CAN VISUAL FIXATION PATTERNS IMPROVE IMAGE FIDELITY ASSESSMENT?

Eric C. Larson, Cuong Vu, and Damon M. Chandler, Member, IEEE
Image Coding and Analysis Lab
Oklahoma State University, Stillwater OK

ABSTRACT
This paper presents the results of a computational experiment designed to investigate the extent to which metrics of image fidelity can be improved through knowledge of where humans tend to fixate in images. Five common metrics of image fidelity were augmented using two sets of fixation data, one set obtained under task-free viewing conditions and another set obtained when viewers were asked to judge image quality. The augmented metrics were then compared to subjective ratings of the images. The results show that most metrics can be improved using eye fixation data, but a greater improvement was found using fixations obtained in the task-free condition (task-free viewing).

Index Terms— image fidelity, fixation, region-of-interest

1. INTRODUCTION
In the past decade, image fidelity assessment has seen great strides of improvement. Researchers are continually searching for algorithms that can predict fidelity ratings in a manner that agrees with human subjective ratings. However, when a person is asked why they rated an image a particular score, they describe the distortions of objects or regions in an image. Therefore, the classification of regions-of-interest (ROIs) in images could be immensely important to correctly assessing image fidelity. Indeed, some researchers have explicitly incorporated ROIs into image fidelity metrics [1][2].

The utility of ROI-based processing has led to an increased interest in predicting human visual fixations. But, it is unclear if eye fixations correspond directly to ROIs in images. Moreover, it is unclear to what extent eye fixation data can improve fidelity assessment or even what type of eye fixation data should be predicted, as viewers change fixation depending on their task when viewing images [3].

In this paper, we augmented five metrics of image fidelity with fixation data collected in [4]: Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM) [5], Visual Information Fidelity (VIF) [6], Visual Signal-to-Noise Ratio (VSNR) [7], and Weighted Signal-to-Noise Ratio (WSNR). Two types of visual fixation data were used: The first set of fixations was collected when the viewers were given no task (i.e., they simply looked at the images).

The second set of fixations was collected when the viewers were asked to assess image fidelity. The visual fixation points were used to segment images into three regions: primary ROI, secondary ROI, and non-ROI. For each metric, the output was computed separately for each of the three regions, and then an overall score of fidelity was obtained by computing a weighted linear combination of the per-region metric outputs. Two different segmentations were made, the first using task-free visual fixations and the second using the tasked fixations. Figure 1 shows example segmentations under no task and tasked fixations. For simplicity, these regions will be addressed as ROIs, noting that it remains an open research topic whether fixations directly correspond to interesting regions in images.

We then asked,
1. Can existing fidelity metrics be improved using eye fixation data?
2. If so, is it more appropriate to use eye fixations obtained under no task viewing conditions or when viewers were asked to assess quality?
3. Can PSNR be augmented using eye fixation data to perform as well as SSIM, VIF, VSNR, or WSNR?
4. Using the tertiary segmentation, can we quantify how important each region is for each metric to correlate highly with human subjective ratings?
This paper is organized as follows: Section 2 explains the methods used to augment the metrics and our evaluation procedures. Section 3 presents the results of the metric augmentation. Finally, section 4 summarizes and concludes the paper.

2. METHODS

2.1. Database
In [4], eye fixation data was collected for ten unique images from the LIVE image database [8]. In the database, the ten unique images have 26 to 29 distorted versions along with differential mean opinion scores (DMOS) that correspond to perceived distortion of each distorted image. The DMOS ratings are part of the LIVE database and were not experimentally verified in this study. The distorted versions of the images were created using varying degrees of Gaussian blurring, additive white noise, JPEG compression, JPEG2000 compression, and simulated packet loss of transmitted JPEG2000 compressed images. Using the ten original images, there were a total of 265 distorted versions with DMOS ratings available.

Fixation data was obtained under two viewing conditions [4]. First, viewers were asked to only look at the ten images, referred hereafter as “no-task condition,” and, secondly, viewers were asked to view the images while judging image quality, referred hereafter as “tasked condition.”

