Perceptually Significant Spatial Pooling Techniques for Image Quality Assessment

Anush K. Moorthy and Alan C. Bovik

Laboratory for Image and Video Engineering (LIVE), Department of Electrical & Computer Engineering, The University of Texas at Austin, USA.

ABSTRACT

Spatial pooling strategies used in recent Image Quality Assessment (IQA) algorithms have generally been that of simply averaging the values of the obtained scores across the image. Given that certain regions in an image are perceptually more important than others, it is not unreasonable to suspect that gains can be achieved by using an appropriate pooling strategy. In this paper, we explore two hypothesis that explore spatial pooling strategies for the popular SSIM metrics.[1,2] The first is visual attention and gaze direction - 'where' a human looks. The second 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. The improvements in correlation between the objective metrics' score and human perception is demonstrated by evaluating the performance of these pooling strategies on the LIVE database[3] of images.

Keywords: I mage quality assessment, subjective quality assessment, structural similarity, visual fixations, quality-based weighting.

1. INTRODUCTION

Image Quality Assessment (IQA) methods can be classified as subjective assessment - evaluation of quality by humans, and objective assessment - evaluation of quality by algorithms. The requirement for objectively assessing quality is obvious when one considers the time involved in conducting a controlled study of quality.

The success of an objective algorithm is defined by its correlation with human perception, and this necessitates the creation of databases with subjective opinion scores defining the human perception of quality. To address this need, the VQEG dataset[4] of videos and the LIVE database[3] were created.

A statistical analysis of full-reference IQA algorithms[5] demonstrated that amongst the algorithms tested, the Multi-Scale Structural SIMilarity Index (MS-SSIM)[2] and the Visual Information Fidelity index (VIF)[6] performed consistently well. The Single-Scale SSIM (SS-SSIM) index[1] remains attractive owing to its extreme simplicity and excellent performance relative to old standards such as the MSE.[7] Even though VIF is based on Natural Scene Statistics (NSS) and is computed in the wavelet domain, its has been demonstrated to be equivalent to the SSIM metrics.[8] Hence, we develop pooling strategies for the SSIM metrics in this paper.

The SS and MS SSIM implementations[1,2] compute local scores at single/multiple scale(s) and then compute a *mean* score at the end from the SSIM maps. Even though in MS-SSIM each level 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.

Intuitively, each region in an image may not bear the same importance as others. Visual importance has been previously 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.

When shown an image, the hypothesis that certain regions may be perceptually more significant than others for the human observer is one that has been previously made and applied toward pooling strategies.[16,17] In

Further author information: (Send correspondence to A. K. Moorthy)
A. K. Moorthy: E-mail: anushmoorthy@gmail.com, Telephone: 1 512 415 0213
A. C. Bovik: E-mail: bovik@ece.utexas.edu, Telephone: 1 512 471 5370

a previous paper exploring the possibility of changing the pooling techniques for SSIM based on perceptual significance,[16] the authors evaluated various pooling strategies including local quality-based pooling. In this paper we further investigate this possibility of local quality-based pooling, albeit in a different manner. Again, the other aspect that we study - evaluation of the effect of visual attention, has been previously explored,[17] where a number of factors that influence visual attention were evaluated to produce an Importance Map (IM).[17,18] This IM was then used to weight the IQA indices, resulting in meagre improvements.

Thus, we believe that there are two hypotheses which may influence human perception of image quality. The first is visual attention and gaze direction - 'where' a human looks and the second 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. By computing a mean score of the local quality indices the existing IQA algorithms do not attempt to compensate either of these hypothesis.

In this paper, we investigate pooling SSIM scores using the concepts of perceptual importance as gauged by a visual fixation predictor - an algorithm which tries to predict human gaze and perceptual importance as gauged by heavily weighting the lowest SSIM map scores. In a previous paper evaluating spatial pooling strategies,[16] the authors recognized that the weighting low quality scores more heavily was intuitively sensible. However, the approach of using a monotonic function of quality for weighting led to the conclusion that this approach would lead only to incremental improvement, at best. By contrast, we demonstrate that significant gains in performance can be obtained using the right strategy.

