A Psychophysical Investigation of Global Illumination Algorithms Used in Augmented Reality

TIMOTHY J. HATTENBERGER, MARK D. FAIRCHILD, GARRETT M. JOHNSON, and CARL SALVAGGIO
Rochester Institute of Technology

The overarching goal of this research was to compare different rendering solutions in order to understand why some yield better results specifically when applied to rendering synthetic objects into real photographs. A psychophysical experiment was conducted in which the composite images were judged for accuracy against the original photograph. In addition, iCAM, an image color appearance model was also used to calculate image differences for the same set of images. Conclusions obtained included the effect of global illumination on the accuracy of the final composite rendering. Also, it was discovered that the original rendering with all of its artifacts is not necessarily an indicator of the final composite image's judged accuracy. Finally, initial results show promise in using iCAM to predict a relationship similar to the psychophysics, which could eventually be used in-the-rendering-loop to achieve photorealism with minimized computation.

Categories and Subject Descriptors: I.3.m [**Computer Graphics**] Miscellaneous

General Terms: Experimentation

Additional Key Words and Phrases: Image difference, psychophysics, global illumination, perception, rendering

ACM Reference Format:

Hattenberger, T. J., Fairchild, M. D., Johnson, G. M., and Salvaggio, C. 2009. A psychophysical investigation of global illumination algorithms used in augmented reality. ACM Trans. Appl. Percpt. 6, 1, Article 2 (February 2009), 22 pages. DOI = 10.1145/1462055.1462057 http://doi.acm.org/10.1145/1462055.1462057

1. INTRODUCTION AND BACKGROUND

Computer graphics and vision research have established a more intimate working relationship in recent years. The convergence of these two fields seems logical as a human observer is typically the final discriminator of the output imagery. This area of research is currently of great interest. Perhaps as interesting is that it spans areas including, but not limited to, computer science, vision science, biology, digital image processing, perception, and psychophysics.

This area of research is predicated upon a definition of realism. James Ferwerda of Cornell wrote a paper [Ferwerda 2003] that is meant to serve as a framework to help define realism in the context of computer graphics. It is important to realize that an "image is a visual *representation* of a scene, in that it 're-presents' selected properties of the scene to the viewer with varying degrees of realism" [Ferwerda 2003]. He presents three types of realism in computer graphics: physical realism,

Author's address: T. Hattenberger, Digital Imaging and Remote Sensing Laboratory, Rochester Institute of Technology, 54 Lomb Memorial Dr., Rochester, NY 14623; email: tim@cis.rit.edu.

photorealism, and functional-realism. The idea of photorealism is intriguing due to the intimate relationship between the human visual system (HVS), image synthesis, and image evaluation.

With synthetic images, the user has complete control over their creation. Therefore, one can use these images and psychophysics to investigate the rendering algorithms or various aspects of the HVS itself. The most obvious experiment is to compare a photograph of the real world scene to the synthetic image. Meyer et al. [1986] used an image synthesis approach that is based on two specific modules: physical and perceptual. Their approach was to achieve accurate light simulation before the image was degraded with a perceptual transform. The results of this study (from 1986) showed promise in achieving photorealism. McNamara [2001] suggested a more robust approach to control the viewing conditions and scene content. They developed a technique for measuring the perceptual equivalence of a graphical scene to a real scene. They began by running several psychophysical experiments where human observers were asked to compare two dimensional target regions of a real physical scene to regions of the synthetic representation of that scene. The results of these experiments showed that the visual response to the real scene was similar to the high-fidelity rendered image [McNamara et al. 1998]. This was extended to comparisons of complete three dimensional objects, which inherently allowed comparisons of scene characteristics such as shadow, object occlusion and depth perception [McNamara and Chalmers 2000]. Rademacher et al. [2001] approached the photorealistic question from a different perspective. He proposed that the key to creating good rendering algorithms is to first understand the perceptual process. They proceeded to measure the visual realism in images using psychophysical experiments. It is clear that they wanted to hone in on which cues (shadow softness, surface smoothness, number of light sources, number of objects, and variety of objects) in an image contribute to the realism, not just that the final output appears realistic. The main results of their work were that sharper shadows were perceived as being less real, diffuse surfaces were rated as being more real than spray-painted (more specular) surfaces. Also, observers realism response did not increase with an increase in the number of objects, variety of object shapes or number of lights. The interesting thing to note is that these first experiments all used images of real scenes, which did indeed look like computer generated images.

There has also been relevant research developing metrics to compare real and synthetic images [Rushmeier et al. 1995]. The researchers began with ideas from image compression and modified them to account for specific perceptual phenomena related to image synthesis and image comparison (i.e., relative luminance variations, nonlinear response of the eye, and spatial frequency sensitivity as a function of luminance). "Ultimately, the biggest challenge is to take insights into human perception and apply them to visual simulation directly, computing only as much as is needed to satisfy the observer" [Rushmeier et al. 1995].

To this point, the images used in the experiment, were either completely real, or completely synthetic. The next step is to investigate some of the same things with augmented reality images, or in other words, images with both real and synthetic components. Selan [2003] performed psychophysical experiments most closely related to this research. He isolated four sources of lighting error in compositing: brightness errors, chromaticity errors, shading directionality errors, and case shadow directionality errors. In the experimenters stimuli, both components composited together were real, as opposed to rendering a synthetic object into a real scene, and only the shadow was rendered.

The research presented in this article takes components of this prior work a step further. First, a more complex scene with more complex objects was used. Second, the final image will be composed of both real and synthetic components rendered using various ray-tracing based global illumination algorithms and composited using a differential rendering [Debevec 1998] technique. In addition the images were evaluated using psychophysics as well as iCAM [Fairchild and Johnson 2004], an image color appearance model. iCAM was developed as a general image appearance model capable of predicting (i.e., rendering)

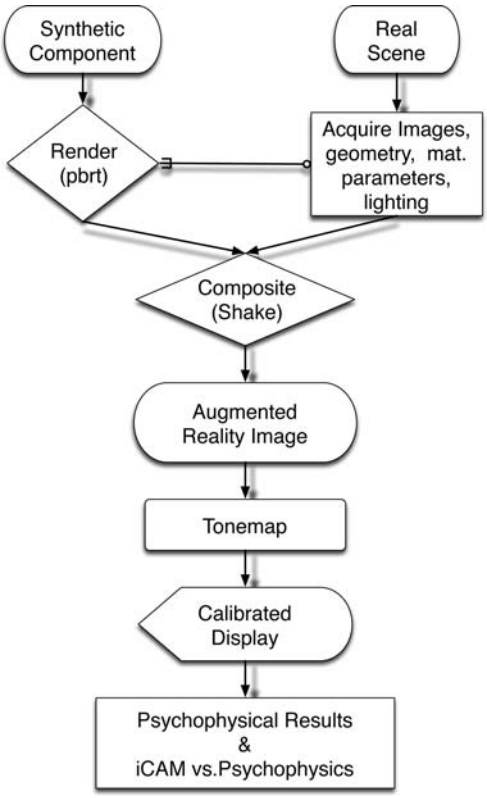

Fig. 1. A high-level representation of the procedure this work follows.

images and measuring differences in various viewing conditions. Unlike Rushmeier et al. [1995] it was not specifically created to measure image differences for rendered images. Therefore, this research aimed to explore iCAM's utility in this capacity and identify inherent strengths and weaknesses as a general model.

2. EXPERIMENTAL APPROACH

This section describes the experimental approach used in this research. Overall the process can be divided into the following areas. First, the scene was constructed in a light booth. This scene was then photographed using special techniques and equipment. In parallel, a virtual model of the scene was built and used to render images. The renderings and photographs were then used in the compositing step to create the stimuli. The stimuli were presented to observers in a psychophysical experiment as well as iCAM. The results of those experiments were analyzed both individually as well as in concert with each other. The flow chart shown in Figure 1 gives the reader a high level of understanding of the procedure used in this work.

2.1 Scene Construction and Image Capture

There were several important requirements that influenced the design of the scene. First, it needed to be a controlled environment. Control in this context refers to the illumination both in terms of its

color and geometry. This requirement was satisfied by constructing the scene in a standard viewing lightbooth. Second, in order to be rendered, the objects comprising the scene needed to be defined both geometrically and in terms of their material composition. As such, the author chose objects with well-behaved BRDFs, close to the idealized Lambertian or perfectly specular BRDFs, and simple in terms of their geometry, with the exception of the cow.

Ultimately, a photograph of the scene would be compared to a photograph of the scene including one rendered object. This cow object must exist both as a physical model, as well as a 3D model in the computer. As this object was the focus for the research, the author wanted an object that was more realistic and intricate than a simple wooden block. The typical process is to choose a real object and then spend a large amount of time modeling that object in a modeling program. This problem was averted by following the opposite path. First, a geometric file was purchased at turbosquid.com. This file was of a cow, and was extremely intricate, including textures. The file was emailed to a company called Stratasys [str], specializing in rapid prototyping. The company uses a process called fused deposition modeling (FDM), which essentially prints a 3D version of the file by building up layer upon layer of extruded plastic. The cow ordered was made of white ABS plastic, relatively strong and nearly opaque. In this case, nearly opaque implied the cow exhibited a fair amount of subsurface scattering, the impact of which will be discussed later. Finally, to complete the collection of objects to construct the scene, a mirror, terra cotta pot, painted wooden spheres and cubes, racquetballs, foam toy blocks, and a clay vase were included.

