Measuring the Perception of Visual Realism in Images

Paul Rademacher [†‡] Jed Lengyel [‡] Edward Cutrell [‡] Turner Whitted [‡]
[†] University of North Carolina at Chapel Hill [‡] Microsoft Research

Abstract. One of the main goals in realistic rendering is to generate images that are indistinguishable from photographs – but how do observers decide whether an image is photographic or computer-generated? If this perceptual process were understood, then rendering algorithms could be developed to directly target these cues. In this paper we introduce an experimental method for measuring the perception of visual realism in images, and present the results of a series of controlled human subject experiments. These experiments cover the following visual factors: shadow softness, surface smoothness, number of light sources, number of objects, and variety of object shapes. This technique can be used to either affirm or cast into doubt common assumptions about realistic rendering. The experiments can be performed using either photographs or computer-generated images. This work provides a first step towards objectively understanding why some images are perceived as photographs, while others as computer graphics.

1 INTRODUCTION

One of the goals in computer graphics research since its inception has been to generate computer images *indistinguishable from photographs*. The most realistic results emerge from special effects studios, which typically forego physically-accurate rendering methods, relying instead on the visual skills of their artists. These artists have a keen understanding of *how an image must look for it to be perceived as real*. They operate in a continual loop of generating images, evaluating them for realism, and then making adjustments as necessary. However, the average practitioner in computer graphics does not precisely understand what makes some images look photographic and others computer-generated, and is unable to create fully-realistic imagery.

If the perceptual criteria by which people evaluate whether an image is real were understood, then one could build new rendering algorithms to directly target the necessary visual cues. Furthermore, one could optimize the rendering budget towards those visual factors that have the greatest impact, without wasting effort on elements that will not noticeably improve the realism of an image.

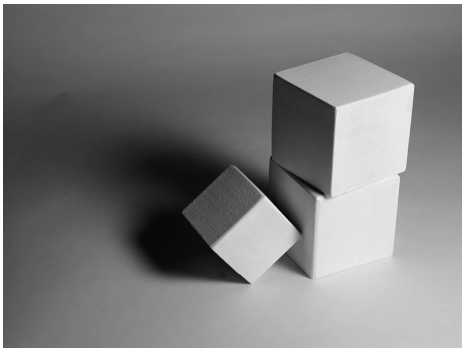

Figure 1. Is this a photograph or CG? What visual factors affect your decision?

In this paper we demonstrate that the perception of visual realism in images can be studied using techniques from experimental psychology. We present an experimental method that directly asks participants whether an image is real (photographic) or not real (CG). We conducted experiments to explore several visual factors, including shadow softness, surface smoothness, number of objects, variety of object shapes, and number of light sources. The resulting data confirmed some common assumptions about realistic rendering, and contradicted others. We found that while shadow softness and surface smoothness played a significant role in determining an image's perceived realism, increasing the number of light sources did not. Also, increasing the number of objects in a scene, or the variety of object shapes, did not increase an image's likelihood to be perceived as real / photographic. Our method can be conducted using exclusively photographs, or using exclusively computer-generated images. We ran duplicate experiments using both photographs and computer-generated images, with similar and consistent results between the two.

2 PREVIOUS WORK

There have been several approaches to the creation of realistic images. One is to analyze, measure, approximate, and simulate the various physical processes that form a real-world image (light transport, surface BRDFs, tone mapping, and more). This approach has met with only limited success. Furthermore, the few projects that *have* created images indistinguishable from specific target photographs (such as the Cornell Box [Patt97]) do not reveal which visual factors a viewer expects in order to perceive the image as real.

Another approach to realism is image-based rendering [Leng98], which has created synthetic images which are nearly indistinguishable from photographs. This is to be expected, of course, since image-based rendering works by rearranging image samples taken directly from photographs. What was it about the original photographs that made them realistic in the first place?

There exists an enormous amount of previous work on human vision and classical perception. [Bruc96] and [Gord97] are good introductions. Work in these fields includes low-level vision, classical psychophysics, object recognition, scene understanding [Hage80], and more. However, the direct question of how people distinguish photographs from computed-generated images has not been raised in the classical perceptual literature. Indirectly, works such as [Parr00], which analyzes the approximate $1/f$ frequency spectrum of natural images, provide clues as to the nature of real-world imagery. An informal essay on how visual realism in computer graphics was given by [Chiu94]. [Barb92] describes limitations of display technologies when trying to simulate *direct* vision.

There are several recent works in the computer graphics literature dealing with human visual perception (e.g., applications of psychophysics in [Ferw98], [Vole00], [Rama99]). [Chal00] describes various image quality metrics. [Rush95] proposed perceptually-based image metrics to differentiate between a pair of images, in order to evaluate the accuracy of synthetic renderings of real-world scenes. In [Thom98], shadows and other visual cues are tested against subjects' ability to discern properties such as object orientation or proximity. [Horv97] measures subjects' response to various settings of image quality, to guide an efficient renderer. [Mcna00] compares computer-generated images with real, physical scenes (viewed directly) to evaluate the perceptual fidelity of the renderings, in a manner similar to [Meye86]. None of these perceptually-based research efforts directly studied the visual causes for the perception of some images as photographic and others as synthetic.

We conclude that while many areas of classical perception, realistic rendering, and perceptually-based rendering have been thoroughly studied, the central problem of determining what about an image tells a person whether it is photographic or computer-generated remains largely unexplored.

