Visual Importance Pooling for Image Quality Assessment

Anush Krishna Moorthy and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—Recent image quality assessment (IQA) metrics achieve high correlation with human perception of image quality. Naturally, it is of interest to produce even better results. One promising method is to weight image quality measurements by visual importance. To this end, we describe two strategies—visual fixation-based weighting, and quality-based weighting. By contrast with some prior studies we find that these strategies can improve the correlations with subjective judgment significantly. We demonstrate improvements on the SSIM index in both its multiscale and single-scale versions, using the LIVE database as a test-bed.**

***Index Terms*—Image quality assessment (IQA), quality-based weighting, structural similarity, subjective quality assessment, visual fixations.**

I. Introduction

IMAGE quality assessment (IQA) is important for many applications. IQA methods fall into two categories: subjective assessment by humans and objective assessment by algorithms designed to mimic human subjectivity. While subjective assessment is the ultimate gauge of image quality, it is time-consuming, cumbersome, and cannot be implemented in systems where a real-time quality score for an image or video sequence is needed. Thus, algorithms which predict subjective image quality accurately and rapidly are of considerable value.

How "well" an algorithm performs is defined by how well it correlates with human perception of quality. To this end databases of images and subjective scores have been assembled, including the VQEG dataset [1] and the LIVE database [2]. In [3], a variety of leading IQA algorithms were tested and their performances were reported using statistical criteria such as the Spearman rank-order correlation coefficient (SROCC), the root mean square error (RMSE) (after nonlinear regression) and the linear correlation coefficient (CC) (after nonlinear regression) between the DMOS scores and the scores predicted by the algorithm. Amongst the algorithms tested, the Multi-Scale Structural SIMilarity Index (MS-SSIM) [4] and the Visual Information Fidelity index (VIF) [5] performed consistently well using of a variety of measures of correlation with the human perception of quality.

The MS-SSIM and the Single-Scale SSIM (SS-SSIM) [6] indices are particularly well-suited for application in real-time systems. SS-SSIM remains quite attractive owing to its extreme simplicity and excellent performance relative to old standards such as the MSE [7]. VIF is an alternate approach to IQA based on Natural Scene Statistics (NSS); however, in [8], it has been demonstrated that the SSIM and VIF metrics are equivalent. Hereafter, the acronym SSIM refers to either MS-SSIM or SS-SSIM. The larger acronyms will be used when there is a need to distinguish them.

In this paper, we explore the possibility of improving the performance of the SSIM metrics, by assigning visual importance weights to the SSIM values. In [6] and [4], a *mean* score is calculated at the end from the SSIM maps. While each level of MS-SSIM is scaled by a different parameter, this scaling reflects the importance of resolution on quality, but does not take into account any factors such as the visual importance of image features.

It is intuitively obvious that each region in an image may not bear the same importance as others. Visual importance has been explored in the context of visual saliency [9], fixation calculation [10], and foveated image and video compression [11]–[15]. However, region-of-interest based image quality assessment remains relatively unexplored. It is the furtherance of this area of quality assessment that motivates this paper.

Under the hypothesis that certain regions in an image may be visually more important than others, methods used to spatially pool the quality scores from the SSIM maps are an appealing possibility for improving SSIM scores. In [16], the effect of using different pooling strategies was evaluated, including local quality-based pooling. It was concluded that the best possible gains could be achieved by using an information-theoretic approach deploying "information content-weighted pooling." In this paper, we further investigate quality based pooling and also consider pooling based on predicted human gaze behavior.

There are two hypotheses which may influence human perception of image quality. The first is visual attention and gaze direction—"where" a human looks. The second hypothesis is that humans tend to perceive "poor" regions in an image with more severity than the "good" ones—and hence penalize images with even a small number of "poor" regions more heavily. Existing IQA algorithms, on the contrary, do not attempt to compensate for this prejudice. By weighting more heavily quality scores from lower scoring regions, such a compensation can be achieved. This idea of heavily weighting the lower scoring regions is a form of visual importance. This was also recognized by the authors of [16], who noted that weighting low

Manuscript received April 30, 2008; revised December 04, 2008. Current version published March 11, 2009. This work was supported in part by the National Science Foundation, in part by Texas Instruments, in part by Agilent, and in part by and Boeing. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Robert Safranek.

The authors are with the Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA (e-mail: anushmoorthy@mail.utexas.edu; bovik@ece.utexas.edu).

Digital Object Identifier 10.1109/JSTSP.2009.2015374

quality regions heavily made intuitive sense. However, their approach, which involved weighting quality scores as a monotonic function of quality, led to the conclusion that quality-weighted pooling yields incremental improvement, at best. By contrast, in our approach, we show that quite significant gains in performance can be obtained using the right strategy.

In this paper, we investigate pooling SSIM scores using the concepts of visual importance as gauged by a visual fixation predictor, and visual importance as gauged by heavily weighting the lowest SSIM map scores. These lowest SSIM map scores are pooled as sample percentiles. We attain improvements in both the single and the multiscale versions of SSIM.

The use of visual importance has been previously explored, albeit differently than the way we approach the issue. The authors in [17] evaluated the use of a number of factors that influence visual attention to produce an Importance Map (IM) [17], [18]. The IM was then used to weight IQA indices, resulting in measurable improvements.

However, in [19], the authors conclude that using visual fixations features do not improve SSIM performance. In [19], ground-truth data is used, while we use fixations generated from an algorithm. However, the authors observe, "It seems that the saliency information and the degradation intensity have to be jointly considered in the pooling function."

The rest of the paper is organized as follows. Section II reviews the Structural Similarity algorithms (both the single and the multiscale versions). Section III reviews the fixation finder that we use in this paper. Section IV reviews the concept of percentile scores. Section V explains how our proposed algorithm functions. We present the results of using our algorithm in Section VI and conclude the paper in Section VII.

II. Structural Similarity Index

The SSIM correlates quite well with human perception of image quality [3]. SS-SSIM and MS-SSIM are space-domain IQA metrics. There also exist non-spatial IQA extensions of SSIM such as the Complex-Wavelet SSIM index (CW-SSIM) [20], [21].

A. Single-Scale SSIM

Consider two aligned-discrete non-negative signals, $\mathbf{x} = \{x_i | i = 1, 2, \ldots, N\}$ and $\mathbf{y} = \{y_i | i = 1, 2, \ldots, N\}$. These can be two image patches from images under comparison, drawn from the same location in both images. Let μ_x, μ_y, σ_x^2, σ_y^2, and σ_{xy} be the means of $\mathbf{x}, \mathbf{y}$, the variances of $\mathbf{x}, \mathbf{y}$ and the covariance between $\mathbf{x}$ and $\mathbf{y}$, respectively.

