Psychophysical Validation of the Visible Differences Predictor for Global Illumination Applications

William L. Martens and Karol Myszkowski
The University of Aizu; Aizu-Wakamatsu 965-8580; Japan

Abstract

The perceptually-based Visual Difference Predictor (VDP) developed by Daly [3] has many potential applications in realistic image synthesis. However, systematic validation and subsequent calibration of the VDP response via human psychophysical experiments should be completed before integrating VDP calculation into image synthesis algorithms such as those in global illumination computations. For example, the VDP local error metric can guide decision making in adaptive mesh subdivision, and in selecting regions of interest for more intensive global illumination computations. In this study, we designed two human psychophysical experiments to test whether VDP predictions match well with subjective reports of visible difference between images under conditions mimicking those in our VDP applications. These experiments showed a good match with VDP predictions for shadow and lighting pattern masking by texture, and in comparisons of the perceived quality of images generated at progressive stages of an indirect lighting solution.

CR Categories and Subject Descriptors: I.3.0 [Computer Graphics]: General; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism.

Additional Keywords: spatial vision, visual perception, masking, image quality, error metrics.

1 INTRODUCTION

Many improvements in the computational efficiency of computer graphics have resulted from capitalization upon human perceptual capacities and insensitivities. This is because of the expected dramatic performance gains attained by focusing computations only on those scene features readily perceived by human observers under given viewing conditions; and conversely, by omitting from image synthesis the computation of scene features that will not make significant differences in image appearance.

Our interests lay in combining perception-driven refinement of global illumination algorithms with realistic rendering techniques. We were motivated by the fact that commonly used energy-based metrics of lighting solution errors inadequately predict the perceived quality of lighting reconstruction for the human observer [7]. For example, mean-squared error metrics of the differences between images poorly correspond to actual perceived differences [3, 4, 5, 7]. Because of difficulties in formulating meaningful perceptual error metrics on the atomic level of every light interaction used by global illumination techniques, we chose to use perception-based measures of differences between rendered images to provide summary information on lighting computation performance.

After an extensive analysis of image difference measures used in a wide range of applications (e.g., evaluating lossy image compression and dithering algorithms, designing CRT and flat-panel displays, and generating computer graphics) we selected the Visible Differences Predictor (VDP) developed by Daly [3]. The VDP takes into account important visual characteristics such as a Weber's-law-like amplitude compression and pattern masking, and includes an advanced Contrast Sensitivity Function model. Also, the Cortex transform used by the VDP is a multi-resolution, pyramid-style, invertible, and computationally-efficient image representation. One of the most outstanding characteristics of the VDP is prediction of local differences between images at the pixel level (as discussed below in reference to Figure 2). This paves the way for many improvements of global illumination algorithms such as reinforcing computations in selected scene regions or guiding adaptive mesh subdivision.

Since the VDP is a general purpose predictor of the differences between images, we chose to test its performance in representative global illumination tasks for which the VDP should provide potential improvements. Our approach was to collect subjective judgments from human observers to pairs of images presented on a high-quality CRT display under controlled viewing conditions. The experimental subjects were requested to rank on a scale from 1 to 9 the perceived global difference between each of a pair of images. Within each session subjects could freely scroll the display and adjust their ranking. The stimuli were presented via HTML forms that allowed this interactive reporting of the subjects' judgments. [1]

[1] In order for the authors to obtain more normative results, under albeit less tight experimental control, we would like to invite the reader to take part in our continuing VDP experiments by visiting our website [1].

2 SHADOW MASKING BY TEXTURE

The first experiment was motivated by an observation that the traditional adaptive mesh subdivision algorithms typically used for lighting-pattern reconstruction often lead to excessive mesh splitting. In most cases, those algorithms have not taken into account the background texture onto which the lighting pattern is cast. In our ongoing research, complex shadow patterns to be used as targets are computed at two stages of adaptive mesh subdivision, and cast upon textures of various sizes and patterns. The reader is directed to our website for a comprehensive visual summary of these images [1]. In this experiment, we quantitatively evaluated the masking of the shadow pattern that occurred when it was cast upon a background texture, as shown in Figures 1a and 1b for two selected texture scales. Seven different texture scales were viewed by the 11 subjects. These subjects generated numerical estimates of global difference between images for each of the seven pairs of images, which pairs differed only in texture scale. The compared pair of shadow patterns computed at subsequent stages of the adaptive mesh subdivision was held constant within a given experimental session. The greatest shadow masking was expected for similar scales of major texture and shadow elements.