3 DESIGNING A PERCEPTUAL TEST OF VISUAL REALISM

The goal of this project is to study the perception of realism in images using techniques adapted from classical human visual perception. The ultimate goal of this line of research is to gain a full understanding of exactly what cues tell observers that an image is photographic or computer-generated, so that rendering algorithms can be built to directly target these cues. While this end result is still far away, our intent in this work is to frame the problem in perceptual terms, and to develop methods by which observers can be objectively tested, and meaningful analyses performed.

3.1 Experimental question and task

Our strategy for finding out what visual cues matter for realism is to ask experimental participants to directly rate a series of controlled images as either "real" or "not real" – but how do we communicate to the participants what we mean by "real"?

One of the difficulties is that it would appear that subjects need a clear definition of what is meant by "realism" in order to properly perform the experiment and not yield invalid data. Yet the reason these experiments are being conducted in the first place is because we *do not* have a clear definition of what makes an image realistic – we want *them* to tell *us* what makes an image look real. The more we tell subjects about our notion of realism or the context under which we are studying it, the more their responses will be biased towards what they are told.

Our solution is to give the participants minimal instructions: they are only told that some of the images are "photographic / real" while others are "computer-generated / not real," and that their job is to differentiate between the two (see Appendix for full written instructions). We therefore present the context of photographs versus CGI, but offer no guidance on how to actually evaluate the two. It is the subjects' job to interpret these keywords and respond accordingly (thereby providing an *operational definition* [Levi94] of realism).

Although one might worry that the variability inherent in these sparse instructions could lead to invalid results, the exact purpose of this experimental method is to see whether different participants converge to similar responses, given only a few keywords.

Another potential concern is whether the subjects' responses are more a reflection of the forced-choice nature of the experiment, rather than of the perceived realism of the images. The design of the experiment does force participants to make a choice, but if a given visual factor has no effect whatsoever on a participant's interpretation of "real / not real," then the resulting responses will be completely uncorrelated with the factor levels. If there *is* a correlation, and one which holds across the majority of participants, then the analysis will yield a statistically-significant result, and we can claim that the visual factor does influence subjects' interpretation of "real" versus "not real." In our experiments we found that some visual factors did correlate with subjects' responses (they measurably influenced subjects' interpretation of "real") while other visual factors did not.

3.2 Controlled image factors

We use a common experimental technique in which subjects are presented with sets of controlled images which vary only along some predetermined dimensions, with all other image factors held as constant as possible [Levi94]. We then analyze the pattern of responses across these dimensions of interest. If there is a statistically-significant change in the response, then we can claim the existence of a *causal relationship* between the visual factor and the subjects' responses (since the images are controlled against extraneous factors).

It is important to note that because of this design, we actually do not mix photographs and computer-generated images in a single experiment. If they were mixed, then unless the CG images exactly matched the corresponding photographs, there would be confounding factors which would interfere with the analysis. For a single experiment, therefore, *the images should be either all photographs or all computer-generated*. That is, they should be from the same source. A consequence of this is that these experiments actually have no notion of "correctness" – it is not appropriate to think of the responses as hits, misses, false positives, false negatives, etc. It only matters how the subjects' response pattern changes across the dimensions of interest.

3.3 Types of images

The images in these experiments consisted of very simple scenes, containing only blocks, spheres, and egg-shapes, in grayscale and without motion. We chose this approach to limit the number of factors to contend with in these early studies. An initial concern when using simple scenes was that the simplicity itself might cause a strong sense of *un*realism, which could obscure any true effect of other visual factors. This proved not to be the case, as we did observe statistically-significant effects based on certain visual factors. The issue of scene simplicity is further addressed in Section 6, where we present a series of experiments on number of objects, variety of object shapes, and number of light sources.

3.4 Experiment methodology

All the experiments followed the same format. A series of images was presented to each subject, who rated each as either "real" or "not real." The images were all of simple objects. They varied according to some visual factors under investigation – either shadow softness, surface smoothness, number of objects, variety of object shapes, or number of lights. For example, in the first experiment the shadows were at one of five possible levels, ranging from very sharp to very soft.

The experimental method asks for only a binary "real / not real" response (rather than a multi-point scale) to simplify the task for each subject (who only needs to maintain a single internal differentiation threshold), and to reduce problems of scale interpretation across subjects.

The proportion of "real" responses for a particular level of a factor is the ***realism response rating*** for that level (which we denote by $\Re$). If we assign the numerical value of one to "real," and zero to "not real," then the $\Re$ value is simply the mean of all numerical responses for a given level. For example, if 37 out of 60 images at the sharpest shadow level were rated as real, then we say that $\Re$ = .62 for sharp shadows. In the analysis stage, we infer the effect of the various visual factors on realism by testing for statistically-significant changes in $\Re$.

All image presentation and data collection was automated, and the order of presentation was fully randomized for each subject at run-time. Subjects ran all their image trials in one sitting (with short breaks). The average completion time was 1¼ hours.

The 21" monitors were set to 1152×864, and each image was 800×600. Subjects sat approximately two feet from the screen, giving a subtended viewing angle of the images of approximately 30 degrees. The experiments were all conducted under controlled illumination.

Subjects all gave informed consent, and were naïve to the study's purpose, non-experts in computer graphics or related visual fields, aged 20 to 50, with normal or corrected-to-normal vision.

3.5 Creating the images

For the photographic experiments, we acquired images with an Olympus 3030Z digital camera, at 800×600. The green channel (least noisy) was used to create a grayscale image. The camera was locked into place for all the images. The objects were wooden cubes and spheres (5 centimeters in height), and 7-cm wooden egg-shapes. They were all painted with white acrylic paint. In all the images, the objects are set against a large draped sheet of white paper.

