Unveiling Relationships between Regions of Interest and Image Fidelity Metrics

Eric C. Larson and Damon M. Chandler*

Image Coding and Analysis Lab, School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078

ABSTRACT

This paper presents the results of two computational experiments designed to investigate whether the success of recent image fidelity metrics can be attributed to the fact that these metrics implicitly incorporate region-of-interest information. Modified versions of four metrics (PSNR, WSNR, SSIM, and VIF) were created by incorporating spatially varying weights chosen to maximize correlation between each metric and subjective ratings of fidelity for images from the LIVE image database. The results reveal that all metrics can benefit from spatially varying weights, especially when the regions are hand-chosen based upon the objects in an image. However, the results suggest that PSNR and VIF would benefit the most from spatial weighting in which the weights are determined based on region of interest information. Additionally, the results show that object based regions follow an intuitive weighting pattern.

1. INTRODUCTION

Computational metrics of image fidelity aim to accurately and efficiently quantify the fidelity of a processed (distorted) image in a manner that agrees with subjective judgments made by a human. Standard approaches to quantifying the fidelity of an image operate based on a variety of factors, including the energy of the distortions (e.g., MSE, PSNR), human-visual-system-based models (e.g., Refs. 1-9), or overarching premises such as structural or information extraction;[10, 11] see Refs. 12 and 13 for reviews. A recent study by Sheikh *et al*.[12] has demonstrated that metrics in this latter category (structural or information extraction), particularly the Structural SIMilarity (SSIM)[10] and Visual Information Fidelity (VIF)[11] metrics, can offer statistically significant improvements in predicting fidelity compared to other widely used metrics.

However, when subjects are asked *why* they assigned a particular fidelity rating to an image, they often attribute their rating to distortion of specific objects or regions within the image, suggesting that perceived fidelity might also be affected by *what* in the image is actually distorted. Indeed, some researchers have explicitly incorporated ROI and/or visual fixation information into metrics of image fidelity[14, 15], although the predictive performance of these metrics has not been evaluated in terms of statistical significance.

In this paper, we present two experiments designed to investigate whether the success of recent image fidelity metrics can be attributed to the fact that these metrics implicitly incorporate region-of-interest information. In Experiment I, we created modified versions of PSNR, SSIM, and VIF by incorporating weights into each metric which were allowed to vary across space (block-based); the weights were chosen to maximize the correlation between each modified metric and subjective ratings from the LIVE image database.

In Experiment II, we examined whether PSNR, WNSR (a weighted signal-to-noise ratio), and SSIM could be augmented with *human drawn* spatial regions of interest. Again, the LIVE image database was used except now each image was divided into three regions: primary ROI, secondary ROI, and non-ROI. Weights were found for each region that maximized correlation to perceived fidelity.

An example of the masks used in each experiment is shown in Figure 1. The block-based mask shows hand chosen weights corresponding to regions of the image considered more important. We were interested

* E.C.L.: E-mail: ericcl@okstate.edu; D.M.C.: E-mail: damon.chandler@okstate.edu

Visual Communications and Image Processing 2008, edited by William A. Pearlman, John W. Woods, Ligang Lu, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6822, 68222A, © 2008 SPIE-IS&T · 0277-786X/08/$18

to see if Experiment I could produce such a mask where blocks containing interesting information had greater weights. Also shown in Figure 1 are the hand-chosen regions of interest used in Experiment II with gray-level signifying the possible importance of each region. We were interested to see if the weights from Experiment II might correspond to the intuitive weighting pattern shown in the region-of-interest mask in Figure 1. We then asked:

1. Can PSNR, modified to allow these correlation-maximizing, spatially varying weights in each experiment, achieve the same level of performance as WSNR, SSIM, and VIF?
2. To what extent can the performances of WSNR, SSIM, and VIF be improved by allowing these spatially varying weights? (See also Ref. 16.)
3. Do the resulting block weights from Experiment I correspond to regions of interest; i.e., do regions which tend to be considered more interesting receive greater weights than other, less interesting regions?
4. Do the weights found using the human drawn maps in Experiment II correspond to common sense weights for the hand segmented regions; i.e., do higher weights come about in regions that are considered primary and secondary ROI?

This paper is organized as follows. Section 2 provides details of methods used in the experiment. The results and analysis of the experiment are presented in Section 3. General conclusions and notes of ongoing experiments are provided in Section 4.

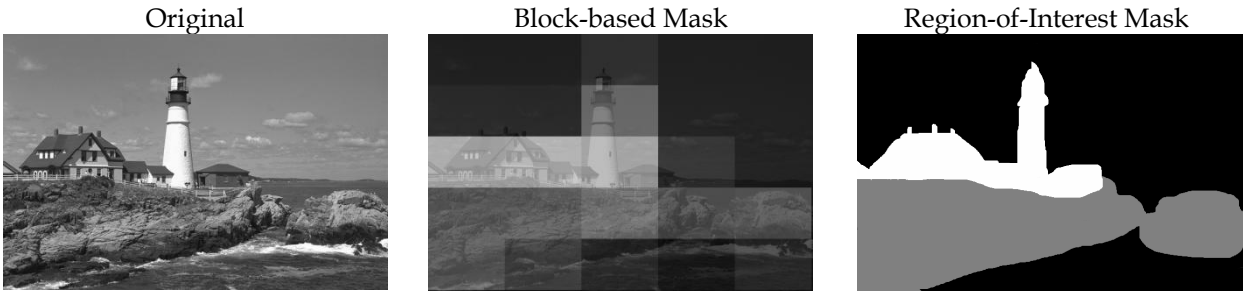

Figure 1. An example image, "lighthouse," from the LIVE image database and the corresponding masks used in Experiment I and Experiment II, respectively. In Experiment I, the aim was to find a set of weights for the Block-Based mask. In Experiment II, the aim was to find and analyze three weights for each region of the Region-of-Interest mask above. In each, the weights were chosen to maximize correlation between outputs of metrics of fidelity and perceived fidelity. We asked if the weighted blocks from Experiment I could be considered indicators of interesting regions, like the mask shown, and if the weights from Experiment II followed an intuitive weighting pattern for the three different areas in the Region-of-Interest Mask.

2. METHODS

2.1. Image Database

Images used in the experiment were obtained from the LIVE image database.[17] The LIVE database contains 29 original images, 26 to 29 distorted versions of each original image, and subjective ratings of fidelity (differential mean opinion score, DMOS values) for each distorted image. The distortions present in the database were: Gaussian blurring, additive white noise, JPEG compression (DCT based), JPEG2000 compression (wavelet based), and simulated data packet loss of transmitted JPEG2000 compressed images.

The DMOS values were computed by averaging z-scores obtained from subjective ratings of fidelity on a continuous linear scale that was divided into five equal regions labeled "Bad," "Poor," "Fair," "Good," and "Excellent." Approximately 20 − 29 human observers rated each distorted image. Note that the DMOS values were provided as part of the LIVE image database; they were not experimentally determined nor verified in the current study.

The images ranged in size from 408×704 pixels to 768×512 pixels: Six of the images were of size 408×704, three of the images were of size 640×512, and 14 of the images were of size 768×512; the remaining six images ranged in size from 608×416 to 608×480. Grayscale versions of the original and distorted images were obtained via a pixel-wise transformation of I = 0.2989 R + 0.5870 G + 0.1140 B, where I, R, G, and B denote the 8-bit grayscale, red, green, and blue intensities, respectively.

2.2 Experiment I

Modified versions of PSNR (used as a control), SSIM, and VIF were created by incorporating into each metric weights which were allowed to vary across space. Specifically, the following steps were employed:

1. The original and distorted images were broken into 25 non-overlapping blocks in which the block dimensions (typically 154x103 or 96x141) were chosen such that the 25 blocks tiled the entire image.
2. Using the distorted images and the metric under evaluation (PSNR, SSIM, or VIF), a measure of the error in each block was calculated.
3. A single output for each distorted image was calculated by scaling each block error by a weight and summing the outputs together via $E_{tot} = \sum_{i=1}^{25} \alpha_i E_i$, where E_{tot} is the total error output, E_i is the error measured in the current block using PSNR, SSIM or VIF, and α_i is the weight applied to the i^{th} block error.
4. For each original image, an optimal set of α_i values was computed by maximizing the correlation between the summed error output (logistically transformed) and the DMOS; the optimization was completed using the Nelder-Mead simplex method[18].