Figure 1c shows the mean standardized ratings (square symbols) of the visible differences between the two shadow patterns when these are cast upon a background texture at 7 different scalings. (The "background texture scale" of 4 and 16 on the abscissa in Figure 1c correspond to the textures shown in Figures 1a and 1b, respectively.) In order to remove individual differences in response bias, both in the subjects' criterion for central tendency and for the response range used, the ratings from a given subject were standardized by subtracting their mean response in the session, and dividing by the standard deviation of their responses. This operation projects the subjects' responses into a common space with mean of zero and standard deviation of one. Standardizing VDP predictions in the same way allows us to compare directly the obtained subjective ratings with predictions based upon an arbitrarily-set percentage threshold for detection of differences at the pixel level. These standardized VDP values (shown in Figure 1c as circular symbols) almost always lay within one standard error of the mean standardized rating for a given background texture scale.

The goodness-of-fit relation between VDP results and subjective ratings can be captured in a single coefficient of determination $R^2 = .83$, that describes the proportion of the variance in subjective ratings accounted for by the variance in the VDP predictions. Note that such a relatively high value can only be observed in the results of experiments in which the modulation in the subjective ratings over the independent variable (in this case, background texture scale) well exceeds the standard error of the means. Though this point is obvious in the context of statistical analysis, it is a point that might fail to be considered during the design phase of psychophysical validation experiments. In a masking study such as that reported here, the variation in the extent of masking should be made great enough to allow the model to be tested, but not so great that the subjective responses are dominated by extreme ratings (in which case the observed coefficient of determination will be artificially inflated, and the validity of the test will be questionable).

