Structural similarity quality metrics in a coding context: exploring the space of realistic distortions

Alan C. Brooks and Thrasyvoulos N. Pappas

EECS Dept., Northwestern Univ., 2145 Sheridan Rd, Evanston, IL 60208, United States

ABSTRACT

Perceptual image quality metrics have explicitly accounted for human visual system (HVS) sensitivity to subband noise by estimating thresholds above which distortion is just-noticeable. A recently proposed class of quality metrics, known as structural similarity (SSIM), models perception implicitly by taking into account the fact that the HVS is adapted for extracting structural information (relative spatial covariance) from images. We compare specific SSIM implementations both in the image space and the wavelet domain. We also evaluate the effectiveness of the complex wavelet SSIM (CWSSIM), a translation-insensitive SSIM implementation, in the context of realistic distortions that arise from compression and error concealment in video transmission applications. In order to better explore the space of distortions, we propose models for typical distortions encountered in video compression/transmission applications. We also derive a multi-scale weighted variant of the complex wavelet SSIM (WCWSSIM), with weights based on the human contrast sensitivity function to handle local mean shift distortions.

Keywords: structural similarity, image quality, human perception, video compression, video coding, error concealment

1. INTRODUCTION

We examine objective criteria for the evaluation of image quality based on both low-level models of visual perception and high-level characteristics of the human visual system (HVS). The term “image quality” is quite general and covers the entire range from reference-free quality evaluation to perceptual image fidelity, i.e., how perceptually close an image is to a given original or reference image. Most existing objective fidelity metrics compare the reference and distorted images on a point-by-point basis, whether this is done in the original image domain, as in mean squared error based metrics such as peak signal to noise ratio (PSNR), or in a transform domain, such as the perceptually weighted subband/wavelet or discrete cosine transform (DCT) domain.[1] The most advanced of these metrics are based on low-level models of the HVS. On the other hand, a recently proposed class of quality metrics, known as Structural SIMilarity (SSIM),[2] accounts for high-level HVS characteristics, and allows substantial point-by-point distortions that are not perceptible, such as spatial and intensity shifts, as well as contrast and scale changes. Our primary goal is to evaluate SSIM metrics and to compare their performance to traditional approaches in the context of realistic distortions that arise from compression and error concealment in video transmission applications. In order to better explore this space of distortions, we also propose models for typical distortions encountered in such applications.

Perceptual image quality metrics have relied on explicit low-level models of human perception that account for sensitivity to subband noise as a function of frequency, local luminance, and contrast/texture masking.[1, 3] Typically, the signal is analyzed into components (e.g., spatial and/or temporal subbands), and the role of the perceptual model is to provide the maximum amount of distortion that can be introduced to each component without resulting in any perceived distortion. This is usually referred to as the *just noticeable distortion* level or *JND*. While these metrics were developed for near-threshold applications, they have also been used in suprathreshold applications.[4, 5] The main idea is to normalize the distortion by the JND.[6, 7] More systematic studies of the suprathreshold case have been conducted by Hemami's group.[8–11] However, while perceptual metrics can successfully account for subband (frequency) dependence of the HVS sensitivity to noise and contrast and luminance masking, they cannot account for imperceptible structural changes, such as spatial shifts, intensity shifts, contrast changes, and scale changes.

A.C.B.: email: alanbrooks@ieee.org; T.N.P.: email: pappas@ece.northwestern.edu

The SSIM metrics[2] are based on high-level properties of the HVS, but employ no explicit model of the HVS. They are derived from assumptions about the high-level functionality of the HVS, and in particular, account for the fact that it is adapted for extracting structural information (relative spatial covariance) from images. Thus, they can more effectively quantify suprathreshold compression artifacts, as such artifacts tend to distort the structure of an image. Even though the SSIM metrics are not based on explicit models or measurements of HVS sensitivities, they implicitly account for important HVS properties such as light adaption and masking, in addition to the perception of image structure.[2] However, while the SSIM metrics have been shown to have a number of desirable properties, they have not been systematically studied in the context of video compression artifacts.

In an increasing number of applications, such as video transmission over band-limited and noisy channels, there is a need to achieve very high compression ratios. In such cases, a certain amount of perceived distortion is unavoidable. Thus, there is an increased need for quantitative objective measures of perceived distortion. In this paper, we examine the performance of SSIM metrics for such suprathreshold video transmission applications and compare their performance to traditional approaches. We compare specific SSIM index implementations both in the image space and the wavelet domain. We also evaluate the effectiveness of the complex wavelet SSIM (CWSSIM), a translation-insensitive SSIM implementation. Our experimental results indicate that structural metrics, and in particular CWSSIM, generally agree with subjective evaluations.