As important as the object materials were the object placements within the booth and relative to each other. Recalling the ultimate objective of this research, the author wanted to accentuate differences in global illumination rendering algorithms. Therefore diffuse-diffuse and diffuse-specular interactions between the cow and other objects were created through careful object placement. First, the cow was placed on a mirror, which created a significant amount of illumination on the underside of the cow, as well as a unique pattern of light on the large white vase in the back of the scene. Secondly, the multi-colored foam blocks were placed almost in direct contact with the camera-side of the cow (see Figure 3). In addition to producing color bleeding on the front of the cow, the object occlusion produces complexity that adds to realism and something an observer might encounter in real life. The entire scene was then geometrically modeled in Blender [Andauer et al. 2004], an open-source 3D modeling and animation package. After the geometry was created, the material parameters of the objects were specified. The surface reflectance characteristics (bidirectional reflectance distribution fucntion [BRDF]) were assumed Lambertian for the foam blocks, racquetballs, light booth interior, dark gray and black wooden blocks, and the cow. The mirror was modeled using a perfect specular BRDF. The vase and red wooden objects were modeled using more realistic BRDFs. They were determined iteratively and visually. Ideally, one would render the image spectrally and discard information just before display. However, this was computationally prohibitive. Therefore, reflectance of the objects was measured spectrally using a spectrophotometer, and tristimulus values calculated as shown in Figure 2. This was an appropriate method as the materials used did not exhibit fluorescence behavior or sharp features in their spectral reflectance functions.

The scene was photographed using a proprietary digital camera with a filter wheel designed to match the XYZ standard observer and high-bit depth sensor. It was colorimetrically calibrated prior to use. Images photographed with this camera and processed with the in-house software package were XYZ float images and imported directly into the compositing software described later.

2.2 Image Rendering

The next step was to render the light booth scene, and use those results to composite the synthetic cow into photograph without the cow. `pbrt` which stands for physically based rendering techniques, was the

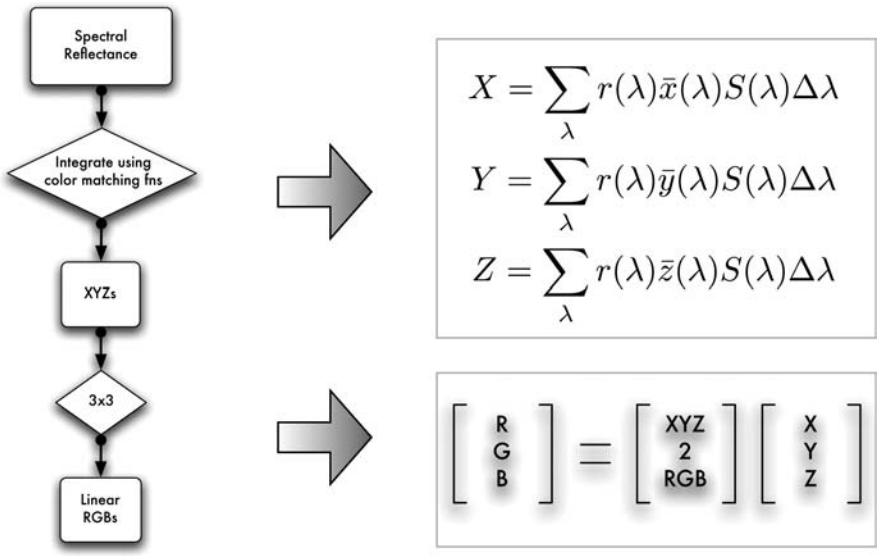

Fig. 2. This figure shows the procedure used to go from spectral measurements to RGB triplets that were then assigned to the materials used in the rendering software.

Fig. 3. Photograph of the scene built inside of the light booth. Notice the reflection from the mirror onto the vase as well as the color bleeding from the foam blocks onto the underside of the cow. These are the very visible indirect illumination effects reproduced in renderings physically through the use of a global illumination algorithm.

software used in rendering the synthetic images. As the name implies, `pbrt` attempts to render "a 2D image from a description of a 3D scene . . . " using the " . . . principles of physics to model the interaction of light and matter" [Pharr and Humphreys 2004]. This is an important statement, as many rendering

systems used to create imagery, especially those in the entertainment industry, are at best loosely based on physics, rather than first principles based.

There are advantages and disadvantages to each type of system. The nonphysics-based systems allow more degrees of freedom for the artist to manipulate the image throughout the entire process. With `pbrt`, however, the user can only manipulate the input, including descriptions of the geometry, surface properties, and color of objects. `pbrt` uses this information as input to physics-based equations. Therefore, even those user specified inputs are expected to be based in reality. The result is a high-quality realistic image, demonstrating real physical phenomenology. The tradeoff, however, is typically large amounts of computation time and fewer degrees of freedom for the user. This type of rendering system is ideal for the author for at least two reasons. First, physical properties can be measured using analytical devices (i.e., spectrophotometers) and those measurements input into `pbrt`. Also, the final solution (including artifacts) is more easily understood by looking to the equations and methods of implementation.

"In general the amount of light that reaches the eye from a point on an object is given by the sum of emitted light and reflected light" [Pharr and Humphreys 2004]. This idea is referred to as the *light transport or rendering* equation and is shown in Equation 1, "...which says that the outgoing radiance $L_o(p, \omega_o)$ from a point p in direction ω_o is the emitted radiance at that point in that direction, $L_e(p, \omega_o)$, plus the incident radiance from all directions on the sphere S^2 around p scaled by the BSDF $f(p, \omega_o, \omega_i)$ and a cosine term:

$$L_o(p, \omega_o) = L_e(p, \omega_o) + \int_{S^2} f(p, \omega_o, \omega_i) L_i(p, \omega_i) \, |cos\theta_i| \, d\omega_i. \tag{1}$$

This equation is almost always too complicated to be solved by anything other than numerical integration techniques. The method used to solve this integral can be changed to one of several options included with `pbrt`. These surface integrators were the primary variable manipulated to create the stimuli for this research. The integrators are introduced at an extremely high level below.

—Direct: As the name implies, this integrator only considers illumination arriving directly at the first intersected surface. The illumination is integrated over the entire hemisphere. The only user adjustable setting is the samples per pixel [spp] that refines the final image.

—Whitted: Similar to direct illumination, the biggest difference is the ability to recursively trace perfectly refracting or reflecting surfaces, such as a glass sphere or mirror, respectively. This implementation assumes a point source light, and again allows the user to specify the samples per pixel.

—Path tracing: A Monte Carlo-based ray tracing algorithm that recursively traces from the camera to through the scene to the light sources, integrating over the hemisphere at each intersection. Path tracing is an unbiased algorithm where the mean value is correct; however, large noise of variance may be present. On the upside, this is predictably lowered with a corresponding increase in the rays cast into the scene (spp).

—Irradiance caching: A biased global illumination algorithm that attempts to solve for the indirect component by precomputing the irradiance values at a select number of locations typically at the areas where there is the highest frequency change in indirect illumination distribution. As the image is being rendered, these precomputed irradiance samples are reused, interpolated, and averaged for intersected points that do not correspond directly with one of the stored samples. If the error is larger than a user-defined value, a new irradiance cache will be computed on-the-fly. The user can specify the samples per pixel and more importantly the error threshold. The problem is that there is not necessarily a 1:1 mapping between the error metric and the final image quality as there is in path tracing. This is one of the problems with a biased algorithm. Additionally, the algorithm decides

Table I. Rendering Settings Used in pbrt to Render the Whitted, Direct, and Path Tracing Images

Image	1	2	3	4	5	6	7
maxdepth	5						
Algorithm	Whitted		Direct		Path		
SPP	1	16	1	16	16	128	1024
Time [s]	7.4	88	103.7	1738	225	1909	16536
Name	Whitted_1	Whitted_16	Direct_1	Direct_16	Path_16	Path_128	Path_1024

Table II. Rendering Settings Used in pbrt to Render the Irradiance Cached Images

Image	8	9	10	11
SPP	16			
maxspeculardepth	5			
maxindirectdepth	5			
maxerror	2.0	2.0	0.02	0.02
nsamples	256	4096	256	4096
Time [s]	2912	3031	4066	40,000
Name	Irrad_2_256	Irrad_2_4096	Irrad_02_256	Irrad_02_4096

where to precompute the irradiance, and, therefore, the settings will tend to be scene dependent. Also, as the algorithm computes irradiance samples, there is an underlying assumption that the surfaces contributing to the indirect component are primarily diffuse.

