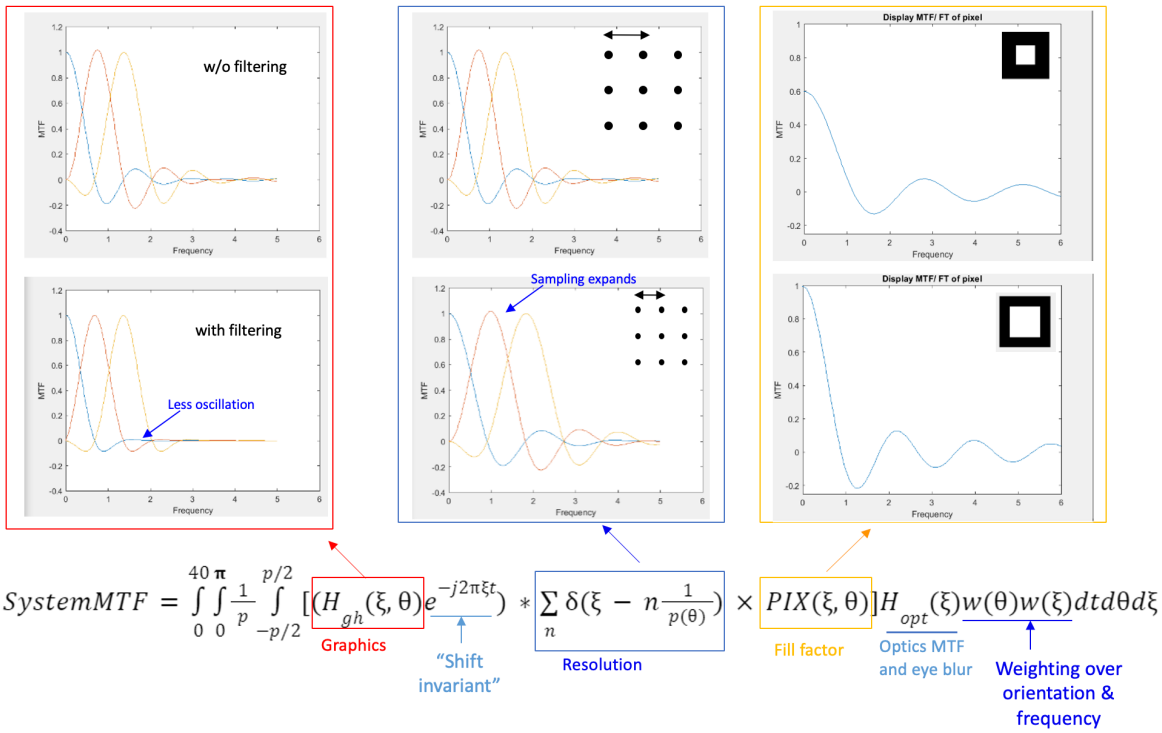

Figure 4. An overview of the system MTF model and formula. Some key contributors it includes are the content source's MTF, contribution from resolution/ PPD, pixel layout/ fill factor, and optics MTF including eye blur. We also demonstrate the change in frequency response when applying a graphic filter, or varying PPD and fill factor.

2) *MTF is quantifiable with a specification and measurable using test metrology.* For example, traditional photographic imaging knowledge states that at about 3-15 cycles/deg we need > 0.2 ideally > 0.35 MTF. MTF metrology is widely available with several different approaches. Below system MTF model provides an additional approach targeted for sampled system.

Figure 4 is an overview of the system MTF model from this work. Some key variables it includes are: the content source/ graphics, resolution/ PPD, pixel layout/ fill factor, and optics MTF including eye blur. The top graphs of Figure 4 also show how the shape of frequency response changes when graphic filter, PPD and fill factor are varied. The unit of x axis' frequency is the inverse of a spatial unit in the sampled image. The product of these curves to generate system MTF can be complex and somewhat unpredictable. Thus a complete and ideally analogue formula is desirable to predict the end performance.