Visual fixation based techniques previously explored,[19] used ground-truth data, and concluded that improvements in SSIM performance was not demonstrated. Using an algorithm for predicting human fixations, we demonstrate that significant improvements for SSIM metrics is seen. Finally, before further discussion we note that the authors[19] 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 2 reviews the Structural Similarity algorithms (both the single and the multi-scale versions). Section 3 reviews the fixation finder that we use in this paper. Section 4 explains how our proposed algorithm functions. We present the results of using our algorithm in Section 5 and conclude the paper in Section 6.

2. STRUCTURAL SIMILARITY INDEX

The SSIM metrics explored in this paper are space-domain metrics. There also exist non-spatial IQA extensions of SSIM such as the Complex-Wavelet SSIM index (CW-SSIM).[20,21]

2.1 Single-scale SSIM

Consider two aligned-discrete non-negative signals, $\mathbf{x} = \{x_i | i = 1, 2, \ldots, N\}$ and $\mathbf{y} = \{y_i | i = 1, 2, \ldots, N\}$. 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 ($s(\mathbf{x}, \mathbf{y})$):[1]

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

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

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

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 $K_1 << 1$ and $K_2 << 1$ 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:

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

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

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

At each coordinate, the SSIM index is calculated within a local window. We use a 11×11 circular-symmetric Gaussian weighting function[1] $w = \{w_i | i = 1, 2, \ldots, N\}$, 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 re-defined 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.

2.2 Multi-scale SSIM

The perceived quality of an image is heavily dependent upon the scale at which the image is analyzed. In case of SS-SSIM, this analysis is performed at the original image scale. Images are naturally multi-scale and in order to evaluate the quality at multiple scales, the Multi-scale SSIM (MS-SSIM)[2] 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 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:

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

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

The exponents α_j, β_j, γ_j are selected such that $\alpha_j = \beta_j = \gamma_j$ and $\sum_{j=1}^{M} \gamma_j = 1$. The specific parameters used in[2] 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 original spatial pooling strategy used[2] was the ensemble mean.

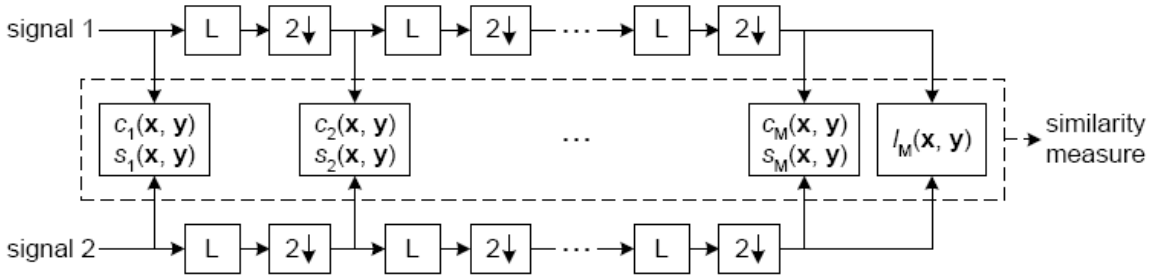

Figure 1. Multi-scale SSIM. L = low-pass filtering, 2↓ = downsampling by 2.

3. PERCEPTUALLY SIGNIFICANT POOLING

3.1 GAFFE

When shown an image or a video sequence the human visual system is streamlined to select and assimilate only those features that are relevant. 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] How the human visual system selects certain regions for scrutiny is of tremendous relevance for many applications, including, but not limited to, quality assessment.