For the CG experiments, we used 3D Studio Max, with raytraced soft shadows. The CG experiments used only blocks (no spheres or egg-shapes), and the texture maps were acquired by orthographically photographing our physical wooden blocks, and normalizing the resulting textures. The background texture map was taken from a photograph of our physical stage. No indirect illumination was computed. Since the CG images were all batch-rendered from the same dataset, the CG version of the experiments had very precise experimental control.

To reduce the dependence on any single spatial arrangement of objects (position and orientation), we used several spatial arrangements in each experiment. For example, in the shadow softness experiment, we placed the objects in a given spatial layout, then gathered the images at each of the five shadow levels (keeping the positions and orientations *constant across shadow levels* for each "scene"). We then rearranged the objects and gathered images again at each shadow level, and so on. Since the "scenes" are orthogonal to the main visual factors under investigations, they do not in any way confound the analysis of the factors, but only reduces bias towards any single spatial arrangement.

The images were all generated with the light source on the right side. As the experiments ran, images were randomly flipped horizontally, so that half of them appeared to have the light source on the right side, and half on the left side (chosen randomly at run-time for each image presentation). This was done to reduce bias towards a particular light direction, and to reduce fatigue on the participants by increasing the image variety. Since the two image directions were evenly and randomly distributed, they have no impact on the analysis (they cancel out).

3.6 Analysis method

The appropriate method of statistical analysis was dictated by two design elements. First, because the response variable was binary, standard linear regression models or analysis of variance (ANOVA) are not appropriate (they are only valid on continuous data from normal distributions). Instead, the correct analysis is *logistic regression* [Wine91], an extension of linear regression for binary data. Logistic regression computes the correlation between a manipulated factor (e.g., level of shadow softness) and a binary response variable ("real" vs. "not real").

Second, because each subject ran many trials (and the responses are therefore not all independent), a *repeated measures* analysis [Wine91] was called for, to take into account the correlation between responses by the same subject. We used the commercial statistics package SUDAAN [Shah96], which handles repeated measures logistic regression designs.

A concern when subjects run many trials is that time-dependent/training effects may emerge. That is, as the experiment progresses, responses could begin to drift towards one end of the response scale. We tested for this by computing the regression between trial number and subjects' responses, and found no presence of time-dependent/training effects.

The subjects' response times were also measured. An analysis showed no correlation between the subjects' response times and their response values. This indicates that subjects did not respond any faster to images they rated as real than to those rated as not real.

3.7 Additional experimental details

A blank gray screen was displayed for approximately one second between images. Subjects chose between "real" and "not real" by pressing one of two keys. They were free to change their responses (visual feedback was given) and they confirmed their current response and advanced to the next image by pressing the spacebar.

To prevent regression to the mean – where responses degenerate as the experiment progresses due to the lack of a fixed reference point – the images were presented in groups of eight. In a first pass the images in each group were only previewed, and in a second pass the subjects actually rated them. This provided an internal reference point for subjects throughout each experiment. At the start of each experiment, a number of images (sixteen) were presented, to allow the subject to become acquainted with the experiment. The responses for these were excluded from analysis.

4 SHADOW SOFTNESS

In this first experiment, we were interested in whether subjects' realism response would change significantly as a result of varying the shadow softness. It is typically taken for granted that very sharp shadows are seen as unrealistic, yet not much is known about how realism is affected when shadows are not perfectly sharp. For example, if softening a shadow makes an image more realistic, does softening it twice as much double the increase in realism?

4.1 Setup: Shadow Softness

There were five levels of shadow softness. The lowest (sharpest) level was created with a focused 300W spotlight, at 2.3 meters from the scene. The next two levels were created with a clear incandescent 300W light bulb, at 2.0 and 1.0 meters from the scene, respectively. The last (softest) two shadow levels were created with the same light bulb, but now diffused, at 1.0 and .2 meters. The resulting penumbral spread angles were .39°, 1.5°, 2.5°, 5.2°, and 10.3°. Close-ups of some shadows from this experiment are shown below. Note that the images are nearly identical except for the shadows.

There were twelve scenes (different spatial arrangements), and each scene was photographed at each of the five shadow levels. That is, there were 12 sets of 5 images, where the five within each set *were nearly-identical except for their shadow softness*. The total number of images presented to each subject was therefore $12 \times 5 = 60$ images.

Because the different shadow levels were generated using lights at different distances, the images varied slightly in brightness and contrast. They were manually adjusted to account for any obvious exposure differences. The remaining differences were small and randomly distributed, and therefore should not affect the analysis. This slight loss of experimental control when using photographs is one of the motivations for performing experiments using computer-generated images (as described in Section 7), which offer precise experimental control.

We measured the penumbra angles for the shadows in all of the images, and averaged these to get a single penumbra angle measurement for each of the five shadow levels, as shown below.

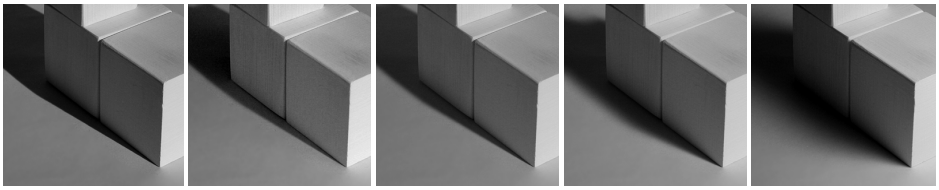

Figure 2. Detail of images from shadow softness experiment. Average penumbra angles for the five shadow levels were .39°, 1.5°, 2.5°, 5.2°, and 10.3°.

