Assessing 3D Scan Quality Through Paired-comparisons Psychophysics

Jacob Thorn[1,2], Rodrigo Pizarro[1,3], Bernhard Spanlang[3], Pablo Bermell-Garcia[1],
Mar Gonzalez-Franco [1*]
[1] Airbus Innovation Laboratories, Airbus Group, United Kingdom
[2] Department of Computer Science, University College London, United Kingdom
[3] Experimental Virtual Environments for Neuroscience and Technology Lab (EVENTLab), Universitat de Barcelona, Spain
*Corresponding author: margon@microsoft.com

ABSTRACT

Consumer 3D scanners and depth cameras are increasingly being used to generate content and avatars for Virtual Reality (VR) environments and avoid the inconveniences of hand modeling; however, it is sometimes difficult to evaluate quantitatively the mesh quality at which consumer available 3D scans should be exported, and whether the object perception might be affected by its shading. We propose using a paired-comparisons test based on psychophysics of perception to do that evaluation. As psychophysics is not subject to opinion, skill level, mental state, or economic situation it can be considered a quantitative way to measure how people perceive the mesh quality. In particular, we compare four different levels of mesh quality (1K, 5K, 10K and 20K triangles). We present two studies within subjects: in one we investigate the influences of seeing an object in a regular screen vs. in a Head Mounted Display (HMD); while in the second experiment we aim at detecting the effects of shading into quality perception. At each iteration of the pair-test comparisons participants pick the mesh that they think had higher quality; by the end of the experiment we compile a preference matrix. The results show a correlation between real and assessed quality, despite participants' reported uncertainty. We also find an interaction with quality and shading, which gains importance for quality perception when the mesh has high definition. Furthermore, we assess the subjective realism of the most/least preferred scans using an Immersive Augmented Reality (IAR) video-see-through setup to compare the real vs the 3D scanned object in the same HMD environment. Results show higher levels of realism were perceived through the HMD than when using a regular monitor, although the quality was similarly perceived in both systems.

Keywords

Virtual reality, mixed / augmented reality, scanners, metrics, mesh geometry models, perception, paired-test, psychophysics.

MM '16, October 15 - 19, 2016, Amsterdam, Netherlands

ACM 978-1-4503-3603-1/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2964284.2967200

1. INTRODUCTION

Content creation and 3D modeling have long been a critical restriction to the VR expansion. In the game industry one approach to tackle this problem has been based on self-content creation, by providing tools for the users, they can generate and share their own content, scenes and avatars [1]. Thus, moving from the pre-modeled content to personalized environments with infinite content combinations. Similar ideas for content creation are appearing for VR setups, e.g. with digital sculpting [2], however some methods can get increasingly complex and require artistic skills from the users. With the appearance of depth sensors and 3D scans, we have seen a new boost in self-content creation. Using this technology users create new content out of real-life objects [3], own look-a-like avatars [4], [5] or complete scenes [6]. The quality of the scenes or objects created may vary across different scanning technologies and researchers have turned to the field of mesh quality evaluation in order to optimize the newly generated meshes so they can be used in scenarios such as Immersive Virtual Environments (IVE).

A classical evaluation approach in the domain of imaging that has been applied to computer rendering is the Visible Difference Predictor (VDP) [7], which assesses the dissimilarities with original inputs to estimate changes in perception [8]. Besides, research on mesh quality evaluation has also focused on geometric criteria or algebraic theories [9]. These approaches are able to quantify the results of meshing processes and influence the specifications of the mesh creation, optimization or smoothing algorithms. However, it is clear that the quality of the mesh is also bound to the fundamental limits of human perception, thus more subjective components have also been explored with questionnaires [9], or with ordered selections [10]. Some more complex approaches have created predictors of the subjective effect that a new rendering or lossy compression technique will have on the participants' perception [9]. Additionally, human perceptual evaluations can also be measured through behavioral responses [11], or even physiological measures [12], [13]; other research has looked at finding thresholds of perception through psychophysics [14], [15]. In fact, psychophysics is a good way to explore the fundamental limits of perception, since it involves the use of innate skills in estimation and sensory mechanisms [16]. For the case of mesh quality perception the psychophysics methodology can be borrowed from colorimetric matching [17]; where subjects determine equivalence classes of spectral content based on matching colors until they are perceptually indistinguishable, at Just-Noticeable-Differences [18] and bellow subjective thresholds. By analogy with the psychophysics of color matching, equivalence between different mesh qualities can be determined via a forced-choice pair-based comparison [19]. This method builds a full ranking of preference ordered mesh qualities instead of providing only absolute values.

