EUROGRAPHICS '01 / Roberto Scopigno, David Duke
(Guest Editors)

Volume xxx, (2001), Number yyy

Visual Perception in Realistic Image Synthesis

Ann McNamara

Department of Computer Science, Trinity College, Dublin 2, Ireland

Abstract

Realism is often a primary goal in computer graphics imagery, we strive to create images that are perceptually indistinguishable from an actual scene. Rendering systems can now closely approximate the physical distribution of light in an environment. However, physical accuracy does not guarantee that the displayed images will have authentic visual appearance. In recent years the emphasis in realistic image synthesis has begun to shift from the simulation of light in an environment to images that look *as real as the physical environment they portray. In other words the computer image should be not only physically correct but also perceptually equivalent to the scene it represents. This implies aspects of the Human Visual System (HVS) must be considered if realism is required. Visual perception is employed in many different guises in graphics to achieve authenticity. Certain aspects of the visual system must be considered to identify the perceptual effects that a realistic rendering system must achieve in order to effectively reproduce a similar visual response to a real scene. This paper outlines the manner in which knowledge about visual perception is increasingly appearing in state-of-the-art realistic image synthesis. After a brief overview of the HVS, this paper is organised into four sections, each exploring the use of perception in realistic image synthesis, each with slightly different emphasis and application. First, Tone Mapping Operators, which attempt to map the vast range of computed radiance values to the limited range of display values, are discussed. Then perception based image quality metrics, which aim to compare images on a perceptual rather than physical basis are presented. These metrics can be used to evaluate, validate and compare imagery. Thirdly, perception driven rendering algorithms are described, these algorithms focus on embedding models of the Human Visual System (HVS) directly into global illumination computations in order to improve their efficiency. Finally, techniques for comparing computer graphics imagery against the real world scenes they represent are discussed.*

1. Some Characteristics of the Human Visual System

Many computer graphics images are produced for viewing by human observers, as opposed to automated inspection. It is therefore important to understand the characteristics and limitations of the Human Visual System (HVS). The HVS is well studied, but *perception* is a complex process. Evidence exists to indicate that features of the HVS do not operate independently, but rather functions overlap making it difficult to describe the perceptual process completely. However, there are many key features which can be modelled, and because many of the techniques described in this paper use or model these features, they are described briefly in this section. A more comprehensive description of the characteristics of the HVS can be found in [29].

Visual acuity is the ability of the HVS to resolve detail in an image. The human eye is less sensitive to gradual and sudden changes in brightness in the image plane but has higher sensitivity to intermediate changes. Acuity decreases with increase in distance. Visual acuity can be measured using a *Snellen Chart*, a standardised chart of symbols and letters. Under low levels of illumination our eyes are very sensitive and can detect small changes in luminance, however acuity for detail and ability to detect colour is poor. In high levels of illumination we have sharp colour vision and good acuity, but luminance differences must be large to be detected.

The Contrast Sensitivity Function (CSF) tells us how sensitive we are to the various frequencies of visual stimuli. If the frequency of visual stimuli is too high we will not be able to recognise the stimulus pattern any more. Imagine an image consisting of vertical black and white stripes. If the stripes are very thin (i.e. a few thousand per millimetre) humans will be unable to see individual stripes. All that we will see is a grey image. If the stripes then become wider and wider, there is a threshold width, after which humans are able to distinguish the stripes. Contrast Sensitivity depends on the size (coarse/fineness) of image features, or the *spatial frequency.*

Contrast is defined as:

$$\frac{l_{max} - l_{min}}{l_{max} + l_{min}}$$

where l_{max} and l_{min} are the maximum and minimum luminance. Human brightness sensitivity is logarithmic, so it follows that for the same perception, higher brightness requires higher contrast. Apparent brightness is dependent on background brightness. This phenomenon, termed simultaneous contrast, is illustrated in Figure 1. Despite the fact that *all* centre squares are the same brightness, they are perceived as different due to the different background brightness.

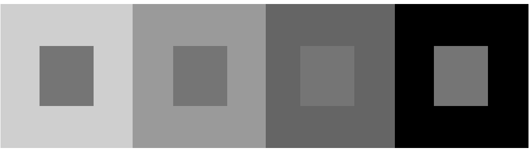

Figure 1: *Simultaneous contrast: the internal squares all have the same luminance but the changes in luminance in the surrounding areas change the* perceived *luminance of the internal squares*

Masking is the phenomenon by which visibility of a particular pattern is reduced by the presence of a second pattern. The HVS exhibits different spatial acuities in response to different colours. It is known that colour spatial acuity is less than monochrome spatial acuity.

The range of luminance we encounter in natural environments (and hence the range of luminance that can be computed by a physically based rendering algorithm) is vast. Over the course of the day the absolute level of illumination can vary by more than a 100,000,000 to 1 from bright sunlight down to starlight. The dynamic range of light energy in a single environment can also be large, in the order of 10,000 to 1 from highlights to shadows. However, typical display media have useful luminance ranges of approximately 100 to 1. This means some mapping function must be used to translate real world values into values displayable by the device in question, be it electronic (CRT) or print media. Initial attempts to develop such a mapping were simple *ad-hoc* methods that failed miserably for high dynamic range scenes. These ad-hoc methods proceeded by employing a linear arbitrary scaling, either mapping the average of a luminance in the real world to the average of the display, or the maximum non-light source luminance to the maximum displayable value. While such a scaling proved appropriate for scenes with similar dynamic range to the display media, it failed to preserve visibility in scenes with high dynamic ranges of luminance. This is due the fact that very bright or very dim values must be clipped to fall within the range of displayable values. Also, using this method all images are mapped in the same manner irrespective of *absolute* value. This means a room illuminated by a single candle could be mapped to the same image as a room illuminated by a search light, resulting in loss of the overall impression of brightness and so losing the subjective correspondence between real and displayed scene. It follows that more sophisticated mappings were required.

2. Tone Mapping Operators

Tone Mapping, originally developed for use in photography and television, addresses the problem of mapping to a display, and is an attempt to recreate the same *perceptual* response in the viewer of a synthetic image as they would have if looking at the real scene. Taking advantage of the HVS sensitivity to *relative* luminance rather than *absolute* luminance allows the overall subjective impression of a real environment to be replicated on some display media, despite the fact that the range of real world luminance often dwarfs the displayable range [37].

Tone Mapping Operators (TMO) can be classified according to the manner in which values are transformed. *Single-scale* operators proceed by applying the *same* scaling transformation for each pixel in the image, and that scaling only depends on the current level of adaptation, and not on the real-world luminance. *Multi-scale* operators take a differing approach and may apply a different scale to each pixel in the image, this time the scaling is influenced by many factors.