—Photon mapping: This is a two-pass technique. The first pass is to propagate no more than a user-specified number of rays into the scene from the light sources. These rays are propagated based on the surface reflectance properties and stored into a 3D data structure called a kd tree. The scene is then ray-traced as usual, starting from the camera and propagating into the scene. The ray stops at the first object it hits, and the photon map is accessed. Based on some user-defined parameters, the photon map is searched for nearby photons, and the solution for the given point solved by averaging these photon within a given search radius. The photon map can be used to solve the indirect illumination component only (in conjunction with the direct surface integrator described above), or it can be used to solve both the direct and indirect. Since photon mapping propagates rays from the source to the scene, it allows for computation of complex illumination phenomena such as caustics in water. One of the major disadvantages is the number of parameters that can be adjusted in order to tune the photon map. This can also be a distinct advantage over other algorithms.

Readers are encouraged to read pbrt's accompanying text, [Pharr and Humphreys 2004], which contains many examples, images, lines of code, as well as the references to the seminal work upon which this system, and global illumination algorithms in general, are based.

It was not the author's intention to determine thresholds of realism for a given rendering method. Rather, the author wanted to investigate coarse increments in several global illumination techniques to see how realism varied within and across algorithms. The goal was to span a large range of rendering times and realism, including the various algorithm artifacts. This made pbrt an excellent research tool as it has a plug-in type architecture to test different techniques. pbrt comes with the integrators described above that were all used for this research. Two renderings for each algorithm were completed, with and without the cow. Therefore, a total of 32 full-scene images were rendered, and two additional images to create a matte used in the compositing process. Tables I through III show the abbreviations and settings used in pbrt to create the renderings.

Table III. Rendering Settings Used in `pbrt` to Render Images Using Photon Mapping

Image	12	13	14	15	16
SPP	16				
maxdepth	5				
directwithphotons	*T*			*F*	*F*
finalgather	*T*	F		*T*	*F*
directphotons	10 million			-	
indirectphotons	1 million				
causticphotons	20,000				
nused	150				
maxdist	1.0	2.0	1.0	2.0	2.0
finalgathersamples	64	—		64	—
Shoot Time [s]	411	411	388	131	114
Render Time [s]	54K	1463K	1214	53K	2395
Name	Photon_d_1.0_fg	Photon_d_2.0_fg	Photon_d_1.0_nfg	Photon_i_fg	Photon_i_nfg

2.3 Compositing

The basic idea of compositing is to layer various image elements into one complete final image. It is used in many motion pictures to combine computer generated elements with live action footage. At this point in the process there are two photographs and 34 rendered images. The idea was to use compositing techniques and differential rendering [Debevec 1998] in order to add the rendered cow to the photograph without the cow. All of the compositing was done in relative XYZ's by normalizing the Y value of the ptfe, an approximate perfect reflecting diffuser, in the scene to be 1 nominally. Essentially, a matte was created using individual renderings of the cow and the block occluding the cows. This matte was then used to extract the cow from the rendering. The extracted cow was then composited onto the photograph of the scene without the cow (see Figure 4). Initially, the shadows and other inter-reflections were also extracted from the rendering and composited into the photograph, however, due to geometric inconsistencies between the photograph and rendering, it was decided to extract the shadows from the photograph and composite those along with the rendering of the cow. Figure 5 shows all of the composite renderings cropped to the area around the cow as well as the photograph.

2.4 Psychophysics

The final step was to compare the images against one another. In this experiment, the observer sat 36″ from a 23″ Apple Cinema LCD monitor with a maximum luminance of approximately 180 $[cd/m^2]$ in a dark room and was presented with three images on screen: the original photograph on top, and the pair of images on the left and right halves of the bottom of the screen. The observer was asked to "choose the image on the bottom of the screen by clicking on it, that is most like the image on the top. This is an accuracy, not preference judgement." After the observer clicked on one of the images, three noise images were displayed for one second, and the next pair was presented. There were 17 total images, 16 that included the rendered and composited cow, and the completely real photograph (the "real photograph" refers to the photograph where the real cow model was present in the booth. This may be referred to as only the "real" photograph or scene later). This implies that there were trials where the photograph was presented along with one of the renderings. Also, all trials were unique in that there were no trials where both images were the same. In total, there were $(n)(n-1)/2$ total trials, where n is 17. Therefore, there were 136 total trials. It should be noted that although the real photograph was the image presented on top for every trial, it was not explicitly stated. Also, the observers were told to look in and around the area of the cow. This was to reduce any noise due to the randomized presentation order. In other words, there were trials where it was very obvious that the only difference

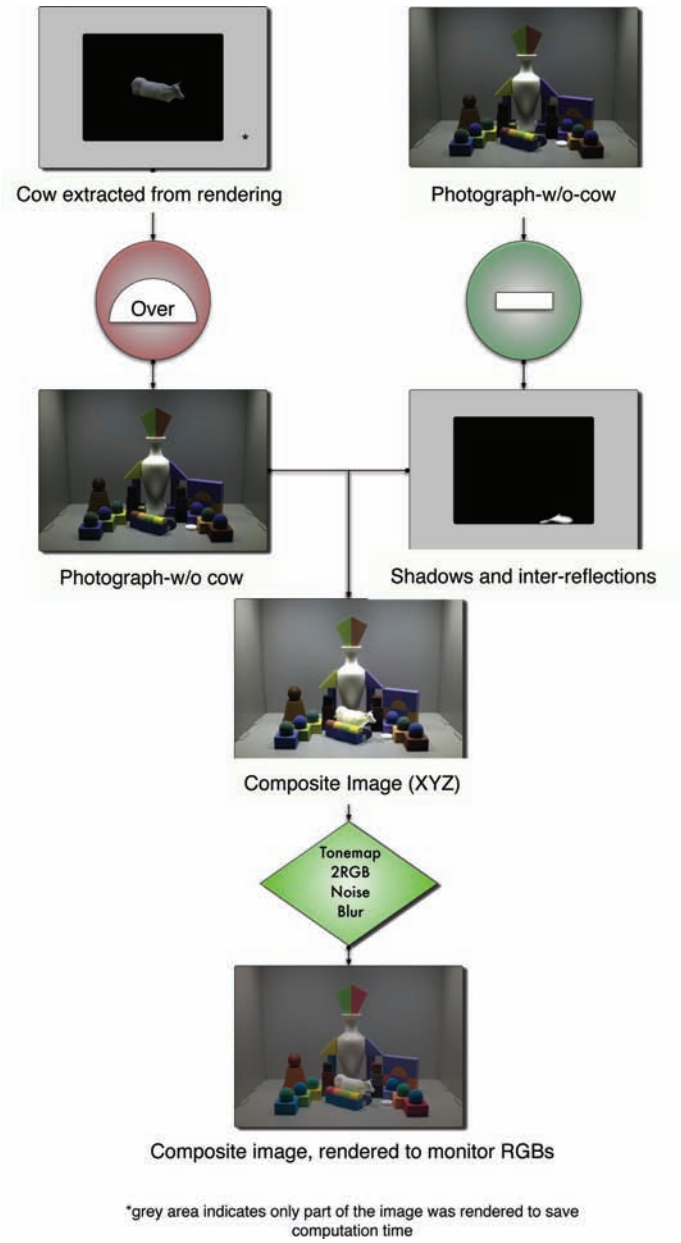

Fig. 4. Flow chart demonstrating compositing process.

between images was the cow, and there were others where this was not the case. After a few trials though, every observer would be focusing their attention around the cow and not in the areas of the image that remained constant.

It should be noted that only a photograph of the original scene, and not the scene itself was used in this experiment. Using images only eliminated any cross-media color appearance differences (luminance