In the current paper we present two experimental studies that explore the feasibility of the pair-based psychophysics approach to research mesh quality perception thresholds as well as to determine the importance of the two influencing factors (display and shading) [20].

2. Materials and Methods

2.1 Procedure

The primary investigation was to determine the extent to which participants were able to distinguish the different qualities of 3D scanned meshes using psychophysics (Figure 1). The secondary purpose of the study was to determine the influence of the use of HMD stereoscopic displays compared to traditional desktop screens (Experiment 1, n=20 participants aged 33.5 ± 8.9 years, 2 females), as well as the importance of shading (Experiment 2, n=21 participants aged 31.9 ± 8.9 years, 3 females). The two experiments were run within subjects, Experiment 2 also had a full factorial design with two factors: mesh quality (with 4 levels corresponding to the four exported models); and rendering mode (with 2 levels: unlit shader vs. Lambert diffuse shader) and was fully provided via HMD.

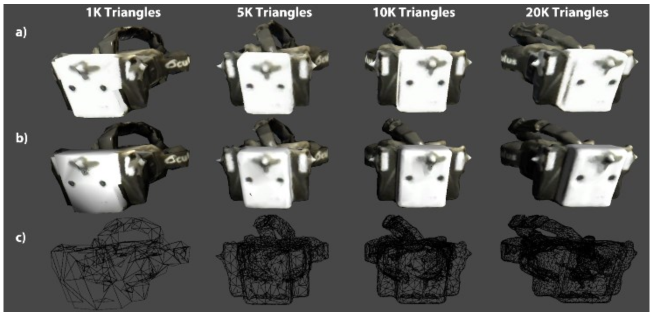

Figure 1. Scanned objects with 1K, 5K, 10K, 20K triangles (from left to right). (a) Unlit objects and (b) Lambert diffuse shaded objects are compared through paired-comparisons psychophysics. (c) Corresponding wireframe.

During the paired-comparisons, the 3D scanned objects were presented in random order appearing side by side and their position was also randomized. At each comparison participants were asked to select "which polygonal mesh had higher quality".

At the end of both experiments, participants were asked to measure the absolute realism of their most and least preferred meshes as described in the Measures section.

2.2 Apparatus

2.2.1 Scan

The models were scanned using an Asus Xtion PRO LIVE RGB and Depth Sensor camera and the software Skanect (see supplementary video). 3D models were trimmed and converted to watertight using the lowest smoothing option, they were also colour textured using the standard colourize settings within Skanect. Models were finally exported to the different resolutions of 1K, 5K, 10K and 20K to be used in the experiment. The scanned object was unknown to the participants.

2.2.2 Display

The mesh quality was tested in a monoscopic display HP LP2065 LCD running at a 1600x1200 pixel resolution, true life 32-bit at 60Hz. And in an Oculus Rift DK1 HMD. The scenario was built in Unity3D and rendered with a resolution of 1024x768, to match it of the HMD. The selections in the force-choice paired comparisons were done with a mouse click in the monoscopic display, and though a wand interaction on the HMD. Participants were allowed to explore the objects by moving the mouse in the monoscopic display and through real-time motion tracking in the HMD. The HMD head tracking was performed with a NaturalPoint Motive motion capture system (24 x Flex 13 cameras) running at 120Hz that streams the head's position and rotation providing a first person perspective to explore the object [21]. Since depth perception could also play a larger role in the selection, objects were equidistant at 20cm in both conditions (see supplementary video).