4.2 Results: ℜ vs. Shadow Softness

The experiment was run with 18 subjects. The graph shows ℜ vs. shadow softness (the proportion of "real" responses for each shadow level). The error bars show the inter-subject variability in ℜ – i.e., the standard error of the set of ℜ values, one from each subject, for the given shadow level.

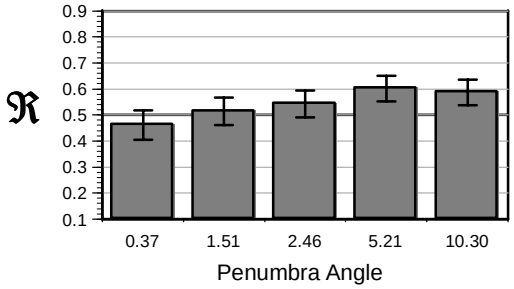

Figure 3. ℜ vs. shadow softness for photographic experiment. (*note:* the x-axis is not evenly scaled). The increase in ℜ rating becomes statistically significant when the shadow penumbra reaches 5.21 degrees. There is no statistical difference between the last two levels of shadow softness.

The first question we ask is whether the subjects' responses varied significantly due to shadow softness. To test this, we fit a repeated measures logistic regression model to the data, using shadow softness as the independent variable, and the binary "Real / Not Real" response as the dependent variable. Shadow softness was found to be a statistically-significant predictor of realism ($\chi^2 = 4.31$, df = 1, p = .0379)[1]. Indeed, the subjects' reported visual realism varied as a result of shadow softness.

Clearly the sharpest shadows (leftmost level) were rated the lowest in realism. This agrees with the common notion in computer graphics that sharp shadows are unrealistic. By performing pair-wise comparisons in a repeated measures logistic regression analysis between the sharpest shadow level and each of the four remaining shadow levels, we found that a statistically significant difference was found beginning at the 4th shadow level (5.21 degrees penumbra). The test was ($\chi^2 = 5.39$, df = 1, p = .0203). This indicates that at 5.21 degrees, we begin to see a measurable change in reported realism. Furthermore, there is no statistical difference between the last two (softest) shadow levels ($\chi^2 = 2.64$, df = 1, p = .1043). From all this we can conclude that in this set of images, perceived realism was maximized with respect to shadow softness in the neighborhood of 5.21 degrees of penumbra angle. Any additional increase in softness had

[1] A p-value of .05 or less denotes a statistically significant effect.

diminishing returns. Rendering soft shadows is an expensive computation, so knowing how people will respond to various shadow qualities can result in significant savings during rendering.

5 SURFACE SMOOTHNESS

It is often said in the computer graphics folklore that for an image to look realistic, "surfaces should not be too smooth." Roughing up the surfaces, for example, was one of the major efforts in the creation of Toy Story [Stre95]. Certainly, with computer graphics it is easy to create surfaces with no surface variation whatsoever – something unlikely to be encountered in real life. Nevertheless, in the real world we *do* find objects with all degrees of surface smoothness. A freshly-painted wall is much smoother than a cork bulletin board, for example – but is a smooth *real-world surface* really seen as less realistic than a rough real-world surface? In this experiment we tested this by comparing the realism response for photographs of smooth-textured objects versus photographs of rough-textured objects.

5.1 Setup: Surface Smoothness

We presented subjects with a series of photographs, where half the images contained smooth-textured cubical blocks, and the other half contained rough-textured blocks. The smooth textures were created by painting the cubes with white spray-paint, which gave a smooth, even coat. The rough blocks were created by painting them white with a rough-bristled brush, which yielded strongly-noticeable brush marks on the surface.

There were thirty scenes, with each scene in both rough-texture and smooth-texture form. The total number of images presented to each subject was therefore $30 \times 2 = 60$ images.

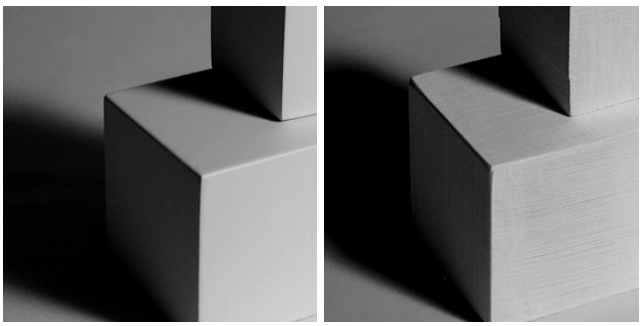

Figure 4. Detail of two images from surface smoothness experiment. The smooth, spray-painted blocks, such as on the left, rated much lower in realism ($\Re$ = .39) than the rough, brush-painted blocks ($\Re$ = .71)

5.2 Results: $\Re$ vs. Surface Smoothness

This experiment was run on 18 subjects. We found that there was a very strong difference in realism for the two types of surfaces. As shown in the graph, the rough-painted blocks rated much higher than the spray-painted ones ($\Re$ = .71 vs. $\Re$ = .39). This effect was stronger than the effect due to shadow softness.

We tested for statistical significance using surface type as the independent variable, and the "real / not real" response as the binary dependent variable. The effect was strongly statistically significant ($\chi^2 = 13.04$, df = 1, p = .0003). This indicates that the smoothness of surface textures was undoubtedly a determinant of realism – which backs up the common graphics folklore that says that surfaces should not be "too smooth."