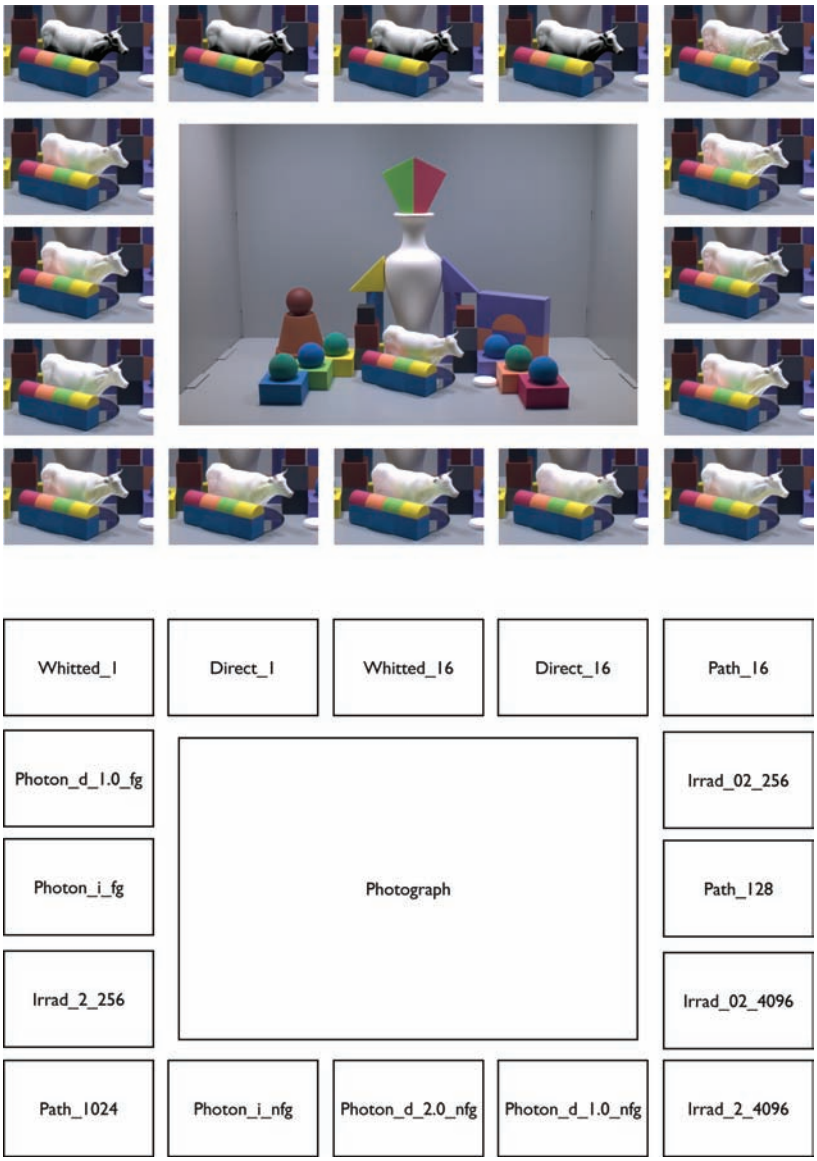

Fig. 5. This figure shows all of the composite images, along with a map of the rendering algorithm used. These images are being shown in RGBs optimized for the Apple Cinema LCD that was used for the psychophysics experiment. Additionally, only the entire photograph is shown due to size constraints.

levels, spatial extent, dynamic range, etc.). In addition, it is impossible to calculate image difference against an original.

Thirty-one observers took part in this experiment comparing the composite images against each other and the real photograph. The observer pool consisted of 15 experts, including the author, and 16 naive. The typical time to complete the experiment ranged anywhere from 10 to 20 minutes, although the exact time taken for an observer to complete the 136 comparisons was not recorded. The observers

were allowed to take as much time as necessary to make a decision for a given trial pair. They were limited only by the brief time the noise images were shown between trials. The error bars used in this research were calculated using the method described in [Montag]. This method is empirical and takes into account not only the number of observations, but also the number of stimuli. In the case of this research, there were 31 observers and 136 pairs, both fairly large numbers, which should lead to smaller 95% confidence intervals (CI).

2.5 iCAM

The final topic that will be discussed is a specific model of human vision capable of predicting image differences, appearance, and overall image quality. This model is termed iCAM (image color appearance model) and was described as a part of Johnson's Ph.D. dissertation [Johnson 2003]. The motivation of his research was to devise a modular framework, that would mimic various properties of the HVS, thus creating a device-independent image quality model, with the ultimate goal "to predict perception" [Johnson 2003].

The first important aspect of this research was to make the model a collection of modules. This allows the use of current color difference research while maintaining the flexibility to add new modules as more research is done. Johnson also presented the idea of a pool of modules, where each module accepts input, and provides output, while acting as a self-contained unit. The strength of this design is that it allows the modules to be linked together to provide an overall metric, and at the same time the individual modules can be used to determine the cause for the image difference.

The image difference functionality of iCAM is based on decades of CIE color difference research. As Johnson points out, though, traditional color difference equations were developed based on uniform color patches, not complex spatially varying stimuli like images. Therefore, applying these equations to images would essentially be treating the individual pixels as separate stimuli, not accounting at all for the coherence in the imagery. It is well known that the HVS keys on certain features in imagery, like edges, in addition to the color difference. The idea then is to preprocess the images spatially in a similar manner as the HVS, and then apply the color difference equations. The first incarnation of this idea was called S-CIELAB, where S stands for spatial, accounting only for the CSF. Johnson extends the S-CIELAB concept to also account for spatial-frequency adaptation, spatial localization, and local and global contrast detection by introducing modules for each. An important distinction of the iCAM model is its use of single-scale spatial filtering with anisotropic filters as opposed to multiscale and multiorientation filtering. "The model predicts spatial frequency adaptation by nomalizing its 2D CSF by the Fourier transform of the spatial adapting stimulus" [Fairchild and Johnson 2005]. Adapting a single contrast sensitivity function has proven adequate for supra-threshold predictions of perceived image differences [Fairchild and Johnson 2005]. In addition, it is much less computationally intensive. This approach has been shown to work for global stimuli, but more work should be done to see if it holds for local stimuli.

Each of these modules contributes to the overall image differences. In some cases, it is also beneficial to reduce the module results into a single number, the image difference. This number is some type of weighted sum of the individual modules, giving a measure of how different a pair of images is in terms of image difference units. In reducing to a single number, information is inevitably lost, yet the result could be a scale of image differences along some variable continuum, which in turn can be compared to a psychophysical scale. Of course, in some situations, this could replace, or at least preprocess the stimuli for a psychophysical experiment, reducing the burden on human observers. Johnson [2003] also points out that "an image difference model is only capable of predicting magnitudes of errors, and not direction." This can be obviated by examining the output of each of the modules individually. Rather than using the traditional color difference equations, which essentially measure a scalar distance, calculations of

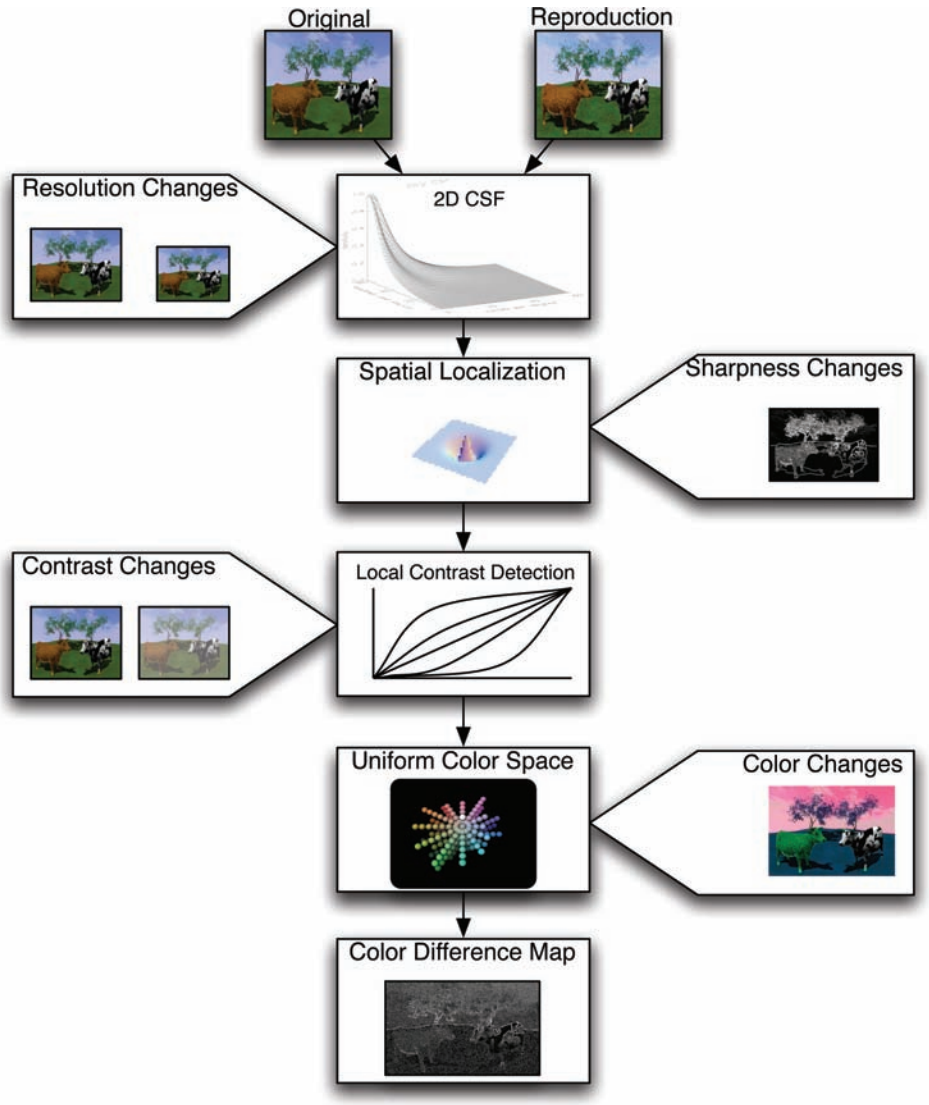

Fig. 6. Image difference flow modules with associated causes of difference [Johnson 2003].

entities like ΔL^* can be performed. Figure 6 indicates specific causes for image differences and the corresponding module that contains this information.