The SSIM index evaluates three terms—luminance ($l(\mathbf{x}, \mathbf{y})$), contrast ($c(\mathbf{x}, \mathbf{y})$), and structure ($s(\mathbf{x}, \mathbf{y})$) [6]

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{1}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{3}$$

where $C_1 = (K_1 L)^2, C_2 = (K_2 L)^2, C_3 = C_2/2$ are small constants, L is the dynamic range of the pixel values, and $K_1 \ll 1$ and $K_2 \ll 1$ are scalar constants. Commonly, $K_1 = 0.01$ and $K_2 = 0.03$. The constants C_1, C_2 and C_3 prevent instabilities from arising when the denominator tends to zero.

The general form of the SSIM index between $\mathbf{x}$ and $\mathbf{y}$ is

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^\alpha \cdot [c(\mathbf{x}, \mathbf{y})]^\beta \cdot [s(\mathbf{x}, \mathbf{y})]^\gamma \tag{4}$$

where α, β, and γ are parameters which define the relative importance of the three components. Usually, $\alpha = \beta = \gamma = 1$, yielding

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{5}$$

At each coordinate, the SSIM index is calculated within a local window. As in [6], we use a 11×11 circular-symmetric Gaussian weighting function $w = \{w_i | i = 1, 2, \ldots, N\}$, with standard deviation of 1.5 samples, normalized to sum to unity ($\sum_{i=1}^N w_i = 1$). The statistics μ_x, μ_y, σ_x^2, σ_y^2, and σ_{xy} are then redefined as

$$\mu_x = \frac{1}{N}\sum_{i=1}^N w_i x_i$$

$$\mu_y = \frac{1}{N}\sum_{i=1}^N w_i y_i$$

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^N w_i (x_i - \mu_x)^2$$

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^N w_i (y_i - \mu_y)^2$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^N w_i (x_i - \mu_x)(y_i - \mu_y).$$

In most implementations, the ensemble mean SSIM index map is used to evaluate the overall image quality, which is a simple form of pooling.

B. Multi-Scale SSIM

The perception of image details is dependent upon a multitude of scale-related factors, including but not restricted to, the sampling density of the image signal, the distance from the image plane to the observer and the perceptual capability of the observer's visual system. In general, the subjective quality of an image also depends on these parameters. Moreover, images are naturally multiscale. To enable evaluation of image quality at multiple resolutions, in [4], the Multi-Scale SSIM (MS-SSIM) index was proposed.

In MS-SSIM, quality assessment is accomplished over multiple scales of the reference and distorted image patches (the signals defined as $\mathbf{x}$ and $\mathbf{y}$ in the previous discussion on SS-SSIM) by iteratively low-pass filtering and downsampling the signals by a factor of 2 (Fig. 1). The original image scale is indexed as

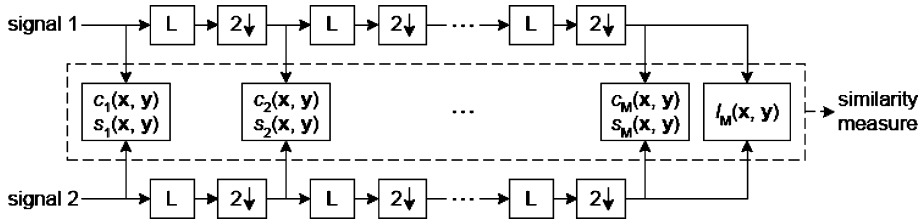

Fig. 1. Multi-scale SSIM. L = low-pass filtering, 2 ↓= downsampling by 2.

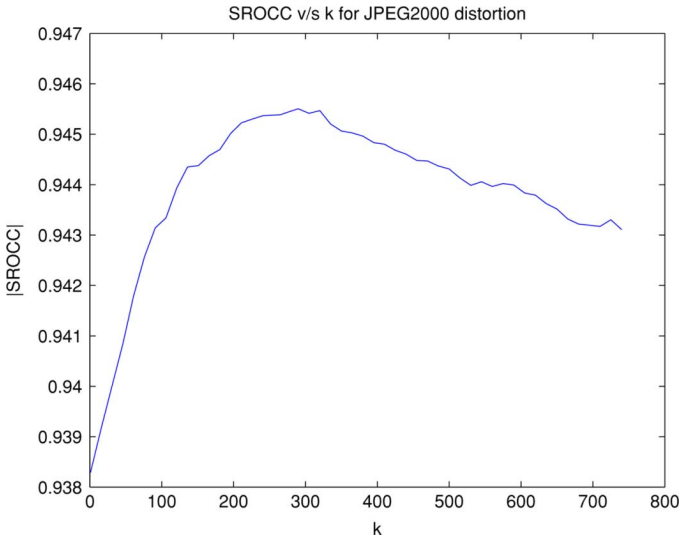

Fig. 2. Plot of $|\mathrm{SROCC}|$ as a function of k for SS-SSIM. The sample points were equally spaced values of k between 1 (indicates no weighting) and 750.

1, the first down-sampled version is indexed as 2 and so on. The highest scale M is obtained after $M-1$ iterations.

At each scale j, the contrast comparison (2) and the structure comparison (3) terms are calculated and denoted $c_j(\mathbf{x},\mathbf{y})$ and $s_j(\mathbf{x},\mathbf{y})$, respectively. The luminance comparison (1) term is computed only at scale M and is denoted $l_M(\mathbf{x},\mathbf{y})$. The overall SSIM evaluation is obtained by combining the measurement over scales

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l_M(\mathbf{x},\mathbf{y})]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(\mathbf{x},\mathbf{y})]^{\beta_j} \cdot [s_j(\mathbf{x},\mathbf{y})]^{\gamma_j} \tag{6}$$

The highest scale used here is $M=5$.

The exponents $\alpha_j, \beta_j, \gamma_j$ are selected such that $\alpha_j = \beta_j = \gamma_j$ and $\sum_{j=1}^{M} \gamma_j = 1$. The specific parameters used in [4] and here are $\alpha_1 = \beta_1 = \gamma_1 = 0.04448$, $\alpha_2 = \beta_2 = \gamma_2 = 0.2856$, $\alpha_3 = \beta_3 = \gamma_3 = 0.3001$, $\alpha_4 = \beta_4 = \gamma_4 = 0.2363$, and $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$, respectively. Again the spatial pooling strategy used in [4] was the ensemble mean.