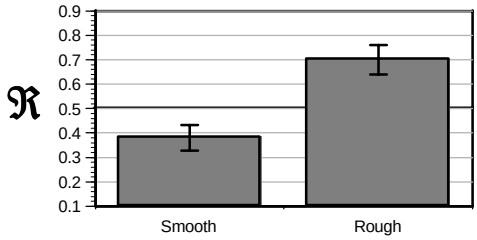

Figure 2. $\Re$ vs. surface smoothness. There was a strong effect due to surface type.

It is worthwhile to point again that this experiment was conducted using *only photographs*. When presented with the question of whether the images were real (photographic), the smooth textures ranked low, even though they were, in fact, physically-real surfaces. This has implications for areas of rendering such as global illumination research, where untextured objects are typically used to report results. This experiment suggests that these untextured images may never look highly realistic – no matter how good the lighting algorithm. As one critiques the results of an advanced lighting algorithm, it is therefore worth remembering that if the surfaces are untextured, then this alone may cause a much stronger decrease in realism than any error in the light transport computation.

As a final point, this experiment only demonstrates that there was a difference in realism between the two surface types, but does not characterize what it was about the rougher surface that made it look more real. As seen in the fields of texture synthesis and BRDF measurement, there are many ways to analyze the properties of surfaces, and it remains as future work to discover exactly which of these are important.

6 NUMBER OF OBJECTS, VARIETY OF OBJECT SHAPES, AND NUMBER OF LIGHT SOURCES

In this set of experiments we looked at what happens to the reported realism as we manipulated three factors: the number of objects in the scene, the variety of object shapes, and the number of light sources. One might expect and assume that the subjects' responses would increase as more objects are added to a scene, the types of objects varied, or the number of light sources increased. But since these increases consume more memory and rendering time, it would be useful to first verify what effect these increases will have on the realism of an image.

6.1 Setup: Number of Objects / Variety of Object Shapes

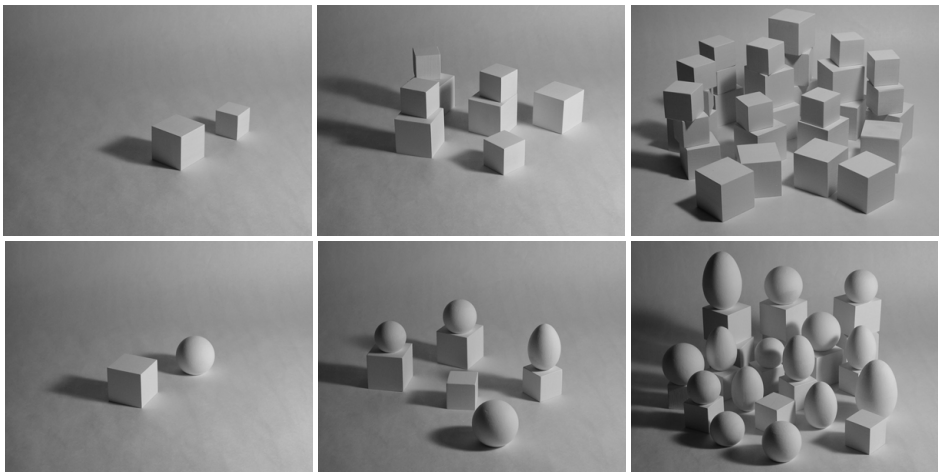

Figure 3. The horizontal axis increases the number of objects, and the vertical axis varies the object shapes (blocks-only above, versus blocks, spheres, and egg-shapes below). There was no statistically-significant difference in perceived realism along either axis.

We tested the effect of increasing the number and types of objects in a scene with a single two-factor experiment. The first factor was the number of objects – each image either contained 2, 4, 8, or 30 objects. The second factor was the variety of object shapes, with two levels: each image consisted either of only cubical blocks, or of half blocks and half curved objects (spheres and egg-shapes). For example, an image might have 8 objects which are all blocks, or it might

have 30 objects with mixed shapes (15 blocks and 15 spheres and egg-shapes). Crossing the two factors yields 4×2 = 8 images. Subjects were shown five different sets of images, each set fully representing the crossed factors (for a total of 4×2×5 = 40 images).

6.2 Setup: Number of Light Sources

Before presenting the results of the previous setup, we describe the setup for the experiment on the number of light sources. There were three levels for the main factor: one light, two lights, and four lights. To create images with accurate exposure and light source control, we radiometrically blended photographs containing a single light each.

The same scene was repeatedly photographed, each time with a single light source placed at four different locations along a 120° arc around the scene. Then, to generate a new image with two light sources, for example, our custom image-assembly utility randomly picked two light source positions, and blended these images to create a single new image that appears to be lit by two lights. The camera was locked into place and operated via remote control to eliminate any camera shake, so that the images would blend well. Also, the aperture and exposure settings were locked across all the original images.

The blend operation was radiometrically correct. We first computed our digital camera's CCD response curve using the ***mkhdr*** software tool ([Diuk98], based on [Debe97]). We then mapped each image into radiometric space (mapped from camera pixel intensities to irradiance), summed in that space (simulating the additive nature of light), adjusted the exposure (multiplied the summed image by either 1/2 or 1/4, depending on the number of lights), and then mapped from radiometric space back to camera space to yield the final image.