The strategy outlined above for Experiment I is explained in more detail below. The PSNR was computed directly via

$$\text{PSNR}=10\log_{10}\left(\frac{255^2}{D}\right) \qquad D=\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}E(m,n)^2 \tag{1}$$

where $E(m,n)$ is difference between pixels at m, n in the distorted and reference images of size $M \times N$.

The SSIM metric was computed on filtered and downsampled versions of the images where the downsampling factor (typically four) was chosen based on the height of each image as described in Ref. 20.

Instead of directly manipulating the VIF algorithm, outside of the current block being processed the reference image and distorted image were zeroed. This method has the advantage of keeping the image large so that the wavelet decomposition (steerable pyramid) remains valid. The image structure, however, changes dramatically which could introduce inconsistencies that the VIF algorithm was not built to handle. But, given that the metric relies on modeling information content that the HVS *can* gather from the original image, the process makes sense to implement in this fashion because zeroing out isolates the available information.

Once the measure of error for each block was found, the correlation-maximizing weights were calculated. The weights were constrained to be greater than zero and sum to one using penalty functions. In addition, the metric was fit logistically each time a new set of weights were guessed.

The logistic transformation polynomial used in Experiment I was:

$$\{f(E_{tot})\} = a \times E_{tot}^{\ 2} + b \times E_{tot} + c \tag{2}$$

Where a, b, and c must be iteratively solved for to minimize the error between the vector output, $\{f(E_{tot})\}$, and the DMOS vector ratings for each of the 29 image sets. The specific equation for the logistic fit was chosen for its simplicity. Other logistic transforms can better fit the data, but increase the search time exponentially. However, we are not interested in attaining the highest correlation from the logistic fit, only the relative correlation improvement once a fit has been applied.

The Nelder Mead Simplex[18] algorithm implemented both the three parameter logistic fit and the 25 parameter alpha search. This algorithm uses a guessed starting point and projects a simplex inside the error space. For example, the three parameter logistic fit above has a corresponding three-dimensional error space

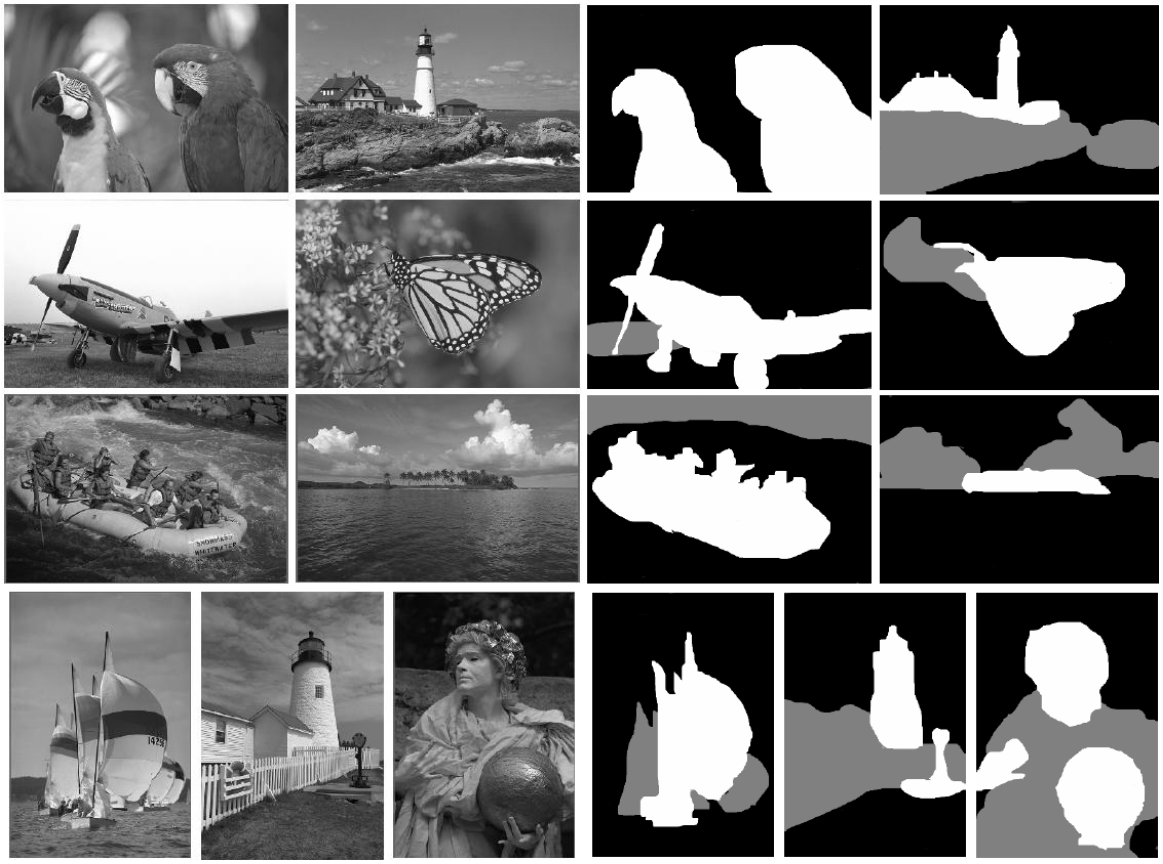

Figure 2. Corresponding masks denoting primary, secondary, and non-ROIs for the representative grayscale images depicted (Experiment II). White pixels denote the primary ROI; gray pixels denote the secondary ROI; black pixels denote the non-ROI.

that is confined using a tetrahedral simplex; each point represents a different set of the a, b, and c parameters. The errors from each of the four points of the tetrahedron are calculated, and the point with the highest error is replaced using an expanding or contracting flip about the centroid of the tetrahedron.

This continues until an error minimizing solution set is found. The algorithm does not guarantee an optimum solution; it is completely dependent on the initial guess. But it does offer a heuristically defined solution that is at least guaranteed to be a local minimum of the error function.

The outer Nelder-Mead search (for the alpha values) is more intensive than the logistic fit. It is defined inside a 25 dimensional error space bounded by a 26 point polyhedron. For each set of alphas guessed, the Nelder-Mead search attempts to converge 50,000 times with a maximum number of 10,000 guesses. For each guessed alpha set, the outputs are logistically fit as explained above. The MSE between $\{f(E_{tot})\}$ and the DMOS scoring is taken as the simplex error. The initial simplex point alpha values are set to 1/25, spread equally across the image.

This optimization is repeated for each of the 29 reference images, resulting in 29 weight masks that maximize correlation to DMOS. No convergence issues were encountered for any alpha set. This process took approximately 34.5 hours running on a 3.0 GHz AMD Athlon X2 machine.

To judge how well the alpha weight masks improve correlation, the PSNR, SSIM, and VIF optimizations were rerun using constant alpha weights of 1/25. The output was fit logistically to the DMOS values. Then, an absolute difference was taken between the maximized correlation and constant-weight (baseline) correlation. Note that baseline PSNR and VIF are not equivalent to the PSNR or VIF of the entire image. The improvement in correlation was calculated by averaging the absolute differences across all 29 images.

2.3 Experiment II

For each of the 29 grayscale original (non-distorted) images, primary and secondary ROIs were selected by the authors by following the criteria specified in Ref. 14: Precedence was given to regions of high contrast; larger regions; regions containing objects in the foreground; regions containing plants, animals, and humans. A primary ROI was defined as the region(s) containing the most interesting portions of the image (e.g., a human face, if present); a secondary ROI was defined as the region(s) containing portions which were less interesting than the primary ROI, but which contained subject matter that was still considered interesting. Largely, this hand segmentation isolated important objects and groups of objects in the images.