III. GAFFE

Although human beings are continuously bombarded with a slew of visual data, the human visual system is streamlined to select and assimilate only those features that are relevant. When shown an image or a video sequence, the human visual system actively scans the visual scene using fixations linked by rapid, ballistic eye moments called saccades [22]. Most visual information is acquired during a fixation and little or no information is gathered during a saccade [23]. Hence, an understanding of how the human visual system selects certain regions for scrutiny is of great interest, not only in the areas of image compression and machine vision, but also in assessing the quality of images.

Given that certain regions in an image are more important than others; specifically, given that, when shown an image, a human tends to fixate at certain points on the image, it is of interest to develop algorithms that attempt to predict where a typical human looks on an average. In [10], this is summarized as the computer/machine vision researcher's dilemma: "How do we decide where to point the cameras next?" In an attempt to answer this question, researchers have produced some successful fixation finding algorithms, such as the one in [9] which seeks regions of high saliency, and the Gaze-Attentive Fixation Finding Engine (GAFFE), which uses image statistics measured at the point of gaze from actual visual fixations [10]. Here we use GAFFE to find points of potential visual importance for deciding IQA weights.

In [10], an experiment to record human eye moments was performed and the gaze coordinates corresponding to the human eye moments were recorded. This experiment was conducted on a subset of images from the van Hateren database of images [24], and these images as well as the eye movement data for each image is available as a part of the DOVES database [25], available online at [26]. The images selected from the van Hateren database were images that contained minimal contextual information, so as not to influence the fixations. Then, the features: luminance, contrast, luminance-bandpass, and contrast-bandpass were calculated at image patches around the humans' fixation points, and compared to the same features generated by randomly selected fixations. The randomly selected fixations were such that the statistics of the eye moments were maintained, even though the fixations themselves did not depend upon the underlying image being analyzed. Further, an eccentricity-based analysis was performed, where each image patch was associated with the length of a saccade. For each feature, and for each image, the ratio of the average patch features at eccentricity e of the observer's fixations and the randomly generated fixation was computed, and averaged across all the images and used in weighting the features. Bootstrapping [27] was then used to obtain the sampling distribution of this ratio to evaluate the statistical significance of the image statistic under consideration.

Given an image, GAFFE selects the center of the image as the first fixation, then foveates the image around this point. The foveated image is then filtered to create a fixation map, using the

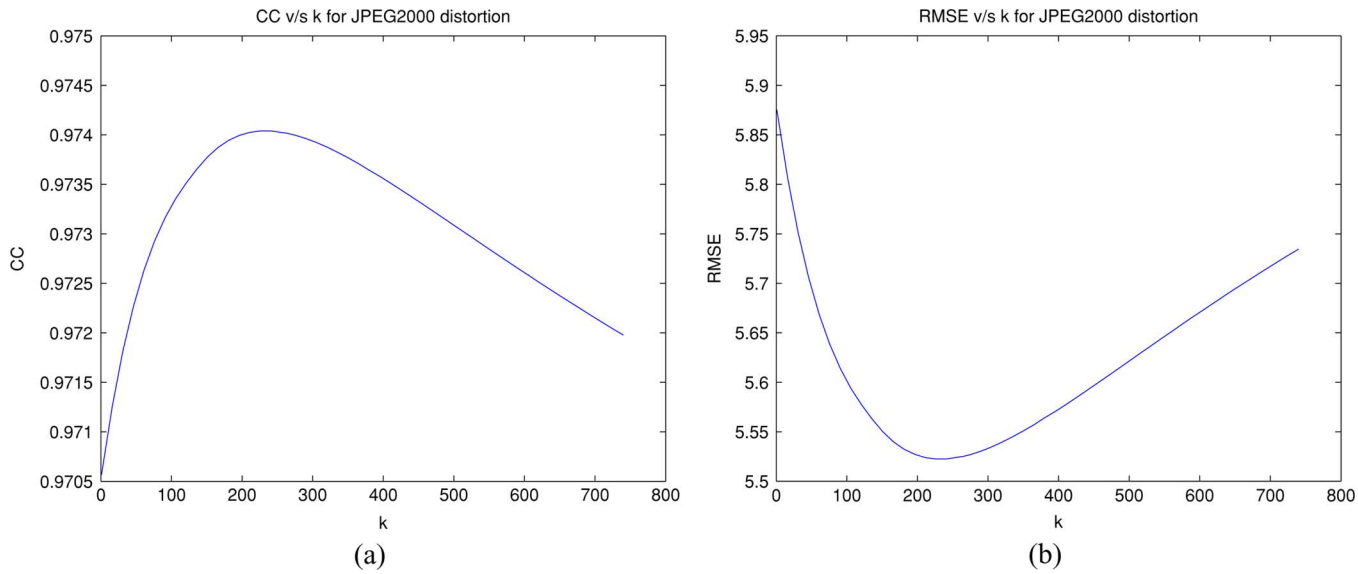

Fig. 3. Plot of (a) CC and (b) RMSE as a function of k for SS-SSIM. The sample points were equally spaced values of k between 1 (indicates no weighting) and 750.

above described features. The four feature maps thus obtained are linearly combined, where each feature is scaled by a factor γ_{feature}. The scaling factors are $\gamma_{\text{luminance}} = 1.04$, $\gamma_{\text{contrast}} = 1.12$, $\gamma_{\text{luminance-bandpass}} = 1.23$ and $\gamma_{\text{contrast-bandpass}} = 1.30$. These weights are normalized to sum to unity. The algorithm uses a greedy criterion to find the maximum value of the weighted selection map as the next fixation point, foveates the image around this point, then repeats the process. An inhibition-of-return mechanism using an inverted Gaussian mask centered at each fixation point is imposed so that fixations do not land very close to each other.

Thus, given an image, GAFFE algorithm outputs a set of vectors that define a set of the points which may correlate well with human fixations. It is important to note however that GAFFE was not designed to account for highly contextual cues, such as facial features, which are often fixation attractors.

A software implementation of GAFFE is available at [28] and the algorithm is explained in detail in [10].

IV. Percentile Scoring

Here, we define terms which will be used through the rest of the paper. The motivation for percentile scores is also explained.

A term that is commonly known amongst statisticians is quartiles [29]. Quartiles denote the lowest 25% values of an ordered set. In an ordered set, the first quartile is the set of the first 25% of the values, the second quartile is the next 25% and so on. Quartiles are actually the 4-quantiles; where quantiles are points taken at regular intervals from a distribution function. Similarly, deciles are the lowest 10% values obtained from an ordered set (the 10-quantiles). Generalizing this, the pth percentile of an ordered set is the lowest $p\%$ values of that set. Given a set, the elements are first ordered by ascending order of magnitude with the lowest $p\%$ values being denoted as the pth percentile. Recall our concept of the visual importance of low-quality image patches as defined in Section I. Our hypothesis suggests that regions of poor quality in an image can dominate the subjective perception of quality. A reasonable approach to utilize the visual importance of low-quality image patches is to more heavily weight the lowest $p\%$ scores obtained from a quality metric. This was done in [16] using several monotonic functions of the SSIM scores as the weights but with desultory effect. By contrast, we obtain substantial improvements by weighting the lowest percentiles heavily, as explained below.