In typical video transmission applications one encounters a variety of distortions due to source coding (quantization) and packet losses (which are concealed by different techniques). In order to isolate the different types of distortion, analyze their severity (as perceived by the HVS), and evaluate how well they correspond to metric predictions, we propose models for typical distortion artifacts such as DCT coefficient quantization, spatial interpolation concealment, temporal replacement concealment, and DC coefficient loss. These models can be generated from still images, which considerably simplifies the computational cost for the simulations. Thus, we can better explore the space of realistic distortions.

Finally, we also derive a multi-scale weighted variant of the complex wavelet SSIM (WCWSSIM), with weights based on the human contrast sensitivity function. We found that this modification is necessary for handling local mean shift distortions. Moreover, by incorporating an explicit model of subband sensitivity to noise, this provides a unification of the two types of approaches described above.

In the following sections, we first review the motivation, development, and theory behind structural metrics and discuss some specific implementations including SSIM and CWSSIM. Then, we describe the results of using structural metrics to assess image quality of a variety of suprathreshold distortions. Finally, we propose a way of modeling error concealment distortion in video compression, evaluate SSIM techniques, and extend CWSSIM to account for perception of block intensity shift distortions.

2. STRUCTURAL APPROACH TO IMAGE QUALITY MEASUREMENT

2.1. SSIM Review

The motivation behind the structural similarity approach for measuring image quality is that the HVS is not designed for detecting imperfections and "errors" in images. Instead, the HVS has evolved so that it can do visual pattern recognition in order to be able to extract the *structure* or connectedness of natural images. Based on this observation, it makes sense that a useful perceptual quality metric would emphasize the structure of scenes over the lighting effects. The idea that image quality metrics can be created on the basis of this philosophy was first explored in Ref. 12 and then modified, implemented, evaluated, and developed in Ref. 2.

The structural similarity approach is mostly *insensitive* to the distortions that lighting changes create: changes in the mean and contrast of an image. On the other hand, the structural approach is *sensitive* to distortions that break down natural spatial correlation of an image, such as blur, block compression artifacts, and noise.

As described in Ref. 13, the structural philosophy can be implemented using a set of equations defining the Structural SIMilarity (SSIM) quality metric in image space. Luminance, contrast, and structure are measured separately. Given two images (or image patches) $\mathbf{x}$ and $\mathbf{y}$ to be compared, *luminance* is estimated as the mean of each image

$$\mu_x = \frac{1}{N}\sum_{n=1}^{N} x_n, \tag{1}$$

contrast is estimated using standard deviation as

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_x)^2}, \tag{2}$$

and *structure* is estimated from the image vector $\mathbf{x}$ by removing the mean and normalizing by the standard deviation

$$\varsigma_x = \frac{\mathbf{x} - \mu_x}{\sigma_x}. \tag{3}$$

Then, the measurements $\mu_x, \mu_y, \sigma_x, \sigma_y, \varsigma_x, \varsigma_y$ are combined using a luminance comparison function $l(\mathbf{x}, \mathbf{y})$, a contrast comparison function $c(\mathbf{x}, \mathbf{y})$, and a structure comparison function $s(\mathbf{x}, \mathbf{y})$ to give a composite measure of structural similarity:

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

where α, β, γ are positive constants used to weight each comparison function.

The comparison functions are given as:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{5}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{6}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\langle \varsigma_x, \varsigma_y \rangle + C_3}{\sigma_x\sigma_y + C_3} = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{7}$$

where $\langle\rangle$ is the inner-product operator defining the correlation between the structure of the two images.

In this paper, we follow the example in Ref. 2 setting $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$ to get the specific SSIM quality metric

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{8}$$

2.2. CWSSIM Review

As suggested in Ref. 14, it is straightforward to implement a structural similarity metric in the complex wavelet domain. As more wavelet-based image and video coding techniques are coming into use, it makes sense to be able to implement image quality metrics in this domain. In addition, if an application requires an image quality metric that is unresponsive to spatial translation, this extension of SSIM can be adapted in a way such that it has low sensitivity to small translations. This requires an overcomplete wavelet transform such as the steerable pyramid,[15] for which *phase information is available.*

Given complex wavelet coefficients $\mathbf{c}_x$ and $\mathbf{c}_y$ that correspond to image patches $\mathbf{x}$ and $\mathbf{y}$ that are being compared, the complex wavelet structural similarity (CWSSIM) is given by:

$$CWSSIM(\mathbf{c}_x, \mathbf{c}_y) = \frac{2|\sum_{i=1}^{N} c_{x,i}c_{y,i}^{*}| + K}{\sum_{i=1}^{N}|c_{x,i}|^2 + \sum_{i=1}^{N}|c_{y,i}|^2 + K}, \tag{9}$$

where K is a small positive constant set to 0.03 in this paper. This equation differs from (8) because the wavelet filters we use are bandpass (i.e., they have no response at zero frequency), thus forcing the mean of the wavelet coefficients to zero ($\mu_x = \mu_y = 0$). This fact acts to cancel the $(2\mu_x\mu_y + C_1)$ and $(\mu_x^2 + \mu_y^2 + C_1)$ terms in (8).

