On pictures and stuff: image quality and material appearance

James A. Ferwerda*
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology, Rochester, NY, USA

ABSTRACT

Realistic images are a puzzle because they serve as visual representations of objects while also being objects themselves. When we look at an image we are able to perceive both the properties of the image and the properties of the objects represented by the image. Research on image quality has typically focused improving image properties (resolution, dynamic range, frame rate, etc.) while ignoring the issue of whether images are serving their role as visual representations. In this paper we describe a series of experiments that investigate how well images of different quality convey information about the properties of the objects they represent. In the experiments we focus on the effects that two image properties (contrast and sharpness) have on the ability of images to represent the gloss of depicted objects. We found that different experimental methods produced differing results. Specifically, when the stimulus images were presented using simultaneous pair comparison, observers were influenced by the surface properties of the images and conflated changes in image contrast and sharpness with changes in object gloss. On the other hand, when the stimulus images were presented sequentially, observers were able to disregard the image plane properties and more accurately match the gloss of the objects represented by the different quality images. These findings suggest that in understanding image quality it is useful to distinguish between quality of the imaging medium and the quality of the visual information represented by that medium.

Keywords: image quality, material appearance

Figure 1. Image properties and object properties in imaging: a) High quality grayscale image of a glossy black car on a wet concrete pad. b) Half-toned, printed, and rescanned version of the image on the left. Note that while the quality of the image on the right is lower, its ability to represent important object properties such as shapes and materials is largely the same as the image on the left.

1. INTRODUCTION

Realistic images are a puzzle because they serve as visual representations of objects while also being objects themselves. When we look at an image we are able to perceive both the properties of the image and the properties of the objects represented by the image. Research on image quality has typically focused improving the properties of images (resolution, dynamic range, frame rate, etc.) while ignoring the issue of whether the image is serving its role as a visual representation. The tacit assumption is that if the image properties are of high quality then the object properties will be represented with high fidelity. This train of thought seems reasonable and has support from the field of signal processing and information theory, however the danger of focusing exclusively on the properties of the image is that we may miss insights and opportunities about image quality that come from distinguishing between the quality of the imaging medium and the quality of the visual information represented by that medium.

Figure 1 illustrates the differences between these two views of image quality. The left panel shows a grayscale photograph of a black sports car parked on a concrete pad. Both the car and the pad show distinct reflections of the

*jaf@cis.rit.edu, http://www.cis.rit.edu/jaf

surrounding environment that indicate that the car is glossy and the concrete pad is wet. The grayscale levels and contrasts in the image also indicate that the car is black (or a dark color) and the concrete pad has a lighter shade. The right panel shows the same scene represented by a half-toned image created to simulate the contrast and sharpness of a typical newspaper print. This image is clearly different than the one on the left, and in conventional terms one would say that its quality is lower. However as a visual representation, this image is largely equivalent to the one on the left in the sense that we can still perceive important properties of the depicted objects such as the shape, shade, and gloss of the black car, and the shade and gloss/wetness of the concrete pad. It is as if we are able to "see through" the limitations of the image to correctly perceive the properties of the objects.

In this paper we describe a series of experiments that investigate how well images of different quality convey information about the properties of the objects they represent. We focus on the effect that two image properties (contrast and sharpness) have on the ability of images to represent the gloss of depicted objects. As described and illustrated above, we find that there are useful distinctions to be made between the quality of the imaging medium and the quality of the visual information represented by that medium.

The purpose of this work is to understand the relationships between the signal properties of images and the fidelity of the visual information those images convey to human observers. Our goals are to learn more about how images work as visual representations, and to develop more meaningful image quality metrics that better predict how well images with different signal properties serve as visual representations of the objects they depict.

2. RELATED WORK

Measuring image quality is an important aspect of image systems development, and a variety of metrics have been developed for this purpose. *Numerical metrics* quantify the distortions in a test image with respect to a real image or a statistical standard. Well known numerical metrics include mean squared error (MSE) and peak signal to noise ratio (PSNR). Although these metrics are easy to compute, they often do not correlate well with human judgments of image quality. For this reason, *perceptual metrics* have been developed that incorporate computational models of human visual processing. Typically in these metrics, visual models are used to represent an observer's responses to reference and test image and then these responses are compared to identify visible differences and compute quality. Popular perceptual metrics include Daly's Visible Differences Predictor (VDP)[1], the Lubin/Sarnoff model[2], the Structural Similarity Metric (SSIM)[3] and derivatives, and the Visual Information Fidelity (VIF) metric[4]. These metrics often do a better job at predicting perceived image quality than the numerical metrics, however for the most part they still treat images as abstract arrays of pixels. There is little accounting for the fact that there are many different kinds of images, and in particular, that realistic images serve as visual representations that provide information about the properties of the real world.