Ultimately, the calculation of an interval scale from the magnitude differences is required in order to compare against the psychophysical experiment and generate a continuum of accuracy for the various rendering algorithms. This implies the use of a magnitude model such as iCAM over other threshold models, whether color or spatial. For example, Daly's [1993] visual differences predictor (VDP), traditionally ignores color altogether and generates a probability prediction of thresholds (e.g., can I see the difference, not the magnitude of the difference).

The composite images were evaluated using iCAM in the image difference configuration. The difference calculation was performed as follows. First the images were individually processed both spatially and

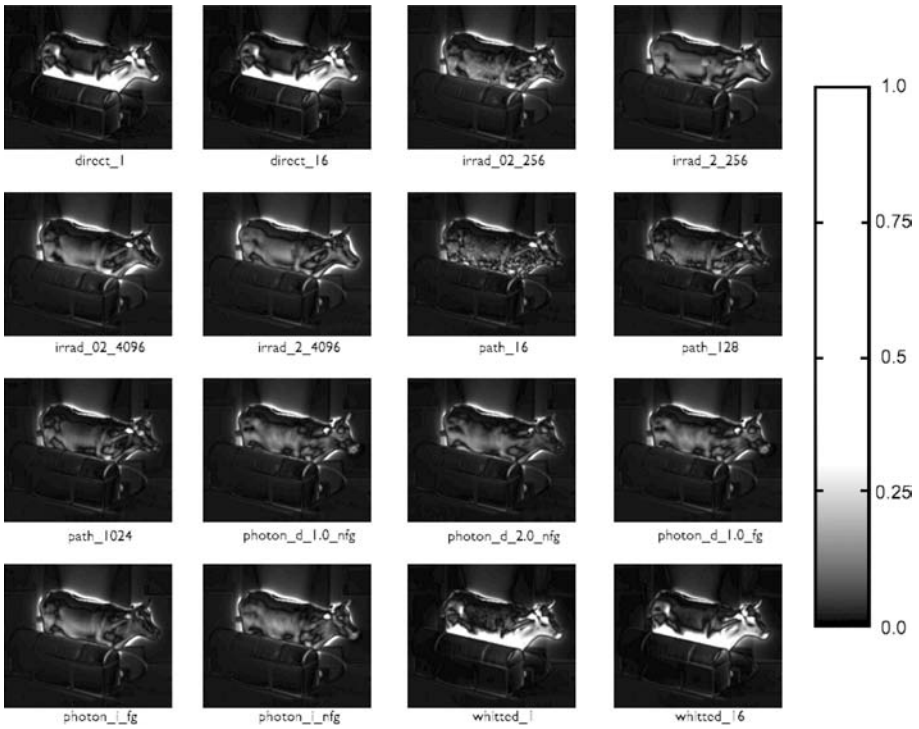

Fig. 7. The image difference maps for each of the renderings as compared to the photograph with the real cow. Maps have been contrast enhanced to accentuate detail.

colorimetrically using the iCAM modules. The resulting images were in the IPT [Ebner and Fairchild 1998] color space. IPT is used because it is a simple to calculate a uniform color space, designed for image processing applications. IPT was specifically designed to correct limitations of the CIELAB space, and to maintain lines of constant hue. Then each of the composite images were subtracted band by band from the photograph. The bands were squared, summed, square rooted (a Euclidean color difference), and then cropped to the area around and including the cow, resulting in an image difference map. Each of the 16 composite images were compared against the original photograph, resulting in image difference maps (see Figure 7). Bright pixels indicate a larger difference between the photograph and the test image. Differences were only calculated for the cow and surrounding area. This was to match the psychophysical experiment where observers were instructed to focus their attention in that area.

Two types of analysis were done on these images. The first was to look at each difference map individually to look for trends. The second was to reduce these maps into a single number so relationships with other variables can be explored. Several different techniques were performed for the latter. They included finding the maximum value, median value, as well as several different percentile values. There are a limitless number of ways to reduce this data; however, only a few make sense. The idea was to choose methods that can be explained at least initially perceptually, with the intent for others to continue this research. In all cases it was determined that the 92nd percentile of the image difference map was a reasonable method to reduce the data. It highlighted the area of the image that observers considered when making a decision. If one looks at those other percentiles, either too much or too little of the cow and surrounding area is included. Other statistics of the image difference images were

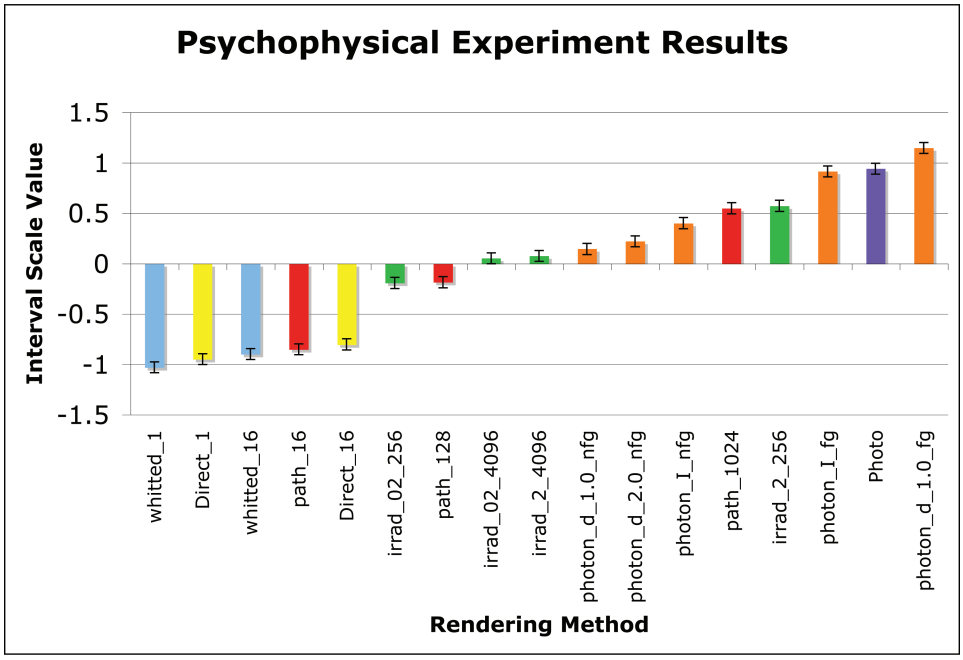

Fig. 8. This plot shows the interval scale plotted against the rendering algorithm for the entire pool of obsevers.

calculated and included the mean and median but were quickly discarded. The rationale was because statistics, such as the mean, would not correspond to real observers who performed the experiment and imply a person is able to average all of the color differences and then make a judgement. This is typically not the behavior observed. Most people search for the first differences they see, which is analogous to the percentile concept. Remember that the observers were asked to look at two images and choose the one that was closest in terms of accuracy to the photograph. Humans do not look at all of the differences for each image and then take an average. They start by looking at the most extreme differences. As they need to examine the images more closely for differences, they are looking at the less extreme color differences, until one image looks worse than the other. These results will be presented below.

3. RESULTS AND DISCUSSION

Figure 8 shows the combined results of naive and expert for the paired comparison experiment. This plot shows the interval scale value for each of the rendering algorithms used as well as the photo. Large negative numbers represent the worst images based on the question asked and increase with increasing quality of the image. There are a couple of important results that can be seen. First is the importance of using a full global illumination algorithm. Direct illumination, and for the purposes of this scene, Whitted surface integrators, were chosen by the observers as less like the original. This result is not a surprise remembering the question for the observers was to choose the image most like the original, and not which they preferred. However, observers consistently chose the noisier results of algorithms such as path tracing, to the smooth, but dark cows created by the direct and Whitted algorithms. The only exception to this is the extremely noisy result of the path tracing integrator using only 16 samples per pixel, [spp], and the direct illumination integrator also using 16 [spp]. The path

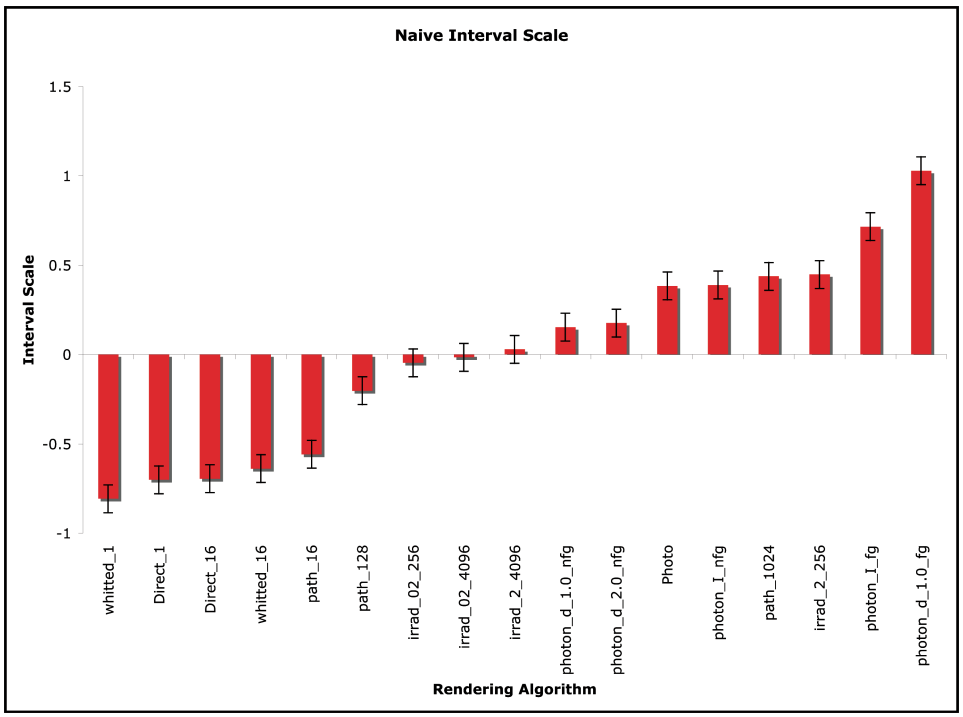

Fig. 9. This plot shows the interval scale plotted against the rendering algorithm for the naive obsevers.

tracing image demonstrates significant amount of noise in capturing the indirect illumination effects, whereas the direct illumination produces a cow, which is dark, but with very smooth tone transitions. As Figure 8 shows, the observers did not think one image was closer to the photograph than the other within the 95% CIs.