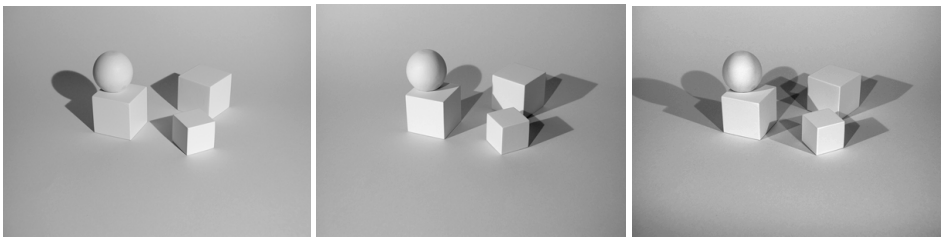

Figure 4. Images from experiment on number of light sources.

Because blending images decreases camera noise, we actually acquired *four* photographs from each of the four light source positions (i.e., 4×4 = 16 photographs per scene). The final images were all created by blending exactly four images out of sixteen (e.g., blending four photographs with the same light position to create an image with "one" light), so they all had the same level of camera noise present.

Note that it is not possible to keep all other factors absolutely constant when we increase the number of light sources. The light on each surface will change, the overall contrast will diminish, and so forth. However, these are all physically-dictated byproducts of increasing the number of lights (the primary factor under investigation), and are accepted since they have a small visual effect compared to the very distinct increase in number of shadows.

Finally, in addition to number of light sources, we also co-varied the shadow softness, to reduce the bias on any particular shadow type. The above process was repeated for each scene, once with a spotlight and once with a diffuse light source.

There were 6 scenes, 2 shadow types per scene (soft and sharp), and 3 numbers of lights per shadow type (1 light, 2 lights, or 4 lights). Thus, this experiment consisted of 6×2×3 = 36 images.

6.3 Results: ℜ vs. Number of Objects / Variety of Object Shapes / Number of Lights

Ten subjects ran the experiment on number of objects and variety of object shapes, and seven subjects ran the experiment on number of light sources. One can immediately see in the graphs below that the realism response did not increase due to either number of objects, variety of object shapes, or number of lights. In fact, the graphs appear to indicate that there was actually a *decrease* in reported realism when the number of objects and the number of lights was increased.

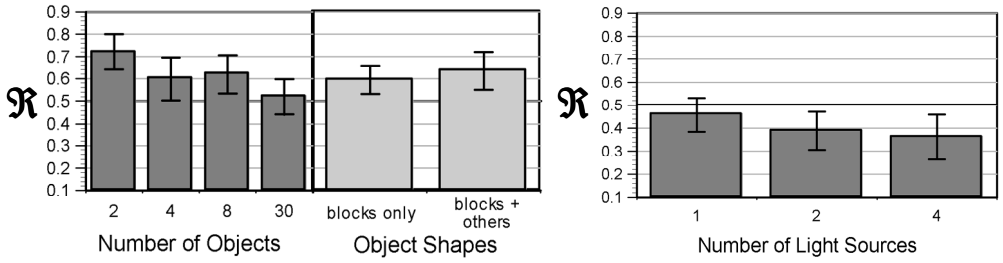

Figure 5. ℜ vs. Number of objects / Variety of object shapes / Number of lights. None of the effects were statistically significant.

We tested for significance in all three cases. For number of light sources, no significant effect was found (χ^2 = .56, df = 1, p =.4546). For variety of object shapes (blocks-only versus blocks, spheres, and egg-shapes), there was also no significant effect (χ^2 = .58, df = 1, p =.4454).

For the number of objects, the results varied depending on how the analysis is performed. If we perform the analysis using the *level number* as the independent variable (taking on the values 1, 2, 3, and 4), then we see a borderline-significant effect (χ^2 = 3.55, df = 1, p = .0597). This is because, as we can see in the graph, the response at the first level is indeed higher than at the last. If, however, we perform the analysis using the actual *number of objects* as the independent variable (taking on the values 2, 4, 8, and 30), then the regression is *not* considered significant (χ^2 = 2.43, df = 1, p = .1193). The interpretation of these two results is that while the low-endpoint case (only two objects) indeed rated higher than the rest, the overall effect is not significant when the large scale of the axis is considered (i.e., there was no significant difference between four objects and thirty objects).

These were unexpected results. Despite the tremendous visual difference between images with only four objects and images with thirty, subjects did not respond any differently to them. Furthermore, subjects were no more convinced by an image with several light sources and shadows than they were by an image with only one, nor were they any more convinced by images that showed a variety of objects types rather than only blocks.

These results have implications for computer graphics rendering. For example, if an image of a simple scene (such as those often found in conference proceedings) appears unrealistic, it is not necessarily because of its simplicity. There may be other factors which are causing the low realism (e.g., sharp shadows or "too smooth" textures), which should be addressed first. Furthermore, these results suggest that in a rendering application, it may be better to spend time on generating proper soft shadows and adequate textures, rather than adding more of the same lights or objects, or simply adding new objects for variety.

6.4 Ramifications of negative results

These negative, non-significant results have important implications for our experimental technique.

One concern before these experiments were conducted was that the subjects might be able to simply "decode" the experimental factors under investigation. For example, if they notice that the only difference between images is the shadow softness, then they may simply give every sharp-

shadowed image one response, and every soft-shadowed image the opposite response. We would still be able to learn something from this, since at least we would know which end of the softness spectrum they considered "real" and which "not real," and what they considered to be the boundary point. However, this decoding is still not ideal, since we want to learn about subjects' true internal perception of the images, and we want each image to be evaluated fairly.