In our further discussion involving percentile scores, assume that a quality map of the image has been found using one of the above discussed SSIM quality metrics, and that these values have been ordered by ascending value.

V. Visual Importance Pooling for SSIM

Here, we incorporate both modes of visual importance considered, fixations and percentile scores, to produce modified versions of the SS-SSIM and MS-SSIM indices. Specifically, we develop SSIM indices that use these features both individually and simultaneously.

Thus, three new versions of SSIM are considered: FIXATION-SSIM or F-SSIM, since GAFFE *fixations* are used to produce the SSIM score weights; Percentile-SSIM or P-SSIM, since the approach uses percentile weighting; and PF-SSIM, which combines the two modes of visual importance weighting to rate images.

A. F-SSIM

Given a set of image-coordinates that may be perceptually important—the fixations—two important decisions are required. First, how many fixations should be used per image? Second, given f fixations per image, what is the scaling factor k by which SSIM values at these fixations should be weighted relative to other pixels?

The number of fixations found in [10] were ten fixations/image (on an average). However, these fixations were generated based on the subjective study carried out, where each image was shown to the subject for 5 s. Conversely, the design of the LIVE

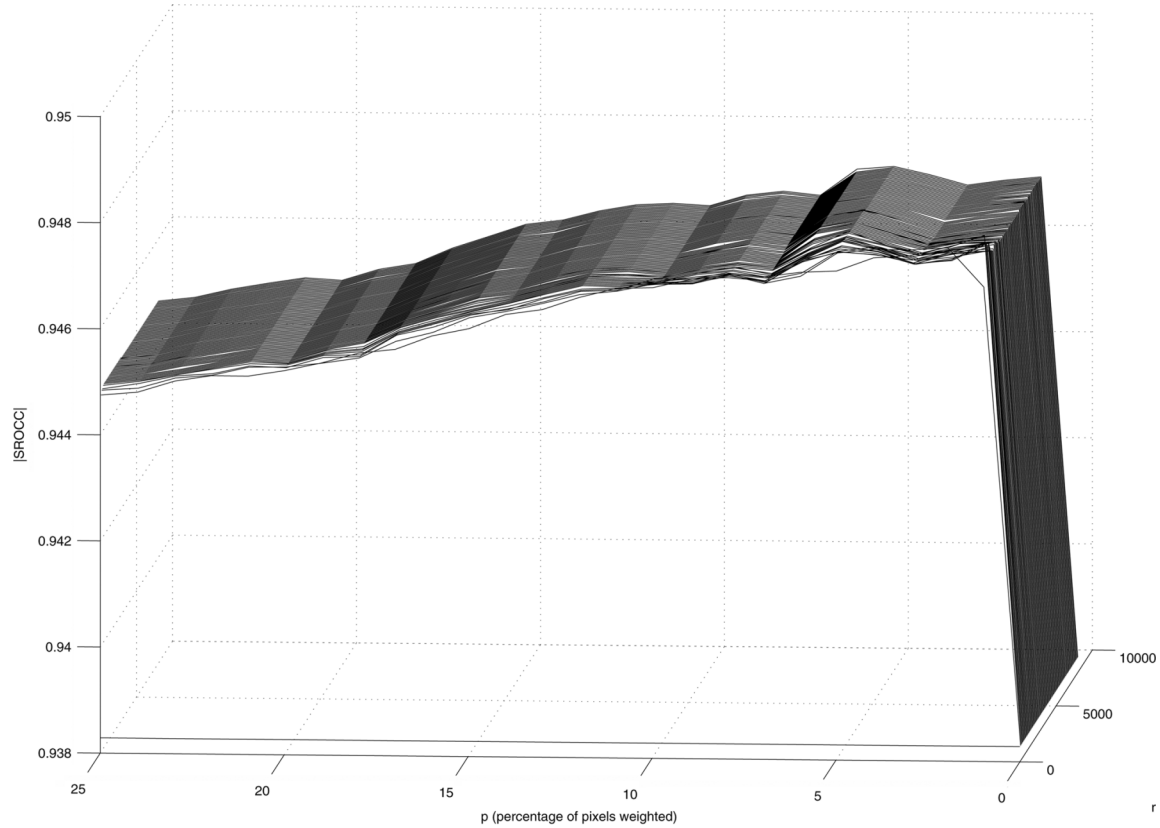

Fig. 4. Plot of $|\mathrm{SROCC}|$ as a function of p and r for SS-SSIM with JPEG2000 distortion. The p(percentage)-axis consists of values of p in the range 0%–25% with a step-size of 1%, the r(weights)-axis consists of values of r in the range 1–8000 with a step size of 100. SROCC value peaks at around 6%. Note that the cases $p = 0$ and $p = 100$ (not shown here) correspond to the original SSIM.

database [2] (which we use as the algorithm test-bed) did not employ time-restrictions during its creation. More specifically, the subjects were allowed to look at the images for as long as they wanted, until they made their decision. Hence, we elected to keep the number of fixations at a constant $f = 10$ (although GAFFE can be programmed to compute any number of fixations). Each fixation is extrapolated by a 11×11 2-D Gaussian function centered at the fixation. Since fixations are recorded at single coordinates and since areas of visual importance may be regional, the Gaussian interpolation used in GAFFE serves to associate the fixations with regions subtending a small visual angle. Each 11×11 region is then scaled by a factor k.

The peak values of the weights applied to the "fixated" regions (the Gaussian centers) relative to the weights of the non-fixated areas is in the ratio $k > 1$. The testing was performed by randomly selecting one of the types of distortion from the LIVE database [2] and simulating various values of k. We found that the value of k that maximizes the correlation between the objective and subjective scores (from the LIVE database) remained approximately the same over various distortion types. In Figs. 2 and 3, we see the absolute value of the Spearman rank ordered correlation coefficient (SROCC) the linear correlation coefficient (CC) and the RMSE plotted as a function of the weighting parameter k for SS-SSIM. Through such empirical testing we found a value $k = 265$ to yield good results, although varying this ratio in the range $125 \leq k \leq 375$ did not the change performance much.