When we look at realistic images we don't see pixels. Rather, we see objects with recognizable shapes, sizes, and materials, at specific spatial locations, lit by distinct patterns of illumination. Ferwerda and Pellacini[5] introduced the term *functional realism* to describe the idea that radically different renderings (e.g. photographs vs. line drawings) could provide equivalent visual information for given tasks. Ramanarayanan et al.[6] built on this insight and developed a new measure of image quality called *visual equivalence*. Visual equivalence is based on the idea that two visibly different images can convey the same information about object and scene properties to a human observer. In experiments they showed that images that were visibly different according to Daly's VDP (due to surface reflection distortions) could be visually equivalent in that they conveyed the same information about the shapes and material properties of the objects they represented. They went on to derive a Visual Equivalence Predictor (VEP) that they applied to develop efficient, high fidelity image synthesis algorithms. Subsequently Rouse et al.[7] investigated similar issues in their work to develop *image utility* metrics.

The common thread in this work is an understanding of the value of distinguishing between images as signals that reproduce patterns of light and images as messages that convey visual information to observers. This distinction provides a new perspective on image quality that we explore in the following sections.

3. EXPERIMENTS

As outlined in Section 1, we have conducted a series of experiments that investigate how well images of different quality convey information about the properties of the objects they represent. In particular we studied how well observers were able to perceive the gloss of objects represented by normal and low quality (low contrast, blurry) images. The stimuli and methods used in the experiments are described in the following paragraphs.

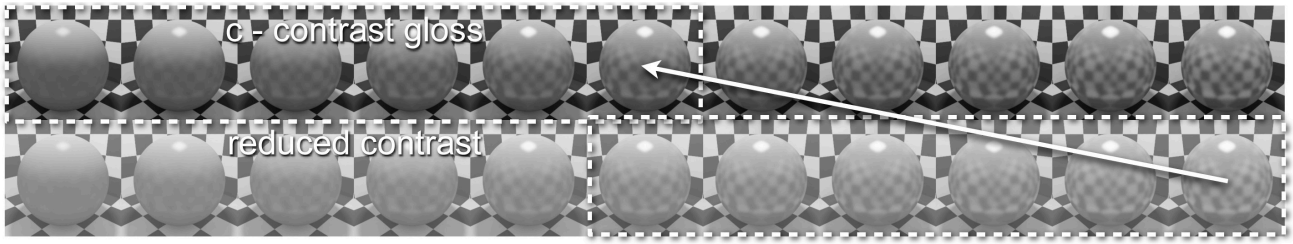

Figure 2. Images used in the experiments, "c" set. The contrast gloss of the balls ranges from 0.019 (low) to 0.190 (high). The top row shows the normal contrast images. The bottom row shows the low contrast images. The reflections in the six glossiest balls in the lower row have approximately the same image contrast as the reflections in the six least glossy balls in the top row.

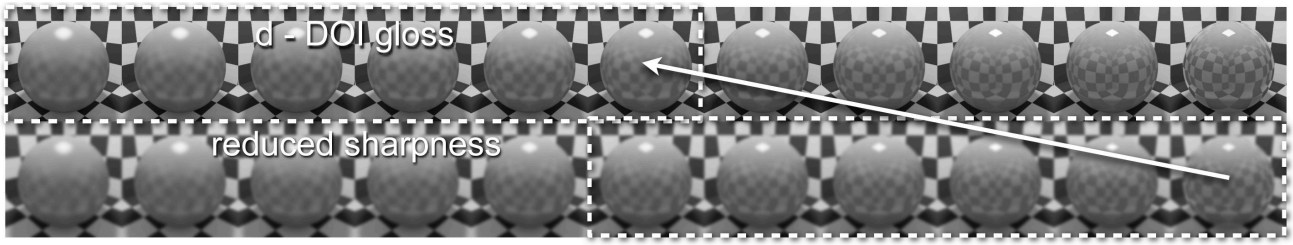

Figure 3. Images used in the experiments, "d" set. The distinctness-of-image gloss of the balls ranges from 0.9 (low) to 1.0 (high). The top row shows the normal sharpness images. The bottom row shows the low sharpness images. The reflections in the six glossiest balls in the lower row have approximately the same edge profiles the reflections in the six least glossy balls in the top row.