A corresponding ternary mask, matched in dimensions to each original image, was then created with pixel values in the range 0−2; pixel values of 2 denoted the primary ROI(s), pixel values of 1 denoted the secondary ROI(s), and pixel values of 0 denoted the non-ROI. Figure 2 shows an example of several images and their corresponding ternary ROI masks.

PSNR, SSIM, and WSNR were used as metrics of fidelity. PSNR and SSIM were used identically to the equations presented in Experiment I, except applied to each section of the ternary mask rather than the block based sections. WSNR uses the contrast sensitivity function (CSF) to weight differences between the Fourier spectrum of the original and distorted images. The WSNR metric requires knowledge of viewing conditions, which was not reported with the LIVE image database. Accordingly, WSNR was computed by assuming parameters which provide a reasonable approximation of typical viewing conditions: A display resolution of 96 pixels/inch (37.8 pixels/cm), and a viewing distance of 19.1 inches (48.5 cm; approximately 3.5 picture heights).

To incorporate ROI information, the distortion measures from each metric were computed separately for each region (i.e., for pixels corresponding to the primary ROI, pixels corresponding to the secondary ROI, and pixels corresponding to the non-ROI). Specifically, let $P_{\text{1st-ROI}}$, $P_{\text{2nd-ROI}}$, and $P_{\text{non-ROI}}$ denote the subset of pixel indices within an image corresponding to each region. Let $N_{\text{1st-ROI}}$, $N_{\text{2nd-ROI}}$, and $N_{\text{non-ROI}}$ denote the respective number of pixels in these regions. The scalar outputs of each metric, regionally defined as $x_{\text{1st-ROI}}$, $x_{\text{2nd-ROI}}$, and $x_{\text{non-ROI}}$, were computed as follows:

For PSNR and WSNR, $x_{\text{1st-ROI}}$, $x_{\text{2nd-ROI}}$, and $x_{\text{non-ROI}}$ were computed via

$$x_{\text{1st-ROI}} = 10\log_{10}\left(\frac{S}{D_{\text{1st-ROI}}}\right), \text{ where } D_{\text{1st-ROI}} = \frac{1}{N_{\text{1st-ROI}}}\sum_{i\in P_{\text{1st-ROI}}} E_i^2 \tag{3}$$

$$x_{\text{2nd-ROI}} = 10\log_{10}\left(\frac{S}{D_{\text{2nd-ROI}}}\right), \text{ where } D_{\text{2nd-ROI}} = \frac{1}{N_{\text{2nd-ROI}}}\sum_{i\in P_{\text{2nd-ROI}}} E_i^2 \tag{4}$$

$$x_{\text{non-ROI}} = 10\log_{10}\left(\frac{S}{D_{\text{non-ROI}}}\right), \text{ where } D_{\text{non-ROI}} = \frac{1}{N_{\text{non-ROI}}}\sum_{i\in P_{\text{non-ROI}}} E_i^2 \tag{5}$$

where S denotes the signal energy, and where E_i denotes a pixel in the error image $\mathbf{E}$ generated by each metric. For PSNR, $S = 255^2$ and $\mathbf{E} = \hat{\mathbf{I}} - \mathbf{I}$; for WSNR, S is the energy of the original image and $\mathbf{E}$ is a corresponding CSF-weighted error image. The error image was isolated using the blackening method described in Experiment I for VIF.

For SSIM, $x_{\text{1st-ROI}}$, $x_{\text{2nd-ROI}}$, and $x_{\text{non-ROI}}$ were computed via

$$x_{\text{1st-ROI}} = \frac{1}{N_{\text{1st-ROI}}}\sum_{i\in P_{\text{1st-ROI}}} \text{SSIM}(I_i, \hat{I}_i) \tag{6}$$

$$x_{\text{2nd-ROI}} = \frac{1}{N_{\text{2nd-ROI}}}\sum_{i\in P_{\text{2nd-ROI}}} \text{SSIM}(I_i, \hat{I}_i) \tag{7}$$

$$x_{\text{non-ROI}} = \frac{1}{N_{\text{non}-\text{ROI}}} \sum_{i \in P_{\text{non}-\text{ROI}}} \text{SSIM}(I_i, \hat{I}_i) \tag{8}$$

where $\text{SSIM}(I_i, \hat{I}_i)$ denotes a point in the SSIM index map between $\hat{\mathbf{I}}$ and $\mathbf{I}$. Note that the customary baseline version of SSIM, referred to as MSSIM in Ref. 10, is given by MSSIM = $1/N \sum_{i=1}^{N} \text{SSIM}(I_i, \hat{I}_i)$, where N denotes the total number of pixels in the image.

The augmented metrics were then taken as a weighted linear sum of $x_{\text{1st-ROI}}$, $x_{\text{2nd-ROI}}$, and $x_{\text{non-ROI}}$ as follows:

$$x = \alpha_{\text{1st-ROI}}\, x_{\text{1st-ROI}} + \alpha_{\text{2nd-ROI}}\, x_{\text{2nd-ROI}} + \alpha_{\text{non-ROI}}\, x_{\text{non-ROI}} \tag{9}$$

where x denotes the augmented metric output; and where $\alpha_{\text{1st-ROI}}$, $\alpha_{\text{2nd-ROI}}$, and $\alpha_{\text{non-ROI}}$ denote the weights constrained such that $\alpha_{\text{1st-ROI}} + \alpha_{\text{2nd-ROI}} + \alpha_{\text{non-ROI}} = 1$ and $\alpha_{\text{1st-ROI}} > 0$, $\alpha_{\text{2nd-ROI}} > 0$, $\alpha_{\text{non-ROI}} > 0$. Note that if $N_{\text{1st-ROI}} = N_{\text{2nd-ROI}} = N_{\text{non-ROI}}$, MSSIM (the customary version of SSIM) is equivalent to using $\alpha_{\text{1st-ROI}} = \alpha_{\text{2nd-ROI}} = \alpha_{\text{non-ROI}} = 1/3$. However, the use of equal weights does not yield expected versions of PSNR nor WSNR. It is also important to note that we are in no way claiming this weighted linear combination to be the optimal way of incorporating ROI information into each metric. Rather, this approach was chosen for its simplicity; it is among the most basic methods of augmenting these metrics to take into account ROI information. The primary goal of this experiment is to determine the values of $\alpha_{\text{1st-ROI}}$, $\alpha_{\text{2nd-ROI}}$, and $\alpha_{\text{non-ROI}}$ which give rise to the best and worst predictions of perceived fidelity.

Because this optimization had less dimensions than in Experiment I (3 as opposed to 25), the logistic transformation could be more rigorous without detrimentally increasing the runtime. For each group of ratings, a logistic function of the form[21]

$$f(x) = \frac{\tau_1 - \tau_2}{1 + \exp\left(\frac{x - \tau_3}{\tau_4}\right)} + \tau_2 \tag{10}$$

was fitted to the data via a Nelder-Mead search[18] to obtain the parameters τ_1, τ_2, τ_3, and τ_4 which minimized the sum-squared error between the transformed metric outputs {f(x)} and the corresponding subjective ratings (DMOS values). This logistic fitting procedure is recommended in the VQEG (Video Quality Experts Group) FRTV I Final Report.[22] This logistic fit was also applied to the baseline (constant weighted) fidelity metrics for comparison purposes. This was also repeated with the objective of *maximizing* the sum-squared error between the transformed metric outputs {f(x)} and the corresponding DMOS values.

The baseline and augmented metrics were applied to each original/distorted image pair in the LIVE database to obtain each metric's corresponding raw rating of fidelity. These raw ratings were then grouped according to distortion type in the following ways:

1. Group *ALL*, consisting of ratings for all 779 distorted images.
2. Group *JP2*, consisting of ratings for the 169 images containing JPEG-2000 compression distortion.
3. Group *JPG*, consisting of ratings for the 175 images containing JPEG compression distortion.
4. Group *NOZ*, consisting of ratings for the 145 images containing Gaussian white noise.
5. Group *BLR*, consisting of ratings for the 145 images containing Gaussian blurring.
6. Group *RAY*, consisting of ratings for the 145 images containing distortion induced via Rayleigh-distributed bit-stream errors in JPEG-2000 compressed streams of the images.