Thus, the F-SSIM index is defined as

$$F - \mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{P} \sum_{j=1}^{Q} \mathrm{SSIM}(x_{ij}, y_{ij}) \cdot w_{ij}}{\sum_{i=1}^{P} \sum_{j=1}^{Q} w_{ij}} \quad (7)$$

where $\mathrm{SSIM}(x_{ij}, y_{ij})$ is the SSIM value at pixel location (i, j); P, Q are the image dimensions and w_{ij} are the SSIM weights.

For MS-SSIM, we reduce the size of the Gaussian mask progressively with the scale. The mask size at a scale M is given by

$$\mathrm{mask}_M = (11 - 2^{M-1}) \times (11 - 2^{M-1}).$$

At each scale, we reduce the number of fixations by a factor of two. Specifically

$$N_{\mathrm{fixations}}^{M} = \left\lceil \frac{10}{2^{M-1}} \right\rceil$$

where $\lceil x \rceil$ is the ceiling function, and M is the scaling index.

The pixels that do not fall under the fixation masks are left untouched: $w_{ij} = 1$.

B. P-SSIM

Here, we follow on the hypothesis that poor quality regions disproportionately affect subjective quality assessment. This suggests that weighting the scores by their rank ordering may produce better results [30]. Many ways of weighting are possible. Here, we consider simple percentile weighting, yet, the

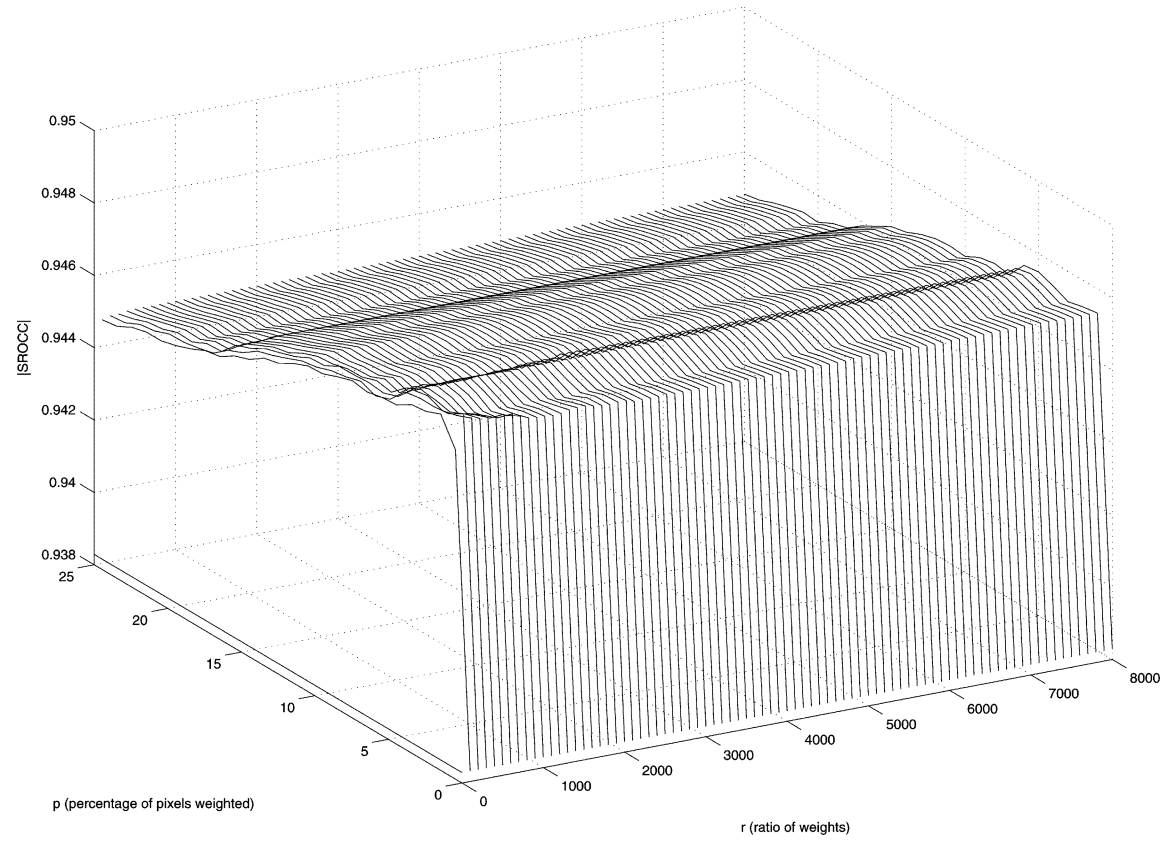

Fig. 5. Plot of $|\mathrm{SROCC}|$ as a function of p and r for SS-SSIM with JPEG2000 distortion. The p(percentage)-axis consists of values of p in the range 0%–25% with a step-size of 1%, the r(weights)-axis consists of values of r in the range 1–8000 with a step size of 100. Note a gradual increase at $p = 6\%$, with r, highest correlations are obtained when $1000 \leq r \leq 8000$. The case $r = 1$ corresponds to the original SSIM.

questions remain—what percentile should be used? and how much should we weight the percentile score by? In order to arrive at a solution we tried values of p from 5%–25% in 1% increments. Rather than using an arbitrary monotonic function of quality (such as the smooth power-law functions used in [16]), we use the statistical principle of heavily weighting the extreme values—in this case, lowest percentiles. Thus, the lowest $p\%$ of the SSIM scores are (equally) weighted. Non-equal weights of the rank-ordered SSIM values are possible, but we have not explored this deeper question [31], [32]. Similar to the approach used for F-SSIM a random subset of distortions from the LIVE database was selected and various values of p (5%–25% in 1% increments) were simulated. In our analysis, we found the value $p = 6\%$ yields good results; however, small perturbations in p do not alter the results drastically. We note that in [30] a similar form of pooling is used for video quality assessment, where only lowest 5% of the spatial scores are pooled together.

Given a SSIM map, we arrange the SSIM values in ascending order of the magnitude and scale the lowest $p\%$ of these values by a factor of r. Again, the ratio r by which these pixels are weighted is $r > 1$. Although we choose $r = 4000$, a variation of this ratio in the range $1000 \leq r \leq 8000$ did not affect the performance much. The pixels that do not fall within the percentile range, are left unchanged $w_{ij} = 1$. We note that this yielded better performance than when $w_{ij} = 0$ for the pixels that do not fall within the percentile range.

These empirical choices are validated by the results seen in Figs. 4–7, where 3-D plots of absolute value of SROCC, CC, and RMSE for SS-SSIM as a function of r and p are seen for JPEG2000 distortion. A clear peak is visible around $p = 6\%$, with the value of the SROCC, CC, and RMSE peaking/attaining a minimum in the range $1000 \leq r \leq 8000$. This trend remains unchanged across distortion types.

