FDPs: Functional Difference Predictors – Measuring Meaningful Image Differences

Fabio Pellacini James A. Ferwerda

Program of Computer Graphics, Cornell University

Abstract

In this paper we introduce Functional Difference Predictors (FDPs), a new kind of perceptual image difference metric that captures how visible image errors affect a user's ability to perform visual tasks in computer graphics applications. To develop the FDP concept, 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 judgments of the relative locations and material properties of objects in the scene. Our results show that under the conditions studied, layout estimation depends only on positional errors in the reflections and material estimation depends only on contrast errors. These results indicate that in different task contexts, large visible image errors may be tolerated without loss in task performance, and that FDPs are better predictors of the relation between image errors and task performance than current Visible Difference Predictors (VDPs). This work represents some initial steps toward developing a significant new class of perceptual metrics for measuring the fidelity of computer graphics images.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Computer Graphics and Realism.

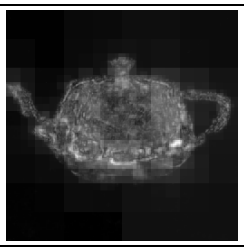
a) Raytraced reflection	b) Environment mapped reflection	c) Bolin and Meyer's [2] VDP result
		

Figure 1: *Image rendered with different algorithms and VDP result predicting areas of visible difference.*

1. Introduction

Measuring the differences between two images is a very important aspect of computer graphics, especially when used to compare the performance of graphics rendering algorithms. In the past, two kinds of metrics have been employed. *Physical metrics* [1] compare images in terms of the numerical differences between their pixel values. If these differences are below a certain threshold, then the images are judged to be *physically equal*. A new trend in computer graphics is to use *psychophysical metrics* that are based on computational models of the human visual system [3,5]. These metrics, generally known as Visual Difference Predictors (VDPs), measure the per-pixel probability that human 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 each other (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, positions, motions, and materials. We use the visual cues provided by images to make judgments about the properties of these objects and to perform meaningful visual tasks [4]. Different rendering methods can affect these visual cues in different ways. This is illustrated in Figure 1.

Figure 1a shows a tabletop scene with a glossy teapot. The reflection in the teapot was rendered using a raytracing algorithm. Figure 1b shows the same scene, but here the reflection was rendered using environment mapping, thus introducing a projective error with respect to the first one. Running a VDP on these images results in Figure 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 the differences in the reflections on the teapots. However, while the images are clearly different, it also appears that the teapot is made out of the same material in both images. If these two images

were to be used in an e-commerce application to show the finish on the teapot, they would be *functionally equivalent*, since the perception of the material is the same. This means that the two images are 'the same' with respect to this particular task.

This simple example shows that although current psychophysical metrics can predict whether two images will be visibly different, they do not predict whether these differences are visually significant. We propose that, for most 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* if 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 a psychophysical experiment we ran to show the major features of functional difference metrics for computer graphics. Inspired by the term VDP, we will use the term Functional Difference Predictors (FDPs) to refer to these new metrics.

2. Related Work

Measuring how image differences affect a person's ability to perform visual tasks has been widely studied in experimental psychology (see [9] for a review). Unfortunately most of these studies are not specific enough, with respect to the visual tasks or image differences introduced, to be used to derive metrics of functional difference for computer graphics.

In the computer graphics literature itself, work 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. Although this work is important, only one task was studied, so broader conclusions about the utility of VDPs in different contexts cannot be drawn. Rademacher et al. [10] presented an experiment to measure the perception of visual realism and its correlation with various visual cues. Although this work is intriguing, our goal differs substantially since our focus is on practical tasks, not on measuring the broad concept of image realism. Finally, Wanger et al. [15] and Rodger and Browse [12] have explored how different visual cues affect subjects' ability to assess the spatial layout and shapes of objects in computer-rendered scenes, but the results of their experiments have not been used to develop perceptual metrics for computer graphics.

3. Functional Difference Predictors

A Functional Difference Predictor (FDP) is a function that takes two images as input and predicts whether differences in the images will affect a user's ability to perform a visual task. FDPs are formulated with respect to a task, since they depend on the visual information required for the task. Thus, there are potentially as many FDPs as there are classes of tasks [14]. In this context it is interesting to note that VDPs are in fact a specific instance of FDPs where the task is detecting contrast differences between images. Thus the FDP framework we are presenting is not in conflict with earlier work, but is rather an important generalization of it.