In summary, Experiment I resulted in 174 sets of data (three sets for each of the 29 images with augmented metrics and three sets for each of the 29 images using the baseline metrics). Experiment II resulted in twelve sets of data (two sets for each of the six distortion groups, one for correlation maximizing and one for correlation minimizing).

Table 1. The average improvement in correlation coefficient (CC) to DMOS for PSNR, SSIM and VIF. SSIM shows a marked improvement over the other metrics. However, the low improvement on all metrics suggests that the weight mask approach is not the ideal way to apply ROI information.

Metric	*Maximized CC*	*Constant Weight CC*	*Avg. Improvement*	*% Avg. Improvement*
PSNR	0.866	0.851	0.01473	1.7%
SSIM	0.899	0.858	0.04139	4.8%
VIF	0.958	0.949	0.00855	0.9%

3. RESULTS AND ANALYSIS

3.1 Experiment I

Improvement. The percent improvement in the augmented and baseline metric was calculated. Table 1 lists the maximized correlation coefficient, baseline (constant weight) correlation coefficient, and average improvement for each metric. All metrics show an improvement, the greatest of which is found for SSIM. Note that these improvements are similar to those found in Ref. 16. Although the performance of PSNR improves, it demonstrates substantially lesser correlations than SSIM or VIF. Figure 3 shows improvements in correlation on a per image basis. Notice that, except for a single image, SSIM has the most improvement, followed by VIF, and then PSNR.

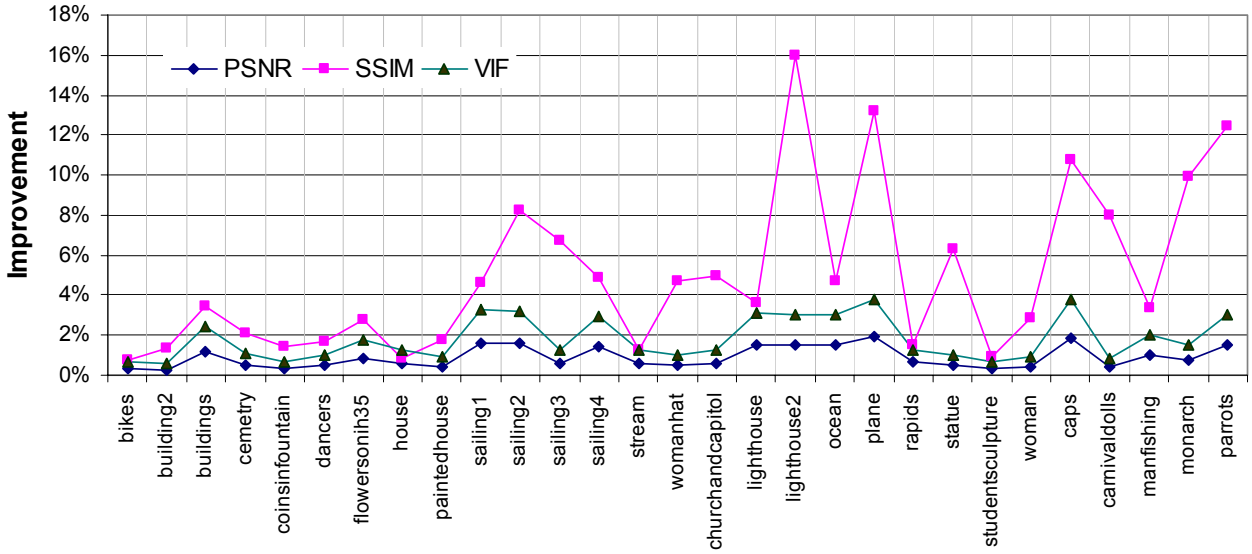

Figure 3. (Experiment I) The average improvement in correlation between fidelity measure and DMOS for each of the uncompressed images in the LIVE database. Table 1 shows that average from the above graph. Except for one image, SSIM shows the greatest room for improvement using block based weight masks.

Statistical Significance. To assess the statistical significance of each augmented metric's performance relative to the baseline version of that metric, an F-test was performed on the prediction errors.[21] Specifically, if the prediction errors (residuals) are assumed to be distributed according to a Gaussian distribution, an F-test can be used to assess whether the residuals from two metrics correspond to the same population, and can thus be used to determine if one metric has significantly larger residuals (greater prediction error) than another metric (see Refs. 21, 22). Let σ_A^2 and σ_B^2 denote the variance of the residuals from the augmented and baseline versions of a metric, respectively. The F statistic is given by $F = \sigma_A^2 / \sigma_B^2$. Values of $F > F_{critical}$ (or $F <$

$1/F_{critical}$) signify that, at a given confidence level, the augmented metric has significantly larger (or smaller)

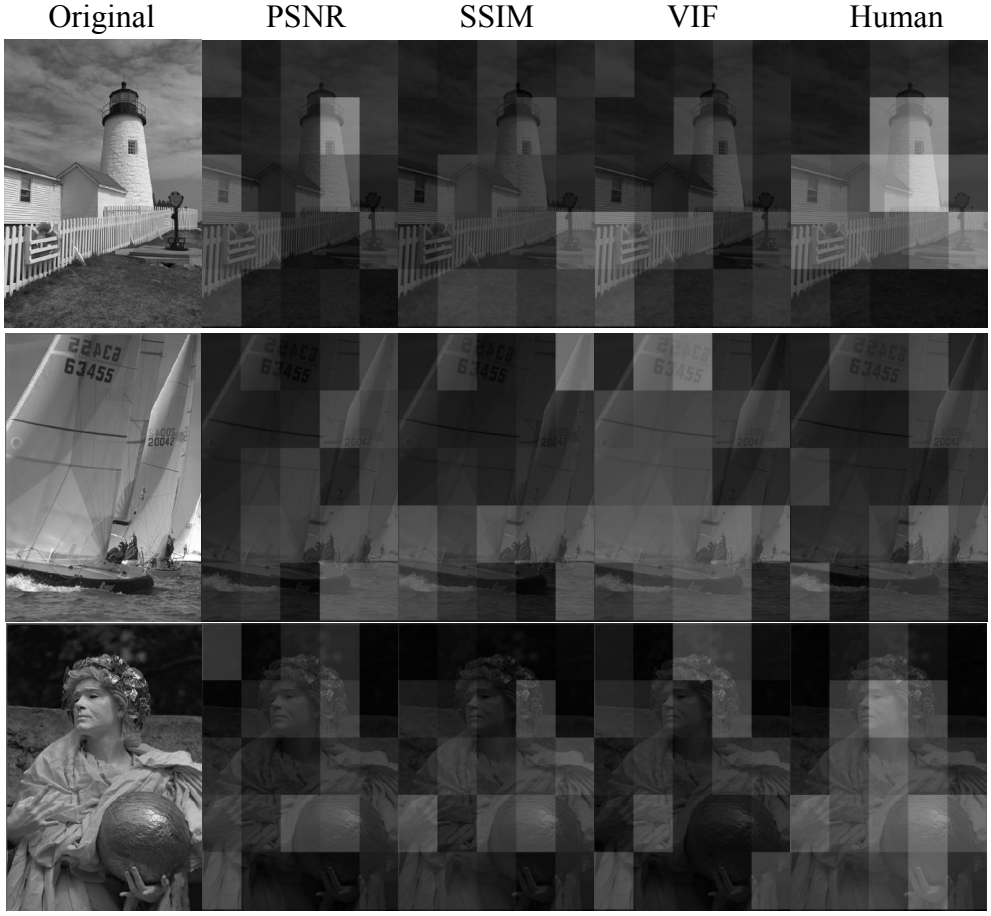

Figure 4. (Experiment I) The original images, correlation-maximizing weights for PSNR, SSIM, and VIF, and principal weights derived from hand segmented regions of interest. Weights have been overlaid on the original images with brighter blocks denoting greater weights.

residuals than the baseline metric, where $F_{critical}$ is computed based on the number of residuals and a confidence level.