For the subjective comparisons with the real object in the stereoscopic condition we implement a video-see-through augmented reality rift [22], [23] (Figure 2), coupled to the HMD with a 3D printed body that holds two Logitech C310 cameras. The lenses were replaced by those of a Genius Widecam F100 to reduce the disparity in FOV between the HMD and the cameras. As a result, the setup features a 90º horizontal FOV and the aspect ratio is 1.33:1 for both the cameras and the Rift (the resulting FOV can be observed in Figure 2b). Although the frame rate of the cameras is less than the one featured by the HMD's (~45Hz and 60Hz respectively) the system is operative in real-time. The camera lenses optical distortion was corrected in real-time with a shader using pre-calculated camera calibrations [24].

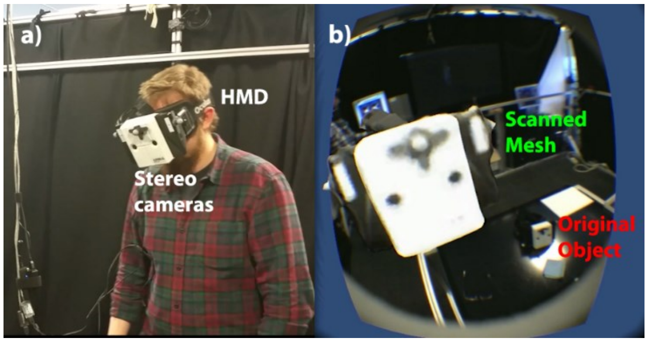

Figure 2. Augmented Reality Stereoscopic Setup: a) head tracked HMD and stereo cameras for video see through; b) the selected mesh can be seen together with the real object.

2.3 Measures

2.3.1 Paired comparisons test

Participants compared the different quality models exported from the scanning software in a forced-choice paired comparison psychometric task [19]. This comparative approach is based on context interpretation rather than abstract rankings and can be used to explore thresholds of perception while reaching statistical stability [19]; indeed paired comparisons can provide reliable rankings for an entire set of elements especially when participants are completely unable to subjectively determine a difference between the options at first glance [19]. The total number of comparisons presented is given by equation (1),

$$c = \frac{n\,(n-1)}{2} \qquad (1)$$

where n is the number of meshes in the set and c is the total number of comparisons between them.

During the experiments, the comparisons are performed twice for each pair, i.e. $2c$. If the participant's choice is not consistent in both comparisons, a third one is presented. This method results in a greater reliability of the final result. Therefore, the maximum number of comparisons c_{max} is given by equation (2).

$$c_{max} = \frac{3n\,(n-1)}{2} \qquad (2)$$

Therefore, in the first experiment with four meshes to be compared (n=4), participants had to do between 12 and 18 comparisons in each condition (stereoscopic and monoscopic). While in the second experiment with the full factorial design comparing four meshes and two shaders (n=8), participants had to perform between 56 and 84 comparisons in total.

While one way to determine confidence is to directly ask participants, this information can also be derived from the number of comparisons they underwent. Peterson and Brown's method accounts for the ambiguity of choices by creating a preference score that includes the number of comparisons that were needed at each step [19]. Therefore, the preference score *ps* between two meshes *A, B* corresponds to the difference in number of times the participant selected each mesh over the number of times the comparison was presented, represented by equation (3),

$$ps = \frac{(t_A - t_B)}{(t_A + t_B)} \qquad (3)$$

where t_A and t_B are the number of times the participant preferred meshes A and B correspondingly. I.e. if two meshes were compared and the same mesh was selected both times the preference score for that comparison would be (2-0)/(2+0)=1, if a third comparison was needed because no clear choice was found, the preference score would reduce and be (2-1)/(2+1)=0.333. The final preference score for a mesh would include the scores from all the comparisons within all the meshes: $\sum ps$. I.e, when comparing 4 meshes the maximum preference score for a mesh would be 3. Additionally Peterson and Brown also propose consistency checks on the decisions through the evaluation of circular triads in a person's choices [19].

2.3.2 *Subjective evaluation, realism and confidence*

In order to have a more absolute evaluation of the realism we select the most and least preferred mesh and asked for a comparison with the real object using the Immersive AR setup. In the monoscopic condition participants undergo a screen to real object comparison. "How much from 1 to 10 does this mesh look like the real object?" (From 1 not at all, 10 looks just like the real object).

