A Perceptual Metric for Production Testing

Yangli Hector Yee & Anna Newman
hyee@pdi.com - anewman@pdi.com
PDI/DreamWorks Feature Animation
1800 Seaport Blvd, Redwood City, CA 94301

1. Abstract

This paper describes a perceptually-based image comparison process that can be used to tell when images are perceptually identical even though they contain some numerical differences. The technique has shown much utility in the production testing of rendering software.

2. Problem Definition and Solution

Regression testing based upon reference image comparison quantifies software stability across versions and configurations. In early versions of the automated test suite for our in-house renderer, a series of reference frames were compared with the output frames of the test suite by finding the numerical L2 difference between pixels. This led to a huge number of false positives that arose from sampling differences that were not visible to the human eye. In effect, this limitation reduced the usefulness of the automated test suite. To quickly and procedurally expose only legitimate rendering defects, a perceptually-based image comparison tool was developed to weed out the false positives. The tool uses a perceptually-based error metric on the reference canonical and the output of the automated test suite to determine if they are perceptually identical. In its current incarnation, the tool allows perceptual thresholds to be tweaked for individual tests or for entire suites of tests. Images that fail the perceptual-metric test are then presented as error-images highlighting the color or luminance difference. Additionally, thumbnail image summaries are presented to the user for quick manual confirmation. The perceptual test increased the utility of the test suite tremendously.

3. Implementation

Our perceptual metric is largely based on the Visible Differences Predictor (VDP) in [Daly93] with some speed improvements from Ramasubramanian et. al. [Rama99]. The first step will be to convert the RGB images into XYZ and CIE L*A*B space. XYZ is a color space where Y represents the luminance of a pixel and X, Z are color coordinates. CIE L*A*B is a perceptually uniform color space where the Euclidean distance between two colors directly corresponds to perceptual distance. L also represents luminance and A, B are color coordinates that are spread out uniformly in the perceptual domain.

Next, the threshold elevation factor F, a measure of increase in tolerance to error, is computed as in [Rama99]. A spatial frequency hierarchy is constructed from the Y channel of the reference image. This step is efficiently computed using the Laplacian pyramid of Burt and Adelson. The pyramid enables us to compute the spatial frequencies present in the image to determine how sensitivity to contrast changes decreases with increasing frequency. Next, we compute the normalized Contrast Sensitivity Function (CSF), multiplied by the masking function given in [Daly93] to obtain the combined threshold elevation factor, F. We compute some of the intermediate variables from the field of view (fov) and the image width.

Finally, we perform the following two tests and mark the images as different if either of the following two tests fails. The first test is performed on the luminance channel, Y. If the difference of luminance between two corresponding pixels (x,y) in the reference and test images is $\Delta Y(x,y) = Y1(x,y) - Y2(x,y)$, then the luminance test fails if:

$$\Delta Y(x,y) > F * TVI(Y(adapt))$$

where TVI is the Threshold vs Intensity function and the adaptation luminance is the average of pixels in a one degree radius from the Y channel of the reference image. The second test is performed on the A and B channels of the reference and test images. The color test fails if:

$$(A\_ref(x,y) - A\_test(x,y))^2 + (B\_ref(x,y) - B\_test(x,y))^2 * color\_scale^2 > F$$

color_scale is a scale factor that turns off the color test in the mesopic and scotopic luminance ranges (night time light levels) where color vision starts to degrade. We use a value of one for adaptation luminances greater than 10.0 cd/m^2. We then ramp color_scale linearly to zero with decreasing adaptation luminances.

4. Implementation Details

There were some implementation details in using the perceptual error metric for Quality Assurance testing of a production renderer that had to be taken into account.

First of all, the threshold elevation factor, F, depends strongly on the frequency content of the image. This in turn is affected by the viewing parameters of the observer, the most important of which is the field of view. We measured a few cinemas in Hollywood and found out that the average front row and back row field of views were 85 degrees and 27 degrees respectively. Using a field of view of 85 degrees was the most conservative and increased the probability that the simulated front row observer will not notice differences between the reference and test images. Another important factor is the width of the image in pixels. We used a value of 1827 for film resolution images. The color_scale factor was added because the perceptual metric was returning false positives in very dark areas where the hue does not matter. The scotopic scaling used no rigorous perceptual data other than the fact that the visual system loses its color sensitivity in the mesopic and scotopic ranges.

5. References

[Daly93] Scott Daly. The visible differences predictor: an algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, pages 179-206, MIT Press, Cambridge, MA, 1993.

[Rama99] Mahesh Ramasubramanian, Sumant N. Pattnaik, Donald P. Greenberg. A perceptually based physical error metric for realistic image synthesis. In *SIGGRAPH 99 Conference Proceedings,* pages 73-82, Los Angeles, CA, 1999.