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 an attempt to answer this question, researchers have produced some successful fixation finding algorithms, which seeks regions of high saliency,[9] and the Gaze-Attentive Fixation Finding Engine (GAFFE), which uses image statistics measured at the point of gaze from actual visual fixations.[10] In this paper, we use GAFFE to predict human fixations, in order to gauge possible regions of importance. In this section we briefly overview the GAFFE approach to predicting human fixations.

First, an experiment to record human eye moments was performed and the gaze coordinates corresponding to the human eye moments were recorded.[10] The 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.[26] The images used were such that they do not influence attention - ones that contained minimal contextual information.

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. 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 above described features. The four feature maps thus obtained are linearly combined, where each feature is scaled by a factor, $\gamma_{feature}$. The scaling factors are $\gamma_{luminance} = 1.04$, $\gamma_{contrast} = 1.12$, $\gamma_{luminance-bandpass} = 1.23$ and $\gamma_{contrast-bandpass} = 1.30$. 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 online[28] and the algorithm is explained in detail in another paper.[10]

3.2 Percentile Pooling

We define the p^{th} percentile of an ordered set[29] as 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 p^{th} percentile. 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.

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.

4. IMPLEMENTATION

We consider three new versions of SSIM based on the pooling strategy: 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.

4.1 F-SSIM

Under the assumption that the GAFFE algorithm can be run on images to produce f fixations per image, the value of f is the first unknown. The number of fixations found in GAFFE[10] were 10 fixations/image (on an average) when an image was shown to the subject for 5 seconds. However the LIVE database,[3] which we use as a test-bed did not place any restriction on the amount of time the subjects could view the image before rating it. Hence, we elected to keep the number of fixations at a constant $f = 10$. 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, which leads us to the second unknown.

In order to explore possible values for the weight applied to the fixated region, relative to the non-fixated areas ($k > 1$), we randomly selected one of the types of distortion from the LIVE database, and then simulated various values of k. The value of k that maximized the Spearman Rank Order Correlation Coefficient (SROCC) was chosen as the weight to be applied to the fixated regions. In Figure 2, we see the absolute value of the SROCC plotted as a function of the weighting parameter k for SS-SSIM. The trend seen in Figure 2 remains consistent across distortion types, and hence we choose $k = 265$, although varying this ratio in the range $125 \leq k \leq 375$ does not change performance much.

Thus, the F-SSIM index is defined as:

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

where $SSIM(x_{ij}, y_{ij})$ 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:

$$mask_M = (11 - 2^{M-1}) \times (11 - 2^{M-1})$$

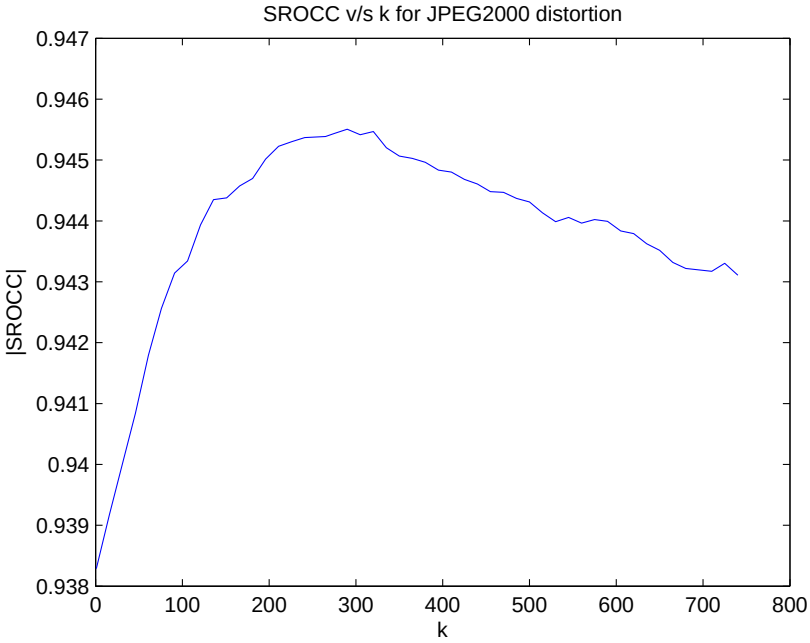

Figure 2. Plot of |SROCC| as a function of k for SS-SSIM. The sample points were equally spaced values of k between 1 (indicates no weighting) and 750.