As was the case for F-SSIM, the implementation differs slightly when incorporated into MS-SSIM. Since the percentiles scores are a measure of a ratio given a set, we did not deem it necessary to reduce the percentile being weighted at each level. We, however, experimented with reducing the weights in the same way as discussed for F-SSIM. We found that such a weighting scheme did not notably improve the results. Indeed, we found that the greatest gains were achieved by weighting only the second level i.e., $M = 2$ of the multi-scale decomposed image set. This corroborates the observation made in [4], where the highest gains relative to SS-SSIM were achieved at $M = 2$.

C. Combined Percentile and Fixation-Based SSIM (PF-SSIM)

Since gains are achieved by using both of the individual concepts of calculated fixations and percentile scores (as will be demonstrated in Section VI), it is natural to consider using them together to further improve on the achieved gains. Hence, in PF-SSIM, first F-SSIM is implemented, then the values are sorted and weighted as described in P-SSIM. The values thus obtained are normalized to lie between 0 and 1. The order of implementation, i.e., F-SSIM followed by P-SSIM or

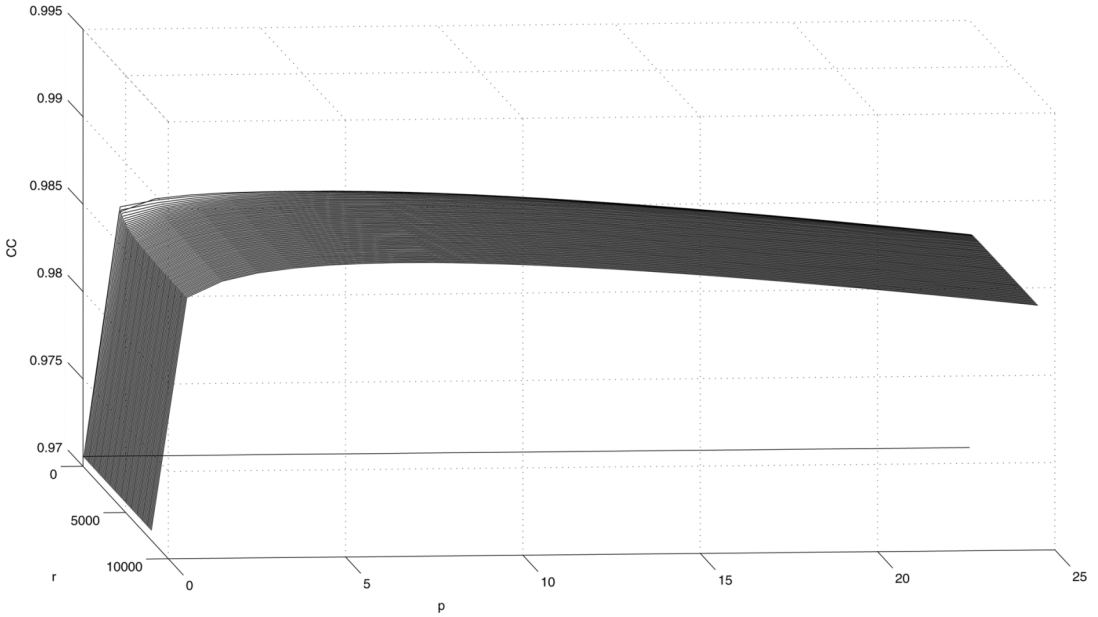

Fig. 6. Plot of CC as a function of p and r for SS-SSIM with JPEG2000 distortion. The p (percentage)-axis consists of values of p in the range 0%–25% with a step-size of 1%, the r(weights)-axis consists of values of r in the range 1–8000 with a step size of 100. CC value attains a maximum at around 6%. Note that the cases $p = 0$ and $p = 100$ (not shown here) correspond to the original SSIM.

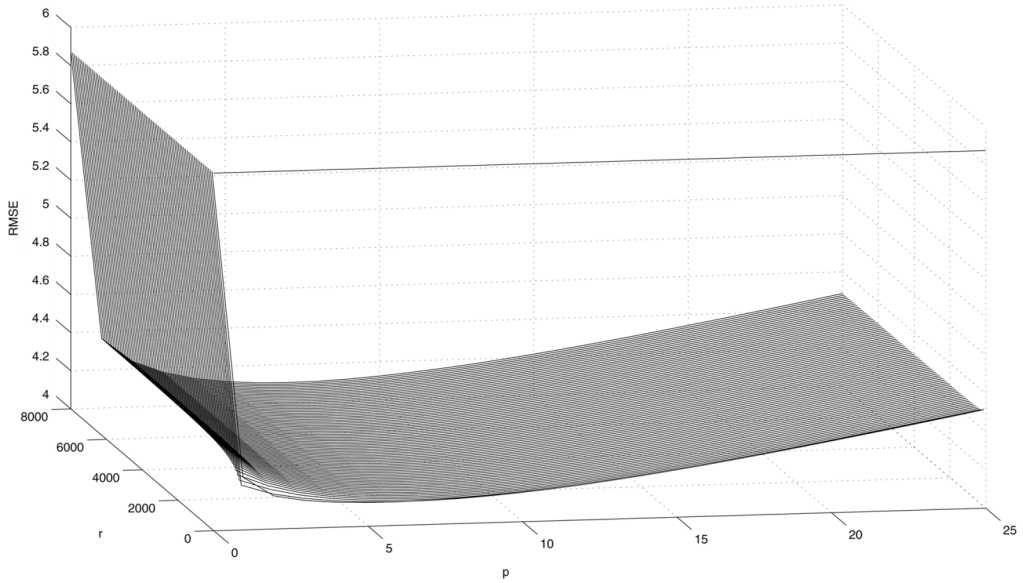

Fig. 7. Plot of RMSE as a function of p and r for SS-SSIM with JPEG2000 distortion. The p(percentage)-axis consists of values of p in the range 0%–25% with a step-size of 1%, the r(weights)-axis consists of values of r in the range 1–8000 with a step size of 100. RMSE value attains a minimum at around 6%. Note that the cases $p = 0$ and $p = 100$ (not shown here) correspond to the original SSIM.

vice-versa does not seem to change the results much, and hence we quote results for the order mentioned above only.

The parameters used for weighting the fixations and the percentile scores are given in Table I.