2.1. Single Scale Tone Mapping Operators

Tumblin and Rushmeier were the first to apply the dynamics of tone mapping to the domain of realistic

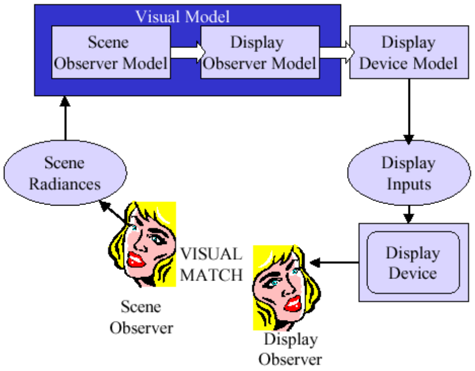

Figure 2: *A block diagram of Tone Mapping*

image synthesis [37]. Using a psychophysical model of brightness perception first developed by Stevens and Stevens [36], they produced a TMO that attempted to match the brightness of the real scene to the brightness of the computed image displayed on a CRT. To achieve this an *observer model* is built which describes how real world and display luminance are perceived, and a *display model* that describes how a frame-buffer value is converted into displayed luminance, Figure 2 [37]. The image is presented to a hypothetical real world observer, who adapts to a luminance $L_{a(w)}$. Applying Stevens' equation, which relates brightness to target luminance, the perceived value of a real world luminance, L_w, is computed as:

$$\beta_w = 10^{\beta(L_{a(w)})}(\pi \times 10^{-4} L_w)^{\alpha(L_{a(w)})}$$

where $\beta(L_{a(w)})$ and $\alpha(L_{a(w)})$ are functions of the real world adaptation level:

$$\alpha(L_{a(w)}) = 0.4 \log_{10}(L_{a(w)}) + 1.519$$

$$\beta(L_{a(w)}) = -0.4(\log_{10}(L_{a(w)}))^2 - 0.218 \log_{10}(L_{a(w)}) + 6.1642$$

Luminances are in cd/m^{-2}. If it is assumed that a display observer viewing a CRT screen adapts to a luminance, $L_{a(d)}$, the brightness of a displayed luminance value can be similarly expressed:

$$\beta_d = 10^{\beta(L_{a(d)})}(\pi \times 10^{-4} L_d)^{\alpha(L_{a(d)})}$$

where $\beta(L_{a(d)})$ and $\alpha(L_{a(d)})$ are as before. To match the brightness of a real world luminance to the brightness of a display luminance, β_w must equal β_d. The luminance required to satisfy this can be determined:

$$L_d = \frac{1}{\pi \times 10^{-4}} 10^{\frac{\beta_{a(w)} - \beta_{a(d)}}{\alpha_{a(d)}}} (\pi \times 10^{-4} L_w)^{\frac{\alpha_{a(w)}}{\alpha_{a(d)}}}$$

This represents the concatenation of the real-world observer and the inverse display observer model. To determine n, the frame buffer value, the inverse display system model is applied to give:

$$n = [\frac{L_d - L_{amb}}{L_{dmax}}]^{\frac{1}{\gamma}}$$

giving

$$\tau_{TUMB}(L_w) = [\frac{10^{\frac{\beta_{a(w)} - \beta_{a(d)}}{\alpha_{a(d)}}} (\pi \times 10^{-4} L_w)^{\frac{\alpha_{a(w)}}{\alpha_{a(d)}}}}{\pi \times 10^{-4}}]$$

Taking a slightly different approach, Ward [40] searched for a linear transform to give a similar result, while keeping computational expense to a minimum. He proposed transforming real world luminance, L_w, to display luminance, L_d, through m, a scaling factor:

$$L_d = m L_w$$

The consequence of adaptation can be thought of as a shift in the absolute difference in luminance required in order for a human observer to notice a variation. Based on psychophysical data collected by Blackwell [4], Ward defines a relationship that states that if the eye is adapted to luminance level L_a, the smallest alteration in luminance that can be seen satisfies:

$$\triangle(L_a) = 0.0594(1.219 + L_a^{0.4})^{2.5}$$

Real world luminance are mapped to the display luminance so the smallest discernible differences in luminance can also be mapped, using:

$$\triangle L(L_{a(d)}) = m \triangle L(L_{a(w)})$$

Where L_{aw} and $L_{a(d)}$ are the adaptation levels to the real world scene and display device respectively. The scaling factor, m, dictates how to map luminance from the world to the display such that a Just Noticeable Difference (JND) in world luminance maps to a JND in display luminance :

$$m = \frac{\triangle L(L_{a(d)})}{\triangle L(L_{a(w)})} = (\frac{1.219 + L_{a(d)}^{0.4}}{1.219 + L_{a(w)}^{0.4}})^{2.5}$$

To estimate the adaptation levels, $L_{a(w)}$ to $L_{a(d)}$, Ward assumes that the adaptation level is approximately half the average radiance of the image, ($L_{a(d)} = L_{dmax}/2$). Substituting in to equation (above) results in values from 0 to L_{dmax}, and dividing by L_{dmax} then gives values in the required range from [0..1]. The scaling factor is then given by:

$$m = \frac{1}{L_{dmax}}[\frac{1.219 + \frac{L_{dmax}}{2}^{0.4}}{1.219 + L_{a(w)}^{0.4}}]^{2.5}$$

where L_{dmax} is typically set to $100 cd/m^{-2}$.

In 1996, Ferwerda et al. [7] developed a model conceptually similar to Ward's, but in addition to preserving threshold visibility, this model also accounted for changes in colour appearance, visual acuity, and temporal sensitivity. Different TMOs are applied depending on the level of adaptation of the real world observer. A *threshold sensitivity function* is constructed for both the real world and display observers given their level of adaptation. A linear scale factor is then computed to relate real world luminance to photopic display luminance. The required display luminance is calculated by combining the photopic and scotopic display luminances using a parametric constant, k, which varies between 1 and 0 as the real world adaptation level goes from top to bottom of the mesopic range.

To account for loss of visual acuity, Ferwerda et al. used data obtained from experiments that related the detectability of square wave gratings of different spatial frequencies to changes in background luminance. By applying a Gaussian convolution filter, frequencies in the real world image which could not be resolved when adapted to the real world adaptation level are removed. Light and dark adaptation are also considered by Ferwerda, by adding a parametric constant, b, to the display luminance, the value of which changes over time.

A critical and underdeveloped aspect of all this work is the visual model on which the algorithms are based. As we move through different environments or look from place to place within a single environment, our eyes adapt to the prevailing conditions of illumination both globally and within local regions of the visual field. These adaptation processes may have dramatic effects on the visibility and appearance of objects and on our visual performance. In order to produce realistic displayed images of synthesised or captured scenes, a more complete visual model of adaptation needs to be developed. This model will be especially important for immersive display systems that occupy the whole visual field and therefore determine the viewer's visual state.

2.2. Multi-Scale Tone Mapping Operators

After careful investigation of the effect tone mapping had on a small test scene illuminated only by a single incandescent bulb, Chiu et al. [3] believed it was incorrect to apply the same mapping to each pixel. By uniformly applying any tone mapping operator across the pixel of an image, incorrect results are likely. They noted that the mapping applied to a pixel should be dependent on the spatial position in the image of that pixel. This means that some pixels having the same intensities in the original images may have differing intensity values in the displayed image. Using the fact that the human visual system is more sensitive to *relative* changes in luminance rather than *absolute* levels, they developed a spatially non-uniform scaling function for high contrast images. First the image is blurred to remove all the high frequencies, and then the result is inverted. This approach was capable of reproducing all the detail in the original image, but reverse intensity gradients appeared in the image when very bright and very dark areas were close to each other. Schlick [35] proposed a similar transformation based on a rational TMO rather than a logarithmic one. Neither of these methods accounted for differing levels of adaptation. Their solutions are based purely on experimental results, and no attempt is made to employ psychophysical models of the HVS.

Larson et al. [12] developed a histogram equalisation technique that used a spatial varying map of foveal adaptation to transform a histogram of image luminances in such away that the resulting image lay within the dynamic range of the display device and image contrast and visibility were preserved. First a histogram of brightness (approximated as a logarithm of real-world luminances) is created for a filtered image in which each pixel corresponds to approximately 1^o of visual field. A histogram and a cumulative distribution function are then obtained for this reduced image. Using threshold visibility data from Ferwerda, an automatic adjustment algorithm is applied to create an image with the dynamic range of the original scene compressed into the range available on the display device, subject to certain restrictions regarding limits of contrast sensitivity of the human eye.

Displaying high dynamic range images on low dynamic range devices without loss of important fine details and textures is difficult. The aim of the Low Curvature Image Simplifier (LCIS) approach introduced by Tumblin and Turk [39] is to preserve visibility of important fine details and textures when displaying high dynamic range images. LCIS is inspired by an artistic approach and proceeds by separating an image into a hierarchy of large features, boundaries and fine detail. The LCIS hierarchy creates progressively simpler images based on a single input parameter. The final output is constructed by strongly compressing the contrasts of the base image and adding back the details with little or no compression. Tumblin [38] gives a complete and comprehensive introduction to all major Tone Mapping Operators.