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

$$N^M_{fixations} = \lceil \frac{10}{2^{M-1}} \rceil$$

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: $w_{ij} = 1$.

4.2 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] Even though many such methods of weighting are possible, we consider the statistical principle of heavily weighting the lower-percentile scores; instead of an arbitrary monotonic function of quality.[16] The unknowns here are what percentile (p) should one use and how much should the percentile scores be weighted by (r). In order to arrive at a solution, we simulated values of p from 5-25% in 1% increments, and found that the value $p = 6\%$ yields good results, however; small perturbations in p does not alter the results drastically. 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] We note that a similar form of pooling has been used for video quality assessment,[30] where only lowest 5% of the spatial scores are pooled together.

Again, the ratio r by which lowest p^{th} percentile of 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. The choices for r are empirical and are made based on results similar to those seen in Figures 3 and 4, where 3-D plots of the absolute value of the SROCC for SS-SSIM as a function of p and r are seen for JPEG2000 distortion. A clear peak is visible around $p = 6\%$. This trend remains unchanged across distortion types.

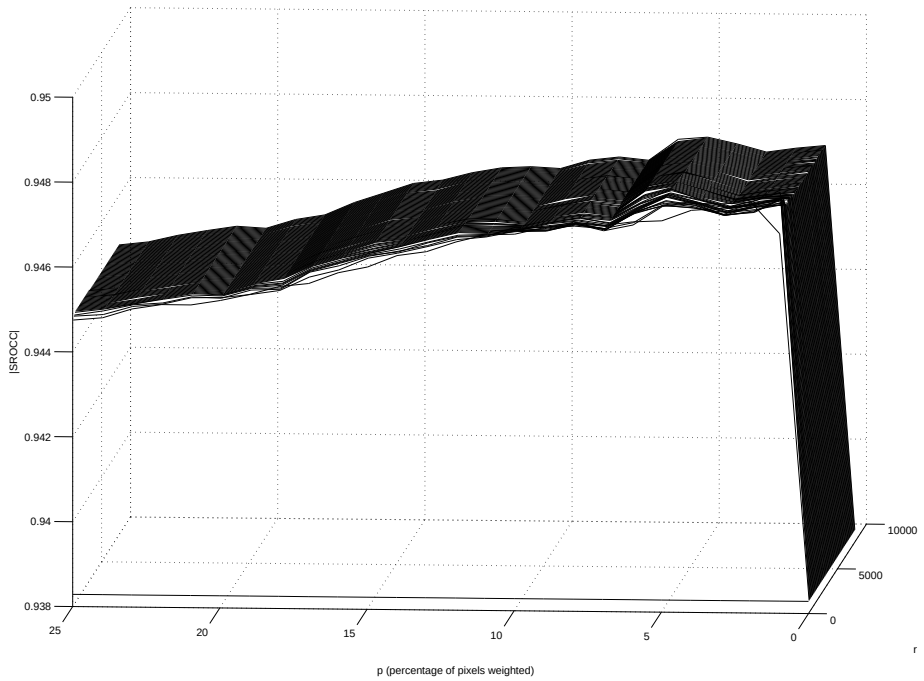

Figure 3. Plot of |SROCC| as a function of p and r for SS-SSIM with JPEG2000 distortion. The p(percentage)-axis consists of values of p in the range 0-25% with a step-size of 1%, the r(weights)-axis consists of values of r in the range 1-8000 with a step size of 100. SROCC value peaks at around 6%. Note that the cases $p = 0$ and $p = 100$ (not shown here) correspond to the original SSIM.