Additionally, in the second experiment we also assessed participants' confidence level during the choices "How often were you certain of the answers or were you guessing?" (From 1 always certain to 10 always guessing).

3. Results

3.1 Experiment 1: Stereoscopy

3.1.1 *Paired-test comparison*

Participants compared the different quality models exported in a regular monitor condition and in the HMD condition, all under the same unlit shading conditions (Figure 1a). In both conditions we find a correlation between the mesh quality and the preference score in the force-choice paired test (Pearson correlation for monoscopic *r*(78)=0.61, p<0.001, and for stereoscopic *r*(78)=0.55, p<0.001), this correlation shows how the preference (number of times each mesh is selected) increases for higher mesh qualities (Figure 3). When looking into the data more in detail, we find that the thresholds of perception were not so clear for participants when comparing smaller quality changes such as 5K to 10K or 10K to 20K (Wilcoxon signed-rank test, Z <-1.3 p>0.1), but bigger mesh qualities where significantly differentiated such as 5K to 20K or 1K to the rest (Z <-2.7, p<0.01), both in the monoscopic and stereoscopic condition. Post-hoc comparisons between the two viewing conditions do not show significant differences (Z=0.37, p=0.7). Regarding confidence, the number of repeated comparisons needed both the monoscopic and the stereoscopic condition were similar (Z=-0.42, p=0.7).

3.1.2 *Subjective realism evaluation*

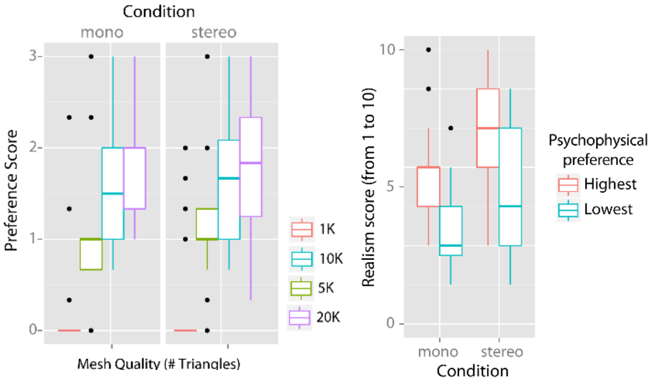

Figure 3. Left: boxplot representing distributions of the psychophysical preference score between the different meshes. Right: subjective realism of the most and least preferred meshes for both conditions. (Quartiles, median, standard deviation, and outliers marked).

Meshes seen in the stereoscopic condition were significantly perceived as more realistic than in the monoscopic condition (Wilcoxon paired signed rank test, Z=-2.5, p=0.01), independently of the quality mesh (Figure 3). Furthermore, as expected, the realism of the perceived highest quality mesh was ranked significantly higher than the one perceived as lowest quality mesh in both conditions (Z =-3.5, p<0.001). This result is a consistency check as when the quality mesh is increased so is the perceived realism of the object.

3.2 Experiment 2: Shading and Confidence

Participants compared the different quality meshes (5K, 10K and 20K) in two rendering conditions with an unlit shader vs. a Lambert diffuse shader, which computationally is not much more expensive than no shading. This experiment was run in the HMD and had a full factorial design, we analysed the results within and between class comparisons covariate by the subjective confidence level as a result to the question "How often were you certain of the answers or were you guessing?" (From 1 always certain to 10 always guessing). Participants were clustered as High Confidence (HC) if scored from 1 to 4, (n=8) and as Low Confidence (LC) if scored from 5 to 8 (n=13), nobody reported values higher than 8. Scores of 5 or higher can arguably be considered LC as it would represent guessing in 50% or more of the cases.

3.2.1 *Within comparisons*

We find that while the HC participants were able to suitably assess the mesh quality in the unlit meshes (significant Pearson correlation *r*(24)=0.53, p=0.007), the LC had a reduced ability to detect higher qualities (no correlation, Pearson p=0.85) (Figure 4).

Those differences were not found for the shaded models, where both HC and LC participants were not able to tell which mesh was better (no correlations, Pearson p>0.48). This can be interpreted as if the thresholds of perception where more diffused in the Lambert shaded condition, the original differences and correlations found in the Unlit condition disappear as if the different qualities of the mesh were harder to distinguish in the Lambert condition.