However, because we have negative, non-significant results for these previous three experiments – despite the strong, obvious visual differences in the images – we have evidence to support the claim that subjects were *not*, in fact, simply decoding or "figuring out" the experimental factors, but rather were responding with a true measure of their perception of realism for each image. Otherwise, we would have seen significant changes in $\Re$ for the previous three experiments, just as we did for shadow softness and surface smoothness.

7 EXPERIMENTS USING COMPUTER GENERATED IMAGES

All the experiments presented thus far have employed photographs exclusively. As explained in Section 3.3, it is not important where the images come from, as long as they only differ along a particular dimension of interest, with all other visual factors held as nearly constant as possible. However, we can clearly achieve a higher level of control using computer-generated images than using photographs. Furthermore, with CGI we can easily manipulate certain dimensions that would be difficult to do with photographs (e.g., secondary illumination).

It would be useful, therefore, to know whether our experimental methodology is valid in the CG case. If, for example, we found that the results from some all-CG experiments, mimicking the photographic experiments above, yielded only responses of $\Re = 0$ (all images rated as "not real"), or had response curves that were qualitatively different than the curves for the photographic cases, then we would lose confidence in the robustness of the experimental method. To test this, we replicated the shadow softness and surface texture experiments using CG images exclusively. We hoped to find the response curves to be similar to those from the photographic cases, allowing for differences in scaling, offset, and noise.

We rendered images using 3D Studio Max, with raytraced soft shadows, and object textures extracted from intensity-normalized orthographic photographs of the wooden cubes. Seven subjects ran the computer-generated experiments. These were different subjects from those that ran the photographic experiments, so there was no crossover effect between the two types of images.

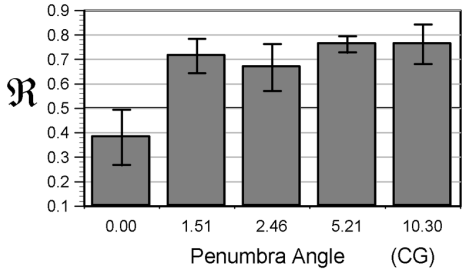

Figure 9. $\Re$ vs. shadow softness, for CG images. Note the sharp increase between the first two levels. This may be because with CG we could achieve a perfect point light source (smaller than our physical spotlight).

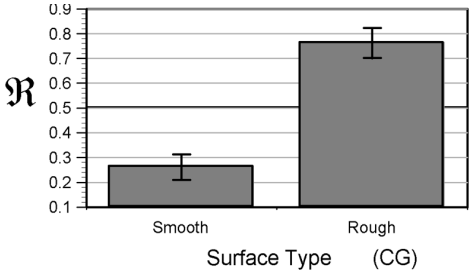

Figure 10. $\Re$ vs. surface texture, for CG images. We used textures obtained from photographs of our wooden blocks. The CG results match the earlier photo-based results closely.

By comparing the CG graphs to their photographic counterparts from the previous sections, we see that the computer-generated version of these experiments yields qualitatively similar data. In the surface texture experiment, the smooth texture is still much lower in $\Re$ than the rough texture. In the shadow softness case, the $\Re$ curve ascends as it did with the photographs. There is

a difference here compared to the photographic case, however, in that the jump between the sharpest and the second-sharpest shadows is much more pronounced for the CG case than for the photographic case. This may be due to the fact that in the CG renderer we were able to create a true point light source, and so the sharpest shadow level in the CG case actually causes a sharper, much more "unrealistic" penumbra than with the photographic spotlight.

We now perform tests for significance, by applying the same repeated measures logistic regression analysis as before. The test yielded significance for both the computer-generated shadow softness experiment ($\chi^2 = 4.92$, df = 1, p=.0265) and for the computer-generated surface smoothness experiment ($\chi^2 = 20.51$, df = 1, $p < .0001$).

Since the all-CG experiments yielded qualitatively-similar data to the photographic experiments, and were statistically significant (as were the photographic experiments), we claim that our experimental methodology indeed yields valid results using only computer-generated images. For completeness, the remaining three photographic experiments should also be replicated in CG form – we leave this for future work.

Of course, if the rendered CG images had some artifacts that were extremely fake-looking, this could have pulled the response curve down to zero, and the effect of the variable under investigation would have been lost. Nonetheless, by having the option of running experiments using only CG images, we open up the possibility of much more complicated experiments than what could be done with photographs – e.g., investigations of global illumination, BRDF models, tessellation / simplification techniques, and more.

8 CONCLUSION

A crucial component for the creation of realistic imagery is an objective understanding of the perceptual criteria by which viewers decide if images are real or not. While much research has been invested into physics-based rendering, the experiments presented in this paper have shown that even *photographs* (which are, by definition, "photo-real") are not all equally realistic. Physics, therefore, is not the only key to realism. Once the graphics community understands how different visual factors determine whether an observer perceives an image as photographic, then new rendering algorithms can be developed to specifically target these visual cues.

In this paper we have presented an early step towards understanding this perceptual process, with an experimental technique that directly asks subjects about the realism of images. The method was shown to be capable of affirming common assumption in graphics, of providing quantitative data , and also of casting into doubt certain common notions about realistic rendering. Furthermore, these experiments can be conducted using either photographs or computer generated images, which greatly expands the range of hypotheses that can be tested.