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

Table 1. 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 M = 1, once-downsampled image M = 2 and so on.

For MS-SSIM, 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 previously made,[2] where the highest gains relative to SS-SSIM were achieved at $M = 2$.

4.3 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; 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 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 1.

5. RESULTS

5.1 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

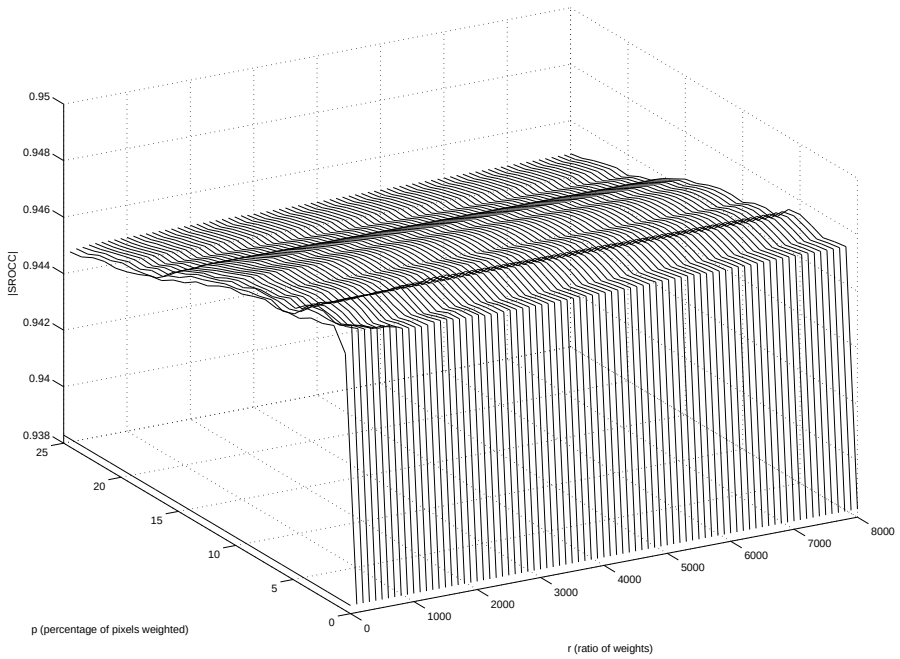

Figure 4. Plot of |SROCC| as a function of p and r for SS-SSIM with JPEG2000 distortion. The p(percentage)-axis consists of values of p in the range 0-25% with a step-size of 1%, the r(weights)-axis consists of values of r in the range 1-8000 with a step size of 100. Note a gradual increase at $p = 6\%$, with r, highest correlations are obtained when $1000 \leq r \leq 8000$. The case $r = 1$ corresponds to the original SSIM.

	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 2. Linear correlation coefficient values(after non-linear regression) - F-SSIM, P-SSIM, PF-SSIM (Single-scale)

: 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 three popular metrics : the Spearman rank ordered correlation coefficient (SROCC), the linear correlation coefficient (CC) (after non-linear regression), and the root mean-squared error (RMSE) (after non-linear regression). The non-linearity 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:

$$Quality(x) = \beta_1 logistic(\beta_2, (x - \beta_3)) + \beta_4 x + \beta_5$$

$$logistic(\tau, x) = \frac{1}{2} - \frac{1}{1 + exp(\tau x)}$$

The results are tabulated in Tables 2 - 7. In all tables, SS = Single-scale, MS = Multi-scale, WN = White Noise, Gblur = Gaussian Blur and FF = Fast Fading.

We calculated all metrics for all distortions for MS-SSIM and SS-SSIM.