3.1 Stimuli

The stimuli used in the experiments are shown in Figures 2 and 3. They are computer graphics renderings of a ball in a checkerboard box with an overhead light source. The materials in the scene were described using the Ward light reflection model[8] that uses three parameters (ρ_d – diffuse reflectance, ρ_s – specular energy, and α – surface roughness) to specify surface reflectance properties. The white and black checks in the checkerboard were perfectly matte with ρ_d's of 0.9 (90%) and 0.03 (3%) respectively, and ρ_s and α values set to 0.0. The balls all had ρ_d values of 0.193 (19.3%) and their ρ_s and α values were varied to produce the different gloss properties seen in the Figures. In previous work Pellacini et al.[9] developed a gloss model that showed that the perceived gloss of similar objects rendered in images was well described by two perceptually uniform parameters, c – contrast gloss, and d – distinctness-of-image gloss, that could be directly related to the parameters of the Ward model though the equations:

$$c = \sqrt[3]{\rho_s + \rho_d/2} - \sqrt[3]{\rho_d}/2$$

$$d = 1 - \alpha$$

This perceptual gloss model was used to describe the gloss properties of the balls in the scene. In the "c" set (Figure 2 top row) the contrast gloss varied in equal steps from 0.019 on the low end to 0.190 on the high end, "d" was fixed at 0.93. In the "d" set (Figure 3 top row) the distinctness-of-image gloss varied in equal steps from 0.9 to 1.0. "c" was fixed at 0.087 The light source in the scene was set to unit intensity. The scene was then rendered at 300x300 pixels in high dynamic range (HDR) image format using the Radiance rendering system[10]. Finally the HDR images were tone-mapped for display as a group using the "highlight compression" algorithm in Photoshop CS3. These images constituted the "normal" image sets used in the experiments.

Two manipulations (contrast reduction, low pass filtering) were used to produce the "low quality" image sets used in the experiments. To create the "low contrast" image set (Figure 2, bottom row), the normal contrast "c" set had its contrast reduced by raising the images' black level in Photoshop. The black level for the entire set was adjusted (-25%, contrast/brightness tool, legacy settings) so that the checkerboard reflection in the glossiest ball in the low contrast set had the same Michaelson contrast as the checkerboard reflection the middle gloss ball in the normal contrast set. This created a rough correspondence between the six most glossy balls in the low contrast set and the six least glossy balls in the normal contrast set that we studied in the experiments. Similarly, the "low sharpness" image set (Figure 3 bottom row) was created by applying a Gaussian blur filter (3.5 pixel width) in Photoshop to the normal sharpness "d" set so that the edge profile of the checkerboard reflection in the glossiest ball in the low sharpness set was the same as the

profile in the middle gloss ball in the normal sharpness set. Similar to above, this created a correspondence between the six most glossy balls in the low sharpness set and the six least glossy balls in the normal sharpness set that we studied in the experiments.