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

To illustrate the major properties of FDPs and to compare them to VDPs, we designed and ran a multipart experiment. Although we would eventually like to measure FDPs for any possible task and any possible image difference, initially we need to restrict our studies to a manageable domain. For the purpose of illustration, we chose to study two tasks that have

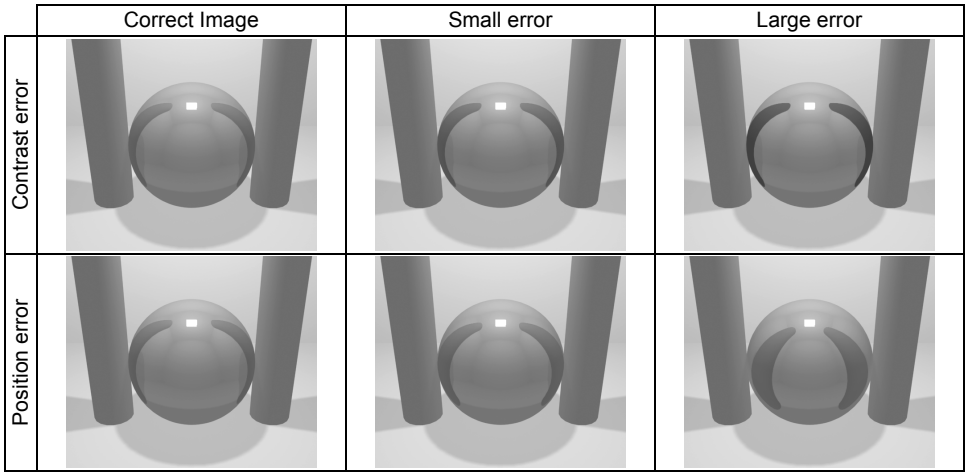


Figure 2: *Examples of stimuli used in the experiment.*

widespread utility: material estimation and spatial layout estimation. Since previous studies have shown that material and layout perception are affected by the characteristics of surface reflections [6,8], we have manipulated this visual cue in our experiment to evaluate how these tasks are affected by rendering errors in the reflections.

4. Experiments

4.1 Stimuli

In order to measure the relationships between physical image differences and functional differences, we presented subjects with a series of image pairs. The environment used to create the images consisted of a sphere and two cylinders enclosed in a box and illuminated by an overhead area light source. All surfaces were achromatic, and the sphere was rendered using mirror-like reflections to simulate a glossy painted surface. Images were generated using 3dsMax™. The Appendix shows the geometric layout of the scene and reports the numerical values of the parameters used to compute the images. Examples of the stimuli used for the contrast and position conditions are shown in Figure 2.

The two images presented were the same except for the reflections in the surface of the sphere. In one image the reflection in the sphere was the correct representation of the environment while in the other it was incorrect either in terms of the contrast or the positions of the reflected cylinders. In each of the *contrast* and *position* conditions, we generated four subsets of images using different environmental parameters. Each of these subsets consisted of three image pairs, where the magnitude of the errors introduced increased within the subset. In the rest of the paper, we will refer to these subsets with the labels C1 to C4 for the contrast condition and P1 to P4 for the position condition.

4.2 Procedure

The subjects were asked to reply to four questions for each image pair. The specific wording of each question is reported in the Appendix. The four questions were related to the 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 visible differences.

In order to measure functional differences in the material estimation task, we simply asked the subjects if the two objects had the same material. For the layout estimation task, we asked the subjects if the relative positions of objects in the scene were the same in the two images. The proportion of negative responses to these two questions is a direct measure of the functional differences in the utility of the images for the respective tasks (i.e. the probability that the subjects' ability to perform the estimation is affected by the difference).

In the *image difference* task we asked the subjects if they could detect any differences between the images. We asked this question to be able to compare the predictions of traditional VDPs and our new FDPs. In the context of the experiment, instead of using a VDP algorithm to measure whether the two images are visually different, we simply asked the subject if the images were the same or not. The proportion of negative responses is a direct measure of the visual difference, i.e. the probability that subjects can detect a difference in the pair. Effectively this question lets us compare the FDP 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 find out if the differences in performance on the functional tasks are a conscious process that depends on being able to correctly identify image errors.