	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 3. Linear correlation coefficient(after non-linear regression) - F-SSIM, P-SSIM, PF-SSIM (Multi-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 4. RMSE (after non-linear regression) - F-SSIM, P-SSIM, PF-SSIM (Single-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 5. RMSE (after non-linear regression) - F-SSIM, P-SSIM, PF-SSIM (Multi-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

Table 6. SROCC values - F-SSIM, P-SSIM, PF-SSIM (Single-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

Table 7. SROCC values - F-SSIM, P-SSIM, PF-SSIM (Multi-scale)

5.2 Performance Metrics

We have chosen to evaluate the performance of the pooling strategies based on three metrics - SROCC, CC and RMSE - which is currently the norm in evaluation of IQA/VQA metrics.[4,5] However, some authors have argued that SROCC operates only on the rank of the data points while assuming an equal spacing between the datapoints and hence images which generate clustered scores, although different in rank, may not differ much in quality.[5] Conversely, it may also be argued that the non-linear regression used is not a sufficient measure of performance, since the scores are dependent upon the degree of fit. In the absence of any other agreed upon statistic, we continue to report results based on the SROCC, CC and RMSE.

5.3 F-SSIM Performance

As mentioned earlier, the number of fixations for GAFFE were fixed at $f = 10$. Having seen improvements in performance (especially for GBLUR and FF), variation in the number of fixations is a topic of interest. As research in the area of gaze-attention evolves, better fixation finding algorithms will probably enhance the results seen here. Finally, we note that temporal saliency has recently been explored for video quality assessment.[33]

The gains in performance using F-SSIM are limited to Gaussian Blur and Fast Fading distortion types - which tend to destroy the structure of perceptually significant features. Gains relative to SS-SSIM were substantially higher than those seen for MS-SSIM - this is to be expected, since MS-SSIM demonstrates extremely good correlation with human perception. The improvement in performance using P-SSIM was more substantial. In fact, the P-SSIM performance for the SS case competes with the more complicated MS-SSIM. This suggests the use of percentile scoring with the simpler SS-SSIM as an alternative to MS-SSIM. Finally, combining P-SSIM and F-SSIM into PF-SSIM produced desultory improvement, if any.

6. CONCLUSION

We found that by incorporating alternative pooling strategies based on perceptual importance, namely, weighting of fixations and percentile scores, an improvement in the performance of IQA metrics is evinced, contradictory to prior studies.[16,19] The improvement in performance was demonstrated using the LIVE database of images.[3] The degree of improvement, in some cases, was dependent upon the type of distortion

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation, Texas Instruments, Agilent and Boeing.

REFERENCES

[1] Wang, Z., Bovik, A. C., Sheikh, H., and Simoncelli, E., "Image quality assessment: From error measurement to structural similarity," *IEEE Signal Processing Letters* **13**, 600–612 (Apr. 2004).

[2] Wang, Z., Simoncelli, E., and Bovik, A. C., "Multi-scale structural similarity for image quality assessment," *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, (Asilomar)* (Nov. 2003).

[3] Sheikh, H. R., Wang, Z., Cormack, L., and Bovik, A. C., "Live image quality assessment database release 2."

[4] "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment."

[5] Sheikh, H. R., Sabir, M., and Bovik, A. C., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing* **15**, 3440–3451 (Nov. 2006).

[6] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE Transactions on Image Processing* **15**(2), 430–444 (2006).

[7] Wang, Z. and Bovik, A. C., "Mean squared error: Love it or leave it? - a new look at fidelity measures." IEEE Signal Processing Magazine (2008). to appear.

[8] Seshadrinathan, K. and Bovik, A. C., "Unifying analysis of full reference image quality assessment," in [*IEEE International Conference on Image Processing*], (Oct. 2008).