TABLE I
TABLE INDICATING THE WEIGHTS FOR F-SSIM, P-SSIM AND PF-SSIM. SS = SINGLE-SCALE, MS = MULTI-SCALE, M = SCALE OF RESOLUTION, EX., ORIGINAL IMAGE $\mathrm{M} = 1$, ONCE-DOWNSAMPLED IMAGE $\mathrm{M} = 2$, AND SO ON

	F-SSIM: k	P-SSIM: p	P-SSIM: r
SSIM(SS)	265	6%	4000
SSIM(MS)	$265/(2^{(M-1)})$	6%	4000

VI. RESULTS

A. Computed Scores

In order to validate the algorithm, the LIVE database of images was used as a test bed. The specific contents of the type of distortions present in the database are: JPEG2000: 227 images, JPEG: 233 images, White Noise: 174 images, Gaussian Blur: 174 images, Fast Fading: 174 images. The database includes DMOS subjective scores for each image.

We present the results for F-SSIM, P-SSIM, and PF-SSIM. The algorithms were evaluated against the DMOS scores using

TABLE II
LINEAR CORRELATION COEFFICIENT VALUES (AFTER NONLINEAR REGRESSION)—F-SSIM, P-SSIM, PF-SSIM (SINGLE-SCALE)

	JP2k	JPEG	WN	GBlur	FF	All data
SSIM (SS)	0.9706	0.9695	0.9508	0.9235	0.9598	0.9444
F-SSIM (SS)	0.9740	0.9700	0.9721	0.9394	0.9703	0.9526
P-SSIM (SS)	0.9853	0.9741	0.9725	0.9749	0.9746	0.9661
PF-SSIM (SS)	0.9847	0.9737	0.9824	0.9756	0.9789	0.9664

TABLE III
LINEAR CORRELATION COEFFICIENT (AFTER NONLINEAR REGRESSION) —F-SSIM, P-SSIM, PF-SSIM (MULTISCALE)

	JP2k	JPEG	WN	GBlur	FF	All data
SSIM (MS)	0.9677	0.9635	0.9787	0.9612	0.9483	0.9488
F-SSIM (MS)	0.9667	0.9628	0.9828	0.9622	0.9508	0.9501
P-SSIM (MS)	0.9695	0.9646	0.9950	0.9773	0.9686	0.9550
PF-SSIM (MS)	0.9695	0.9659	0.9938	0.9670	0.9677	0.9554

TABLE IV
RMSE (AFTER NONLINEAR REGRESSION)—F-SSIM, P-SSIM, PF-SSIM (SINGLE-SCALE)

	JP2k	JPEG	WN	GBlur	FF	All data
SSIM (SS)	5.8754	6.4217	9.7779	8.3553	6.2989	7.5988
F-SSIM (SS)	5.5270	6.5683	8.1382	7.5784	5.3560	7.0352
P-SSIM (SS)	4.1085	6.2347	7.6510	5.1345	5.3878	5.9646
PF-SSIM (SS)	5.3923	6.4385	7.1554	5.1740	5.2548	5.9383

TABLE V
RMSE (AFTER NONLINEAR REGRESSION)—F-SSIM, P-SSIM, PF-SSIM (MULTI-SCALE)

	JP2k	JPEG	WN	GBlur	FF	All data
SSIM (MS)	6.2269	6.4982	4.5092	5.9967	6.9240	7.3040
F-SSIM (MS)	6.2439	6.5577	4.0574	5.9264	6.8444	7.2145
P-SSIM (MS)	5.9744	6.3994	2.1873	4.5234	5.3715	6.8559
PF-SSIM (MS)	5.9800	6.2847	2.1034	5.3068	5.0508	6.8245

TABLE VI
SROCC VALUES—F-SSIM, P-SSIM, PF-SSIM (SINGLE-SCALE)

	JP2k	JPEG	WN	GBlur	FF	All data
SSIM (SS)	0.9383	0.9280	0.9704	0.9312	0.9552	0.9149
F-SSIM (SS)	0.9454	0.9288	0.9725	0.9606	0.9703	0.9287
P-SSIM (SS)	0.9474	0.9293	0.9833	0.9728	0.9620	0.9354
PF-SSIM (SS)	0.9545	0.9365	0.9853	0.9747	0.9656	0.9402

three popular metrics: the Spearman rank ordered correlation coefficient (SROCC), the linear correlation coefficient (CC) (after nonlinear regression), and the RMSE (after nonlinear regression). The nonlinearity chosen to fit the data is a five-parameter logistic function (a logistic function with an added linear term, and constrained to be monotonic) given by

$$\text{Quality}(x) = \beta_1 \text{logistic}(\beta_2, (x - \beta_3)) + \beta_4 x + \beta_5$$
$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x)}$$

where x is the score obtained from the objective metric.

The results are tabulated in Tables II–VII. In all tables, SS = Single-scale, MS = Multi-scale, WN = White Noise, Gblur = Gaussian Blur, and FF = Fast Fading.

We calculate all metrics for all distortions, for both MS-SSIM and SS-SSIM.

TABLE VII
SROCC VALUES—F-SSIM, P-SSIM, PF-SSIM (MULTI-SCALE)

	JP2k	JPEG	WN	GBlur	FF	All data
SSIM (MS)	0.9469	0.9304	0.9768	0.9670	0.9543	0.9420
F-SSIM (MS)	0.9470	0.9296	0.9812	0.9696	0.9600	0.9460
P-SSIM (MS)	0.9555	0.9376	0.9830	0.9733	0.9613	0.9464
PF-SSIM (MS)	0.9553	0.9389	0.9824	0.9735	0.9644	0.9469

B. Performance Metrics

It was noted in [3] that the SROCC operates only on the rank of the data points while assuming an equal spacing between the datapoints. Images which generate clustered scores, although different in rank, may not differ much in quality—since the scatter is only indicative of measurement noise. Further, since each point is treated with the same importance as the other points, data sets which exhibit saturation are not good candidates for evaluation by the SROCC. While we have included SROCC values for completeness, it may be argued as in [3] that the RMSE and the CC (after nonlinear regression) are better choices for measurement of quality across data-sets. However, it may also be argued that the nonlinear regression may affect results owing to the regression procedure. Even though all three metrics have certain drawbacks, we continue to evaluate IQA algorithms based on these metrics for want of a better analysis technique.

C. F-SSIM Performance

We used GAFFE as the fixation finding algorithm. Prior work closest in concept to ours used recorded visual attention [19]. The use of GAFFE was owing to its performance and reasonable computational expense. As explained earlier, the number of fixations were fixed at $f = 10$. Having said this, we still believe that the variation of the number of fixations given a fixation finding algorithm is a topic of interest, as are other fixation-finders or ROI-finders, such as image "saliency" [9]. We note that temporal saliency has recently been explored for video quality assessment [33].