3. Perceptually Based Image Quality Metrics

Reliable image quality assessments are necessary for the evaluation of realistic image synthesis algorithms. Typically the quality of the image synthesis method is evaluated using image to image comparisons. Often

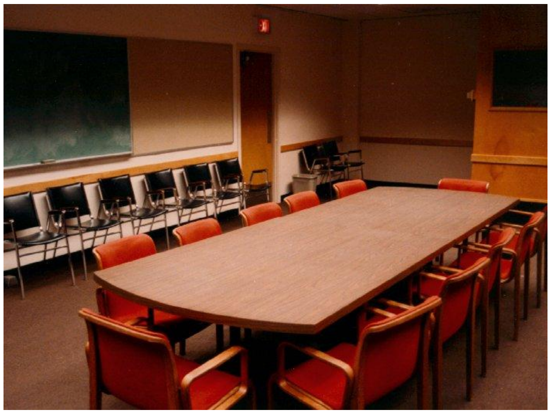

Figure 3: *Photograph of a Conference Room*

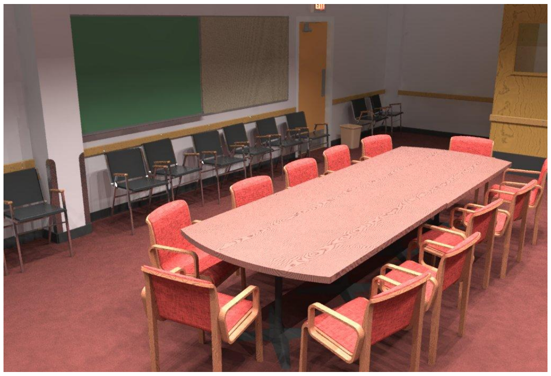

Figure 4: *Photo-Realistic Rendering of the above Conference Room*

comparisons are made with a photograph of the scene that the image depicts, as shown in Figures 3,4 [13].

Several image fidelity metrics have been developed whose goals are to predict the amount of differences that would be visible to a human observer. It is well established that simple approaches like mean squared error do not provide meaningful measures of image fidelity, Figure 5. The image on the left has been slightly blurred, while the image on the right has deliberate scribbles. The Root Mean Square Error (RMSE) value for blurred image is markedly higher than the RMSE for the image on the right. However, a human observer might indicate a higher correlation between the two images. This illustrates that the use of RMSE is not sufficient [30], [33], [8]. Clearly, more sophisticated measures which incorporate a representation of the HVS are needed. It is generally recognised that more meaningful measures of image quality are obtained using techniques based on visual (and therefore subjective) assessment of images, as after all most final uses of computer generated images will be viewed by human observers.

The following image comparison metrics were derived from [6, 9, 17] in a study which compared real and synthetic images, Figure 6, by Rushmeier et al. [33]. Each is based on ideas taken from image compression techniques. Image compression techniques seek to minimise storage space by saving only what will be visible in an image (similar to the goal of perceptually driven rendering where the aim is to minimise rendering times by computing only what will be visible in the image). Rushmeier et al. hoped to obtain results from comparing two images using these models that were large if large differences between the images exist, and small when they are almost the same. These suggested metrics include some basic characteristics of human vision described in image compression literature. First, within a broad band of luminance, the eye senses relative rather than absolute luminances. For this reason a metric should account for luminance variations, not absolute values. Second, the response of the eye is non-linear. The perceived “brightness” or “lightness” is a non-linear function of luminance. The particular non-linear relationship is not well established and is likely to depend on complex issues such as perceived lighting and 3-D geometry. Third, the sensitivity of the eye depends on the spatial frequency of luminance variations. The following methods attempt to model these three effects. Each model uses a different Contrast Sensitivity Function (CSF) to model the sensitivity to spatial frequencies.

Model 1 After Mannos and Sakrison: [17].

This model is adapted from a study in image compression which attempted to derive a numerically based measure of distortion which corresponds to the *subjective* evaluation of the image by a human observer, in order to simulate the optimum encoding technique. First, all the luminance values are normalised by the mean luminance. The non-linearity in perception is accounted for by taking the cubed root of each normalised luminance. A FFT is computed of the resulting values, and the magnitude of the resulting values are filtered with a CSF to an array of values. Mannos and Sakrison [MaSa74] proposed a model of the human CSF.

$$A(f) = 2.6 \cdot [0.0192 + 0.114\sqrt{f}]e^{-(0.114\sqrt{f})^{1.1}}].$$

where f is the spatial frequency of the visual stimuli given in cycles per degree (cpd). Finally, the distance between the two images is computed by finding the Mean Square Error (MSE) of the values for each of the two images. This technique therefore measures similarity in Fourier amplitude between images.

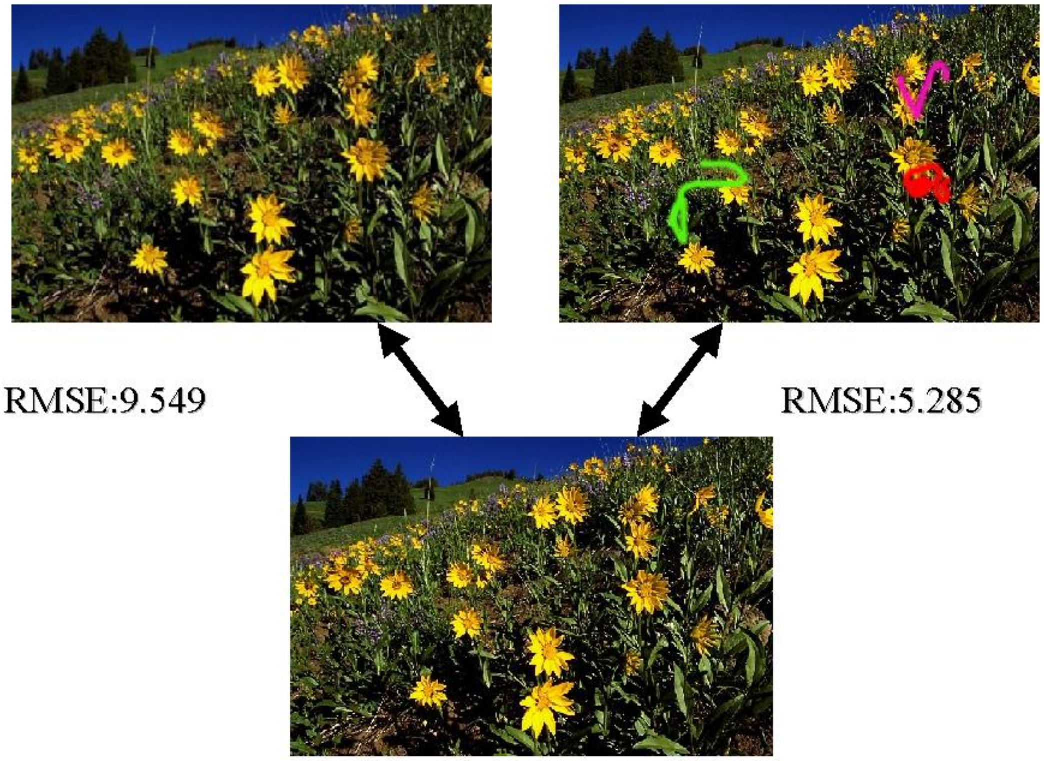

Figure 5: *Comparing top images to the image on the bottom using RMSE* [30]

Model 2 After Gervais et al: [9].
The original purpose of this model was to identify confusion among letters of the alphabet. Even though this problem is quite different to image comparison, Rushmeier et al. justify using this model as it includes the effect of phase as well as magnitude in the frequency space representation of the image. Once again the luminances are normalised by dividing by the mean luminance. A FFT is computed, producing an array of phases and magnitudes. These magnitudes are then filtered with an anisotropic CSF filter function constructed by fitting splines to psychophysical data. The distance between two images is then computed using methods described in [9].