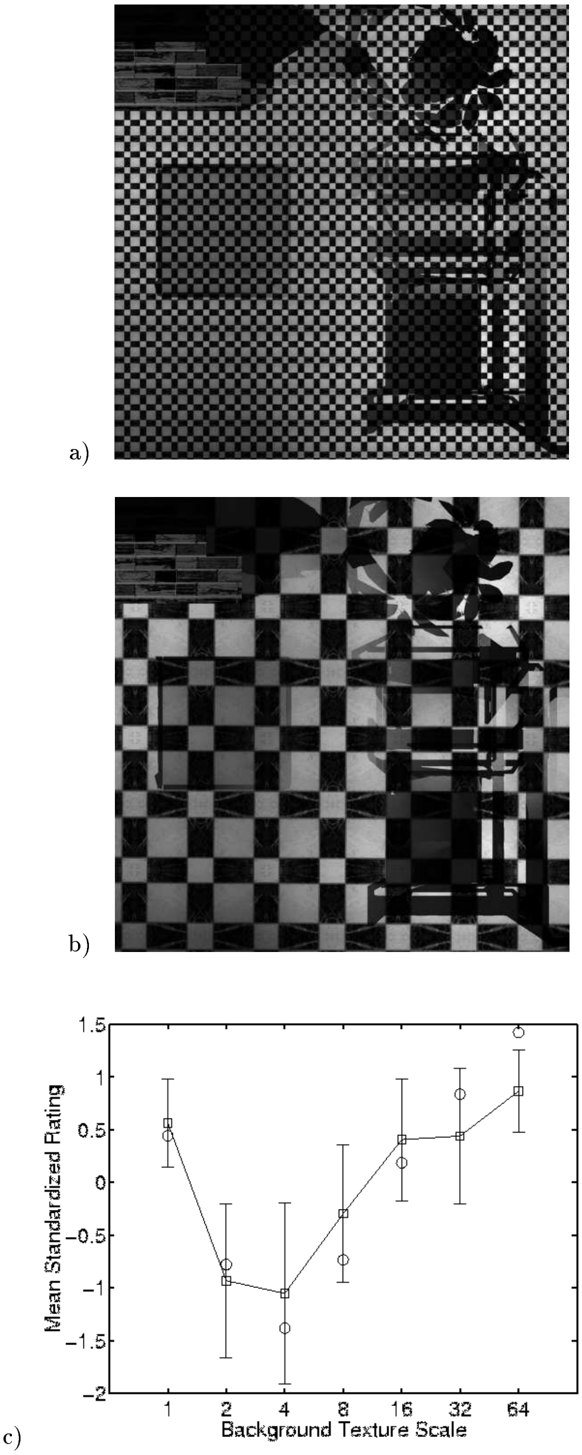

Figure 1: Shadow masking by texture: a) and b) test image examples for various texture scales, c) the mean standardized ratings (squares) from 11 human subjects for 7 texture scales shown with corresponding VDP predictions (circles).

3 ESTIMATING PROGRESSION OF INDIRECT LIGHTING SOLUTIONS

In this experiment we compared VDP responses with those obtained from human subjects for a series of image pairs resulting from successive refinement in a progressive hierarchical radiosity solution. The focus was upon quantitative measures of the "perceptual" convergence of global illumination solutions. For the subject, this meant comparing the final image, or fully-converged solution, with images taken from the sequence of images generated during successive stages of computation. As in the above experiment, the VDP also was used to compute predictions of the perceivable differences between those same pairs of images, this time generated using hierarchical progressive radiosity. (only ten images were chosen from the complete sequence). We are ultimately interested in determining whether the VDP can be used to decide upon the calculation stopping condition, using its prediction of the point in the sequence at which the change in image quality would become negligible for the human observer. This experiment does not test for that possibility, since images were compared to the final image, rather than the preceding image.

Figures 2a through 2d show two representative stages of refinement in the progressive hierarchical radiosity solution, and the related VDP results (see figure caption for details). [2] Figure 2g shows the obtained standardized mean ratings (square symbols) for the differences between such images and the image resulting from the final converged solution shown in Figure 2f. Even better agreement was observed in this experiment than that in the previous experiment (as indicated by the high coefficient of determination, $R^2 = .93$, for the relation between VDP results and subjective ratings). As before, the standardized VDP values (circular symbols) almost always lay within one standard error of the mean standardized rating. This means that as the progressive radiosity solution converged, close agreement between the VDP predictions and the subjective judgments was maintained.

Ranking images taken from the final stages of computation (70–200 minutes) was considerably more difficult than ranking those taken from the intermediate stages. This naturally resulted from the fact that these images did indeed appear quite similar to one another. It is not surprising that a similar pattern of results was observed in the corresponding VDP responses, since the image generated after 70 minutes of computation already exhibited negligible differences in comparison to the final convergence. Even though more stages of convergence were included in this experiment than were strictly required to validate the model, the observed match between VDP-predicted and subjectively-reported image differences throughout the final stages of computation serves to legitimize the use of the VDP values to determine a stopping point in radiosity computations at a stage in those computations well before a conservative criterion energy error value has been reached. Such psychophysical validation of VDP use suggests that decisions could be based upon what might be termed "perceptual" convergence in image quality rather than "physical" convergence.

[2] On our website [2] we posted images in non-lossy format matching those used in our experiments. Also, we provide a visualization of the VDP responses for our test data. Once again, the reader is encouraged to visit our website to participate in our ongoing experiments regarding visible differences. We also encourage other researchers to utilize these same images in tests of different visible differences predictors.