The supplement in Ref. 23 lists the F statistics for Experiment I resulting from an F-test performed on the residuals from each metric's augmented version versus its baseline version at 95% confidence. Also listed in the supplement is the sample skewness and kurtosis of each metric's residuals, which can be used to gauge the Gaussianity of the residuals (the Gaussian density has a skewness of zero and a kurtosis of 3; commonly, kurtosis values between 2 − 4 are deemed Gaussian,[21] though see Ref. 24 for a more formal test using the JB statistic).

Only two of the improvements out of the 174 data fits are statistically significant. However, the two significant improvements have high kurtosis values and are therefore not well represented using a Gaussian density. This result is not altogether unexpected. Weighting the error space using blocks is a primitive way of incorporating ROI information. But even without statistically significant improvements, it is interesting to study the selected blocks that maximize correlation to perceived fidelity; namely, to investigate if greater weights correspond to regions of interest in the images.

Weight Interpretation. To determine to what degree the weights correspond to regions of interest, new *principal* weights were calculated from the hand-segmented regions from Experiment II. Using the ternary masks, it made sense to use a biased percentage of the primary, secondary, and non-ROI for each of the 25 blocks in each image to determine the principal block weight. Let P refer to the percentage of the primary ROI, secondary ROI, or non-ROI in the processed image block and C be a scaling constant for each percentage. The principal weight, I_{wght}, is determined as follows:

$$I_{wght} = C_{1stROI} \times P_{1stROI} + C_{2ndROI} \times P_{2nd\_ROI} + C_{nonROI} \times P_{nonROI} \quad (11)$$

The C constants are chosen intuitively. A block containing no ROI should receive less importance than a block containing ROI. Also, a block containing more secondary ROI than primary ROI should receive less importance than a block having the opposite relationship. With this in mind, the scaling constants were chosen intuitively as $C_{ROI} = 0.75$, $C_{sec\_ROI} = 0.25$, and $C_{non\_ROI} = 0$.

Three original images and versions of the images onto which the optimal weights have been overlaid are shown in Figure 4. Also shown in Figure 4 (rightmost column) are the weights derived based on the ternary region-of-interest masks. Observe that none of the metrics yield weights that are perfectly matched to the principal weights. The average mean-squared-error between the correlation maximizing weights and the principally assigned weights are 0.058, 0.065, and 0.059 for PSNR, SSIM, and VIF, respectively. The average correlation coefficient between the correlation maximizing weights and the subjectively determined weights are 0.133, 0.079, and 0.149 for PSNR, SSIM, and VIF, respectively. These data tend to suggest that although SSIM can benefit the most from block based spatial weighting, PSNR and VIF might benefit the most from spatial weighting based on region of interest information. We are currently in the process of extending the experiment to include other recent metrics of image fidelity, and perform a more rigorous optimization strategy using stochastic optimization procedures.

3.2 Experiment II

Improvement. To investigate the improvement of the metrics from baseline, separate fits of Equation (10) were applied to data from Groups *ALL*, *JP2*, *JPG*, *NOZ*, *BLR*, and *RAY*; and, separate correlation-maximizing and correlation-minimizing values of $\alpha_{1st\text{-}ROI}$, $\alpha_{2nd\text{-}ROI}$, and $\alpha_{non\text{-}ROI}$ were computed for each group. These data are listed in Tables 2-7 along with the rank-order correlation coefficients and root mean squared error values for each metric. Notice that, even when hand segmentations are used, augmented PSNR cannot perform better than baseline SSIM except when the distortions are limited to Gaussian white noise.

Table 2. (Experiment II) Correlation coefficient, rank-order correlation coefficient, and RMSE between subjective ratings for the LIVE database and transformed metric outputs for Group *ALL* (all 779 distorted images).

Metric	*Version*	$\alpha_{1st\text{-}ROI}$	$\alpha_{2nd\text{-}ROI}$	$\alpha_{non\text{-}ROI}$	*Corr. Coeff.*	*Rank-Ord. CC*	*RMSE*
PSNR	Baseline	–	–	–	0.822	0.820	9.127
	Corr. Max.	0.610	0.390	0.000	0.843	0.841	8.651
	Corr. Min.	0.000	0.000	1.000	0.743	0.741	10.782
WSNR	Baseline	–	–	–	0.862	0.863	8.192
	Corr. Max.	0.772	0.228	0.000	0.875	0.877	7.790
	Corr. Min.	0.000	0.000	1.000	0.767	0.770	10.329
SSIM	Baseline	–	–	–	0.903	0.900	6.925
	Corr. Max.	0.479	0.465	0.056	0.912	0.916	6.622
	Corr. Min.	0.000	0.000	1.000	0.832	0.831	8.935

Table 3. (Experiment II) Correlation coefficient, rank-order correlation coefficient, and RMSE between subjective ratings for the LIVE database and transformed metric outputs for Group *JP2* (169 images containing JPEG-2000 compression distortion).

Metric	*Version*	$\alpha_{1st\text{-}ROI}$	$\alpha_{2nd\text{-}ROI}$	$\alpha_{non\text{-}ROI}$	*Corr. Coeff.*	*Rank-Ord. CC*	*RMSE*
PSNR	Baseline	–	–	–	0.896	0.889	7.193
	Corr. Max.	0.610	0.390	0.000	0.910	0.903	6.730
	Corr. Min.	0.000	0.306	0.694	0.838	0.834	8.832
WSNR	Baseline	–	–	–	0.896	0.893	7.179
	Corr. Max.	0.898	0.102	0.000	0.923	0.918	6.223
	Corr. Min.	0.003	0.000	0.997	0.762	0.755	10.482
SSIM	Baseline	–	–	–	0.956	0.952	4.737
	Corr. Max.	0.508	0.394	0.097	0.957	0.952	4.719
	Corr. Min.	0.001	0.000	0.999	0.866	0.868	8.093

Table 4. (Experiment II) Correlation coefficient, rank-order correlation coefficient, and RMSE between subjective ratings for the LIVE database and transformed metric outputs for Group *JPG* (175 images containing JPEG compression distortion).

Metric	*Version*	$\alpha_{1st\text{-}ROI}$	$\alpha_{2nd\text{-}ROI}$	$\alpha_{non\text{-}ROI}$	*Corr. Coeff.*	*Rank-Ord. CC*	*RMSE*
PSNR	Baseline	–	–	–	0.860	0.841	8.160
	Corr. Max.	0.610	0.390	0.000	0.879	0.855	7.630
	Corr. Min.	0.098	0.003	0.899	0.802	0.797	9.555
WSNR	Baseline	–	–	–	0.899	0.876	6.996
	Corr. Max.	0.744	0.256	0.000	0.906	0.883	6.784
	Corr. Min.	0.003	0.001	0.996	0.831	0.820	8.888
SSIM	Baseline	–	–	–	0.943	0.911	5.339
	Corr. Max.	0.235	0.537	0.228	0.944	0.909	5.279
	Corr. Min.	0.001	0.000	0.999	0.753	0.892	10.628

Table 5. (Experiment II) Correlation coefficient, rank-order correlation coefficient, and RMSE between subjective ratings for the LIVE database and transformed metric outputs for Group *NOZ* (145 images containing Gaussian white noise).

