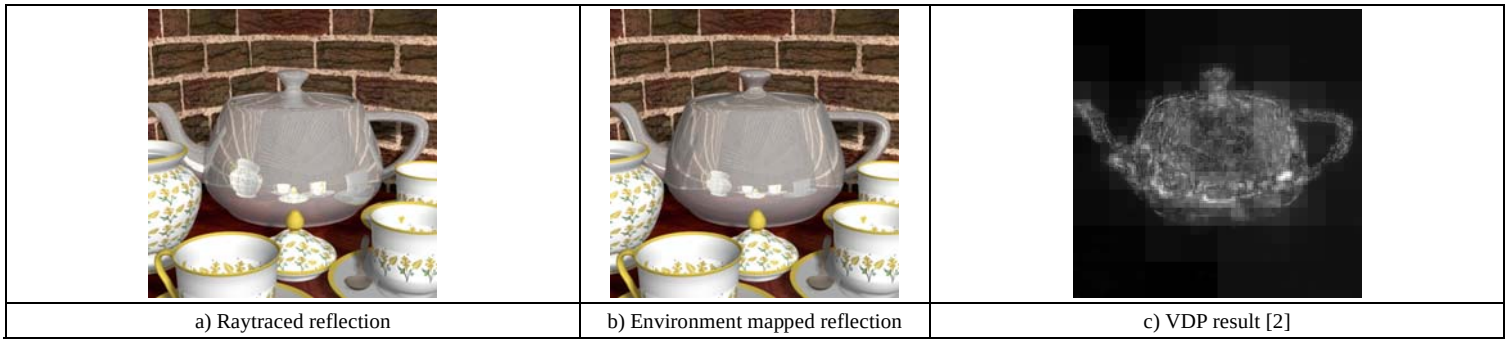
Functional Difference Predictors (FDPs): Measuring Meaningful Image Differences

James A. Ferwerda
Program of Computer Graphics
Cornell University

Fabio Pellacini
Pixar Animation Studios
Emeryville, CA



a) Raytraced reflection | b) Environment mapped reflection | c) VDP result [2]

Fig 1. Images rendered with different reflection algorithms and VDP map showing areas of visible difference.

Abstract **- In this paper we introduce Functional Difference Predictors (FDPs), a new class of perceptually-based image difference metrics that predict how image errors affect the ability to perform visual tasks using the images. To define the properties of FDPs, we conduct a psychophysical experiment that focuses on two visual tasks: spatial layout and material estimation. In the experiment we introduce errors in the positions and contrasts of objects reflected in glossy surfaces and ask subjects to make layout and material judgments. The results indicate that layout estimation depends only on positional errors in the reflections and material estimation depends only on contrast errors. These results suggest that in many task contexts, large visible image errors may be tolerated without loss in task performance, and that FDPs may be better predictors of the relationship between errors and performance than current Visible Difference Predictors (VDPs).**

INTRODUCTION

Measuring the differences between images is a very important aspect of computer graphics, especially when comparing the performance of graphics rendering algorithms. In the past, two kinds of metrics have been used. *Physical metrics* [1], compare images in terms of the numerical differences between their pixel values. A new trend in computer graphics is to use *perceptual metrics* that are based on computational models of human vision [3,5]. These metrics, generally formulated as Visible Difference Predictors (VDPs), measure the probability that observers will be able to detect differences in pixel contrasts between images. VDPs have been widely used by graphics researchers to compare rendering algorithms and to determine when images produced by different algorithms will be visually indistinguishable from one another (see [2,7,11] for recent reviews).

However when we look at images, we do not see pixels. Rather, we see objects with distinct shapes, sizes, locations, motions, and materials. We use the visual information provided by images to make judgments about the properties of these objects and to perform meaningful visual tasks [4]. Different rendering methods can affect this information in different ways. This is illustrated in Fig. 1.

Fig. 1a shows a tabletop scene with a glossy teapot. The reflection in the teapot surface was rendered using a raytracing algorithm. Fig. 1b shows the same scene, but here the reflection was rendered using environment mapping, a fast but approximate technique that introduces projective errors in the reflection with respect to the ray-traced version. Running a VDP on these images produces Fig. 1c, where the probability that an observer will detect differences in the images is proportional to the grayscale values. The VDP correctly predicts that observers can see differences in the reflections on the teapots. However, while the images are visibly different, there are many respects in which they are also similar. For example, it seems clear that the two teapots are made out of the same material. If these images were used in an e-commerce application to show the finish on the teapot, they would be of equal fidelity with respect to that task since the appearance of the material is the same in both images.