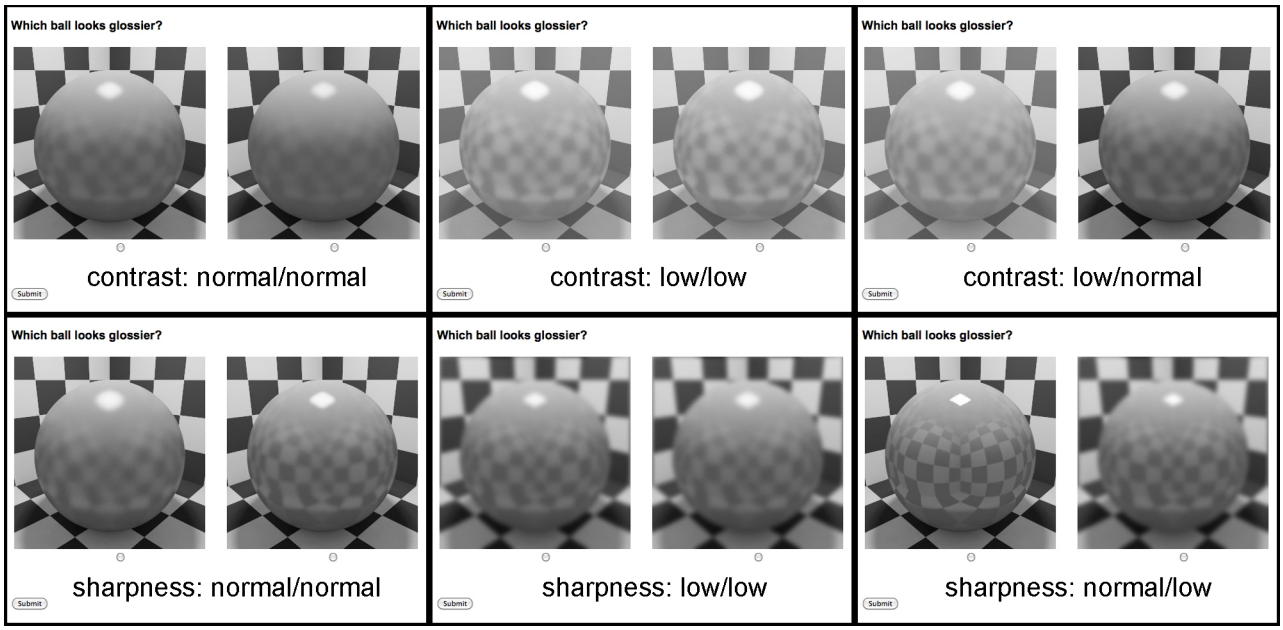

Figure 4. Sample trial screens from Experiment 1: pair comparison method. Top row shows normal/normal, low/low, and low/normal pair comparisons from the "c" condition. Bottom row shows normal/normal, low/low and normal/low pair comparisions from the "d" condition.

3.2 Experiment 1: Methods

Using these image sets we performed a gloss scaling experiment to investigate the effects of image contrast and sharpness on gloss perception. The experiment used a standard full-way pair comparison design[11] and was conducted online using the interface shown in Figure 4. On each trial the observer saw a pair of images randomly selected from either the "c" or "d" sets. The observer's task was to indicate "Which ball looks glossier?" by clicking on one of the radio buttons below each image. Within each "c" and "d" set all image combinations were tested (e.g. normal/normal, normal/low, low/low). Left/right display of each pair was randomized across each trial. Figure 4 shows sample pairs from both the "c" and "d" conditions. Given the 22 images in each set (11 normal, 11 low) this design resulted in (n*n-1)/2 = (22*21)/2*2 = 231 trials/observer which took observers approximately 30 minutes to complete. The "c" and "d" conditions were run as separate studies.

Forty observers participated in the experiment (twenty in each condition), which was conducted online using the Amazon Mechanical Turk system[12]. Observers performed the task using their own computers in their own locations. Constraints were placed on the observer pool to assure that each observer performed a complete set of trials, that the observers had high performance ratings, and that no observers were outliers relative to others in the pool. While running experiments using the Turk system offers no control over viewing conditions and little control over observer characteristics, if the observers perform consistently it suggests that the findings will be robust in real-world applications. In line with system conventions, observers were paid a small amount for their participation.

3.3 Experiment 1: Results and Discussion

The data were analyzed using standard Thurstonian scaling techniques[13] to derive interval scales of perceived gloss as functions of the gloss parameters (c,d) and image types (normal contrast/sharpness, low contrast/sharpness). To do this, first the raw binary judgments from each trial were collated across observers to calculate the frequencies with which any ball was judged to be glossier than any other ball. These frequencies were then converted to Z-scores probabilities in a standard normal distribution and the average Z-scores for each ball/image combination were used to define the perceived gloss values for that visual stimulus on an interval scale. The results of Experiment 1 are summarized in Figures 5 and 6.

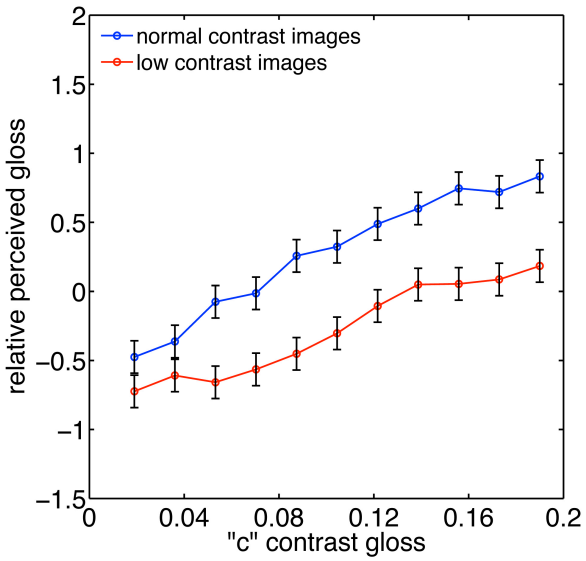

Figure 5. Results of Experiment 1, pair comparison method, "c" condition. Note that for both the normal contrast and low contrast images the perceived gloss values of the balls increase with contrast gloss "c", but that at each "c" level the perceived gloss values for balls shown in the low contrast images are systematically lower than the same balls shown in normal contrast images.