Eighteen subjects participated in the experiment. The subjects were the first author, 6 graduate students and researchers in our computer graphics lab, 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 delivered on paper. Each image pair was printed side-by-side at the top of a page and each of the task questions were printed below. The images were tone-mapped and color corrected for the printing process with the same procedure used in [8]. The subjects replied to the task questions by marking checkboxes below each question. 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.

5. Discussion

Figure 3 summarizes the results of the experiment by reporting the data relative to each image subset: series C1 to C4 contain contrast differences, while series P1 to P4 contain 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 analyzed the data using binomial distribution statistics to compute mean and variance measures and we used the logistic regression method when testing for correlation [17]. Chi square tests for statistical significance were also performed for all data points; unless otherwise mentioned, all data points were better than random guesses within a confidence interval of 0.02.

5.1 Image difference task

The results of the first question regarding whether the images were visibly different are shown in Figure 3a. Here visible difference is expressed in terms of the probability that an observer will detect differences between the images. Note that in each series, the smallest physical differences were always just noticeable (i.e. barely above the standard psychophysical discrimination threshold) and the larger physical differences were clearly visible.

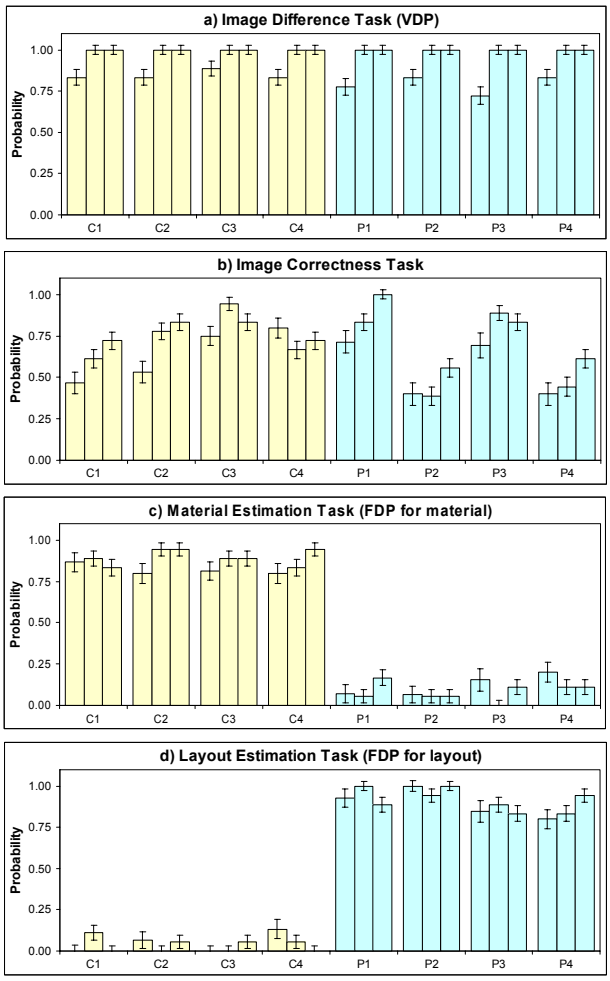


Figure 3: *Experiment results. Yellow: contrast series. Blue: position series. Error bars are twice the standard error.*

5.2 Image correctness task

The results for the second question, in which the subjects were asked to indicate which image was correct, are shown in Figure 3b, and are expressed as the probability of selecting the image that is correctly reflecting the environment. The graph shows that a subject's ability to choose the correct image depends on the magnitude of the physical image differences. Performance was better than random in a statistically significant way for only the largest physical differences. This means that when the differences are small, subjects are simply guessing; on the other hand, when the differences are large, subjects can consistently detect which of the images is the correct one. Logistic regressions showed that the correlation between subjects' performance and difference magnitude is statistically significant (contrast p=0.06, layout p=0.07).

5.3 Material estimation task

The results for the material estimation task are shown in Figure 3c. Here the graph indicates the probability that the objects in the two images were judged to be made of different materials. The data shows that the task is clearly affected by the contrast differences introduced into the object reflections. When these differences were present, subjects consistently judged the objects to be made of different materials. On the other hand, differences in the positions of objects in the reflections did not affect the subjects' judgments.

5.4 Layout estimation task

The spatial layout estimation task presented a similar pattern of results. In this case we have an inversion of the subjects' behavior, where responses are affected by errors in the position of the reflections and not affected by errors in contrast.