This simple example shows that while current perceptual metrics can predict whether two images will be visibly different, they do not predict whether these differences will be visually significant. We propose that in many applications, the most meaningful way to compare images is to determine if their differences affect the task the user is trying to perform. We will say that two images are *functionally equivalent* with respect to a task when the user's ability to perform the task is the same using either image.

In the remainder of this paper, we will first define how images can be functionally equivalent or different; we will

then describe an experiment we conducted to define the properties of functional difference metrics for computer graphics. Inspired by the term VDP, we call these new metrics Functional Difference Predictors (FDPs).

RELATED WORK

Measuring how image differences affect a subject's ability to perform visual tasks has been widely studied in experimental psychology (see [9] for a review). Unfortunately, most of the image manipulations and tasks that have been explored are so reductionistic that the results of these studies are not directly applicable to the problem of developing functional difference metrics for computer graphics. However, the experimental methodologies developed do provide an essential foundation for this work.

In the computer graphics literature itself, research in this area is just beginning. Watson et al. [16] and Rushmeier et al. [13] have studied the correlation between VDP measures and subjects' ratings of shape in the context of geometric compression. Rademacher et al. [10] conducted an experiment to measure the perception of visual realism and its correlation with various visual cues. Finally, Wanger et al. [15] and Rodger and Browse [12] have explored how different visual cues affect subjects' abilities to assess the spatial layout and shapes of objects in computer-rendered scenes.

FUNCTIONAL DIFFERENCE PREDICTORS

A Functional Difference Predictor (FDP) is an operator that takes two images as input and calculates whether differences in the images will affect a user's ability to perform a visual task. FDPs are formulated with respect to tasks, since they assess the fidelity of the visual information required for the task. Therefore there are potentially as many FDPs as there are classes of tasks [14]. With this in mind, it is interesting to note that VDPs are in fact a specific instance of FDPs where the task is detecting contrast differences between images. This means that the FDP framework we are developing is not in conflict with earlier work, but is rather a significant generalization of it.

In order to quantify differences in user's abilities to perform visual tasks, we need to be able to measure these abilities. Different tasks might require different measures. For example in some tasks speed of performance might be important, while in others accuracy might be paramount. Since the FDP operator does not depend on the specific form of this measure, to simplify our experiment we restricted our attention to tasks that have binary outcomes such as yes/no and same/different type judgments. In this case, the FDP calculates the probability that a user will perform the task differently using different images.

EXPERIMENTS

To define the major properties of FDPs and to compare them to VDPs, we conducted a multipart psychophysical experiment. Although we would eventually like to develop FDPs for a wide range of tasks and image differences, initially we needed to restrict our studies to a manageable domain. We chose to study two tasks that have widespread utility in both computer graphics and real world applications: material and spatial layout estimation. Since previous studies [6,8] have shown that material and layout perception are affected by the characteristics of surface reflections, we manipulated this visual cue to determine how errors in rendering reflections affect task performance.

A. Stimuli

To measure the relationships between physical image differences and functional differences in performance, we presented subjects with pairs of computer generated images. The scene model used to generate the images consisted of a sphere and two cylinders in a box illuminated by an overhead area light source. All materials were achromatic, and the sphere was rendered using specular reflections to simulate a glossy surface. Images were generated using 3dStudioMax™. The Appendix shows the geometric layout of the scene and provides numerical values for the parameters used to generate the images. Examples of the images are shown in Fig. 2.

On each trial the two images presented were the same except for the reflections in the surfaces of the spheres. In one image the reflection was a correct representation of the scene while in the other it was incorrect either in terms of the contrasts or positions of the reflected cylinders. For each of four *contrast* and four *position* conditions, we generated sets of images using different scene parameters. Each of these sets consisted of three image pairs, where the magnitudes of the errors increased within the set. In the rest of the paper, we will refer to these sets with the labels C1 to C4 for the contrast conditions and P1 to P4 for the position conditions.