Model 3 After Daly: adapted from [6].
Described in more detail in Section 3, this model combines the effects of adaptation and non-linearity into a single transformation, which acts on each pixel individually. This is in contrast to the first two models, in which each pixel has significant global effect in the normalisation by contributing to the image mean. Each luminance is transformed by an amplitude non-linearity value. An FFT is applied to each transformed luminance and then they are filtered by a CSF (computed for a level of 50 cd/m^2). The distance between the two images is then computed using MSE as in model 1.

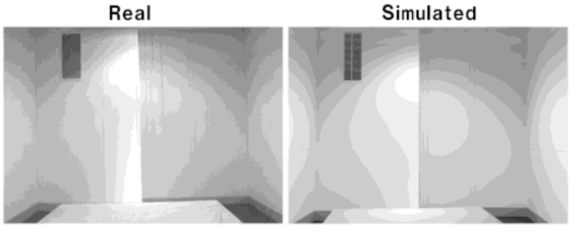

Figure 6: *NIST Comparison using a Conference Room*

In 1998, Li and Meyer conducted a comprehensive study that compared two of the more successful image quality models, outlined here:

Daly's Visible Differences Predictor

The **V**isible **D**ifferences **P**redictor (VDP) is a perceptually based image quality metric proposed by Daly [6]. The VDP takes a psychophysically based approach to construct a model of human vision. Two images serve as input to the VDP, and a difference map is produced as output. This difference map predicts the probability of detection of differences between the two images. Figure 7 gives a block diagram of the components of the predictor. The main stages are an initial non-linearity, frequency domain weighting with the human contrast sensitivity function CSF, and a series of detection mechanisms.

To account for adaptation and the non-linear response of retinal neurons, a non-linear response function is applied to each image. Daly assumed that adaptation is a function of each pixel individually. The model used for adaptation estimates the relationship between brightness sensation and luminance. At low levels of luminance a cube-root power law is applied, while at higher luminance levels it approximates the logarithmic dependence.

The next stage involves converting the image to the frequency domain. The transformed data is weighted with the CSF i.e. the scaled amplitude for each frequency is multiplied by the CSF for that spatial frequency. This data is then normalised (by dividing each point by the original image mean) to give local contrast information.

The image is then divided into 31 independent streams. It is known that the HVS has specific selectivities based on orientation (6 channels) and spatial frequency (approximately one octave per channel). Each of the five overlapping spatial frequency bands is combined with each of the six overlapping orientation bands to split the image into thirty channels. Along with the orientation-independent base band this gives a total of 31 channels. At this point the individual channels are transformed back into the spatial domain.

A mask, which is a function of image location in the image, is associated with each channel. The presence of masking information at a specific location, spatial frequency and orientation increases the threshold of detectability for a signal with those characteristics. A threshold elevation map for each channel is computed as a function of the mask contrast. Finally, mutual masking is applied between the two sets of threshold elevation maps from both input images to produce a single threshold elevation map per channel.

Contrasts of corresponding channels in one image are subtracted from those of the other images, and the difference is scaled down by threshold elevation. The scaled contrast differences are used as the argument to a psychometric function to compute a detection probability. The psychometric function yields a probability of detection of a difference for each location in the image, for each of the 31 channels. The detection probabilities for all of the channels are combined using the assumption of independent probabilities, giving an overall signed detection probability for each location in the image.

Sarnoff Visual Discrimination Model

The Sarnoff VDM [15] focuses more attention on modelling the physiology of the visual pathway. Therefore the VDM operates in the spatial domain (as opposed to the frequency domain approach of VDP). The main components of the VDM include spatial resampling, wavelet-like pyramid channelling, a transducer for JND calculations and a final refinement step to account for CSF normalisation and dipper effect simulation. The VDM also takes as input two images along with a set of parameters for viewing conditions, and here the output is a map of JND's. The overall structure of the VDM is shown in figure 8.

To account for the optics of the eye and mosaic structure of the retina, a single Point Spread Function (PSF) is used to predict the foveal performance of the two dimensional optics of the eye (it is assumed the PSF is circularly symmetric). The effect of the PSF convolution is blurring of the input images. A spatial resampling, at a rate of 120 pixels per degree, is then applied to account for the fixed density of the cones in the fovea. This resampling is essential in a spatial domain approach since the extraction of the different frequency bands is dependent on the resampling kernels and the resampling rates. If the original image is too big, and the local image quality cannot be assessed in a single glance, then the image can be subdivided into smaller blocks.

A Laplacian pyramid stores a wavelet representation of the resampled input images and a quadrature mirrored pair of convolution kernels records information along each of the four orientations. On completion of this stage, the raw luminance signal has been converted into units of local contrast. Due to the use of a spatial domain convolution approach, the peak frequency of each level has to be a power of two. The seven bandpass levels have peak frequencies from 32 to 0.5 cycles per degree, where each level is separated from its neighbours by one octave. A steerable pyramid is used to perform the decomposition, to increase performance. This is a multi-scale, multi-orientation,

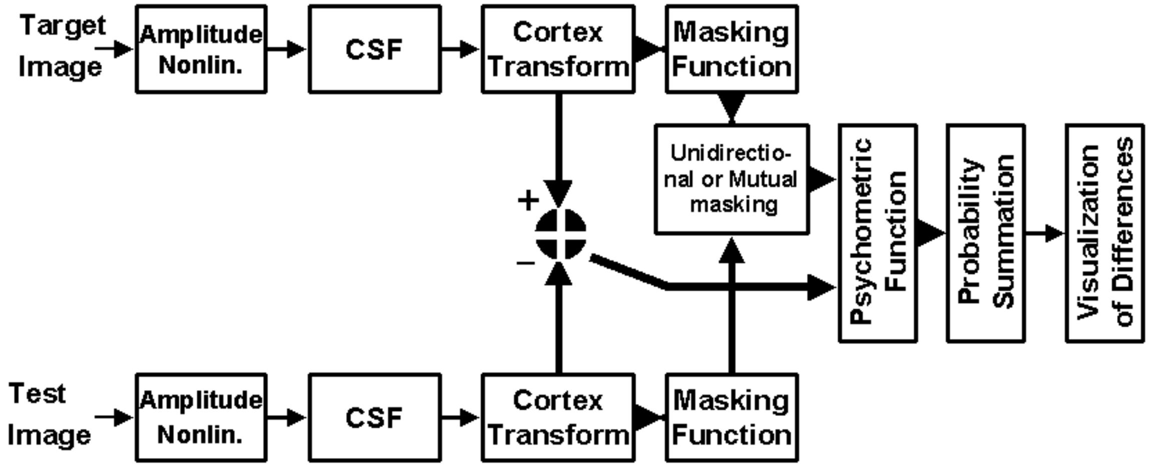

Figure 7: *A block diagram of the Visible Difference Predictor* [5]

image transform with both frequency and orientation components. The last step in the decomposition process is computation of a phase-independent energy response by squaring and summing odd phase and even phase coefficients. They are determined by convolving the quadrature mirror pair filters with a certain frequency band.

The energy measure is normalised by the square of the reciprocal of the CSF, then a transducer is used to refine the JND map by taking the spatial masking dipper effect into account. The dipper shape reflects on characteristic of the contrast discrimination function. This stage involves the transformation by a sigmoid non-linearity. Finally the model includes a pooling stage in which transducer outputs are averaged over a small region by convolving with a disc-shaped kernel.