6. Implications of the results

Several important implications of the experimental results

should be mentioned. First, logistic regression found no correlations between the magnitude of the errors introduced in 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 subject's ability to perform the tasks does not 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 assumption commonly used in VDPs that there are monotonic relations between suprathreshold error magnitude and task performance.**

Second, we found no correlations between the subjects' responses in the image correctness task and their performance in material and layout tasks ($p > 0.41$ in all cases). Interestingly **this suggests that although subjects were able to correctly identify the errors in the images *regardless* of the task, these errors may or may not affect their performance *depending* on the task.**

Finally, we found no correlations between the subjects' responses in the image difference task and their performance in the material and layout estimation tasks ($p > 0.45$ in all cases). This shows that VDPs do not accurately predict how subjects' performance in these tasks will be affected by visible image differences. What this suggests is that **for many meaningful tasks, VDPs are just too conservative: while they can predict if two images will be visibly different, they cannot predict whether the images will be functionally different with respect to a given task.**

7. Formulating FDP metrics

Based on the findings of our experiments, we can now formulate FDP metrics for graphics applications. As 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 rendering engine in the application is trying to provide the high quality images at interactive rates, but there are not sufficient computational resources and algorithmic 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 rendering 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 users is moving the objects in the scene. Note that use of the FDPs does not preclude use of traditional VDPs, and in fact the FDP can work in concert with a VDP to focus the computational effort involved in applying the VDP to situations where it has been determined (by the FDP) that some class of visible errors have negative impact on user performance.

We should emphasize clearly that, while 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, but this work demonstrates the value of the FDP concept.

8. Conclusions and future work

This paper introduces Functional Difference Predictors (FDPs), a new kind of perceptually-based image difference metric for computer graphics. *FDPs capture the effect of image differences on users' abilities to perform visual tasks*. In this sense, *FDPs capture meaningful image differences since these are directly correlated with the purpose the images where intended for*. Unlike VDPs that predict whether images will be visibly different, FDPs measure whether images will be *functionally different*, affecting a user's ability to perform a visual task.

In our experimental studies, we have introduced a general methodology for measuring functional differences between images and the results have shown that FDPs are superior to VDPs at predicting how rendering errors will affect a user's ability to perform different tasks. Although our initial experiment only looked at two tasks, material estimation and spatial layout estimation, we believe that our approach is valuable because the ability to perform these tasks is important in a variety of applications, and since the same experimental methodology can be used to explore a wide range of other meaningful visual tasks.

Although we feel these results are promising there is clearly much more work to be done to fully develop the FDP concept and our studies should be considered as just the first steps toward this goal. In future work, we hope to develop Functional Difference Predictors 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 number of tasks and image 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. Single formulations of the FDP should suffice for each class. Similarly the image errors produced by common graphics 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.

We would also like to extend the FDP concept to include both realistic and non-realistic rendering styles. This would allow us to assess the impact of rendering style on a user's ability to perform the task they're trying to do. For example, using the methods described, we should be able to

quantitatively measure when technical illustration-like renderings are superior to realistic images and vice versa. This could lead to fast but visually reliable rendering methods where the rendering style is designed (with reference to experimental results), to support the task at hand.

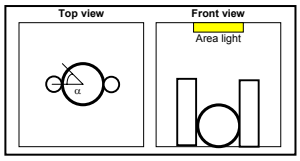
9. Acknowledgements

We would like to thank Steve Westin and James Cutting for their useful comments and all of our test subjects for their patience. This work for partly supported by the NSF Science and Technology Center for Computer Graphics and Scientific Visualization (ASC-8920219) and performed using equipment generously donated by Intel Corporation.

10. Bibliography

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.

11. Appendix – Experiment details



The figure on the right shows the spatial layout of the scene used to generate the images for the experiment. The scene is composed of a box-shaped room containing a reflecting sphere and two cylinders, placed in contact with the sphere, and illuminated by an overhead area light. The room has a diffuse albedo of 0.7, while the sphere has an albedo of 0.26 and a reflective coefficient of 0.3. The following table reports the diffuse albedo of the cylinders and the angle α indicated in the previous figure for each image (when the values are changing within the series, the values reported are the ones used for the correct image and the three variations).

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 list reports the questions and the possible answers as formulated 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, is the relative position of the spheres and the cylinders the same in both images?”. Possible answer: Yes/No.