B. Procedure

The subjects were asked four questions about each image pair. The specific wording of the questions is given in the Appendix. The four questions were related to four visual tasks we asked the subjects to perform. The *material estimation* and *layout estimation* tasks where designed to measure functional differences between the images. The *image difference* and *image correctness* tasks were designed to measure more traditional visual differences.

In the *material estimation task*, we asked the subjects if the spheres in both images were made of the same material. In the *layout estimation task*, we asked if the relative positions of objects were the same in both images. The proportion of negative responses to these two questions is a direct measure

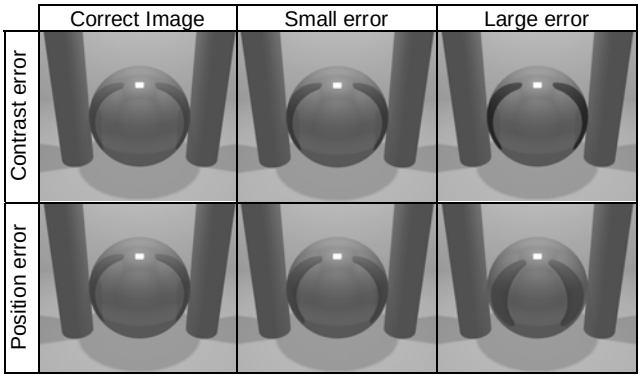


Fig. 2. Examples of stimuli used in the experiment.

of the difference in utility of the images for the respective task (i.e. the probability that the subjects' performance will be affected by the difference).

In the *image difference task* we asked subjects if they could see any differences between the images. Our goal here was to compare the predictions our new FDPs and traditional VDPs. However, rather than using a VDP algorithm to measure image differences, we simply asked the subjects if the images were the same or not. The proportion of negative responses to this question is a direct measure of the visible differences (i.e. the probability that subjects can detect a difference between the images). Effectively this question lets us benchmark our FDPs against the best possible VDP: the human observer.

Finally, in the *image correctness task*, we asked the subjects if they could tell which image was correctly reflecting the surrounding environment. We asked this question to explore whether differences in performance on the material and layout estimation tasks depend on being able to correctly identify image errors.

Eighteen subjects participated in the experiment. The subjects were the second author, 6 computer graphics graduate students and researchers, and 11 graduate students and researchers in other engineering fields. All subjects had normal or corrected to normal vision, and with the exception of the author, were naïve to the purpose and methods of the experiment.

The experiment was conducted on paper. Each image pair was printed side-by-side at the top of a page and each of the four questions were printed below. The images were tone-mapped and color corrected for the printing process using the procedure described in [8]. The subjects replied to the questions by marking checkboxes on each sheet. Each subject was shown each image pair only once. The pairs were presented in random order, and the horizontal positions of the correct and incorrect images were randomized and balanced across subjects.

RESULTS

Fig. 3 summarizes the results of the experiment. Each column represents the results for one image set. Column series C1 to C4 show the results for images with contrast differences, and series P1 to P4 show results for images with positional differences. Within each series, the results are graphed in order of increasing magnitude of physical image difference.

Since the subjects' responses were binary variables, we used binomial distribution statistics to compute mean and variance measures and used the logistic regression method when testing for correlations [17]. Chi square tests for statistical significance were also performed for all data points. Unless otherwise mentioned, the confidence intervals on all data points are 0.02 or less.

A. *Image difference task*

The results for the image difference task are shown in Fig. 3a. Here visible difference is expressed as the probability that subjects reported seeing the images as different. Note that in each series, the smallest images differences were always just noticeable (i.e. at or above the standard 75% discrimination threshold) and larger differences were always clearly visible.

B. *Image correctness task*

The results for the image correctness task are shown in Fig 3b. Here performance is expressed as the probability that the subjects selected the image that was a correct representation of the scene. The graph shows that in contrast to the subjects' high level of performance on the image difference task, here performance was much more mixed and often below threshold, suggesting that subjects were frequently unsure about which image was an accurate representation of the scene.