The final important result shown in the graph is the image created using photon mapping to solve both direct and indirect illumination with final gathering was chosen (outside of the 95% CIs) by observers to be a better reproduction of the photograph, than the photograph itself! The composite image using photon mapping does indeed look a lot like the original, with the exception that it is a bit brighter. The author proposes the following explanation for this result. The observers were unaware of the fact that the image on the top was the original photograph, or that it never changed. Additionally, they were unaware that the original photograph, the same as the image at the top of the screen, was randomly being presented in the test pairs. Perhaps then the observers, particularly the naive, were assuming that every image presented to them was in some way manipulated from the original image. This caused them to switch their criteria from image accuracy to preference equating brighter with better, and choose the image with the brighter cow. This is still valid, because in cases where they are very different or very similar, the observer is being asked to make a comparison where two images may be equidistant along two different perceptual axes. They must choose one image or the other, resulting in a 50% likelihood of either.

In order to explore this further, the data was divided into naive and expert observers and analyzed again. The following plots in Figures 9 through 11, show the results when analyzed this way. The first plot shows the naive results plotted versus the interval scale, with recalculated errorbars. There are a couple of interesting results. First, the photograph is ranked even lower on the scale, implying

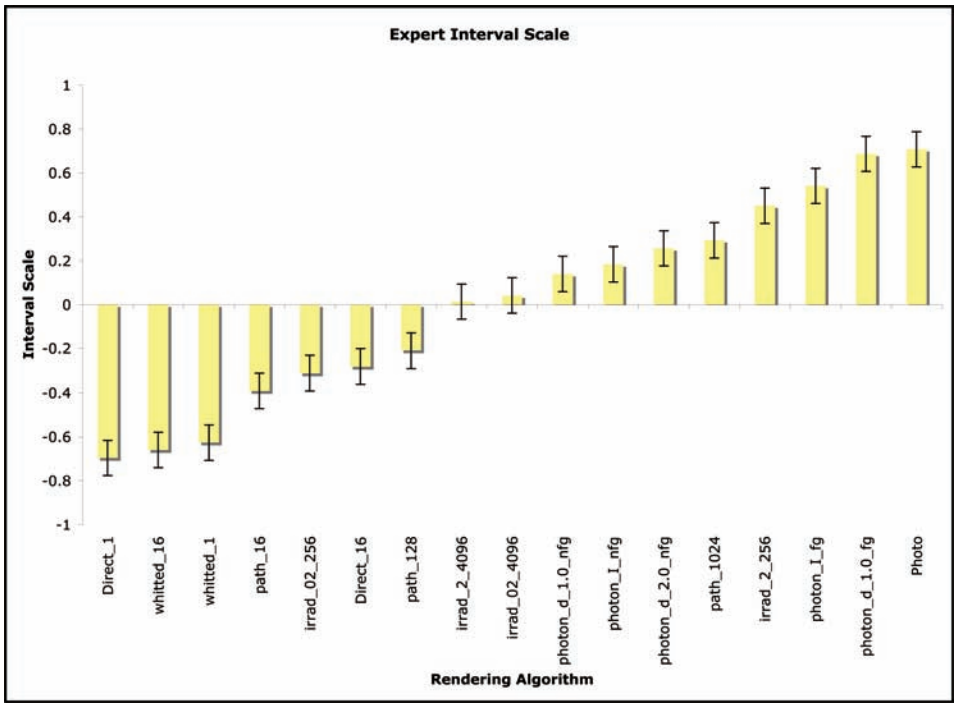

Fig. 10. This plot shows the interval scale plotted against the rendering algorithm for the expert obsevers.

all rendering algorithms above it are at least the same visually as the photograph. Also, the naive observers tended rank all of the dark (i.e., no indirect illumination) as the worst in terms of accuracy to the original photograph with the exception of the noise path tracing rendering (path_16), which is equally bad as whitted_16 and direct_16.

The next plot (Figure 10) shows the results of the psychophysical experiment for the expert observers only. Interestingly, the experts disagree with the naive observers in that the path tracing rendering with 16 [spp] is a lot worse than the direct with 16 [spp], however it is still within the error bars. Also, the experts ranked the irrad_02_256 as being in the low grouping for accuracy. Both the path_16 and irrad_02_256 displayed a significant amount of high-frequency noise in the cow. Also, the experts ranked the photograph at the top, which one would expect. However it is still within the errorbars of the next best rendering using photon mapping and final gathering, the same image the naive group judged as most accurate. In the case of the experts, they chose the smoothest photon-mapping image as opposed to the naive.

In general, one would expect that iCAM should predict a large image difference where the psychophysics scale value is small as well as the converse. Several different analyses were completed using iCAM. The first is a general plot (Figure 12), similar to the ones in the preceding section. However, keep in mind that low image difference values imply a closer match to the original photograph. Of course in this computational example, the photograph will always receive an image difference of exactly 0. Additionally, as one might expect all of the Whitted and direct integrator images have a significantly larger image difference than all of the other algorithms. Other than that, the only thing that can be said about this plot is that it appears that the irradiance caching and photon-mapping algorithms yield images with a larger difference than the path tracing in general. It seems then that iCAM calculates

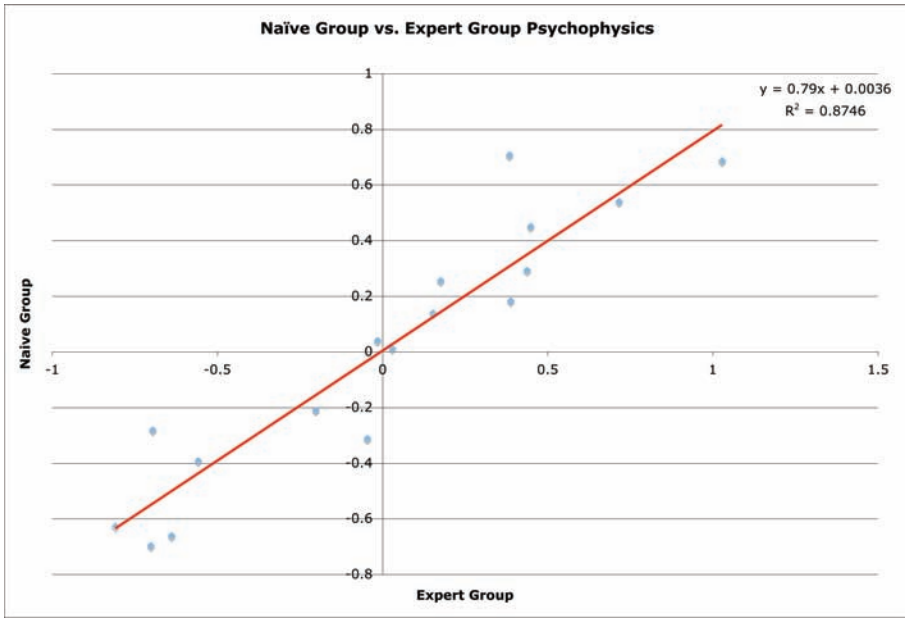

Fig. 11. This plot shows the naive interval scale plotted against the expert interval scale.