Once the JND difference map has been computed for each channel, the final stage involves putting together the contributions from each channel. This leads to the concept of a space of multiple dimensions. There are 28 channels involved in the summation, seven pyramid levels times four different orientations. For each spatial position the final JND distance can be regarded as the distance between the 28-dimensional vectors.

Meyer and Li concluded that although both methods performed comparably, the Sarnoff VDM was deemed slightly more robust producing better JND maps and required less re-calibration than the Daly VDP. Despite this both have been successfully incorporated into global illumination algorithms to produce favourable results [26, 27, 1].

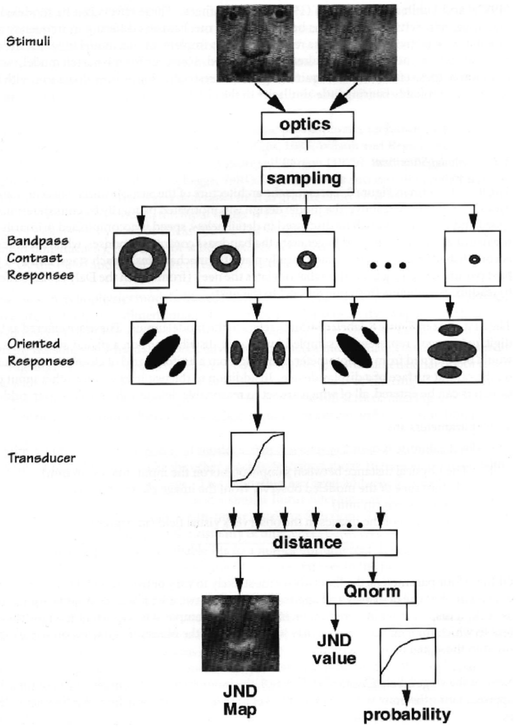

Figure 8: *A block diagram of the Visual Discrimination Model (VDM)*

The main contribution of this study was the independent verification of the major features of each model. Meyer and Li do agree however, that psychophysical experiments involving a large set of images would be needed for a complete evaluation, to investigate the performance of models under a wider range of conditions.

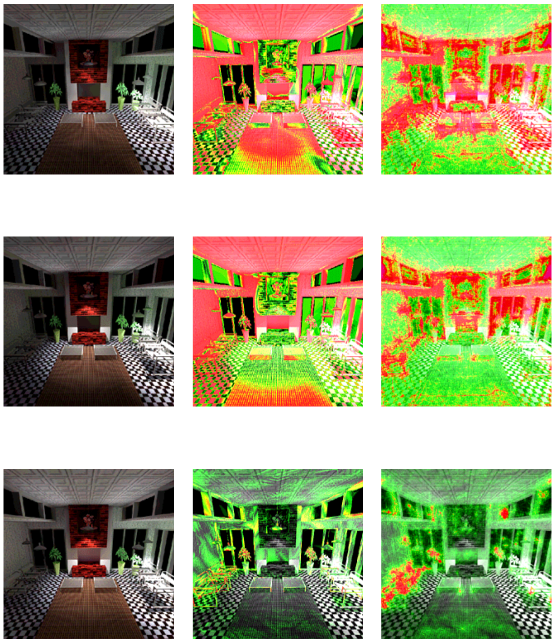

Figure 9: *Perceptual convergence of the image quality* [26]

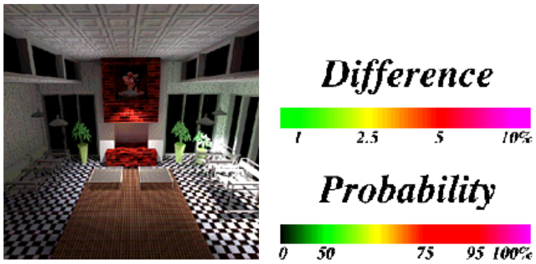

Figure 10: *Fully Converged Image, and Perceptual Scales* [26]

Myszkowski[26] realised the VDP had many potential applications in realistic image synthesis. He completed a comprehensive validation and calibration of VDP response via human psychophysical experiments. He subsequently used the VDP local error metric to steer decision making in adaptive mesh subdivision, and in isolating regions of interest for more intensive global illumination computations, Figures 9, 10. The VDP was tested to determine how close VDP predictions come to subjective reports of visible differences between images by designing two human psychophysical experiments. Results from these experiments showed a good correspondence between human observations and VDP results.

These perception based image quality metrics have demonstrated the success of implementing a visual model, in spite of the fact that knowledge of the visual process is as yet incomplete.

4. Perceptually driven rendering

Even for realistic image synthesis there may be little point spending time or resources to compute detail in an image that would not be detected by a human observer. By eliminating any computation spent on calculating image features which lie below the threshold of visibility, rendering times can be shortened leading to more efficient processing. Because the chief objective of physically based rendering is *realism*, incorporating models of HVS behaviour into rendering algorithms can improve performance, as well as improving the quality of the imagery produced. So by taking advantage of the limitations of the human eye, just enough detail to satisfy the observer can be computed without sacrificing image quality. Several attempts have been made to develop image synthesis algorithms that detect threshold visual difference and direct the algorithm to work on those parts of an image that are in most need of refinement.

Raytracing produces an image by computing samples of radiance, one for each pixel in the image plane. Producing an anti-aliased image is difficult unless very high sampling densities are used. Mitchell [25] realised that deciding where to do extra sampling can be guided by knowledge of how the eye perceives *noise as a function of contrast and colour*. Studies have shown that the eye is most sensitive to noise in intermediate frequencies [34]. While frequencies of up to 60 cycles per degree (cpd) can be visible, the maximum response to noise is at approximately 4.5 cpd, so sampling in regions with frequency above this threshold can be minimised, without affecting the visual quality of the image. Mitchell begins by sampling the entire image at low frequency then uses an adaptive sample strategy on the image according to the frequency content. This results in a non uniform sampling of the image, which enables aliasing noise to be channelled into high frequencies where artefacts are less conspicuous. However, non-uniform sampling alone doesn't eliminate aliasing, just changes its characteristics to make it less noticeable. Mitchell applies two levels of sampling. To decide whether the high sampling density should be invoked the variance of samples could be used [14], but this is a poor measure of visual perception of local variation. Instead Mitchell chooses to use contrast to model the non-linear response of the eye to rapid variations in light intensity:

$$C = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$

As each sample consists of three separate intensities for red, green and blue, three separate contrasts can be computed for each of them. These three contrasts are tested against separate thresholds, 0.4, and 0.3 and 0.6 for red, green and blue respectively, and supersampling is done if any one exceeds the threshold. The contrast metric is then used to determine when the high sampling density should be invoked. This test is most sensitive to green in accordance with the human eye's response to noise as a function of colour. Multi stage filters are then used to reconstruct the non-uniform samples into a digital image. Although this idea has the beginnings of a perceptual approach, it is at most a crude approximation to the HVS. Only two levels of sampling are used and it doesn't account for visual masking [†].

The HVS exhibits different spatial acuities in response to different colours. Evidence exists that colour spatial acuity is less than monochrome spatial acuity. Exploiting this *poor colour spatial acuity* of the HVS, Meyer and Liu [23] developed an adaptive image synthesis algorithm that uses an opponents processing model of colour vision [22] comprising chromatic and achromatic colour channels. Using a Painter and Sloan [28] adaptive subdivision, a k-D tree representation of the image is generated. Areas of the image containing high frequency information are stored at the lower levels of the tree. They then modified a screen subdivision raytracer to limit the depth to which the k-D tree must be descended to compute the chromatic colour channels. The limit is determined by psychophysical results describing the colour spatial frequency. They achieved a modest saving in computational effort and showed, using a psychophysical experiment, that decreasing the number of rays used to produce the chromatic channels had less of an effect on image quality than reducing the number of rays used to create the achromatic channels. This was the first work to attempt to minimise the computation of colour calculations, as opposed to just decreasing costly object intersection calculations.