The improvements afforded by F-SSIM were not across the board, and indeed were limited to the Gaussian blur and Fast Fading distortion types. These distortions tend to destroy the structure of perceptually significant features such as edges. This was true for both SS-SSIM and MS-SSIM, although the gains relative to SS-SSIM were much more substantial. This is to be expected, since MS-SSIM has a very high correlation with human subjectivity that may be pushing reasonable expected limits of algorithm performance.

The improvement in performance using P-SSIM was more substantial. Indeed, the improvement afforded by single-scale P-SSIM is so significant that it competes with standard MS-SSIM! This suggests that using percentile scoring in combination with simple SSIM is a viable alternative to the more complex MS-SSIM. Yet, using P-SSIM for MS-SSIM affords even better gains.

Finally, combining P-SSIM and F-SSIM into PF-SSIM produced desultory improvement, if any. While both approaches are individually successful, P-SSIM appears to be more so, and the benefits of combining them is not obvious. However, this may change as the state-of-the-art in "fixation-finding" evolves, as that field remains in a nascent state.

VII. Conclusion

We found that by visual importance weighting of the computed fixations and by using error percentile scores, better agreement with subjective scores can be produced for IQA metrics, in contradiction to prior studies in [16] and [19]. The increase in correlation between the modified IQA metrics and the subjective DMOS scores by the use of these concepts was demonstrated to be significant for both single-scale and multiscale SSIM, using the LIVE database of images. The degree of significance depended, in some cases, on the type of distortion.

References

[1] "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment," [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

[2] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, Live Image Quality Assessment Database Release 2. [Online]. Available: http://live.ece.utexas.edu/research/quality

[3] H. R. Sheikh, M. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[4] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, Nov. 2003, pp. 1398–1402.

[5] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Signal Process. Lett.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[7] Z. Wang and A. C. Bovik, "Mean Squared Error: Love it or Leave it?—A New Look at Fidelity Measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, 2009.

[8] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1200–1203.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[10] U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Gaffe: A gaze-attentive fixation finding engine," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 564–573, Apr. 2008.

[11] W. Geisler and J. Perry, "A real-time foveated multi-resolution system for low-bandwidth video communication," in *Proc. SPIE Human Vis. Electron. Imag.*, 1998, vol. 3299, pp. 294–305.

[12] L. Itti, "Automatic foveation for video compression using a neurobiological model for visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[13] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1397–1410, Oct. 2001.

[14] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[15] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 977–992, Jul. 2001.

[16] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *IEEE Int. Conf. Image Process.*, Sep. 2006, pp. 2945–2948.

[17] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors," in *Proc. Int. Conf. Image Process.*, 1998, pp. 414–418.

[18] A. Maeder, J. Diederich, and E. Niebur, "Limiting human perception for image sequences," *SPIE—Human Vis. Electron. Imag.*, vol. 2657, pp. 330–337, 1996.

[19] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barbba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *Proc. IEEE Int. Conf. Image Process. ICIP'07*, 2007, vol. 2, pp. 169–172.

[20] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'05)*, Mar. 2005, pp. 573–576.

[21] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. Markey, "Complex wavelet structural similarity: A new image quality index," *IEEE Trans. Image Process.*, submitted for publication.

[22] A. L. Yarbus, *Eye Movements and Vision*, 2nd ed. New York: Plenum Press, 1967.

[23] D. Burr, M. C. Morrone, and J. Ross, "Selective suppression of the magnocellular visual pathway during saccadic eye movements," *Nature*, vol. 371, no. 6497, pp. 511–513, Oct. 1994.

[24] J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," in *Proc. R. Soc. London B*, 1998, pp. 265:359–265:366.

[25] I. van der Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Doves: A database of visual eye movements," *Spatial Vis.*, pp. 161–177, 2009.

[26] I. van der Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Doves: A Database of visual eye movements," [Online]. Available: http://live.ece.utexas.edu/research/doves

[27] B. Efron, *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC, 1994.

[28] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "Gaffe: A gaze-attentive fixation finding engine," [Online]. Available: http://live.ece.utexas.edu/research/gaffe

[29] H. A. David, *Order Statistics*. New York: Wiley, 2003.

[30] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, no. 3, pp. 312–313, Sep. 2004.

[31] A. C. Bovik, T. S. Huang, and D. C. Munson, "A generalization of median filtering using linear combinations of order statistics.," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 6, no. 31, pp. 1342–1350, Dec. 1983.

[32] H. G. Longbotham and A. C. Bovik, "Theory of order statistic filters and their relationship to linear fir filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 2, pp. 275–287, 1989.

[33] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer.*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

Anush Krishna Moorthy received the B.E. degree in electronics and telecommunication from the University of Pune, Pune, India, in 2007. He is currently pursuing the M.S. degree in electrical engineering at The University of Texas at Austin.

He is currently the Assistant Director of the Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin. His research interests include image and video quality assessment, image and video compression, and computational vision.

Alan Conrad Bovik (S'80–M'81–SM'89–F'96) is the Curry/Cullen Trust Endowed Chair Professor at The University of Texas at Austin, where he is the Director of the Laboratory for Image and Video Engineering (LIVE). His research interests include image and video processing, computational vision, digital microscopy, and modeling of biological visual perception. He has published over 500 technical articles in these areas and holds two U.S. patents. He is the author of *The Handbook of Image and Video Processing* (Academic, 2005), *Modern Image Quality Assessment* (Morgan & Claypool, 2006), *The Essential Guide to Image Processing* (Academic, 2009), and *The Essential Guide to Video Processing* (Academic, 2009).

Dr. Bovik has received a number of major awards from the IEEE Signal Processing Society, including the Education Award (2007), the Technical Achievement Award (2005), the Distinguished Lecturer Award (2000), and the Meritorious Service Award (1998). He is also a recipient of the Hocott Award for Distinguished Engineering Research at the University of Texas at Austin; he received the Distinguished Alumni Award from the University of Illinois at Champaign-Urbana (2008), the IEEE Third Millennium Medal (2000), and two journal paper awards from the international Pattern Recognition Society (1988 and 1993). He is a Fellow of the Optical Society of America and a Fellow of the Society of Photo-Optical and Instrumentation Engineers. He has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996–1998; Editor-in-Chief, the IEEE Transactions on Image Processing, 1996–2002; Editorial Board, Proceedings of the IEEE, 1998–2004; Series Editor for Image, Video, and Multimedia Processing, Morgan & Claypool Publishing Company, 2003-present; and Founding General Chairman, First IEEE International Conference on Image Processing, held in Austin, TX, in November, 1994. He is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial, and academic institutions.