C. *Material estimation task*

The results for the material estimation task are shown in Fig. 3c. Here performance is expressed as the probability that the subjects' saw the spheres as being made of different materials. The graph shows that this task is clearly affected by contrast differences in the reflections. When contrast differences were present, subjects consistently saw the spheres as being made of different materials. On the other hand, positional

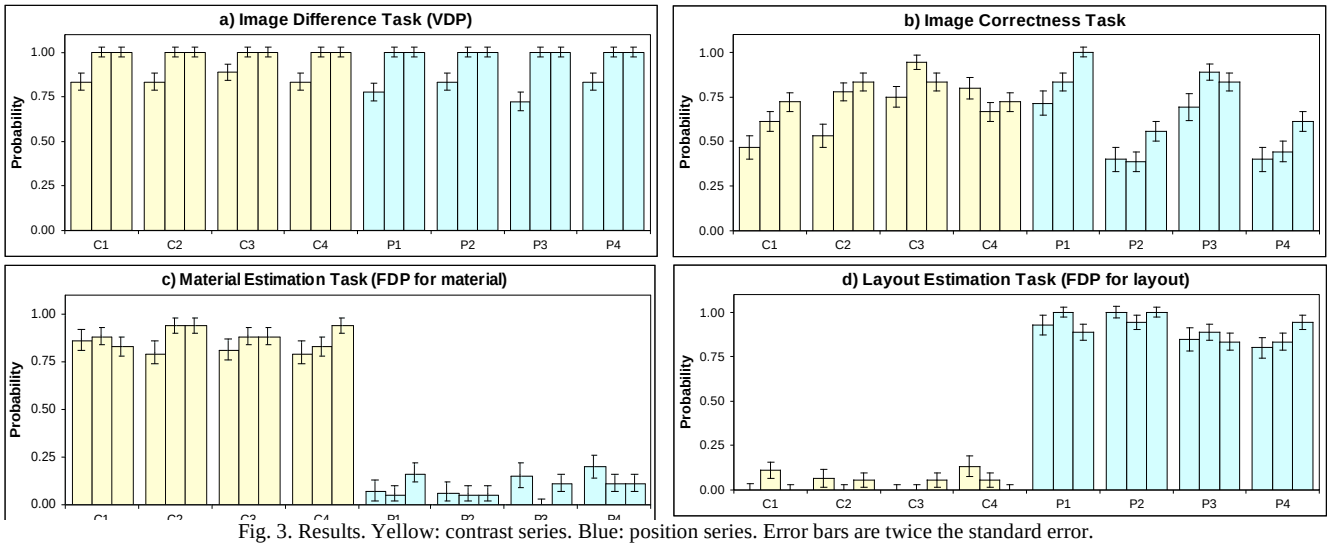


Fig. 3. Results. Yellow: contrast series. Blue: position series. Error bars are twice the standard error.

differences in the reflections did not produce differences in material appearance.

D. Layout estimation task

The results for the layout estimation task are shown in Fig 3d. Here performance is expressed as the probability that the subjects' saw the objects as being in different locations in the two images. Here the results are reversed with respect to the material estimation task. Differences in the positions of the reflections caused the layouts to appear different, but contrast differences in the reflections had no effect on layout appearance.

DISCUSSION

Several implications of the experimental results should be mentioned. First, logistic regressions found no correlations between the magnitude of the errors introduced into the images and the subjects' performance in the material and layout tasks ($p > 0.56$ in all cases). It is interesting to note that under these conditions, although the range of error magnitudes introduced is large, varying from just visible to completely objectionable, the subjects' abilities to perform the tasks seems not to depend on the magnitude of the errors. It appears that subjects are either totally affected or totally unaffected by the errors, depending on the type of error introduced. This result was surprising, and is counter to the tacit assumption underlying many visible difference metrics: that suprathreshold error magnitude and task performance are monotonically related.

Second, we found no correlations between the subjects' performance in the image difference task and their performance in the material and layout estimation tasks ($p > 0.45$ in all cases). This indicates that the visibility of image differences is not a good predictor of subjects' performance on the latter tasks. In particular this suggests that, for many visual tasks, VDPs may be too conservative as image difference metrics, for while they can predict whether two images will be visibly different, they cannot predict whether these differences will have any meaningful affect on task performance.