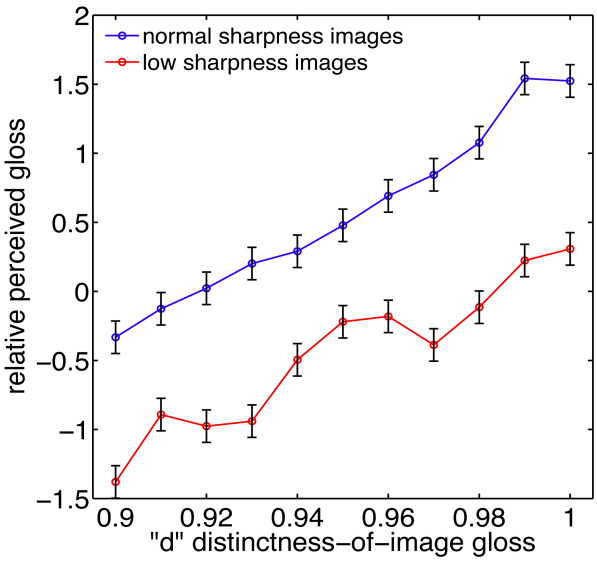

Figure 6. Results of Experiment 1, pair comparison method, "d" condition. Note that for both the normal sharpness and low sharpness images the perceived gloss values of the balls increase with distinctness-of-image gloss "d", but that at each "d" level the perceived gloss values for balls shown in the low sharpness images are systematically lower than the same balls shown in normal sharpness images.

Figure 5 shows the results for the "c" (normal/low contrast) condition. The graph plots perceived gloss as a function of contrast gloss "c" for balls presented in normal and low contrast images. Note first that both curves increase across the range, indicating that observers perceive gloss differences across the range in both the normal and low contrast images. The variance in scale values represented by the error bars indicates that observers are able to perceive approximately 4 JNDs in gloss across the range in the normal images and 3 JNDs in the low contrast images. While the similarity of the curves suggests that observers are getting equivalent information from both the normal and low contrast images, note that at each "c" gloss level the perceived gloss values for the balls presented in the low contrast images are systematically lower than the ratings for the balls presented in the normal contrast images. In fact, the scale value for the glossiest ball (c=0.190) shown in a low contrast image is approximately the same as the mid-gloss ball (c=0.105) shown in a normal contrast image, with rough correspondence across the normal/low contrast pairs tested. This result suggests that while observers are able to perceive gloss in the balls presented in both the normal and low contrast images, to some degree they are conflating the image contrast with the gloss contrast, and are judging the balls seen in the low contrast images as having lower gloss than the balls shown in the normal contrast images.

Figure 6 shows the results for the "d" normal/low sharpness condition, which are similar to those of the previous study. As before, the perceived gloss values for balls presented in the both the normal and low sharpness images increase across the "d" range, indicating that observers are perceiving changes in gloss across the range in both kinds of images. Here discrimination is a bit better with approximately 5 JNDs across the range for the normal images and 4 JNDs for the low sharpness images. Also similar to before, the gloss scale for balls shown in the low sharpness images is systematically lower than the scale for the balls shown in the normal contrast images, with rough correspondence between the six most glossy balls presented in the low sharpness images and the six least glossy balls presented in the normal images. Again this suggests that while observers are perceiving object gloss though both sets of images, they are also influenced by image sharpness when judging distinctness-of-image "d" gloss.

Taken together the results of Experiment 1 seem to confirm the conventional view that the signal properties of images (contrast, sharpness) exert a strong influence on the messaging properties of images (ability to represent the gloss of imaged objects), While it is clear from the experiment that sensitivity to gloss differences is reduced in the "low quality" conditions (4 vs. 3 JNDs across the "c" range and 5 vs. 4 JNDs for the "d" range), it is less clear that the experimental method used (simultaneous pair comparison) provides an unbiased answer to the question of how well observers can

"see through" image properties to perceive the object properties the image represents. It is at least plausible that the simultaneous pair comparison method might emphasize the surface differences between the images (the contrast or blur) and which could influence the observers' gloss judgments. An alternate experimental method, widely used in the color appearance literature is sequential matching[14], in which observers are presented with a reference stimulus which is then removed and replaced with test stimuli which must be matched in some aspect to the reference. The important feature of this method is that the sequential presentation of the reference and test stimuli forces the observer to perform the task based on his or her internal representation of the reference stimulus rather than on the surface features of the images. To investigate if this change of method changes the results, we conducted a second experiment that is described in the following sections.