We spend a few steps to explain the system MTF model. In a sampled display or imaging system, the output of the image is

$$img(x) = (f(x) \times \sum_{-n}^{n} \delta(x - np)) * pixel(x) \quad (1)$$

where $f(x)$ is the input signal for the content, p is the pitch distance of the display pixels, $pixel(x)$ is the pixel spatial function (e.g., the shape function shown in the last column of Figure 4). From (1) we cand define the PSF of the sampled system:

$$psf(x) = (psf_{gp}(x) \times \sum_{-n}^{n} \delta(x - np)) * pixel(x) \quad (2)$$

In a sampled system, the requirement of shift invariance when calculating MTF is violated. As shown in Figure 5a, depending on where the signal lands relative to the display pixels, the reproduced signal on the sampled display can be very different. A spatially averaged impulse response and a corresponding MTF component that is inherent in the sampling process by assuming that the scene being imaged/displayed is randomly positioned with respect to the sampling site. The new PSF will be:

$$psf_{avg}(x) = \frac{1}{p}\int_{-p/2}^{p/2}(psf_{gp}(x - t) \times \sum_{-n}^{n} \delta(x - np)) * pixel(x)dt \quad (3)$$

We then perform Fourier transform of the PSF and calculate the MTF:

$$MTF_{avg}(\xi) = |\frac{1}{p}\int_{-p/2}^{p/2}((H_{gp}(\xi)e^{-j2\pi\xi t}) * \sum_{-n}^{n} \delta\left(\xi - n\frac{1}{p}\right) \times PIX(\xi))|dt \quad (4)$$

where $H_{gp}(\xi)$ is the graphics/content's MTF; $PIX(\xi)$ is the Fourier transform of the pixel function. The integral can be done and the equation simplifies to:

$$MTF_{avg}(\xi) = |PIX(\xi))\sum_{-n}^{n} H_{ph}(\xi - n/p)sinc(p\xi - n)| \quad (5)$$

We take n of -1, 0 and 1 (sometimes we need to consider more terms such as +/-2 and +/-3) and the equation becomes:

$$MTF_{avg}(\xi) = |PIX(\xi)(H_{gp}\left(\xi - \frac{1}{p}\right) sinc(p\xi - 1) +$$

$$H_{gp}(\xi)sinc(p\xi) + H_{gp}\left(\xi + \frac{1}{p}\right) sinc(p\xi + 1))| \quad (6)$$

We further find the MTF also depends on the orientation of the feature. For example, as shown in Figure 7a, the capital letter "E" only has horizontal lines and vertical lines, and small letter "e" has a different set of orientation distributions. We added a parameter angle θ which affects the pitch function p and the *PIX* function in the formula above. Now we have:

$$MTF_{avg}(\xi, \theta) = |PIX(\xi, \theta)[H_{gp}(\xi - 1/p)(sinc(p(\theta)\xi - 1) +$$

$$H_{gp}(\xi)sinc(p(\theta)\xi) + H_{gp}(\xi + 1/p)sinc(p(\theta)\xi + 1)]| \quad (7)$$

The final expression for system MTF on displaying text content when adding two weight function is shown below:

$$SystemMTF = \int_0^{40} \int_0^{\pi} \frac{1}{p} \int_{-p/2}^{p/2} [(H_{gh}(\xi, \theta)e^{-j2\pi\xi t}) * \sum_{-n}^{n} \delta(\xi -$$

$$n\frac{1}{p(\theta)}) \times PIX(\xi, \theta)]H_{opt}(\xi)w(\theta)w(\xi)dtd\theta d\xi \quad (8)$$

Where $w(\theta)$ is the angle dependency of the text content (for example, English text has more vertical content than Chinese, Japanese and Korean, Figure 6b). $w(\xi)$ is the frequency weight function, which represents what frequencies affect the text quality. For example, the low frequencies from 0 cycles/deg to 5 cycles/deg form the skeleton of text while the high frequencies determine how sharp or clear the texts appear. Text quality instead of text legibility is the focus. A preliminary study indicates the frequency around 10-20 cycles/deg (centered at 15 cycles/deg) is dominating in influencing text quality. A future study will give a more complete frequency dependency function.