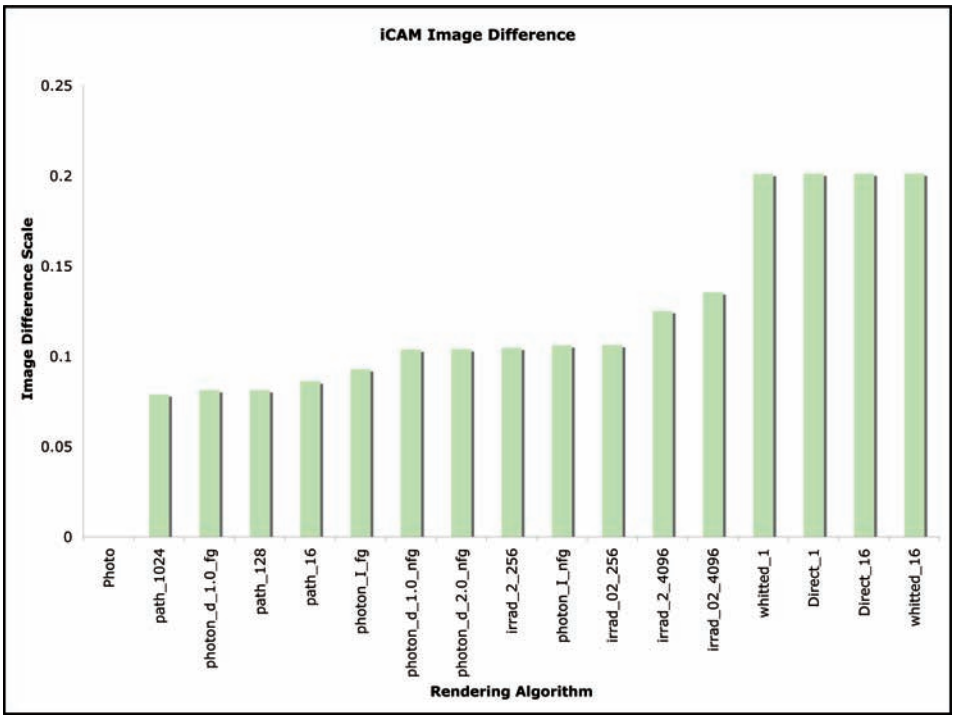

Fig. 12. This plot shows the iCAM image difference scale value for each of the rendering algorithms. Higher values indicate a larger difference from the original photograph.

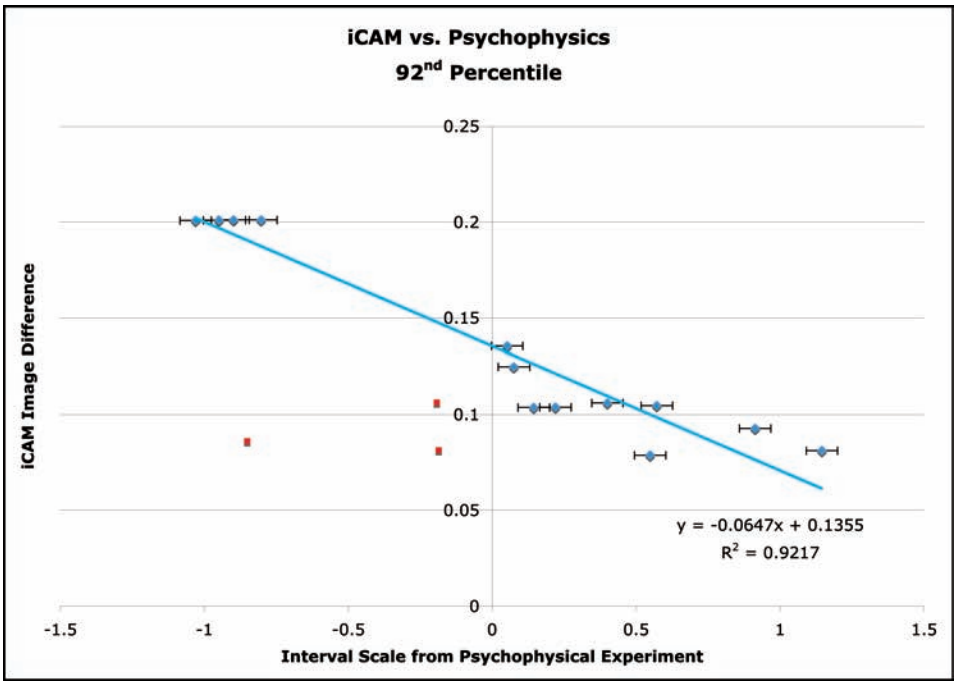

Fig. 13. This plot shows iCAM image differences vs. the psychophysical experiment results, with the three outlier data points, shown in red, removed from the data fit.

a smaller image difference for the noisy unbiased path tracing algorithms rather than the biased algorithms. The spatial noise is weighted less than the absolute color difference.

The next logical thing to do is to compare the iCAM results against the psychophysics results. Recall that iCAM inherently produces an image difference map, yet in the plots a single number is used. In all cases it was determined that the 92nd percentile of the image difference map was a reasonable method to reduce the data. It highlighted the area of the image that observers considered when making a decision. Other statistics of the difference images were calculated and included the mean and median, but were quickly discarded. The rationale was because statistics, such as the mean, would not correspond to real observers who performed the experiment, and imply a person able to average all of the color differences and then make a judgement. This is typically not the behavior observed. Most people search for the first differences they see, which is more analogous to the percentile idea. Again, the observers were asked to look at two images and choose the one that was closest in terms of accuracy to the photograph. Humans do not look at all of the differences for each image and then take an average. They start by looking at the most extreme differenes. As they need to examine the images more closely for differences, they are looking at the less extreme color differences, until one image looks worse than the other.

The plot (Figure 13), shows iCAM versus the paired comparison interval scale for all observers combined. Recall that an inverse relationship (negative slope), is desired. In general, the plot shows this relationship, with an $r^2 = 0.5517$ if all data points are included. It is apparent that there are three significant outliers, shown in red. These three correspond with irradiance caching (*error = 0.02, 256 [spp]*), path tracing (*16 [spp]*) and path tracing (*128 [spp]*). If these three images are removed from the data set, a much stronger relationship exists with a $r^2 = 0.9217$ as well as a higher slope. It is obvious

Fig. 14. This image shows the final composited image demonstrating significant high-frequency noise, created using the path tracing algorithm.

that the common characteristic of these three outlier images (see Figure 14) is a significant amount of high frequency noise. Referring back to the plot, according to the psychophysics, these three images scored low on the interval scale. One would expect a large image difference value calculated by iCAM which is not the case. Although there was high-frequency noise, the absolute color difference relative to the photograph in those areas was small, resulting in an overall lower iCAM image difference value based on the 92nd percentile. Recall that iCAM computes an image difference map, which is then reduced to a single number using the 92nd percentile statistic. This procedure of reducing the difference map to a single value does not explicitly include any spatial information, such as high-frequency noise. This is an extremely important as one of the major advantages of using iCAM versus a simple color difference equation is the incorporation of the spatial dimension.

4. CONCLUSION AND FUTURE WORK

The goal of this research was to begin to explore various global illumination rendering algorithms, specifically those based on ray-tracing, as applied to the rendering synthetic objects into real photographs. These algorithms in conjunction with augmented reality were analyzed through the use of psychophysical experiments and iCAM, a computational model of human vision. Through all of this, hopefully one could learn something about image synthesis, the human visual system and perception, and perhaps the confluence of the two. The ultimate goal of this work was not to judge reality, or look for thresholds of reality within a given rendering algorithm. If it were, obviously finer steps of settings within a given algorithm would have to be tested. The author initiated the study into factors that affect realism, using a global illumination ray tracing technique when applied to a composite image.

In terms of image rendering, several things were learned, specifically when applied to rendering synthetic objects into real photographs. First, it seems that global illumination algorithms will perform better than one that does not account for indirect illumination, except in the presence of significant noise of variance. This may not seem like a significant effect, until one considers the stimuli for the experiment. Consider the images rendered with the direct illumination integrator as compared to the path integrator. Both the expert and naive observers ranked the noise path tracing image lower than the direct integrator with 16 [spp]. However, iCAM calculates the greatest image difference for all four algorithms that do not consider indirect illumination. There were trials where the observer was presented with two images

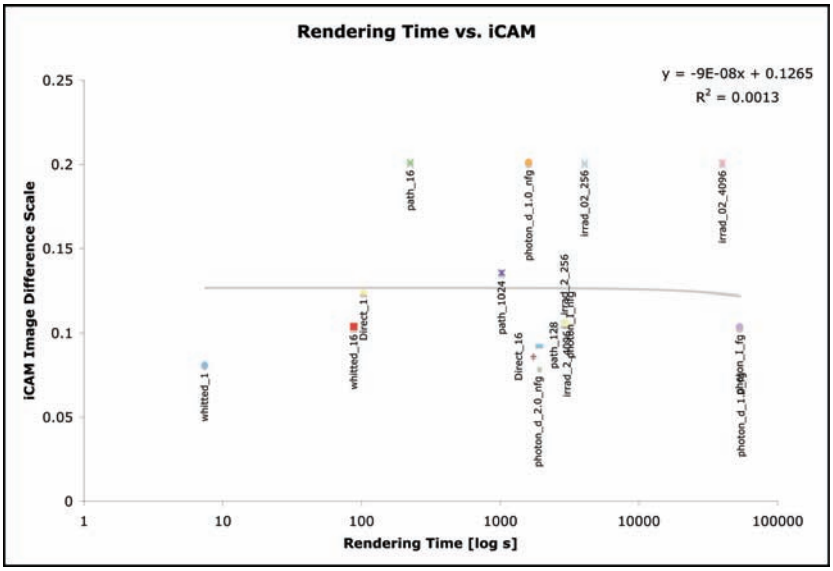

Fig. 15. This plot shows iCAM image differences versus the rendering time.

that may have been equally "wrong," leading them to judge a preference rather than accuracy. This perhaps indicates that there are more axes for decision making than originally considered. The work of Stokes et al. [2004] could be used as a baseline to study these different axes, separating the direct, diffuse, glossy, and specular components for example.