[9] Itti, L., Koch, C., and Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(11), 1254–1259 (1998).
[10] Rajashekar, U., Bovik, A. C., and Cormack, L. K., "Gaffe: A gaze-attentive fixation finding engine," *IEEE Transactions on Image Processing* **17**, 564–573 (Apr. 2008).
[11] Geisler, W. and Perry, J., "A real-time foveated multi-resolution system for low-bandwidth video communication," in [*Human Vision and Electronic Imaging, SPIE Proc.*], **3299**, 294–305 (1998).
[12] Itti, L., "Automatic foveation for video compression using a neurobiological model for visual attention," *IEEE Transactions on Image Processing* **13**, 1304–1318 (Oct. 2004).
[13] Wang, Z. and Bovik, A., "Embedded foveation image coding," *IEEE Transactions on Image Processing* **10**, 1397–1410 (Oct. 2001).
[14] Wang, Z., Lu, L., and Bovik, A., "Foveation scalable video coding with automatic fixation selection," *IEEE Transactions on Image Processing* **12**, 243–254 (Feb. 2003).
[15] Lee, S., Pattichis, M. S., and Bovik, A. C., "Foveated video compression with optimal rate control," *IEEE Transactions on Image Processing* **10**, 977–992 (July 2001).
[16] Wang, Z. and Shang, X., "Spatial pooling strategies for perceptual image quality assessment," in [*IEEE international conference on Image Processing*], (Sept. 2006).
[17] Osberger, W., Bergmann, N., and Maeder, A., "An automatic image quality assessment technique incorporating higher level perceptual factors," in [*Proceedings, International Conference on Image Processing*], 414–418 (1998).
[18] Maeder, A., Diederich, J., and Niebur, E., "Limiting human perception for image sequences," *SPIE - Human Vision and Electronic Imaging* **2657**, 330–337 (1996).
[19] Ninassi, A., Meur, O. L., Callet, P. L., and Barbba, D., "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," in [*Image Processing, 2007. ICIP 2007. IEEE International Conference on*], **2**, 169–172 (2007).
[20] Wang, Z. and Simoncelli, E. P., "Translation insensitive image similarity in complex wavelet domain," in [*Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*], 573–576 (Mar. 2005).
[21] Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., and Markey, M., "Complex wavelet structural similarity: a new image quality index." IEEE Transactions on Image Processing. to be published.
[22] Yarbus, A. L., [*Eye movements and vision*], Plenum Press, New York, second ed. (1967).
[23] Burr, D., Morrone, M. C., and Ross, J., "Selective suppression of the magnocellular visual pathway during saccadic eye movements," *Nature* **371**, 511–513 (Oct. 1994).
[24] van Hateren, J. H. and van der Schaaf, A., "Independent component filters of natural images compared with simple cells in primary visual cortex," in [*Proc.R.Soc.Lond. B*], 265:359–366 (1998).
[25] van der Linde, I., Rajashekar, U., Bovik, A. C., and Cormack, L. K., "Doves: A database of visual eye movements," in [*Spatial Vision*], (2008).
[26] van der Linde, I., Rajashekar, U., Bovik, A. C., and Cormack, L. K., "Doves: A database of visual eye movements."
[27] Efron, B., [*An Introduction to the Bootstrap*], Chapman & Hall/CRC (1994).
[28] Rajashekar, U., van der Linde, I., Bovik, A. C., and Cormack, L. K., "Gaffe: A gaze-attentive fixation finding engine."
[29] David, H. A., [*Order Statistics*], John Wiley & Sons Inc. (2003).
[30] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting* , 312–313 (Sept. 2004).
[31] Bovik, A. C., Huang, T. S., and Munson, D. C., "A generalization of median filtering using linear combinations of order statistics.," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **6**, 1342–1350 (Dec. 1983).
[32] Longbotham, H. G. and Bovik, A. C., "Theory of order statistic filters and their relationship to linear fir filters," *IEEE Transactions Acoustics Speech and Signal Processing* **37**(2), 275–287 (1989).
[33] Wang, Z. and Li, Q., "Video quality assessment using a statistical model of human visual speed perception," *Journal of the Optical Society of America* **24**, B61–B69 (Dec. 2007).