Bolin and Meyer [2] took a frequency based approach to raytracing, which uses a simple vision model, making it possible for them to control how rays are cast in a scene. Their algorithm accounts for the *contrast sensitivity, spatial frequency* and *masking properties* of the HVS. The contrast sensitivity response of the eye is non-linear. So, when deciding where rays should be cast, the algorithm deems a luminance difference at low intensity to be of greater importance than the same luminance difference at high intensity. The spatial response of the HVS is known to be less for patterns of pure colour than for patterns that include luminance differences. This means that it is possible to cast fewer rays into regions with colour spatial variations than are cast in regions with spatial frequency variations in luminance. Finally, it is known that the presence of high spatial frequency can mask the presence of other high frequency information (masking). When used in conjunction with a Monte Carlo raytracer, more rays are spawned when low frequency terms are being determined than when high frequency terms are being found. Using this strategy, the artefacts that are most visible in the scene can be eliminated from the image first, then noise can be channelled into areas of the image where artefacts are less conspicuous. This technique is an improvement on Mitchell's method because the vision model employed accounts for *contrast sensitivity, spatial frequency* and *masking.*

Despite the simplicity of the vision models used in these approaches, the results are promising, especially as they demonstrate the feasibility of embedding HVS models into the rendering systems to produce more economical systems without forfeiting image quality. Fuelled by the notion that more sophisticated models of the HVS would yield even greater speedup, several researchers began to introduce more complex models of the HVS into their global illumination computations.

Myszkowski [26] applied a more sophisticated vision model to steer computation of a Monte Carlo based raytracer. Aiming to take maximum advantage of the limitations of the HVS, his model included *threshold sensitivity, spatial frequency sensitivity and contrast masking.* A perceptual error metric is built into the rendering engine allowing adaptive allocation of computation effort into areas where errors remain above perceivable thresholds and allowing computation to be halted in all other areas (i.e. those areas where errors are below the perceivable threshold and thus not visible to a human observer). This perceptual error metric takes the form of Daly's [6] Visible Differences Predictor (VDP), discussed in Section 3.

Bolin and Meyer [1] devised a similar scheme, also using a sophisticated vision model, in an attempt to make use of all HVS limitations. They integrated a simplified version of the Sarnoff Visible Discrimination Model (VDM) into an image synthesis algorithm to detect threshold visible differences and, based on those differences direct subsequent computational effort to regions of the image in most need of refinement. The VDM takes two images, specified in CIE XYZ colour space, as input. Output of the model is a Just Noticeable Difference (JND) map. One JND corresponds to a

[†] The presence of high spatial frequency in an image can mask the presence of other high frequency information

75% probability that an observer viewing the two images would detect a difference [16]. They use the upper and lower bound images from the computation results at intermediate stages and used the predictor to get an error estimate for that stage. The image quality model is used to control where to take samples in the image, and also to decide when enough samples have been taken across the entire image, providing a visual stopping condition. A more comprehensive description of the VDM is given in Section 3.

Applying a complex vision model at each consecutive time step of image generation requires repeated evaluation of the embedded vision model. The VDP can be expensive to process due to the multi-scale spatial processing involved in some of its components. This means that in some cases the cost of recomputing the vision model may cancel the savings gained by employing the perceptual error metric to speed up the rendering algorithm. To combat this, Ramasubramanian [31] introduced a metric that handles luminance-dependent processing and spatially-dependent processing independently, allowing the expensive spatially-dependent component to be *precomputed*. Ramasubramanian developed a physical error metric that predicts the *perceptual* threshold for detecting artefacts in the image. This metric is then used to predict the sensitivity of the HVS to noise in the indirect lighting component. This enables a reduction in the number of samples needed in areas of an image with high frequency texture patterns, geometric details, and direct lighting variations, giving a significant speedup in computation.

Using *validated* image models that predict image fidelity, programmers can work toward achieving greater efficiencies in the knowledge that resulting images will still be faithful visual representations. Also in situations where time or resources are limited and fidelity must be traded off against performance, perceptually based error metrics could be used to provide insights into where computation could be economised with least visual impact.

In addition to TMOs being useful for rendering calculated luminance to the screen, they are also useful for giving a measure of the perceptible difference between two luminances at a given level of adaptation. This function can then be used to guide algorithms, such as discontinuity meshing, where there is a need to determine whether some process would be noticeable or not to the end user.

Gibson and Hubbold [10] have used features of the *threshold sensitivity* displayed by the HVS to accelerate the computation of radiosity solutions. A perceptually based measure controls the generation of view independent radiosity solutions. This is achieved with an *a-priori* estimate of real-world adaptation luminance, and uses a TMO to transform luminance values to display colours and is then used as a numerical measure of their perceived difference. The model stops patch refinement once the difference between successive levels of elements becomes perceptually unnoticeable. The perceived importance of any potential shadow falling across a surface can be determined, this can be used to control the number of rays cast during visibility computations. Finally, they use perceptual knowledge to optimise the element mesh for faster interactive display and save memory during computations. This technique was used on the adaptive element refinement, shadow detection, and mesh optimisation portions of the radiosity algorithm.

Discontinuity meshing is an established technique used to model shadows in radiosity meshes. It is computationally expensive, but produces meshes which are far more accurate and which also contain fewer elements. Hedley [11] used a perceptually informed error metric to optimise adaptive mesh subdivision for radiosity solutions, the goal being to develop scalable discontinuity meshing methods by considering visual perception. Meshes were minimised by discarding discontinuities which had a negligible *perceptible* effect on a mesh. They demonstrated that a perception-based approach results in a greater reduction in mesh complexity, without introducing more visual artefacts than a purely radiometrically-based approach.

5. Comparing Real and Synthetic Scenes

There is a fundamental problem with the image quality metrics describe in section 3 from the point of view of *validation*. Although these methods are capable of producing images based on models of the HVS, there is no standard way of telling if the images “capture the visual appearance” of scenes in a meaningful way. One approach to validation could compare observers’ perception and performance in real scenes against the predictions of the models. This would enable calibration and validation of the models to assess the level of fidelity of the images produced.

Using perceptual data we can compare and validate existing rendering algorithms, allowing us to demonstrate to the world just how useful and reliable the images we create can be. *Psychophysics* is one approach to evaluating, comparing and validating synthetic imagery to *real* images occurring in our physical surroundings.

While image quality metrics have been successfully incorporated into global illumination algorithms to guide computations more efficiently, metrics can also be useful to validate and compare rendering techniques. As the goal of realistic image synthesis is to generate representations of a physical scene, simulations should therefore be compared to the real world scenes.

Using a simple five sided cube as their test environment, Meyer et al. [24] presented an approach to image synthesis comprising separate physical and perceptual modules. They chose diffusely reflecting materials to build a physical test model. Each module is verified using experimental techniques. The test environment was placed in a small dark room. Radiometric values predicted using a radiosity lighting simulation of a basic scene are compared to physical measurements of radiant flux densities in the real scene. Then the results of the radiosity calculations are transformed to the RGB values for display, following the principles of colour science. Measurements of irradiation were made at 25 locations in the plane of the open face for comparison with the simulations. Results show that irradiation is greatest near the centre of the open side of the cube. This area provides the best view of the light source and other walls. In summary, there is good agreement between the radiometric measurements and the predictions of the lighting model.