It would not be appropriate to say any global illumination is better none at all in every situation. Obviously it will depend on the scene content. This scene was designed to accentuate the indirect interactions through the use of the mirror and the diffuse cow next to the brightly colored blocks. So, in this experiment, the indirect effects were very important. In addition, it must be reiterated that only a single scene with a single object was tested, thus global conclusions cannot be drawn. In addition, most objects were diffuse, or specular. More scenes, objects, and materials need to be tested to look for more general trends.

Secondly, the experiments concluded that rendering time alone is not a direct indicator of the most accurate match to an original. It is generally true that more samples (i.e., more time) are required to achieve a better rendering especially for unbiased algorithms such as path tracing. However, when biased and unbiased algorithms are pooled, this is no longer the case. Stated differently, Figure 15 shows algorithms that take roughly the same time to render, but vary wildly relative to the iCAM image difference.

Related to this result is the fact that the most refined or "tuned" rendering will always rank the best. Again, this is true both for iCAM and the psychophysics. Clarifying, this result is most likely the case for the augmented reality application only. In other words, let us assume one looks at the entire rendering with all of the artifacts, and then extracts one object such as the cow and composites that into a photograph and compares the two images (see Figure 16). The cow does not necessarily appear as bad as the entire rendering because the artifacts are not as pronounced. This could be because of the material or lighting or a number of other things. The converse is probably also true that some rendered objects show much more of the artifacts than the original rendering, thus making the final composite appear worse.

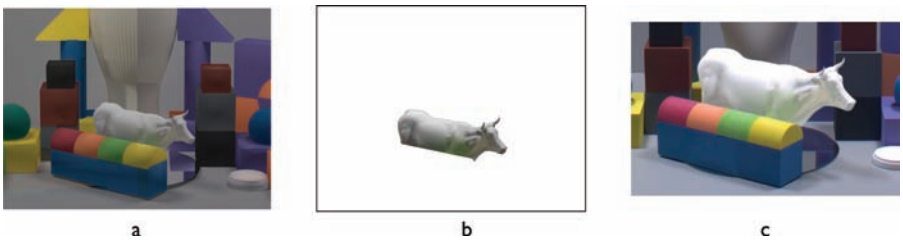

Fig. 16. (a) The irrad_2_256 rendering, notice the artifacts on the vase and cow. (b) The extracted cow, the artifacts are harder to see, unless one compares to the rendering. (c) The final composite with some clipping applied, making the artifacts harder yet to see.

One of the most promising results was the correlation between iCAM and the psychophysical experiment. The expected relationship was present, with some outliers however. The outliers all tended to be the renderings that had high-frequency artifacts, whereas the other images contained lower frequency artifacts. In other words, it seems as though the human observers judged these images using different criteria than the other 13 stimuli. Of course, this may also be indicative of how the iCAM image difference maps were reduced to a single number.

A more general compositing process could be implemented, as described in the background and theory. More specifically, the effects of the object on the scene and scene on the object could be generically included. Also, more rendering algorithms could be used, including local illumination shaders that use a hack for the ambient term. It would be interesting to see where along the continuum these algorithms would fall. It is also possible to implement some of these algorithms and other shaders in graphics processing units (GPUs), allowing even more possibilities of realtime rendering and interaction with the real scene. Ideally, it would be interesting to use this research to extract a baseline "threshold for reality." This could be used with iCAM in the rendering loop to produce images that are believed to within an acceptable accuracy to an original. Of course, all of these results would then be analyzed psychophysically. This is of course more tractable when the image rendering is on the order of minutes (GPU), rather than days with a ray-tracer.

Perhaps the most important recommendation is to continue research in reducing the image difference maps to a single number. The strength of iCAM is that it produces a map, of image differences. In other words, it calculates color differences, spatially, on complex spatial stimuli, images. It seems counterintuitive to discard that spatial information in order to determine a relationship with a psychophysical experiment. This research clearly points out the need for more study into the reduction of the map into a single number. iCAM does perform spatial processing and results in the two-dimensional image difference map, which is in and of itself valuable. The author's need to reduce that data to plot relationships against psychophysical experiments resulted in the "loss" of spatial data in some sense. The author believes all of information is there at various stages of the iCAM image processing. Parameters could be derived at these various steps and a multivariable equation derived that reduces the difference map, including the color and the spatial characteristics. The outliers in the iCAM / psychophysical relationship may not be outliers at all, but just not completely described by the 92nd percentile statistic.

REFERENCES

STRATASYS, INC. Rapid Prototyping, Cad Plastic Prototyping, Digital Prototypes, Cad Plastic Prototype Engineering. http://ww.bookmarksync.com/details/1a1b171b1c1e1d.

ANDAUER, C., BASTIONI, M., ESTEVEZ, A. C., DAMBEKALNS, K., FINDEISS, F., HEIZER, A., ET AL. 2004. Blender Documentation Volume I - User Guide. http://www.blender.org/documentation/htmlI/.

DALY, S. 1993. The visible differences predictor: An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*. MIT Press, 179–206.

DEBEVEC, P. 1998. Rendering synthetic objects into real scenes: bridging traditional and imagebased graphics with global illumination and high dynamic range photography. In *Proceedings of the ACM Special Interest Group on Graphics and Interactive Techniques (SIGGRAPH'98).* ACM, New York.

EBNER, F. AND FAIRCHILD, M. D. 1998. Ebner. Development and testing of a color space (ipt) with improved hue uniformity. In *Proceedings of the 6th Color Imaging Conference*. ACM, New York, 8–13.

FAIRCHILD, M. D. AND JOHNSON, G. M. 2004. The icam framework for image appearance, image differences, and image quality. *J. Electron. Imag. 13*, 126–138.

FAIRCHILD, M. D. AND JOHNSON, G. M. 2005. On the salience of novel stimuli: Adaptation and image noise. In the *Proceedings of the IS&T SID 13th Color Imaging Conference*. Society for Imaging Science and Technology, Springfield, VA, 333–338.

FERWERDA, J. A. 2003. Three varieties of realism in computer graphics. In *Proceedings of SPIE Human Vision and Electronic Imaging*. International Society for Optical Engineering, Bellingham, WA.

JOHNSON, G. M. 2003. Measuring images: Differences, quality and appearance. Ph.D. thesis, Rochester Institute of Technology.

MCNAMARA, A. 2001. Visual perception in realistic image synthesis. In *Proceedings of European Association for Computer Graphics (EUROGRAPHICS '01).* ACM, New York.

MCNAMARA, A. AND CHALMERS, A. 2000. Comparing real and synthetic scenes using human judgements of lightness. In *Proceedings of the EUROGRAPHICS Workshop on Rendering.* Springer-Verlag, London, UK, 207–218.

MCNAMARA, A., CHALMERS, A., TROSCIANKO, T., AND REINHARD, E. 1998. Fidelity of graphics reconstructions: A psychophysical investigation. In *Proceedings of the 9th Eurographics Rendering Workshop*. ACM, New York.

MEYER, G. W., RUSHMEIER, H. E., COHEN, M. F., GREENBERG, D. P., AND TORRANCE, K. E. 1986. An experimental evaluation of computer graphics imagery. *ACM Trans. Graph. 5*, 1, 30–50.

MONTAG, E. 2006. Empirical formula for creating error bars for the method of paired comparison. *J. Electron. Imag.* 15, 1, 010502-1-3.

PHARR, M. AND HUMPHREYS, G. 2004. *Physically Based Rendering from Theory to Implementations*. Morgan Kaufman, San Francisco, CA.

RADEMACHER, P., LENGYEL, J., CUTRELL, E., AND WHITTED, T. 2001. Measuring the perception of visual realism in images. In *Proceedings of the 12th Eurographics Workshop on Rendering*. Springer-Verlag, London, UK, 235–248.

RUSHMEIER, H. E., WARD, G., PIATKO, C., SANDERS, P., AND RUST, B. 1995. Comparing real and synthetic images: Some ideas about metrics. In *Proceedings of the 6th Eurographics Workshop on Rendering*. Springer-Verlag, London, UK.

SELAN, J. A. 2003. Merging live video with synthetic imagery. M.S. thesis, Cornell University.

STOKES, W., FERWERDA, J., WALTER, B., AND GREENBERG, D. 2004. Perceptual illumination components: A new approach to efficient, high quality global illumination rendering. In *Proceedings of the ACM Special Interest Group on Graphics and Interactive Techniques (SIGGRAPH'04).* ACM, New York, 742–749.

Received March 2007; revised July 2007, November 2007; accepted December 2007