FORMULATING FDP METRICS

Based on the findings of our experiments, we can now formulate FDP metrics for graphics applications. As we mentioned earlier, there will be separate formulations for each task. We write:

$$FDP_{\text{material estimation}} = \begin{cases} 1 & \textit{for contrast errors} \\ 0 & \textit{for position errors} \end{cases}$$

$$FDP_{\text{layout estimation}} = \begin{cases} 0 & \textit{for constrast errors} \\ 1 & \textit{for position errors} \end{cases}$$

We can apply these metrics in graphics applications in the following ways. Imagine that a user is trying to model a three-dimensional scene. The application is trying to provide high quality images at interactive rates, but computational resources are not sufficient, and rendering shortcuts need to be taken. If it can be determined that the user is adjusting object material properties (either through automatic mode tracking or manual user preference settings), then the decision can be made to preserve reflection contrasts at the expense of introducing positional errors in the reflections (e.g. by the use of environment mapping techniques). Similar, (though in this case opposite) rendering decisions can be made if the user is moving the objects in the scene.

It should be noted that the use of FDPs does not preclude the use of traditional VDPs in perceptually-based rendering, and in fact FDPs can work in concert with VDPs to focus the computational effort involved in applying VDPs to situations where it has been determined that visible errors will negatively impact user performance.

It should also be emphasized that although for the tasks and errors we studied, the FDPs have simple binary formulations, FDP metrics for other tasks and errors may be more complex functions which might depend on the magnitudes of the errors as well as other factors. Further study is clearly necessary to develop with a general formulation for FDP metrics.

CONCLUSIONS AND FUTURE WORK

This paper has introduced Functional Difference Predictors (FDPs), a new class of perceptually-based image difference metrics. Unlike VDPs that predict whether images will be visibly different, FDPs predict whether images will be *functionally different*, affecting a user's ability to perform a visual task.

In our experimental studies, we have introduced a new methodology for measuring functional differences between images and the results of the experiments have shown that in the cases studied, FDPs are superior to VDPs at predicting how rendering errors will affect a user's ability to perform a visual task. Although our initial experiment only looked at two tasks, material and layout estimation, we believe that our our methodology can be used to explore and develop FDPs for a wide range of meaningful visual tasks.

Although we feel our initial results are promising, there is clearly much more work to be done to fully develop FDP metrics and our studies are only a small first step toward this goal. In future work, we hope to develop FDPs for other classes of visual tasks and other kinds of image errors. At first glance this might seem unachievable since there are potentially an infinite variety of tasks and errors. However we believe that the problem is tractable, because recent perception research [12,14,15] has shown that visual tasks can be organized into classes in terms of the visual information that is essential for the task and the information that is marginal or irrelevant. A single formulation of an FDP should suffice for each class. Similarly the image errors produced by common graphics or imaging algorithms can be organized into a small number of classes (e.g. noise, projective distortions, etc.) and FDPs can be tailored to the error classes produced by particular algorithms.

In the context of computer graphics, we would also like to extend the FDP framework to include both photorealistic and non-realistic rendering styles. This would allow us to assess the impact of image realism on a user's ability to perform the

task they are trying to do. For example, using the methods described, we should be able to quantify when technical-illustration-like renderings are superior to realistic images and vice versa. This should enable the development of efficient but high-fidelity rendering methods where the rendering style is optimized the task at hand.

ACKNOWLEDGEMENTS

We would like to thank Steve Westin and James Cutting for their useful comments, all of our test subjects for their efforts and patience, and Ben Watson for his VDP processing expertise. This work was supported by NSF grants ASC-8920219, IIS-0113310, and the Cornell Program of Computer Graphics, and was performed on equipment generously donated by Intel Corporation.

REFERENCES

[1] Arvo, J., Torrance, K. and Smits, B. 1994. A Framework for the Analysis of Error in Global Illumination Algorithms. In *Proceedings of SIGGRAPH 1994*, 75-84.

[2] Bolin, M.R. and Meyer, G.W. 1998. A Perceptually Based Adaptive Sampling Algorithm. In *Proceedings of SIGGRAPH 1998*, 299-310.

[3] Daly, S. 1993. The visual difference predictor: An Algorithm for the assessment of visual fidelity. *Digital Image and Human Vision*, MIT Press.

[4] Gibson, J.J. 1979. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associated.

[5] Lubin, J. 1995. A visual discrimination model for imaghing system design and evaluation. *Vision Models for target detection and recognition*, World Scientific, E. Peli editor.