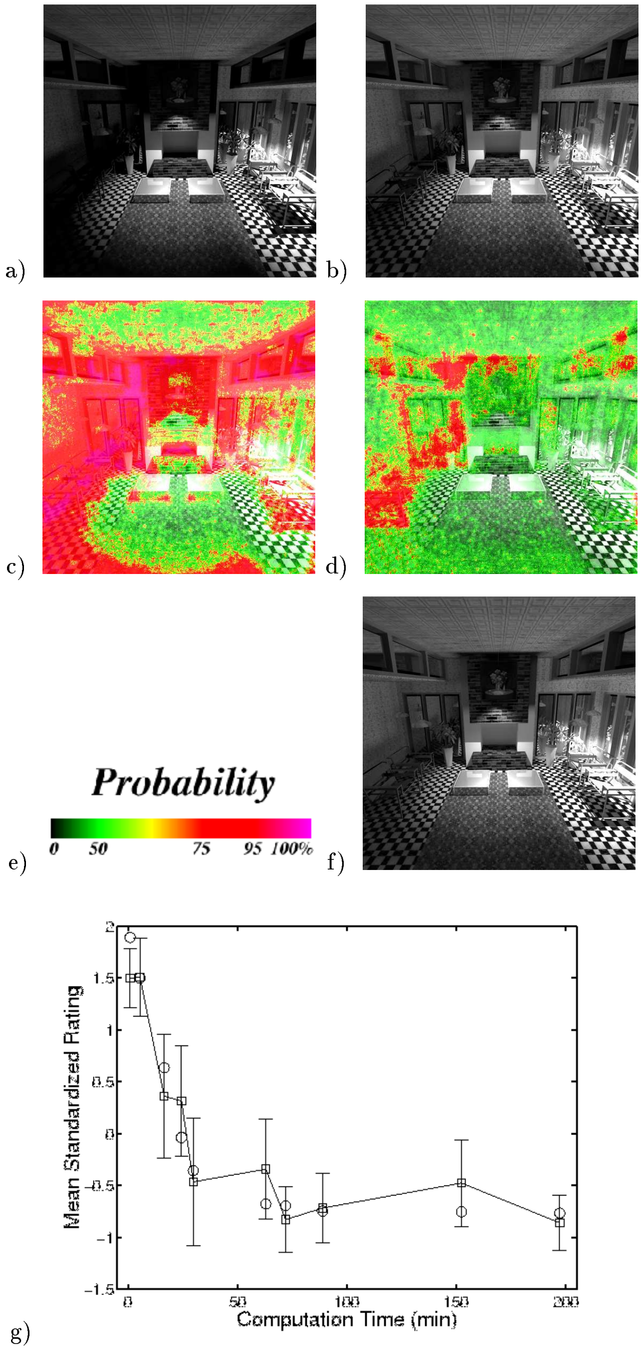

Figure 2: Comparison of various stages in "perceptual" convergence of the indirect lighting simulation. The progressive radiosity solution after **a** 1 and **b** 30 minutes of computation. The corresponding VDP predicted differences **c** and **d** (refer to [2] for color images) in respect to the fully converged solution **f**. Color scales for encoding the probabilities of detecting the differences in **c** and **d** are shown in **e**. The standardized mean ratings (squares) at each of 10 cumulative computation times are shown in **g** along with corresponding VDP predictions (circles).

4 DISCUSSION

Global illumination applications such as these need refined methods for answering the question "how much is enough?" It is likely that the search for better perceptually-based metrics for comparing images will continue, since many applications can benefit from such tools (in terms of validity, efficiency, reliability, etc.). The questions we would like to raise for discussion are concerned with the context within which such "perceptually-informed" tools are developed and used. Here we only attempt to outline these important questions, since we cannot hope to provide adequate answers at this stage of research.

The first questions we wish to raise are the most basic: How can we best apply the results of VDP to given applications? We would also like to inquire into the criteria that would allow us to say when VDP will be useful, and when it will not be so useful. To begin to answer these questions, it is most important to understand what VDP is and what it is not. VDP is not a model of image quality, but only a predictor of differences in image appearance. If image quality is the issue, the VDP user must establish a reference image that sets the standard for high quality, and then regard deviations from this standard as reductions is quality. This approach makes sense in testing lossy compression algorithms, for example, but it could fail in an application in which the goal is the synthesis of a high quality image. In the shadow-masking experiment described in this paper, the application of VDP results was straightforward, since we were concerned only with the change in visibility of details within a target image. The application motivating our second experiment, however, was more problematic. For example, we cannot guarantee that a human viewer would declare the final image obtained at full convergence to be the best standard of reference. Also, there is the possibility of reaching early local minima in the change in VDP output during the progression, though such was not apparent in this particular study.