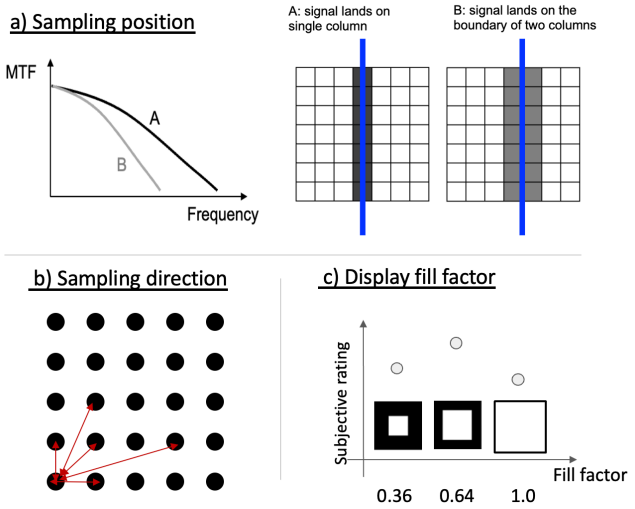

Figure 5. Example factors considered in the math model of system MTF metric: a) sampling nature of a display and sensor; b) sampling direction; c) pixel structure and layout. In this case of a low-resolution design, users prefer a moderate but not the highest fill factor.

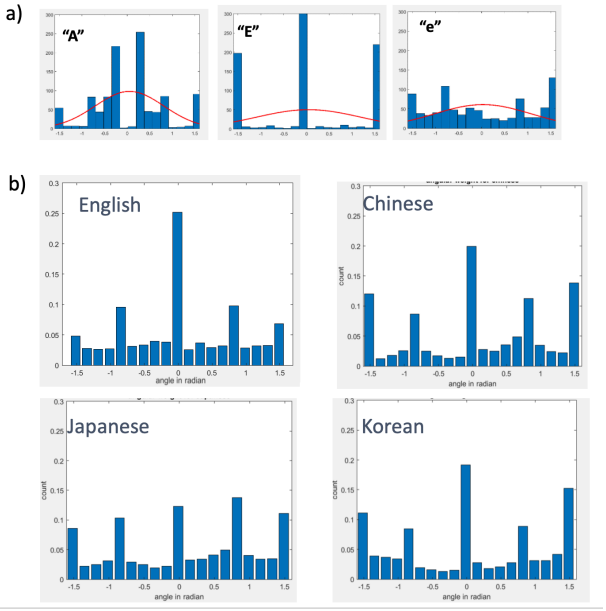

Figure 6. Text orientation weighting functions. a) example analysis of orientation weights for strokes in letter "A", "E" and "e". This letter-wise analysis was done across 3 popular fonts and then combined with usage frequency of each letter. b) comparison of orientation weights for four different languages.

User Study for Validation of the Model

In order to correlate the MTF model with the user preference, we conducted a user study to evaluate text quality over different display resolutions. We used a high resolution 8K monitor (280PPI, Dell UP3218K) and a viewing distance of approximately 2500 mm to achieve a 480 PPD native resolution. Multiple display pixels can be combined to emulate the pixel structure of a VR/AR headsets (e.g., using 10x10 native pixels to emulate 1x pixel in VR) as shown in Figure 7. High-resolution content is processed with the display pixel layout to generate a pixel-by-pixel map. By varying the virtual pixel size (e.g., 10x10, 15x15, 20x20) and the viewing distance, we can emulate multiple resolutions on a single display. Font size was chosen to be equivalent to 10-12 pt at 400mm viewing distance (standard monitor), or 0.28-0.34 deg x-height. In other words, text legibility is not a problem (there are >11 pixels per letter height). Participants sat in a chinrest at a specified distance away from the display to simulate the desired PPDs. The display was calibrated to match typical VR headset brightness (~100 nits). Display fill factor was fixed at 100% for this study. Participants ran the study in a dark room and were verified to have corrected vision with a questionnaire, prior to participation. We used a two-alternative forced choice experimental paradigm, presenting pairs of text samples with different PPDs and asking participants to choose the preferred sample.