2.2. Metric Augmentation
In each viewing condition, the eye fixation data was used to segment the image into three regions as follows: The non-ROI was defined as the region in the image where the viewer never looked. The secondary ROI was defined as areas that the viewer looked, but the number of fixations in the region was less than the average fixation per pixel in the image. The primary ROI was defined as the region where the number of fixations was greater than the average fixation per pixel. This method is only one means of clustering eye fixations into regions. The method employed here was chosen because it is simple and intuitive.

For each region, PSNR, SSIM, VIF, VSNR, and WSNR were used to evaluate the error between the original image region and the distorted versions of the region. For each metric, this resulted in three regional fidelity measures for each of the images. The three fidelity measures were then combined linearly to form an overall error measure according to

\[ E_{\text{tot}} = \alpha_{1\text{-ROI}} E_{1\text{-ROI}} + \alpha_{2\text{-ROI}} E_{2\text{-ROI}} + \alpha_{\text{non-ROI}} E_{\text{non-ROI}} \]  

(1)

where \( E_{1\text{-ROI}}, E_{2\text{-ROI}}, \) and \( E_{\text{non-ROI}} \) denote the metric outputs of each corresponding region; and where \( \alpha_{1\text{-ROI}}, \alpha_{2\text{-ROI}}, \) and \( \alpha_{\text{non-ROI}} \) denote weights constrained such that the sum, \( \alpha_{1\text{-ROI}} + \alpha_{2\text{-ROI}} + \alpha_{\text{non-ROI}} = 1 \), and \( \alpha_{1\text{-ROI}} > 0, \alpha_{2\text{-ROI}} > 0, \alpha_{\text{non-ROI}} > 0 \). The choice of \( \alpha \) values is discussed in Section 2.3. In this way, the contribution of error from each region was isolated.

The metric output of each region for PSNR and SSIM is straightforward. However, VIF, VSNR, and WSNR use frequency-based decomposition to evaluate fidelity and therefore lose some (or all) of the spatial information in the image. This makes it impossible to perfectly evaluate the fidelity of a non-continuous and oddly shaped region. Instead, for each region, the original and distorted images were masked (set to zero) in areas outside of the region. Then, the metrics are used to evaluate fidelity of the masked images. This has the advantage of keeping the images large for wavelet decomposition, but also introduces erroneous frequency components in the images. The process is not perfect but can be considered an adequate means of incorporating visual fixation data into each metric.

2.3. Evaluation
To judge improvement, a correlation between each metric and subjective DMOS ratings was desired. However, before a correlation could be taken we needed to combine the regional error measures into a single output using a set of three \( \alpha \) weights specified in equation (1), and apply a logistic transformation to each linearly combined metric output.

One might expect that the primary ROI receives the greatest weight. A common sense set of \( \alpha \) weights might be \( \alpha_{1\text{-ROI}} > \alpha_{2\text{-ROI}} > \alpha_{\text{non-ROI}} \). However, the “ideal” \( \alpha \) weights are unknown. Therefore, many \( \alpha \) sets were applied, with each \( \alpha \) allowed to increase from 0.0 to 1.0 in increments of 0.01 while maintaining the equality constraint of summing to one. This is also used as a means of quantifying region importance to the metric output.

For each metric, this produces many \( E_{\text{tot}} \) vectors, where the elements of each vector are fidelity measures for each of the 265 images selected in the database. The only difference between each vector is the weights chosen. Each vector is then fit logistically to the DMOS rating according to

\[ f(E_{\text{tot}}) = r_1 + r_2 E_{\text{tot}} + r_3 E_{\text{tot}}^2 \]  

(2)