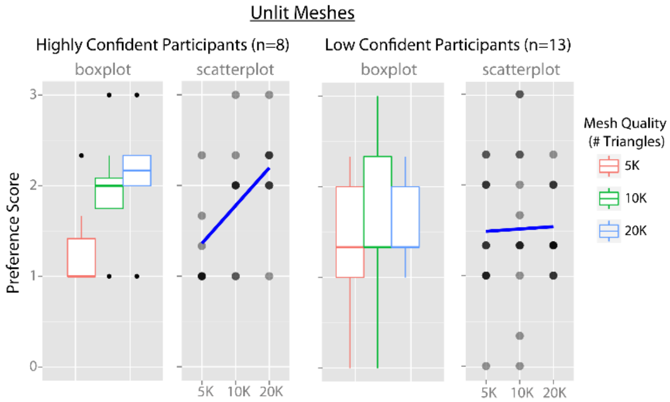

Figure 4. Left: HC preferences for the unlit comparisons. Right: LC results for the unlit comparisons.

3.2.2 *Between comparisons*

When considering all the between comparisons (unlit vs lit and different qualities) we did not find significant differences in preference between the two modes of shading per mesh quality, (Unlit and Lambert). No differences nor significant correlations were found when clustering the HC and LC participants' preferences (Figure 5). This can be interpreted as if participants were not taking shading into account to judge the mesh quality of 3D objects, although they might find higher realism in the Lambert shaded condition.

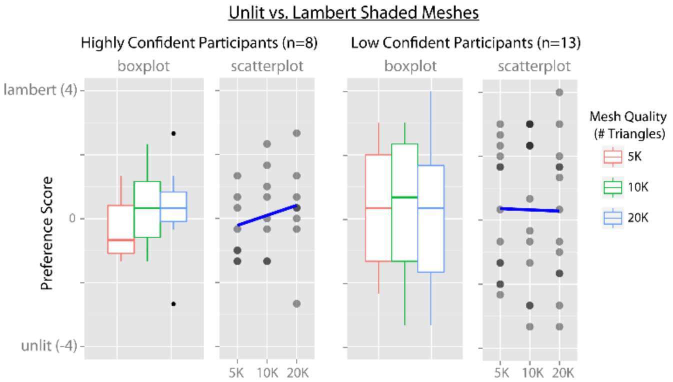

Figure 5. Left: HC participants' results for the Unlit vs. Lambert. Right: LC participants' results for the Unlit vs. Lambert. The boxplot represents the preference for the different mesh qualities. The Y axis goes from preferred Unlit to preferred Lambert, Zero means no preference.

4. Discussion

In the first experiment exploring mesh quality perception under two levels of immersion (monoscopic and stereoscopic viewing) we found a significant correlation between the quality of the mesh and the frequency of selection for both modalities. I.e. participants were equally good at estimating the quality of the meshes in both conditions. These results are consistent with previous studies comparing stereo to mono when performing different tasks (see recent review of over 70 papers [25]). Indeed, our study would be included into the "Finding/Identifying/Classifying Objects" group; over half of the papers in this group found that the stereoscopic view did not directly increase participant performance. Other tasks such as judging distances beneficiated more of stereoscopic viewing [25]. However, participants in our study consistently rated meshes in the stereoscopic view as more realistic than in the monoscopic view. This indicates that although they were not able to differentiate between the mesh qualities in both conditions, the stereoscopic view still made the meshes look more realistic overall. This is consistent with previous research that found clinicians assessing whether the optic disc of patients was presenting signs of glaucoma performed equally well on both viewing modes but had a slight preference towards the stereoscopic condition as being more realistic [26].