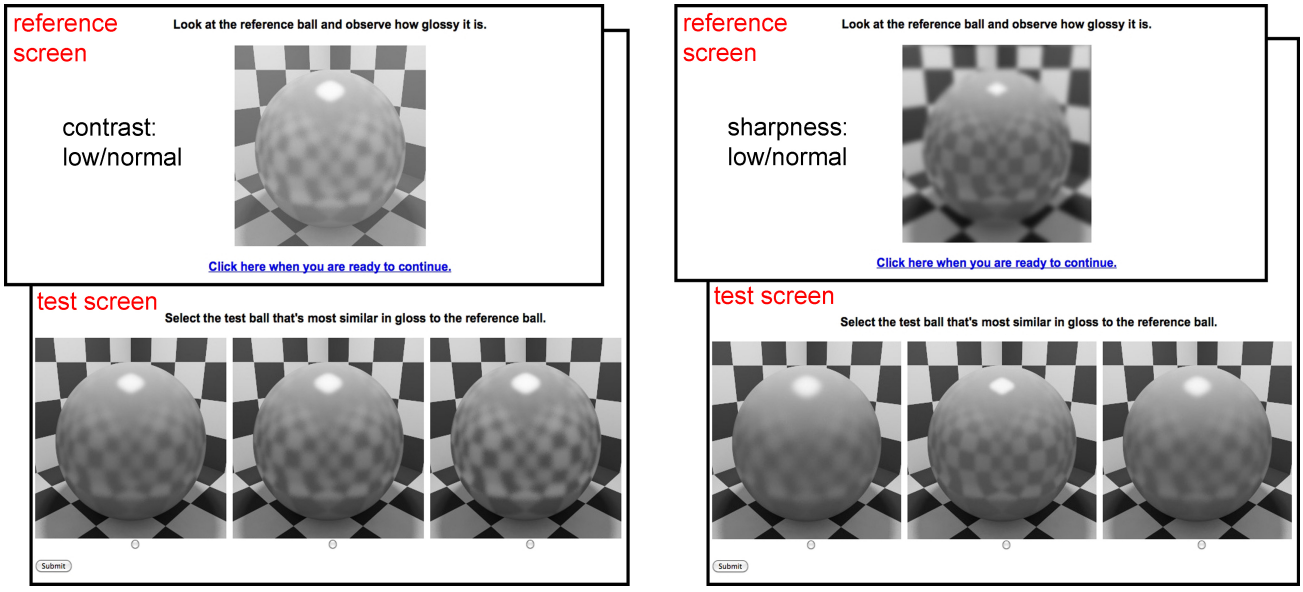

Figure 7. Sample trial screens from Experiment 2: sequential matching method. The left panel shows a trial from the "c" condition. The right panel shows a trial from the "d" condition. At the start of a trial the reference screen appeared observers viewed the object/image and then clicked the link. The reference screen then disappeared and the test screen appeared. The observers' task was to select the test ball with the same gloss as the reference ball. Both normal image vs. normal image and low quality vs. normal image trials were presented.

3.4 Experiment 2: Methods

The stimuli used in Experiment 2 were the same as those used in Experiment 1. Experiment 2 was also delivered using the Amazon Mechanical Turk system. The forty observers who participated in Experiment 2 were also recruited, screened, tested, and compensated in the same ways as those in Experiment 1, but they were not the same observers who participated in Experiment 1.

The interface used in Experiment 2 is shown in Figure 7. The left and right halves of the Figure show sample trials from the "c" and "d" conditions that were run as separate studies. In each study, each trial was composed of a pair of screens shown to the observer in sequence. At the start of a trial a reference screen appeared which showed one of the balls from the either the normal or low quality image sets in the upper half of the screen. Only the six most glossy balls in each set were used as reference balls for reasons that will be explained below. Observers were instructed to "Look at the reference ball and observe how glossy it is." then "Click (here) when you are ready to continue." On clicking, the reference screen disappeared and was replaced with a test screen that showed three balls selected from either the normal or low quality image sets. The gloss values of the test balls were: 1) the same as the reference ball; 2) two or three steps lower than the reference ball (presentations were split across observers and averaged); and 3) six steps lower than the reference ball. The left/center/right positions of the test ball images were randomized. The observers were instructed to "Select the test ball that's most similar in gloss to the reference ball." by clicking on one of the radio buttons beneath each image. The observers then clicked a "Submit" button and the next trial sequence began.