Metric	*Version*	$\alpha_{1st\text{-}ROI}$	$\alpha_{2nd\text{-}ROI}$	$\alpha_{non\text{-}ROI}$	*Corr. Coeff.*	*Rank-Ord. CC*	*RMSE*
PSNR	Baseline	–	–	–	0.986	0.985	2.681
	Corr. Max.	0.691	0.279	0.000	0.987	0.985	2.610
	Corr. Min.	0.000	0.106	0.864	0.985	0.983	2.752
WSNR	Baseline	–	–	–	0.971	0.968	3.810
	Corr. Max.	0.828	0.172	0.000	0.973	0.969	3.654
	Corr. Min.	0.003	0.000	0.997	0.942	0.941	5.374
SSIM	Baseline	–	–	–	0.959	0.969	4.551
	Corr. Max.	0.394	0.310	0.296	0.972	0.978	3.756
	Corr. Min.	0.001	0.000	0.999	0.946	0.951	5.181

Table 6. (Experiment II) Correlation coefficient, rank-order correlation coefficient, and RMSE between subjective ratings for the LIVE database and transformed metric outputs for Group *BLR* (145 images containing Gaussian blurring).

Metric	*Version*	$\alpha_{1st\text{-}ROI}$	$\alpha_{2nd\text{-}ROI}$	$\alpha_{non\text{-}ROI}$	*Corr. Coeff.*	*Rank-Ord. CC*	*RMSE*
PSNR	Baseline	–	–	–	0.784	0.780	9.769
	Corr. Max.	0.513	0.411	0.077	0.837	0.834	8.597
	Corr. Min.	0.000	0.372	0.628	0.663	0.710	11.773
WSNR	Baseline	–	–	–	0.857	0.858	8.099
	Corr.Max.	0.791	0.121	0.088	0.883	0.889	7.376
	Corr. Min.	0.002	0.946	0.053	0.597	0.660	12.611
SSIM	Baseline	–	–	–	0.945	0.955	5.228
	Corr.Max.	0.609	0.328	0.063	0.961	0.959	5.133
	Corr. Min.	0.000	0.089	0.910	0.720	0.750	10.907

Table 7. (Experiment II) Correlation coefficient, rank-order correlation coefficient, and RMSE between subjective ratings for the LIVE database and transformed metric outputs for Group *RAY* (145 images containing distortion induced via bit-stream errors in JPEG-2000 compressed streams).

Metric	*Version*	$\alpha_{1st\text{-}ROI}$	$\alpha_{2nd\text{-}ROI}$	$\alpha_{non\text{-}ROI}$	*Corr. Coeff.*	*Rank-Ord. CC*	*RMSE*
PSNR	Baseline	–	–	–	0.885	0.893	7.665
	Corr. Max.	0.678	0.281	0.041	0.913	0.913	6.707
	Corr. Min.	0.003	0.000	0.997	0.789	0.790	10.106
WSNR	Baseline	–	–	–	0.902	0.915	7.110
	Corr. Max.	0.733	0.267	0.000	0.928	0.930	6.137
	Corr. Min.	0.001	0.005	0.995	0.796	0.794	10.112
SSIM	Baseline	–	–	–	0.948	0.951	5.143
	Corr. Max.	0.419	0.345	0.237	0.950	0.963	4.352
	Corr. Min.	0.003	0.000	0.997	0.789	0.887	9.949

To further quantify the predictive performance of each metric in Experiment II, the correlation coefficient, rank-order correlation coefficient, and root mean squared error were computed between the transformed metric outputs {f(x)} and DMOS values for the baseline and correlation-maximizing and correlation-minimizing metrics. Figures 5(a)-5(c) depict plots of {f(x)} vs. DMOS for Group *ALL* obtained when using the baseline metrics (i.e., when supplying no explicit ROI information).

In each graph, the horizontal axis denotes the transformed metric outputs {f(x)}, and the vertical axis denotes DMOS; each data point corresponds to a particular distorted image. The solid line in each graph denotes the best-fitting straight line; values of *R* denote correlation coefficients between {f(x)} and DMOS. Figures 5(d)-5(i) depict plots of {f(x)} vs. DMOS for Group *ALL* obtained when using the augmented metrics. These correlation-maximizing and correlation-minimizing weights are provided on each graph and are also listed in Table 2 which additionally lists the rank-order correlation coefficients and root mean squared error values for each metric.

Statistical Significance. Table 8 lists the F statistic resulting from an F-test performed on the residuals from each metric's augmented version versus its baseline version for Experiment II. Two augmented metric versions were used: (1) the augmented version which makes use of $\alpha_{1st\text{-}ROI}$, $\alpha_{2nd\text{-}ROI}$, and $\alpha_{non\text{-}ROI}$ found to *maximize* correlation; and (2) the augmented version which makes use of $\alpha_{1st\text{-}ROI}$, $\alpha_{2nd\text{-}ROI}$, and $\alpha_{non\text{-}ROI}$ found to *minimize* correlation. Also listed in Table 8 is the sample skewness and kurtosis of each metric's residuals, which can be used to gauge the Gaussianity of the residuals. Values of $F > F_{critical}$ (or $F < 1/F_{critical}$), which are

shown in boldface, signify that with 95% confidence the augmented metric has significantly larger (or smaller) residuals than the baseline version on the corresponding set of images.