Meyer et al. then proceeded by transforming the validated simulated values to values displayable on a television monitor. Twenty participants were asked to differentiate between a real environment and the displayed image, both of which were viewed through the back of a view camera. They were asked which of the images was the real scene. Nine out of the twenty participants (45%) indicated that the simulated image was actually the real scene, i.e. selected the wrong answer, revealing that observers would have done just as well by simple guessing. Although participants considered the overall match and colour match to be good, some weaknesses were noticed in the sharpness of the shadows (a consequence of the discretisation in the simulation) and in the brightness of the ceiling panel (a consequence of the directional characteristics of the light source). The overall agreement lends strong support to the perceptual validity of the simulation and display process. This was the first attempt to compare real and simulated scenes side by side, using human observers.

Although the results of the study are encouraging, there are some drawbacks with this approach: The scene under examination was very simple, the methodology for comparison itself was not inherently controlled, and the results suggest that the participants could have simply guessed. To really investigate the differences between a real environment and its synthetic representation, a more robust approach is required.

By conducting a series of psychophysical experiments McNamara et al. [19, 18, 21, 20] demonstrated how the fidelity of graphical reconstructions of a real scene can be assessed. The study was based on the simple task of lightness perception.

McNamara et al. [19] began by building an experimental framework to facilitate human comparison between real and synthetic scene. They ran a series of psychophysical experiments in which human observers were asked to compare simple two dimensional target regions of a real physical scene with regions of the computer generated representation of that scene. The comparison involved lightness judgments in both the generated image and the real scene. Results from these experiments showed that the visual response to the real scene and a high fidelity rendered image was similar. They then extended this work to investigate comparisons using three dimensional objects as targets, rather than simple regions. This allows examination of scene characteristics such as shadow, object occlusion and depth perception.

The test environment was a five sided box shown in figure 11. Several objects that were placed within the box for examination.

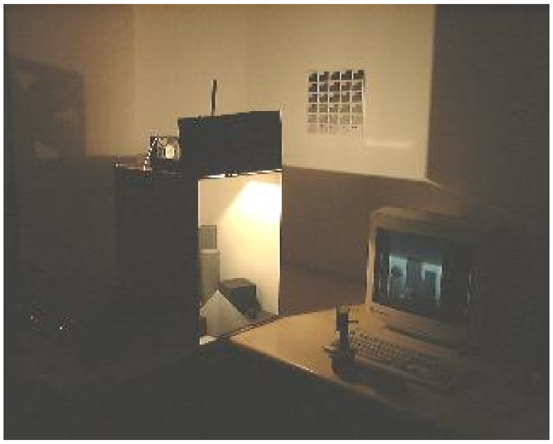

Figure 11: *The test environment showing real environment and computer image.*

Ten images were considered for comparison to the real scene, they included a digital photograph, a series of Radiance [40] images, and a couple of Renderpark images [32] as shown in figures 12, 13,14, 15,16.

Each participant was presented with a series of images, shown in figures 12 through 16, in a random order, in addition to the real environment. Participants were not explicitly informed which image was the physical environment. The images presented were the real scene, the photograph and the nine rendered images. There were seventeen different objects in the test environment, subjects were also asked to match

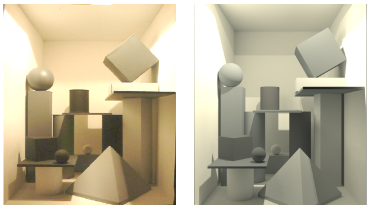

Figure 12: *Digital Photograph (left) Radiance Two Ambient Bounces (right)* [18]

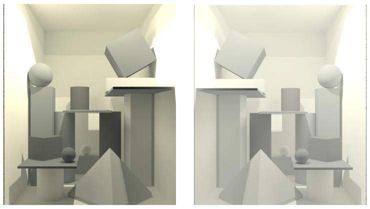

Figure 13: *Radiance Eight Ambient Bounces (left) brightened (right)* [18]

each of the five sides of the environment (floor, ceiling, left wall, back wall and right wall) giving a total of twenty-two matches. Participants were asked to judge the lightness of target objects in a random manner.

In summary, the results show that there is evidence that the Two Ambient Bounces image, the Default image, the Controlled Error Materials image, the Raytraced image and the Radiosity image are perceptually degraded compared to the photograph. However, there is no evidence that the others images in this study are perceptually inferior to the photograph. From this they conclude that the Eight Ambient Bounces image, the brightened Eight Ambient Bounces image, the Tone Mapped image and the Controlled Error Illumination image are of the same perceptual quality as a photograph of the real scene.

The results from such psychophysical studies are becoming increasingly important for realistic image synthesis as these results provide a *perceptual*, rather than mere physical, match between an original scene and its computer generated counterpart. This information can then be used for image evaluation, as well as for comparison of various global illumination simulation algorithms and ultimately can be used to improve the efficiency of such algorithms.

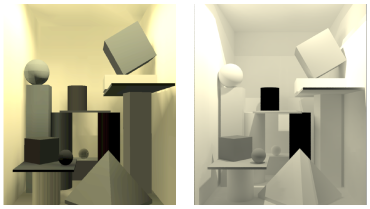

Figure 14: *Radiance Default (left) Estimated Materials (right)* [18]

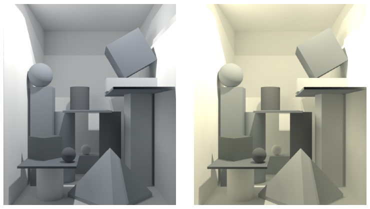

Figure 15: *Radiance Estimated Light Source (left) Tone Mapped (right)* [18]

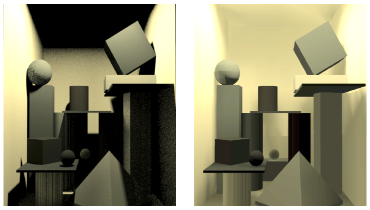

Figure 16: *Renderpark Raytraced (left) Radiosity (right)* [18]

6. Summary

Some of the applications of visual perception in computer graphics were explored. For many applications computer imagery should not only be physically correct but also perceptually equivalent to the scene it represents. Knowledge of the HVS can be employed to greatly benefit the synthesis of realistic images at various stages of production. Global illumination computations are costly in terms of computation. There is a great deal of potential to improve the efficiency of such algorithms by focusing computation on the features of a scene which are more conspicuous to the human observer. Those features that are below perceptual visibility thresholds have no impact on the final solution, and therefore can be omitted from the computation, increasing efficiency without causing any perceivable difference to the final image. Perceptual metrics involving advanced HVS models can be used to determine the visible differences between a pair of images. These metrics can then be used to compare and evaluate image quality. They can also be used within the rendering framework to steer computation into regions of an image which are in most need of refinement, and to halt computation when differences in successive iterations of the solution become imperceptible.

Future applications will require perceptual accuracy in addition to physical accuracy. Without perceptual accuracy it is impossible to assure users of computer graphics that the generated imagery is anything like the scene it depicts. Imagine a visualisation of an architectural design, without perceptual accuracy it is difficult to guarantee the architect that the visualisation sufficiently represents their design, and that the completed building will look anything like the computer representation. This chapter discussed how knowledge of the HVS is being incorporated at various stages in the image synthesis pipeline. The problem is that much of the data used has been obtained from specific psychophysical experiments which have been conducted in specialised laboratory environments under reductionistic conditions. These experiments are designed to examine a single dimension of human vision, however, evidence exists to indicate that features of the HVS do not operate individually, but rather functions overlap and should be examined as a whole rather than in isolation. TMOs map computed radiance values to display values in a manner that preserves perception of the original scene. TMOs produce a perceptual match between the scene and the image in the hopes that the image may be used predictively.

There is a strong need for the models of human vision currently used in image synthesis computations to be *validated* to demonstrate their performance is comparable to the actual performance of the HVS.

References

1. M. R. Bolin and G.W. Meyer, *A perceptually based adaptive sampling algorithm*, ACM SIGGRAPH '98 Conference Proceedings, 1998, pp. 299–310.