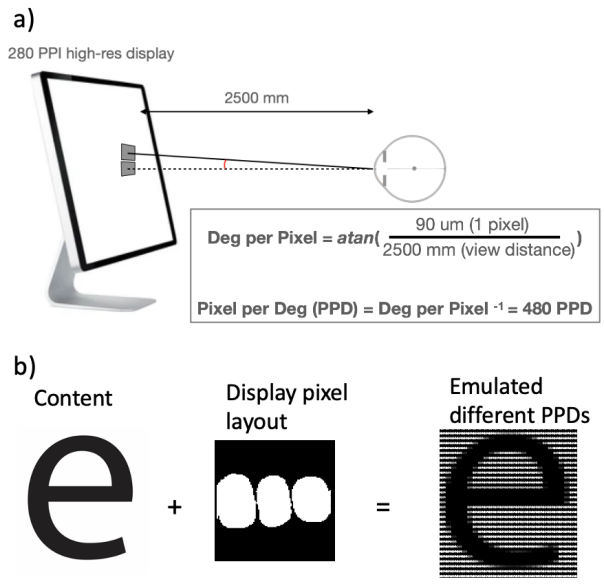

Figure 7. a) Setup for the user study on text quality ratings. The native resolution of the display is 480 PPD when viewing distance is 2500mm. b) process of emulating different PPDs and pixel layout.

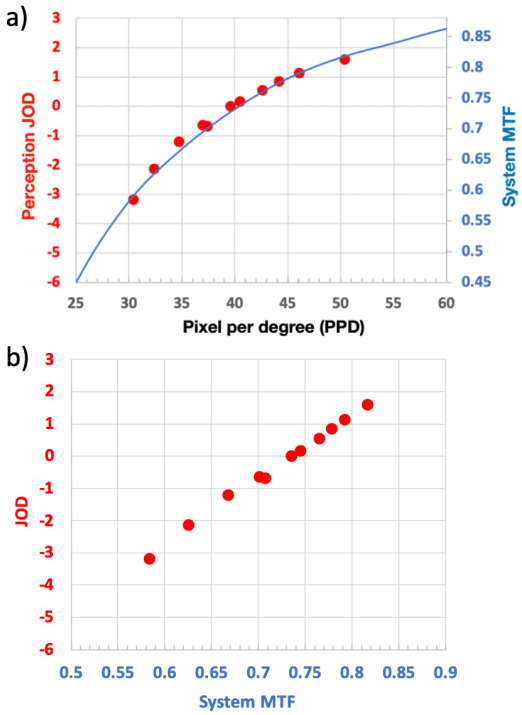

Figure 8. a) the JOD result vs. PPD and the calculated system MTF vs. PPD are overlayed on top of each other. b) the same data as in Figure 8a but plotted as JOD vs. system MTF.

Figure 8 shows the result of the user study with varied angular resolution of 30-50 PPD. Preference differences are quantified in just-objectionable-difference (JOD) units [7], which maps to the probability of observers preferring one PPD over another (Table 1). JOD units are similar in concept to just-noticeable-different (JND) units, but are better suited for describing image quality and preference across multiple dimensions, as described in [7]. In this case, a difference of 1 JOD corresponds to 75% of observers preferring one resolution over the other. A positive JOD indicates a preference for one resolution, whereas a negative JOD indicates preference for the other (Table 1). A difference of 0 JOD indicates equal preference between the two conditions.