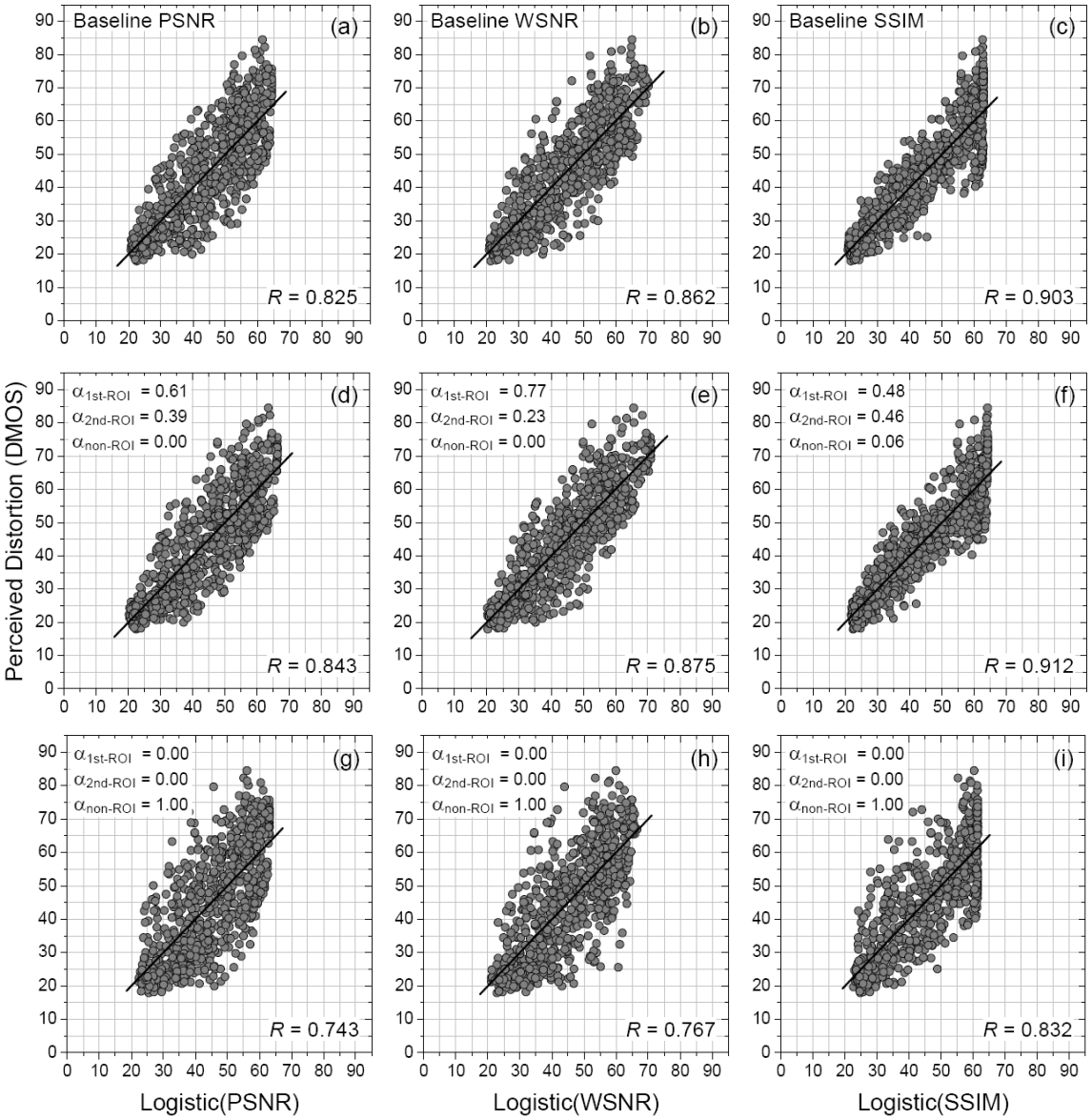

Figure 5. Logistic-transformed metric outputs plotted against corresponding DMOS values for all 779 distorted images of the LIVE image database (Experiment II). Graphs in the first, second, and third columns correspond to PSNR, WSNR, and SSIM, respectively. Graphs in the first row correspond to baseline versions of the metrics. Graphs in the second and third rows correspond to augmented metrics obtained by using the correlation-maximizing and correlation-minimizing weights, respectively. The solid line in each graph denotes the best-fitting first-order polynomial. Observe that when the weights are adjusted to maximize correlation, $\alpha_{\text{1st-ROI}}$ corresponds to the greatest of the three weights, whereas when the weights are adjusted to minimize correlation, $\alpha_{\text{non-ROI}}$ corresponds to the greatest of the three weights.

Table 8. (Experiment II) F statistic for each augmented metric's residuals tested against the corresponding baseline metric's residuals, and the sample skewness and kurtosis of each metric's residuals (italicized values denote non-Gaussian residuals based on the JB statistic[24]). Values of $F > F_{critical}$ (or $F < 1/F_{critical}$), shown in boldface, signify that with 95% confidence, the augmented metric has significantly larger (or smaller) residuals than the corresponding baseline version.

Measure	*Metric*	*Version*	*ALL*	*JP2*	*JPG*	*NOZ*	*BLR*	*RAY*
$1/F_{critical}$	–	–	0.889	0.775	0.779	0.760	0.760	0.760
$F_{critical}$	–	–	1.125	1.290	1.284	1.317	1.317	1.317
F Statistic	PSNR	Baseline	1	1	1	1	1	1
		Corr. Max.	0.903	0.875	0.875	0.948	0.774	0.765
		Corr. Min.	**1.403**	**1.508**	**1.372**	1.054	**1.452**	**1.738**
	WSNR	Baseline	1	1	1	1	1	1
		Corr. Max.	0.901	**0.751**	0.940	0.920	0.829	**0.745**
		Corr. Min.	**1.598**	**2.132**	***1.614***	**1.989**	**2.425**	**2.023**
	SSIM	Baseline	1	1	*1*	1	1	1
		Corr. Max.	0.914	0.992	*0.978*	**0.681**	**0.716**	0.964
		Corr. Min.	**1.665**	**2.918**	**3.963**	1.297	**4.496**	**3.622**
Skewness	PSNR	Baseline	-0.10	0.31	0.15	-0.10	-0.31	0.22
		Corr. Max.	-0.26	0.07	-0.40	-0.08	-0.65	-0.26
		Corr. Min.	-0.25	-0.10	0.03	-0.14	-0.39	0.12
	WSNR	Baseline	-0.23	0.05	-0.39	-0.23	0.27	0.49
		Corr. Max.	-0.24	0.17	-0.50	-0.35	-0.19	0.35
		Corr. Min.	-0.26	-0.25	-0.71	0.43	-0.54	0.33
	SSIM	Baseline	0.04	-0.12	-0.63	0.08	0.11	0.01
		Corr. Max.	-0.24	0.14	-0.73	-0.02	-0.10	-0.13
		Corr. Min.	-0.46	-0.61	-0.39	-0.42	-0.47	-0.41
Kurtosis	PSNR	Baseline	2.66	3.22	3.41	2.96	2.84	3.00
		Corr. Max.	2.79	3.31	3.59	3.10	3.03	2.91
		Corr. Min.	2.75	3.02	3.40	2.80	2.63	2.76
	WSNR	Baseline	3.36	3.53	3.95	2.56	3.08	3.48
		Corr. Max.	3.35	3.75	3.68	2.79	2.60	3.41
		Corr. Min.	3.58	3.29	*4.26*	3.57	2.80	3.82
	SSIM	Baseline	3.42	2.98	*4.65*	2.13	2.93	2.91
		Corr. Max.	3.39	2.92	*4.40*	2.24	2.81	2.52
		Corr. Min.	3.32	3.72	2.44	2.61	2.52	2.66

The results of Table 8 reveal that the improvements in predictive performance achieved by using the correlation-maximizing weights are statistically significant for only a very limited number of metric/distortion-type combinations. For the data of Group *ALL*, none of the improvements are significant; though, the improvements would be significant at confidence levels of 92%, 92%, and 89%, for PSNR, WSNR, and SSIM, respectively. This bodes well with the results from Experiment I that SSIM has less to gain from spatially varying regions based upon interest.

Weight Interpretation. To investigate the relationship between the weights $\alpha_{1st\text{-}ROI}$, $\alpha_{2nd\text{-}ROI}$, and $\alpha_{non\text{-}ROI}$ and the predictive performance of each metric, correlation coefficients between DMOS and the logistic-transformed augmented metrics {f(x)} for Group *ALL* were computed as a function of the weights. The resulting correlation-coefficient surfaces are depicted in Figure 6. Figure 6(a) depicts the results for PSNR; Figure 6(b) depicts the results for WSNR; Figure 6(c) depicts the results for SSIM.

In each graph, the z axis denotes correlation coefficient; the x and y axes represent $\alpha_{1st\text{-}ROI}$ and $\alpha_{non\text{-}ROI}$, respectively; $\alpha_{2nd\text{-}ROI}$ is given by $1 - \alpha_{1st\text{-}ROI} - \alpha_{non\text{-}ROI}$. Observe from these data that correlation between DMOS and the augmented metrics follows an intuitive pattern, increasing when $\alpha_{1st\text{-}ROI} > \alpha_{2nd\text{-}ROI} > \alpha_{non\text{-}$

ROI. The values of $\alpha_{1st\text{-}ROI}$, $\alpha_{2nd\text{-}ROI}$, and $\alpha_{non\text{-}ROI}$ which were found to *maximize* the correlation between {f(x)} and DMOS were as follows: For PSNR, $\alpha_{1st\text{-}ROI} = 0.61$, $\alpha_{2nd\text{-}ROI} = 0.39$, and $\alpha_{non\text{-}ROI} = 0$. For WNSR, $\alpha_{1st\text{-}ROI} = 0.77$, $\alpha_{2nd\text{-}ROI} = 0.23$, and $\alpha_{non\text{-}ROI} = 0$. For SSIM, $\alpha_{1st\text{-}ROI} = 0.48$, $\alpha_{2nd\text{-}ROI} = 0.46$, and $\alpha_{non\text{-}ROI} = 0.06$. Again, similarity between the primary and secondary weights reveals that SSIM shows less promise for spatial weighting based upon interest alone. The values of $\alpha_{1st\text{-}ROI}$, $\alpha_{2nd\text{-}ROI}$, and $\alpha_{non\text{-}ROI}$ which were found to *minimize* the correlation between {f(x)} and DMOS were $\alpha_{1st\text{-}ROI} = 0$, $\alpha_{2nd\text{-}ROI} = 0$, and $\alpha_{non\text{-}ROI} = 1$ for all three metrics.

It is open question how particular values of these weights ultimately translate to the perceptual contributions of the corresponding regions toward overall fidelity. Observe from these data, however, that when the weights are adjusted to maximize correlation, $\alpha_{1st\text{-}ROI}$ corresponds to the greatest of the three weights, followed by $\alpha_{2nd\text{-}ROI}$, and then $\alpha_{non\text{-}ROI}$. When the weights are adjusted to minimize correlation, $\alpha_{non\text{-}ROI}$ corresponds to the greatest of the three weights.