Table 1. The correlation coefficients between the DMOS rating and outputs of the five different metrics using no segmentation, and segmented regions from eye fixations obtained in the no task and tasked conditions. NOTE: the logistic correlation coefficients shown here were calculated using only 10 images from [8], not the entire database.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Original</th>
<th>No Task</th>
<th>Tasked</th>
<th>No Task Improve</th>
<th>Tasked Improve</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.8242</td>
<td>0.8379</td>
<td>0.8287</td>
<td>0.0137</td>
<td>0.0045</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.8335</td>
<td>0.8679</td>
<td>0.8367</td>
<td>0.0344</td>
<td>0.0032</td>
</tr>
<tr>
<td>VIF</td>
<td>0.7887</td>
<td>0.8681</td>
<td>0.8179</td>
<td>0.0794</td>
<td>0.0292</td>
</tr>
<tr>
<td>VSNR</td>
<td>0.8660</td>
<td>0.8682</td>
<td>0.8760</td>
<td>0.0022</td>
<td>0.0100</td>
</tr>
<tr>
<td>WSNR</td>
<td>0.8722</td>
<td>0.8818</td>
<td>0.8760</td>
<td>0.0096</td>
<td>0.0038</td>
</tr>
</tbody>
</table>
where each $I_J$ is chosen to minimize the sum squared error between $f(E_{\text{tot}})$ and the DMOS rating. The $I_J$ values are found using non-linear optimization. Once transformed, the correlation between the DMOS and the metric is taken.

3. RESULTS

Table 1 shows the original correlation coefficient and both augmented correlation coefficients for each of the five metrics. Under the no-task condition, the correlation coefficient improves in all metrics. VIF had the greatest improvement and WSNR had the greatest correlation.

Under the tasked condition, the correlation coefficient improves in all metrics. VSNR showed the best improvement, and VSNR and WSNR showed the greatest correlations (when the task of the viewer is to assess image fidelity). The large performance increase in VIF agrees with results from [9], which found that VIF had the most to gain from ROI-based spatial weighting.

Also of note from Table 1 is that the correlation of each metric is higher under the no-task condition except for VSNR. The difference between tasked and no-task correlation is slight for both VSNR and WSNR, and consequently, may be unique to this image set. Larger differences, however, are present in the other metrics, indicating that the no-task condition is possibly better suited to augmenting image fidelity using spatial weighting.

PSNR performs better than VIF when augmented with visual fixation data under the no task and tasked conditions. However, original VIF performs worse than original PSNR for this image set, so the distinction is not meaningful. PSNR performs worse than SSIM, VSNR, and WSNR across all categories. Also, augmented PSNR performs the worst of all the augmented metrics. An $F$-statistic was calculated to assess statistical significance of each metric. The data can be seen in [10]. However, no improvement was found to be significant. This finding extends the conclusions from [11].

Figures 2 and 3 show the $I_J$ values used and the obtained correlation to DMOS for each metric under the no-task and tasked conditions, respectively. The $I_J$ values are shown as independent variables that contribute to the $z$-axis correlation coefficient. The second ROI weight is not shown but is set implicitly by the equality constraint $I_{2}\text{-ROI} = 1 - I_{1}\text{-ROI} - I_{\text{non-ROI}}$. Under no task, PSNR, SSIM, VIF, and WSNR follow common-sense weighting patterns. VSNR shows a counter-intuitive weighting pattern (i.e., preferring the non-ROI).

Under tasked conditions, all metrics follow a common sense weighting pattern. VSNR prefers the secondary ROI the most, but only slightly more than the primary ROI. The odd patterns present in VSNR indicate that the choice of region importance is not readily predictable from fixation data (or at least using the clustering method and augmentation method presented in this paper).
A computational experiment was presented that segmented images based upon eye fixation data and augmented existing image fidelity metrics with the segmentation regions. It was shown that,

1. Existing fidelity metrics can be positively augmented using fixation data, with SSIM and VIF showing the greatest improvements (for common sense weighting).
2. The no task fixation condition showed the greatest improvements for all metrics except VSNR.
3. Under no task conditions, the primary region of eye fixation corresponds to the most important region for PSNR, SSIM, and VIF. For VSNR, the non-ROI is the most important region.
4. PSNR can be augmented to perform better than original VIF, but not SSIM, VSNR, nor WSNR (under this image set). When all metrics are augmented, PSNR has the worst performance.

Ultimately, the best way to augment metrics using ROI information and how to cluster eye tracking data in the most meaningful manner for image fidelity assessment remains an open question. We are currently working on both avenues of research.

5. REFERENCES