The reference and test ball gloss values were selected with the following rationale. The normal reference vs. normal test trials were created to establish a baseline for performance in the sequential matching task. The expectation was that

observers would most often select the test ball with the same gloss as the reference ball, but that there would be some variance, because in most cases the three test balls were just barely noticeably different in gloss. The low quality vs. normal trials were created to test the hypothesis that observers would be more accurate in judging the gloss properties of the balls if the images were presented sequentially rather than side-by-side. In these trials the test images present the observer with three choices. In one case they could choose a ball that has the same gloss properties as the reference ball (material match), in another case they could choose a ball whose image has the same contrast or sharpness as the reference ball (image match), and in the final case they could choose a ball/image with properties midway between the material and image matches.

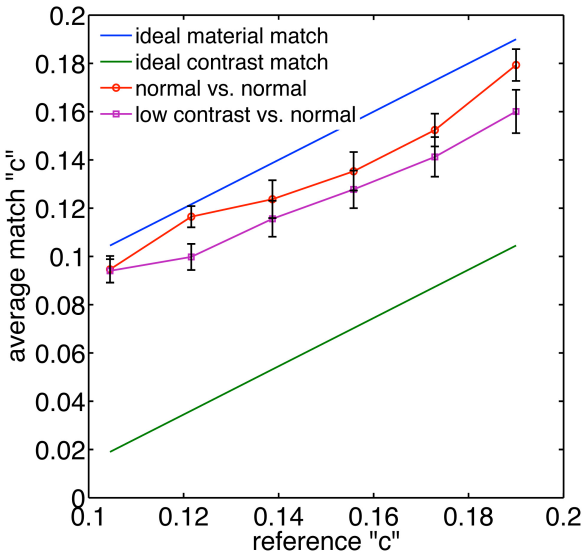

Figure 8. Results of Experiment 2: sequential matching method, "c" condition. Matching "c" values for objects with different reference "c" values are plotted as functions of image contrast. Note that both the normal image vs. normal image trials and low contrast image vs. normal image trials tend toward producing material matches rather than image contrast matches.

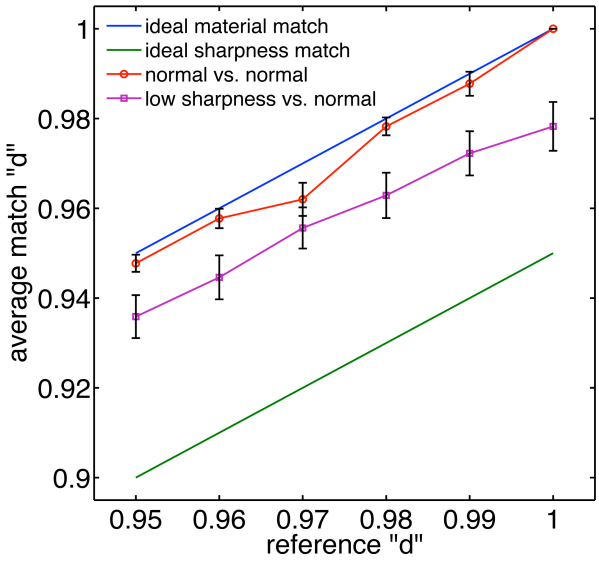

Figure 9. Results of Experiment 2: sequential matching method, "d" condition. Matching "d" values for objects with different reference "d" values are plotted as functions of image sharpness. Note that both the normal image vs. normal image trials and low sharpness image vs. normal image trials tend toward producing material matches rather than image sharpness matches.

3.5 Experiment 2: Results and Discussion

For each of the six reference balls tested, the frequencies with which the observers selected the material match, the image match or the mid-value match were tallied and averaged to estimate the perceived gloss of each reference ball in the normal and low image quality conditions. The results are summarized in Figures 8 and 9.

Figure 8 shows the results for the "c" condition. The graph plots the average matching "c" gloss as a function of the actual "c" gloss of the reference ball for the normal and low contrast images. As stated earlier, the reference balls were the six glossiest in the set ("c" range = 0.105 to 0.190). The "c" values of the test balls spanned the range from zero to six steps below the reference ball. To increase comprehension of the results two lines are plotted on the graph. The upper line indicates the matching value that would be obtained if observers made an ideal material match between the reference and test balls, the lower line indicates the matching value that would be obtained if observers made a perfect image contrast match. The upper curve shows the result for the normal reference vs. normal test trials. Note that while the average matching gloss values obtained are close to being material matches, in each case the average matching gloss is significantly lower than the ideal case. This result is not unexpected since the three test images spanned a small JND range and any error in selection would lower the resulting average. Still, the normal vs. normal curve indicates that in this baseline condition, observers strongly tend toward making material matches under the sequential matching method. The more interesting result is shown by the lower curve. This curve shows the average matching values obtained in the low contrast vs. normal trials. Here while the average match values are lower than in the normal vs. normal condition, in most cases the differences are not significant, and more importantly, the matching values are much closer to the material match line than the image contrast match line. As proposed earlier, this finding suggests that when observers are forced to judge gloss based on their internal representations of a stimulus through sequential presentation of the reference and

test images, rather than on direct side-by-side comparison, their judgments tend toward the veridical (matching object gloss rather than image contrast).