2. M.R. Bolin and G.W. Meyer, *A frequency based ray tracer*, ACM SIGGRAPH '95 Conference Proceedings, 1995, pp. 409–418.

3. K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, and K. Zimmerman, *Spatially Nonuniform Scaling Functions for High Contrast Images*, Proceedings of Graphics Interface '93 (San Francisco, CA), Morgan Kaufmann, May 1993, pp. 245–253.

4. Technical Committee 3.1 CIE 19/2.1, 1981, An Analytical Model for Describing the Influence of Lighting Parameters upon Visual Performance.

5. S. Daly, *The visible difference predictor: an algorithm for the assessment of ima ge fidelity*, In A. B. Watson Editor, Digital Images and Human Vision, MIT Press, 1993, pp. 179–206.

6. S. Daly, *The Visible Differences Predictor: An algorithm for the assessment of image fidelity*, Digital Image and Human Vision (A.B. Watson, ed.), Cambridge, MA: MIT Press, 1993, pp. 179–206.

7. J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg, *A model of visual adaptation for realistic image synthesis*, Computer Graphics **30** (1996), no. Annual Conference Series, 249–258.

8. A. Gaddipatti, R. Machiraju, and R. Yagel, *Steering image generation with wavelet based perceptual metric*, Computer Graphics Forum (Eurographics '97) **16** (1997), no. 3, 241–251.

9. J. Gervais, Jr. L.O. Harvey, and J.O. R.s, *Identification confusions among letters of the alphabet*, Journal of Experimental Psychology: Human Perception and Perfor mance, vol. 10(5), 1984, pp. 655–666.

10. S. Gibson and R. J. Hubbold, *Perceptually-driven radiosity*, Computer Graphics Forum **16** (1997), no. 2, 129–141.

11. D. Hedley, *Discontinuity meshing for complex environments*, Ph.D. thesis, Department of Computer Science, University of Bristol, Bristol, UK, August 1998.

12. G. Ward Larson, H. Rushmeier, and C. Piatko, *A Visibility Matching Tone Reproduction Operator for High Dynamic Range Scenes*, IEEE Transactions on Visualization and Computer Graphics **3** (1997), no. 4, 291–306.

13. G. Ward Larson and R. Shakespeare, *Rendering with radiance: The art and science of lighting visualization*, Morgan Kaufmann, San Francisco, CA, 1998, ISBN 1-55860-499-5.

14. M. E. Lee, R. A. Redner, and S. P. Uselton, *Statistically optimized sampling for distributed ray tracing*, Computer Graphics (SIGGRAPH '85 Proceedings) **19** (1985), no. 3, 61–67.

15. J. Lubin, *A visual discrimination model for imaging system design and development*, Vision models for target detection and recognition (Peli E., ed.), World Scientific, 1995, pp. 245–283.

16. J. Lubin, *A human vision model for objective picture quality measurements*, Conference Publication No. 447, IEE International Broadcasting Convention, 1997, pp. 498–503.

17. J. L. Mannos and D. J. Sakrison, *The effects of a visual criterion on the encoding of images*, IEEE Transactions on Information Theory **IT-20** (1974), no. 4, 525–536.

18. A. McNamara, A. Chalmers, T. Troscianko, and I. Gilchrist, *Evaluating images using human lightness judgements*, Proceedings of the 11th Euro-

graphics Rendering Workshop, Springer Verlag, June 2000.

19. A. McNamara, A. Chalmers, T. Troscianko, and E. Reinhard, *Fidelity of graphics reconstructions: A psychophysical investigation*, Proceedings of the 9th Eurographics Rendering Workshop, Springer Verlag, June 1998, pp. 237–246.

20. A. McNamara and Alan Chalmers, *Image quality metrics*, SIGGRAPH 2000 Image Quality Metrics Course Notes, ACM SIGGRAPH, July 2000.

21. A. McNamara, Alan Chalmers, and Tom Troscianko, *Evaluating image quality metrics v human evaluation*, Visual Proceedings, Technical Sketch at ACM SIGGRAPH 2000, 2000.

22. G. W. Meyer, *Wavelength selection for synthetic image generation*, Computer Vision, Graphics, and Image Processing **41** (1988), no. 1, 57–79.

23. G. W. Meyer and A. Liu, *Color spatial acuity control of a screen subdivision image synthesis algorithm*, Human Vision, Visual Processing, and Digital Display **1666** (1992), no. 3, 387–399.

24. G. W. Meyer, H. E. Rushmeier, M. F. Cohen, D. P. Greenberg, and K. E. Torrance, *An Experimental Evaluation of Computer Graphics Imagery*, ACM Transactions on Graphics **5** (1986), no. 1, 30–50.

25. D. P. Mitchell, *Generating antialiased images at low sampling densities*, Computer Graphics **21** (1987), no. 4, 65–72.

26. K. Myszkowski, *The visible differences predictor: Applications to global illumination problems*, Rendering Techniques '98 (Proceedings of Eurographics Rendering Workshop '98) (New York, NY) (G. Drettakis and N. Max, eds.), Springer Wien, 1998, pp. 233–236.

27. K. Myszkowski, A. B. Khodulev, and E. A. Kopylov, *Validating global illumination algorithms and software*, Visual Proceedings, Technical Sketch at ACM Siggraph'97, 1997, p. 156.

28. J. Painter and K. Sloan, *Antialiased ray tracing by adaptive progressive refinement*, Computer Graphics (SIGGRAPH '89 Proceedings) (Jeffrey Lane, ed.), vol. 23,3, July 1989, pp. 281–288.

29. Stephen Palmer, *Vision science: From photons to phenomenology*, Bradford Books/MIT Press, Cambridge, MA, 1998, ISBN 1-55860-499-5.

30. J. Prikryl, 1999, http://www.cs.kuleuven.ac.be/ graphics/SEMINAR/program.html.

31. M. Ramasubramanian, S.N. Pattanaik, and D.P. Greenberg, *A perceptually based physical error metric for realistic image synthesis*, Proceedings of SIGGRAPH 99 (August 1999), 73–82.

32. Renderpark, 1999, http://www.cs.kuleuven.ac.be/g̃raphics/.

33. H. Rushmeier, G. Ward, C. Piatko, P. Sanders, and B. Rust, *Comparing real and synthetic images: Some ideas about metrics*, Eurographics Rendering Workshop 1995, Eurographics, June 1995.

34. D. J. Sakrison, *On the role of the observer and a distortion measure in image transmission*, IEEE Trans. Commun. **25** (1977), no. 11, 1251–1267.

35. C. Schlick, *An inexpensive BRDF model for physically-based rendering*, Computer Graphics Forum **13** (1994), no. 3, C/233–C/246.

36. S.S. Stevens and J.C. Stevens, *Brightness function: Effects of adaptation*, Journal of the Optical Society of America, vol. 53, March 1963, pp. 375–385.

37. J. Tumblin and H. E. Rushmeier, *Tone reproduction for realistic images*, IEEE Computer Graphics and Applications **13** (1993), no. 6, 42–48.

38. Jack Tumblin, *Three methods of detail-preserving contrast reduction for displayed images*, Phd, Georgia Institute of Technology, http://www.cc.gatech.edu/gvu/people/jack.tumblin/, 1999.

39. Jack Tumblin and Greg Turk, *LCIS: A boundary hierarchy for detail-preserving contrast reduction*, Siggraph 1999, Computer Graphics Proceedings (Los Angeles) (Alyn Rockwood, ed.), Addison Wesley Longman, 1999, pp. 83–90.

40. G. J. Ward, *The RADIANCE lighting simulation and rendering system*, Proceedings of SIGGRAPH '94 (Orlando, Florida) (A. Glassner, ed.), Computer Graphics Proceedings, Annual Conference Series, July 1994, pp. 459–472.