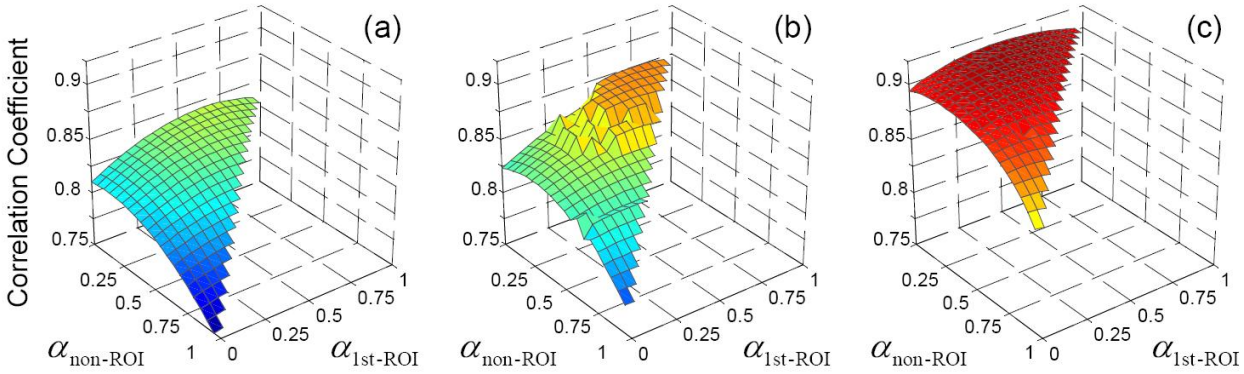

Figure 6. Correlation coefficients between DMOS and the logistic-transformed augmented metrics as a function of the weights $\alpha_{1st\text{-}ROI}$, $\alpha_{2nd\text{-}ROI}$, and $\alpha_{non\text{-}ROI}$ (Experiment II). (a) Correlation coefficient between DMOS and logistic-transformed, augmented PSNR; (b) Correlation coefficient between DMOS and logistic-transformed, augmented WSNR; (c) Correlation coefficient between DMOS and logistic-transformed, augmented SSIM. The z axis denotes correlation coefficient; the x and y axes represent $\alpha_{1st\text{-}ROI}$ and $\alpha_{non\text{-}ROI}$, respectively; $\alpha_{2nd\text{-}ROI}$ is given by $1 - \alpha_{1st\text{-}ROI} - \alpha_{non\text{-}ROI}$. The standard deviations of correlation coefficients are 0.023, 0.023, and 0.013, for PSNR, WSNR, and SSIM, respectively. The ranges of correlation coefficients are approximately 0.75 − 0.84, 0.78 − 0.88, and 0.84 − 0.91, for PSNR, WSNR, and SSIM, respectively.

It is important to note that the weighting method employed here is but one way of incorporating ROI information into these metrics. The central goal of this experiment was to discover the weights $\alpha_{1st\text{-}ROI}$, $\alpha_{2nd\text{-}ROI}$, and $\alpha_{non\text{-}ROI}$ which, within the limitations of weighted linear sum given by Equation (9), gave rise to the minimum and maximum correlations. It is quite possible that one may obtain different findings on statistical significance by augmenting the metrics in different ways. However, we believe that the relative weightings presented here can provide important insights into how future studies might better incorporate ROI information into these and other metrics.

4. CONCLUSIONS

This paper presented two experiments designed to investigate whether the success of recent image fidelity metrics can be attributed to the fact that these metrics implicitly incorporate region-of-interest information. Experiment I modified versions of PSNR, SSIM, and VIF by incorporating block based spatially varying weights chosen to maximize correlation between each metric and DMOS values from the LIVE image database. Experiment II examined from a computational standpoint the relative contributions of primary, secondary, and non-ROIs toward overall perceived fidelity. The results of both experiments revealed that:

- PSNR cannot achieve the same level of performance as SSIM or VIF using spatially varying weights. It can achieve the same level of performance as WSNR.
- WSNR, SSIM, and VIF can partially be improved using spatial weighting, although the improvements are not statistically significant.
- Within the limitations of the weighted linear model presented, the best correlation with subjective ratings of fidelity was achieved when PSNR, WSNR, and SSIM considered mainly the primary and secondary ROIs and largely ignored the non-ROI.
- The resulting block weights from Experiment I correspond to regions of interest for PSNR and VIF, but not for SSIM.
- The weights for regions considered the primary ROI, secondary ROI, and non-ROI follow an intuitive weighting pattern.

The results indicate that the success of SSIM and VIF are not due to an implicit incorporation of region-of-interest information. Overall, we hope that the results of these experiments can provide important insights into how future studies might further explore the relationship between regions of interest and perceived fidelity.

ACKNOWLEDGMENTS

We would like to acknowledge Hamid Sheikh and colleagues for the contribution of the LIVE image database without which this work would not have been possible.

REFERENCES

1. J. L. Mannos and D. J. Sakrison, "The Effects of a Visual Fidelity Criterion on the Encoding of Image," *IEEE Trans. Info. Theory 20*, 525–535 (1974).
2. F. Lukas and Z. Budrikis, "Picture Quality Prediction Based on a Visual Model," *IEEE Transactions on Communications 30*, 1679–1692 (1982).
3. N. Nill, "A Visual Model Weighted Cosine Transform for Image Compression and Quality Assessment," *IEEE Transactions on Communications 33,* 551–557 (1985).
4. S. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," *Digital Images and Human Vision*, A. B. Watson, ed., pp. 179–206 (1993).
5. P. C. Teo and D. J. Heeger, "Perceptual image distortion," *Proc. SPIE* 2179, 127–141 (1994).
6. S. J. P.Westen, R. L. Lagendijk, and J. Biemond, "Perceptual image quality based on a multiple channel HVS model," *Intl. Conf. Acoustics, Speech, and Signal Processing 4*, 2351–2354 (1995).
7. J. Lubin, "A visual discrimination model for imaging system design and evaluation," *Vision Models for Target Detection and Recognition,* E. Peli, ed. (World Scientific, 1995), pp. 245–283.
8. C. J. van den Branden Lambrecht, "A working spatio-temporal model of the human visual system for image representation and quality assessment applications," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* pp. 2291–2294 (1996).
9. S. Winkler, "Visual quality assessment using a contrast gain control model," *IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, pp. 527–532 (1999).
10. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.* 13, 600–612 (2004).
11. H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality," *IEEE Transactions on Image Processing*, Vol. 15, No. 2, pp. 430-444, 2006
12. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms,"

IEEE Transactions on Image Processing, Vol. 15, No. 11, pp. 3440-3451, 2006.
13. R. Dosselmann, X. D. Yang, "Existing and Emerging Image Quality Metrics", *IEEE Canadian Conference on Electrical and Computer Engineering 2005 (CCECE'05)*, Saskatoon, SK, May 2005.
14. W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors," *Proc. IEEE Int. Conf. on Image Processing 3*, 414–418 (1998).
15. M. Carnec, P. L. Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in ICIP 2003, vol. 2, 2003, pp. 185–188.
16. Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," IEEE International Conference on Image Processing, Atlanta, GA, Oct. 8-11, 2006.
17. H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at LIVE." Online. http://live.ece.utexas.edu/research/quality/.
18. J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *J. Comput. 7*, 308–313 (1965).
19. C. Moore, S. Yantis, and B. Vaughan, "Object-based visual selection: evidence from perceptual completion," *Psychological Science 9*, 104–110 (1998).
20. SSIM website: http://www.cns.nyu.edu/~zwang/files/research/ssim/index.html.
21. H. R. Sheikh, M. F. Sabir, , and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," IEEE Trans. Image Process.
22. VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," (2003), http://www.vqeg.org.
23. http://vision.okstate.edu/VCIP2008_F_stat_Supplement.html
24. A. K. Bera and C. M. Jarque, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," Econ. Letters 6, 255–259 (1980).