In a second experiment we analysed the effect of shading on the quality perception of a mesh with different levels of detail. We introduced an additional subjective assessment to compare high confident and low confident results. Results showed a decreased discernibility of mesh quality in the Lambert shaded condition that was not present in the Unlit comparisons. This perceptual glitch was particularly interesting for the case of the High Confident participants: their ability to discern the higher quality mesh significantly reduced only in the Lambert shading condition. Our results are in the same line as results obtained with VDPs [7] and visual equivalences [8]. Certain light directions can affect the quality perception and then higher and lower quality meshes cannot be distinguished well under shading [27]. In sum improved shading tends to mask lower definition and participants subjectively perceive objects as if they had the same quality as the high definition ones. However, participants did not show preference for the shaded models over the unlit ones. Other research has also shown that people are not actually so concerned when objects and scenes are Not Photorealistic (NPR) in Virtual Reality [28]. Suggesting that object rendering styles do not influence basic object identification, and mesh quality assessment could arguably qualify as basic object identification.

Our findings suggest that lower polygonal resolutions can achieve similar levels of quality perception when using shading. Indeed, the discernibility of the mesh quality dropped significantly when using shading; up to a point that even highly confident participants were unable of distinguishing the quality of meshes of 20K, 10K and 5K triangles.

5. Conclusions

Up to date several approaches have been presented to objectively and subjectively evaluate mesh quality [9], which are basic to the future of 3D scanning and self-content creation for VR. However, current objective techniques based on mathematical approaches do not always replicate the actual human perception of the meshes [9], while the subjective approaches have been mostly restricted to questionnaires that can be inadequate to measure differences at the level of perceptual thresholds [9]. The present approach based on psychophysics can help the community in that sense, and provide a more adequate methodology to discover the boundaries in human mesh quality perception [16].

Future research might include: more computationally expensive shading aimed at more realism; or the evaluation of the exported meshes with VDPs and mathematical methods to see how they compare to the psychophysiological results.

In general, we believe the findings and the methodology here presented are of great interest to the Multimedia and VR communities, as they can be used to help generate simpler meshes that are optimized for real-time rendering with fewer triangles and still be perceived as high quality.

6. REFERENCES

[1] J. Togelius, R. De Nardi, and S. M. Lucas, “Towards automatic personalised content creation for racing games,” *Proc. 2007 IEEE Symp. Comput. Intell. Games, CIG 2007*, pp. 252–259, 2007.

[2] H. Chen and H. Sun, “Real-time Haptic Sculpting in Virtual Volume Space,” in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, 2002, pp. 81–88.

[3] F. Bruno, S. Bruno, G. De Sensi, M. L. Luchi, S. Mancuso, and M. Muzzupappa, “From 3D reconstruction to virtual reality: A complete methodology for digital archaeological exhibition,” *J. Cult. Herit.*, vol. 11, no. 1, pp. 42–49, 2010.

[4] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi, “3D Scanning Deformable Objects with a Single RGBD Sensor,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 493–501, 2015.

[5] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev, “3D self-portraits,” *ACM Trans. Graph.*, vol. 32, pp. 1–9, 2013.

[6] M. Niessner, M. Zollhöfer, S. Izadi, and M. Stamminger, “Real-time 3D Reconstruction at Scale Using Voxel Hashing,” *ACM Trans. Graph.*, vol. 32, no. 6, pp. 169:1–169:11, 2013.

[7] S. J. Daly, “Visible differences predictor: an algorithm for the assessment of image fidelity,” in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, 1992, pp. 2–15.

[8] G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala, “Visual equivalence: towards a new standard for image fidelity,” *ACM Trans. Graph.*, vol. 26, no. 3, p. 76, 2007.

[9] D. Berjón, F. Morán, and S. Manjunatha, “Objective and subjective evaluation of static 3D mesh compression,” *Signal Process. Image Commun.*, vol. 28, no. 2, pp. 181–195, 2013.

[10] S. Silva, C. Ferreira, J. Madeira, and B. S. Santos, “Perceived quality of simplified polygonal meshes: Evaluation using observer studies,” *Proc. Ibero-American Symp. Comput. Graph. SIACG06*, no. August 2015, pp. 169–178, 2006.

[11] M. Gonzalez-Franco, D. Perez-Marcos, B. Spanlang, and M. Slater, “The contribution of real-time mirror reflections of motor actions on virtual body ownership in an immersive virtual environment,” in *2010 IEEE Virtual Reality Conference (VR)*, 2010, pp. 111–114.