In Figure 8a, the result of JOD vs. PPD values and the calculated system MTF vs. PPD values are overlayed on top of each other. Figure 8b plots the same data but correlates JOD vs. system MTF. Some interesting conclusions can be drawn: 1) The shape of the system MTF correlates to the user data in this range, indicating that the model may be well-suited for this use-case 2) A 0.05 MTF change corresponds to ~1 JOD, highlighting the sensitivity of the human visual system; 3) an increase of resolution from 25 to 30 PPD greatly improves user perceived text quality (3 JOD), but an improvement of 50 to 60 PPD is still meaningful (1 JOD).

These results link the system MTF metric to user preference of text quality across different display resolutions. Future work will explore how the metric predicts experimental results for other system parameters, such as fill factor, subpixel layout, or rendering algorithms.

JOD	Population % who prefers one PPD over the other
-3	3%
-2	9%
-1	25%
0 (no preference)	50%
1	75%
2	91%
3	97%

Table 1 Interpretation of JOD. JOD scale is linked to a probability of selecting one condition over another.

Conclusion

The high-resolution experience in VR and AR depends on many factors – PPD, pixel layout, fill factor, rendering, optics and content itself. In this work we presented a system MTF method to capture all the complex contributors, match user experience quantitatively, and guide practical system design.

The system MTF can be defined in a similar way of traditional MTF for imaging systems. We demonstrate its application on perceived text quality, where two weight functions depending on text orientation and frequency were incorporated into the above model.

The system MTF model correlates well the perceptual study results and can be used to predict the perceptual performance of a VR or AR system.

References

[1] R. Vollmerhausen, D. Reago, R. Driggers, Analysis and Evaluation of Sampled Imaging Systems, SPIE, 2010

[2] S. Smith, The scientist and Engineer's Guide to Digital Signal Processing, California Technical Publishing, 1997

[3] "Retina Display", Wikipedia, Wikimedia Foundation, 20 February 2022, https://en.wikipedia.org/wiki/Retina_display

[4] "Subpixel Rendering", Wikipedia, Wikimedia Foundation, 20 February 2022, https://en.wikipedia.org/wiki/Subpixel_rendering

[5] G. Boreman, Modulation Transfer Function in Optical and Electro-Optical Systems, SPIE, 2021

[6] D. Williams, "What is an MTF? ... and why you should care?"

[7] M. Pérez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise and R. K. Mantiuk, "From Pairwise Comparisons and Rating to a Unified Quality Scale," in *IEEE Transactions on Image Processing*, vol. 29, pp. 1139-1151, 2020.

Author Biography

Jiawei Lu is a Ph.D. candidate in the field of biomedical imaging instrumentation at Tkaczyk Lab, Rice University. He received his master's degree in Optical Science at the University of Arizona in 2018. Jiawei's research interests include the development of advanced hyperspectral imaging systems, as well as the fabrication of micro-optics using diamond machining and two-photon polymerization. He contributed to this work when he worked as an optical design engineer intern at Meta.

Trisha Lian is a research scientist at Reality Labs. Before joining Meta, she received her PhD in Electrical Engineering at Stanford University, working with Professor Brian Wandell. Her current research focuses on developing imaging system simulations alongside models of the human visual system, to better understand visual artifacts in head mounted displays.

Jerry Jia is a system engineer and human perception specialist at Meta Reality Lab in California US. He advocates for experience-centric product design and an integration of hardware system, algorithms, and human vision in product development of VR/AR. Prior to joining Meta, he was a key inventor and engineer for Apple's True Tone display feature. He received his Ph.D. in Material Science and Engineering from Columbia University in the City of New York, B.S. in Physics, B.A. in Philosophy from Peking University in Beijing.

.