Other questions we wish to raise concern how to evaluate the adequacy of a particular VDP deployment. Perhaps we are relatively confident in the image processing component of the VDP model, but the final stage of processing in which predictions are generated is not so easily fit to a given application. There is a large gap between generating a map of pixel-to-pixel differences, and predicting the global response of a human viewer confronted with a question about the images of interest. This issue presents itself most clearly when one attempts to execute experiments designed to validate VDP results against human performance; but the issue is relevant to the deployment of VDP for a particular application regardless of the user's interest in learning what the human viewer might report in tasks representative of those implied by the deployment. It is worth questioning, for example, whether summing of detection probabilities over an entire image is most appropriate.

Finally, we should briefly discuss how our stimulus selection played an important role in the design of our psychophysical validation experiments. Selecting stimulus pairs that were fairly difficult to discriminate served to keep the viewers' responses within a range where the 70% detection heuristic remained reasonable (and thus we justified utilizing this heuristic for detecting differences at the level of individual pixels to provide the link between VDP model-transformed images and the global prediction of human perceptual outcome). In the masking study reported here, the variation in the extent of masking was great enough to allow us to test the goodness-of-fit between VDP results and subjective responses. If the subjective responses had been too extreme, the validity of the test would have been questionable, since a high coefficient of determination is expected when the discrimination is relatively easy for the subject to make. To discuss further the proper design and interpretation of psychophysical validation experiments is beyond the scope of this paper. Our intention here was to contrast the perceptual-basis of a metric with the psychophysical validation of that metric.

- is work. the

5 CONCLUSIONS

In summary, we observed good agreement between VDP responses and subjective responses for the cases we investigated. Because VDP predictions of shadow masking matched subjective judgments well, shadow reconstruction based on adaptive mesh subdivision may be well guided by these predictions. Our preliminary attempts to exploit this potential resulted in substantial mesh simplification for such applications without substantial degradation of the image quality [6]. The results from our second experiment showed that VDP responses could replace human judgments in evaluating progressive radiosity solution convergence, due to the observed close agreement between VDP and subjective responses. One immediate application that suggests itself is to base decisions regarding stopping conditions on VDP responses [6]. We succeeded in our primary goal to validate VDP results against human performance for tasks representative of those demanded by particular global illumination applications. We also hope that the discussion presented in this paper will serve our secondary purpose, which was to contrast the perceptual-basis of a metric with the psychophysical validation of that metric.

REFERENCES

[1] `http://www.u-aizu.ac.jp/labs/csel/experiment/` - the HTML forms for our psychophysical experiments.

[2] `http://www.u-aizu.ac.jp/labs/csel/vdp/` - visualization of the VDP responses and preliminary results of the VDP applications.

[3] S. Daly. The Visible Differences Predictor: An algorithm for the assessment of image fidelity. In A.B. Watson, editor, *Digital Image and Human Vision*, pages 179–206. Cambridge, MA: MIT Press, 1993.

[4] A. Gaddipatti, R. Machiraju, and R. Yagel. Steering image generation with wavelet based perceptual metric. *Computer Graphics Forum (Eurographics '97)*, 16(3):241–251, September 1997.

[5] B. Girod. What's wrong with mean-squared error? In A.B. Watson, editor, *Digital Image and Human Vision*, pages 207–220. Cambridge, MA: MIT Press, 1993.

[6] K. Myszkowski. The Visible Differences Predictor: Applications to global illumination problems. In *Rendering Techniques '98 (Proceedings of Eurographics Rendering Workshop '98)*, New York, NY, 1998. Springer Wien. To appear.

[7] H. Rushmeier, G. Ward, C. Piatko, P. Sanders, and B. Rust. Comparing real and synthetic images: some ideas about metrics. In *Sixth Eurographics Workshop on Rendering*, pages 82–91. Eurographics, June 1995.