[12] M. González-Franco, T. C. Peck, A. Rodríguez-Fornells, and M. Slater, “A threat to a virtual hand elicits motor cortex activation.,” *Exp. brain Res.*, vol. 232, no. 3, pp. 875–87, Mar. 2014.

[13] G. Padrao, M. Gonzalez-Franco, M. V Sanchez-Vives, M. Slater, and A. Rodriguez-Fornells, “Violating body movement semantics: Neural signatures of self-generated and external-generated errors,” *Neuroimage*, vol. 124 PA, no. January, pp. 174–156, 2016.

[14] K. J. Blom, A. I. Bellido Rivas, X. Alvarez, O. Cetinaslan, B. Oliveira, V. Orvalho, M. Slater, and A. I. Bellido, “Achieving participant acceptance of their avatars,” *Presence Teleoperators Virtual Environ.*, vol. 23, no. 3, 2014.

[15] M. Gonzalez-Franco and P. A. Chou, “Non-linear modeling of eye gaze perception as a function of gaze and head direction,” in *Int’l Workshop on Quality of Multimedia Experience (QoMEX), IEEE*, 2014.

[16] S. S. Stevens, *Psychophysics*. Transaction Publishers, 1975.

[17] A. Blake and H. Bülthoff, “Does the brain know the physics of specular reflection?,” *Nature*, vol. 343, no. 6254, pp. 165–168, 1990.

[18] I. Cheng and P. Boulanger, “A 3D Perceptual Metric using Just-Noticeable-Difference,” pp. 97–100, 2005.

[19] G. L. Peterson and T. C. Brown, “Economic Valuation by the Method of Paired Comparison, with Emphasis on Evaluation of the Transitivity Axiom,” *Land Econ.*, vol. 74(2), no. May, pp. 240–261, 1998.

[20] J. a Ferwerda, F. Pellacini, and D. P. Greenberg, “A psychophysically-based model of surface gloss perception,” *Hum. Vis. Electron. Imaging*, vol. 4299, pp. 291–301, 2001.

[21] B. Spanlang, J.-M. Normand, D. Borland, K. Kilteni, E. Giannopoulos, A. Pomes, M. Gonzalez-Franco, D. Perez-Marcos, J. Arroyo-Palacios, X. N. Muncunill, and M. Slater, “How to Build an Embodiment Lab: Achieving Body Representation Illusions in Virtual Reality,” *Front. Robot. AI*, vol. 1, no. November, pp. 1–22, Nov. 2014.

[22] W. Steptoe, S. Julier, and A. Steed, “Presence and discernability in conventional and non-photorealistic immersive augmented reality,” in *ISMAR*, 2014, pp. 213–218.

[23] M. Gonzalez-Franco, J. Cermeron, K. Li, R. Pizarro, J. Thorn, P. Hannah, W. Hutabarat, A. Tiwari, and P. Bermell-Garcia, “Immersive Augmented Reality Training for Complex Manufacturing Scenarios,” *arXiv.org*, p. 1602.01944, 2016.

[24] Z. Zhang, “A flexible new technique for camera calibration,” *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 22, no. 11, pp. 1330–1334, 2000.

[25] J. P. McIntire, P. R. Havig, and E. E. Geiselman, “Stereoscopic 3D displays and human performance: A comprehensive review,” *Displays*, vol. 35, no. 1, pp. 18–26, Jan. 2014.

[26] B. Parkin, G. Shuttleworth, M. Costen, and C. Davison, “A comparison of stereoscopic and monoscopic evaluation of optic disc topography using a digital optic disc stereo camera.,” *Br. J. Ophthalmol.*, vol. 85, no. 11, pp. 1347–1351, 2001.

[27] B. E. Rogowitz and H. E. Rushmeier, “Are image quality metrics adequate to evaluate the quality of geometric objects?,” *Photonics West 2001-Electronic Imaging*, pp. 340–348, 2001.

[28] P. Zimmons, A. Panter, and D. Ph, “The Influence of Rendering Quality on Presence And Task Performance in a Virtual Environment at Chapel Hill at Chapel Hill,” vol. 2003, pp. 2–3, 2003.