As more visual factors are investigated using this experimental method and future techniques for measuring the perception of realism in images, we will eventually have a full understanding of what it really means for an image to look like a photograph or to look like computer graphics.

REFERENCES

[Barb92] Christopher Barbour and Gary Meyer. Visual Cues and Pictorial Limitations for Computer Generated Photorealistic Images. In *The Visual Computer*, vol 9, pp. 151-165. 1992.

[Bruc96] Vicki Bruce, Patrick Green, and Mark Georgeson. Visual Perception: Physiology, Psychology, and Ecology. East Sussex, UK, 1996. Psychology Press

[Chal00] Alan Chalmers, Scott Daly, Ann McNamara, Karol Myszkowski, and Tom Troscianko. Image Quality Metrics. *SIGGRAPH 2000 Course Notes #14*. July, 2000. ACM.

[Chiu94] Kenneth Chiu and Peter Shirley. Rendering, Complexity, and Perception. In *Proc of the 5th Eurographics Rendering Workshop*. SpringerWien, New York, NY. 1994.

[Debe97] Paul E. Debevec and Jitendra Malik. Recovering High Dynamic Range Radiance Maps from Photographs. In *Proc of SIGGRAPH 97*, August 1997. ACM.

[Diuk98] H.P Duiker, Tim Hawkins, and Paul Debevec. Mkhdr. *www.debevec.org/FiatLux/mkhdr*

[Ferw98] James Ferwerda. Visual Models for Realistic Image Synthesis. Ph.D. thesis, Cornell, 1998.

[Gord97] Ian Gordon. Theories of Visual Perception. John Wiley & Sons, New York, NY, 1997.

[Hage80] Margaret Hagen. The Perception of Pictures. New York, 1980. Academic Press.

[Horv97] Eric Horvitz and Jed Lengyel. Perception, Attention, and Resources: A Decision-Theoretic Approach to Graphics Rendering. In *Proc of Thirteenth Conf on Uncertainty in AI*, pp. 238-249. Providence, 1997.

[Leng98] Jed Lengyel. The Convergence of Graphics and Vision. In *IEEE Computer*, July 1998.

[Levi94] Gustave Levine and Stanley Parkinson. Experimental Methods in Psychology. Hillsdale, New Jersey, 1994. Lawrence Erlbaum Associates.

[Mcna00] Ann McNamara, Alan Chalmers, Tom Troscianko, and Iain Gilchrist. Comparing Real & Synthetic Scenes using Human Judgement of Lightness. In *Proc of Eurographics Workshop on Rendering*. Springer-Verlag 2000.

[Meye86] Gary Meyer, Holly Rushmeier, Michael Cohen, Donald Greenberg, and Kenneth Torrance. An Experimental Evaluation of Computer Graphics Imagery. In *Transactions on Graphics*, *5 (1)*, pp. 30-50. New York, 1986. ACM.

[Parr00] Alejandro Parraga, Tom Troscianko, David Tolhurst. The Human Visual System Is Optimized For Processing The Spatial Information In Natural Visual Images. In *Current Biology, 10*, pp. 35-38. 2000.

[Patt97] Sumanta Pattanaik, James Ferwerda, Kenneth Torrance, and Donald Greenberg. Validation of Global Illumination Solutions Through CCD Camera Measurements. In *Proc of 5th Color Imaging Conf, Soc for Imaging Sci and Tech*, pp. 250-253, 1997.

[Rama99] Mahesh Ramasubramanian, Sumanta Pattanaik, and Donald Greenberg. A Perceptually Based Physical Error Metric for Realistic Image Synthesis. In *Proc of SIGGRAPH 99*. New York, 1999. ACM.

[Rush95] Holly Rushmeier, G. Larson, C. Piatko, P. Sanders, and B. Rust. Comparing Real and Synthetic Images: Some Ideas About Metrics. In *Proc of Eurographics Rendering Workshop 1995*. SpringerWien, New York, NY. 1995

[Shah96] Babubhai Shah, Beth Barnwell, and Gayle Bieler. SUDAAN User's Manual, Release 7. Research Triangle Institute, RTP, NC.

[Stre95] Rita Street. Toys Will be Toys: Toy Story. Cinefex, issue 64. 1995.

[Thom98] William Thompson, Peter Shirley, Brian Smits, Daniel Kersten, and Cindee Madison. Visual Glue. University of Utah Technical Report UUCS-98-007, March 12, 1998 .

[Vole00] Vladimir Volevich, Karol Myszkowski, Andrei Khodulev, and Edward Kopylov. Using the Visual Differences Predictor to Improve Performance of Progressive Global Illumination Computation. In *Transaction on Graphics*, 19(1), pp. 122-161. New York, 2000. ACM.

[Wine91] B. J. Winer, Donald Brown, and Kenneth Michels. Statistical Principles in Experimental Design, 3rd ed. New York, 1991. McGraw-Hill.

APPENDIX: INSTRUCTIONS TO SUBJECTS

These are the written instructions that were provided to each subject at the beginning of an experimental session. Aside from user-interface instructions for entering their responses, no other guidance was given.

> Today we are interested in gathering some information about how people perceive images. In the tasks that follow, you will see a number of images and we will ask you to evaluate what you see. There is no "right" or "wrong" answer to any response; we just want to know what you think. As you look at these images, try not to "think too much" about what you see. Go with your first impression.
>
> In this experiment we will show you a number of images, one shown right after the other. Some of these images are photographs of real objects, and others are computer-generated. For each image, we want to know whether you think it is real or not real. Sometimes it may be a close call, but just do the best you can.