[6] Madison, C., Thompson, W., Kersten, D., Shirley, P and Smits, B. 2001. Use of interreflection and shadow for surface contact. *Perception & Psychophysics, Psychonomic Society Publications*, 63(2), 187-194.

[7] Myszkowski, K., Tawara, T., Akamine, H. and Sidel, H.P. 2001. Perception-Guided Global Illumination Solution for Animation Rendering. In *Proceedings of SIGGRAPH 2001*, 221-230.

[8] Pellacini, F., Ferwerda, J.A. and Greenberg, D.P. 2000. Toward a Psychophysically-based Light Reflection Model for Image Synthesis. In *Proceedings of SIGGRAPH 2000*, 55-64.

[9] Palmer, S.E. 1999. Vision Science. MIT Press.

[10] Rademacher, P, Lengyel, J., Cutrell, E. and Whitted, T. 2001. Measuring the Perception of Visual Realism in Images. In *Rendering Techniques 2001*, 235-248.

[11] Ramasubramanian, M., Pattanaik, S.N. and Greenberg, D.P. 1999. A Perceptually Based Physical Error Metric for Realistic Image Synthesis. In *Proceedings of SIGGRAPH 1999*, 73-82.

[12] Rodger, J.C. and Browse R.A. (2000). Choosing rendering parameters for the effective communication of 3D shape. *IEEE Computer Graphics and Applications*, 20(2), 20-28.

[13] Rushmeier, H., Rogowitz, B., and Piatko, C. 2000. Perceptual issues in substituting texture for geometry. In Human Vision and Electronic Images V, *Proc. Of SPIE*, 3959, 372-383.

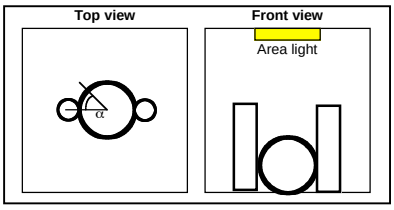
[14] Schrater, P. R., & Kersten, D. 1999. Statistical structure and task dependence in visual cue integration. *Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing, and Sampling*. Fort Collins, Colorado, June 1999.

[15] Wanger, L.R., Ferwerda, J.A. and Greenberg, D.P. 1992. Perceiving Spatial Relationships in Computer-generated Images. *IEEE Computer Graphics and Applications*, 12(3), 44-58.

[16] Watson, B., Friedman, A., McGraffey, A. 2001. Measuring and Predicting Visual Fidelity. In *Proceedings of SIGGRAPH 2001*, 213-220.

[17] Winer, B.J., Brown, D. and Michels, K. 1991. *Statistical Principles in Experimental Psychology*. McGraw-Hill.

APPENDIX



The figure on the right shows the spatial layout of the scene used to generate images for the experiment. The scene model consisted of a sphere and two cylinders in a box illuminated by an overhead area light source. The room surfaces had a diffuse reflectance of 0.7, while the sphere had a diffuse reflectance of 0.26 and a specular coefficient of 0.3. The following table reports the diffuse reflectances of the cylinders and the angle α indicated in the previous figure, for each image (when values change within a series, they are the ones used for the correct image and the three variations respectively).

Series label	Cylinders albedo	Cylinders angle (α) in degrees
C1	0.5, 0.67,0.83,1	0
C2	0.5, 0.33, 0.17, 0	0
C3	1, 0.83, 0.67, 0.5	0
C4	0.25, 0.42, 0.58,0.75	0
P1	0.5	0, 11.25, 22.5, 33.75
P2	0.5	0, -11.25, -22.5, -33.75
P3	0.5	45, 50.62, 56.25, 61.87,
P4	0.5	45, 39.38, 33.75, 28.13

The following questions were asked in the experiment:
1. Image difference task. “Are the two images the same?”. Possible answer: Yes/No.
2. Image correctness task. “Which of these two spheres correctly reflects the environment?”. Possible answer: Left/Right.
3. Material estimation task. “Looking at the reflections on the two spheres, are the spheres made of the same material?”. Possible answer: Yes/No.
4. Layout estimation task. “Looking at the reflections on the two spheres, are the relative positions of the spheres and the cylinders the same in both images?”. Possible answer: Yes/No.