Figure 9 shows the results for the "d" condition. Similar to the results for the "c" condition, the average matches for the normal vs. normal trials are close to material matches and the matching values for the low sharpness vs. normal conditions while significantly below ideal material matches are still closer to this standard than to the image sharpness match standard. Similar to above this suggests that while observers are influenced by the properties of the stimulus images, when they are forced (or allowed) to form an internal representation of the reference stimulus, they tend to choose test images that match in terms of object features (gloss) rather than image features (sharpness).

4. CONCLUSION

In this paper we have described a series of experiments that explore the notion of image quality, focusing on the relationships between the signal properties of images and their role as visual representations. In particular we studied how well observers were able to perceive the gloss of objects represented by normal and low quality (low contrast, blurry) images. We found that different experimental methods produced differing results. Specifically, when the stimulus images were presented using simultaneous pair comparison, observers appeared to be influenced by the surface properties of the images and conflated changes in image contrast and sharpness with changes in object gloss. On the other hand when the stimulus images were presented sequentially, observers seemed able to disregard the image plane properties to some degree and more accurately match the gloss of the objects represented by the different quality images. As suggested in the introduction, when observers perform tasks where they base their judgments on the visual information they receive from images instead of the visual signals presented by the images, it is as if they are able to "see through" the limitations of the images and more faithfully perceive the object properties.

This work is very preliminary, and there is much more that can be done, especially with respect understanding other image distortions and their relationships to other properties of the objects, scenes, and events that are represented by images, but the findings suggest that in understanding image quality that there are useful distinctions to be made between the quality of the imaging medium and the quality of the visual information represented by that medium. The goal of this work is to understand the relationships between the signal properties of images and the fidelity of the visual information those images convey to human observers. Through this approach we should be able to learn more about how images work as visual representations, and to develop more meaningful image quality metrics that better predict how well images with different signal properties serve as visual representations of the things they depict.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation through award IIS-0113310 to the author.

REFERENCES

[1] Daly, S. (1993) The visible differences predictor: an algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, A. B. Watson, Ed. MIT Press, 179–206.

[2] Lubin, J. (1995) A visual discrimination model for image system design and evaluation. In *Visual Models for Target Detection and Recognition*, E. Peli, Ed., World Scientific Publishers, Singapore, 207–220.

[3] Wang, Z., Bovik, A. C., Sheikh H. R., and Simoncelli, E. P. (2004) Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612.

[4] Sheikh, H.R. and Bovik, A.C. (2006) Image information and visual quality. IEEE Transactions on Image Processing 15, 430–444.

[5] Ferwerda, J.A and Pellacini, F. (2003) Functional difference predictors (FDPs): measuring meaningful image differences. Asilomar Conference on Signals, Systems, and Computers, 1388-1392.

[6] Ramanarayanan, G., Ferwerda, J.A., Walter, B.J. and Bala, K. (2007) Visual Equivalence: towards a new standard for image fidelity. ACM Transactions on Graphics, 26(3), (SIGGRAPH '07), 1-11.

[7] Rouse, J.D., Pepion, R., Hemami, S.S., and Le Callet, P. (2009) Image utility assessment and a relationship with image quality assessment. Proceedings SPIE Electronic Imaging (Human Vision and Electronic Imaging XIV), 7420.

[8] Ward, G.J. (1992) Measuring and modeling anisotropic reflection. Proceedings SIGGRAPH ’92, 265-272.

[9] Pellacini, F., Ferwerda, J.A. and Greenberg, D.P. (2000) Toward a psychophysically-based light reflection model for image synthesis. Proceedings SIGGRAPH ’00, 55-64.

[10] Ward, G.J. (1994) The RADIANCE Lighting Simulation and Rendering System. Proceedings SIGGRAPH '94, 459-472.

[11] Gescheider, G.A. (1997) *Psychophysics: The Fundamentals.* 3rd edition, Lawrence Erlbaum and Assoc., New York.

[12] Amazon Mechanical Turk (MTurk) https://www.mturk.com

[13] Guilford, J.P. (1954) *Psychometric Methods*. Mcgraw-Hill, New York.

[14] Fairchild, M.D. (2013) *Color Appearance Models*, 3rd edition, John Wiley, New York.