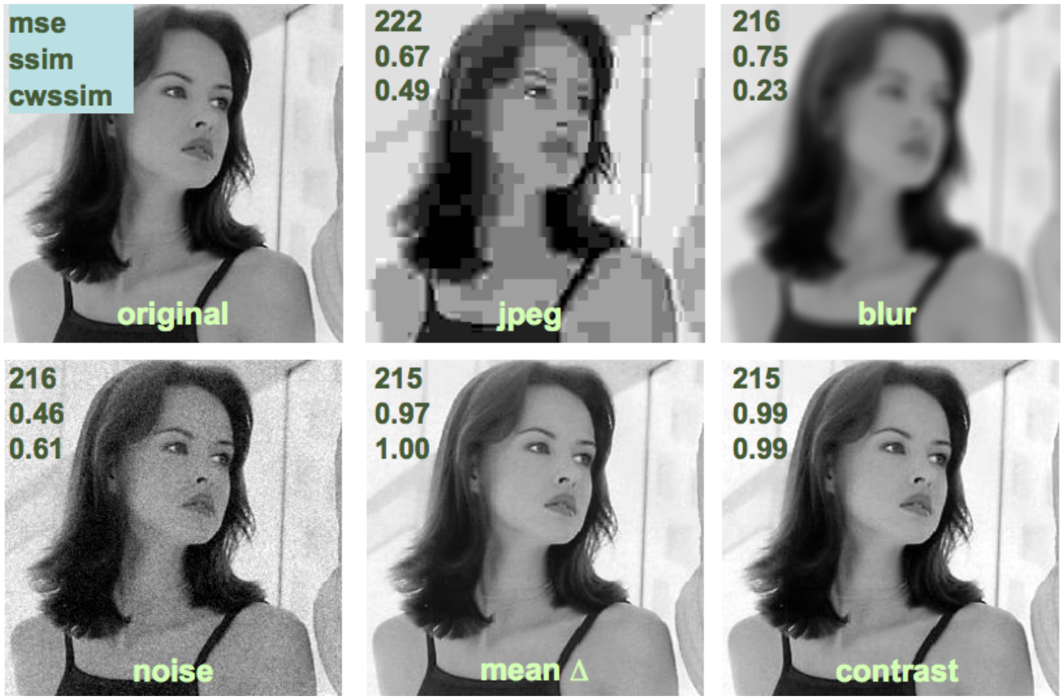

Figure 1. The effect of suprathreshold distortions on MSE, SSIM, and CWSSIM metrics. SSIM and CWSSIM correspond well with the intuitive idea that changes in lighting (mean and contrast) have very little effect on image quality.

Wang and Simoncelli note that the wavelet coefficient phase is the key factor that determines the results of CWSSIM: "the structural information of local image features is mainly contained in the relative phase patterns of the wavelet coefficients".[14] Linear and uniform phase changes correspond to lighting (brightness and contrast) distortions to which CWSSIM is not sensitive because the *structure* is not perturbed. Phase changes that vary irregularly from one coefficient to the next produce structural distortion and therefore low CWSSIM.

3. USING SSIM TO ASSESS IMAGE QUALITY

3.1. Suprathreshold Effectiveness

Structural similarity can distinguish between structural and non-structural distortions, giving results that agree with perception for very strongly distorted images (suprathreshold distortions). The structural similarity metric gives a result ranging from 0.0 to 1.0, where zero corresponds to a loss of all structural similarity and one corresponds to having an exact copy of the original image. Images with lighting-related distortions give high SSIM while other distortions result in low similarity.

An example of the effectiveness of structural similarity in measuring suprathreshold distortion is depicted in Figure 1. The original image is shown in the upper left, then five distorted images with equal mean squared error (MSE), including JPEG compression (DCT coefficient quantization), blur, Gaussian white noise, mean shift, and contrast stretch distortions. Both structural similarity metrics, SSIM and CWSSIM, compute image quality measurements that correspond well with the idea that change in lighting (mean shift, contrast stretch) has little impact on image quality, while changes that affect local relationships between pixels severely degrade image quality.

Perceptual image quality metrics based on the HVS's sensitivity to just-noticeable distortions often do not provide meaningful measurements at suprathreshold levels of distortion.[5] SSIM-based image quality metrics conveniently avoid this problem by focusing on the top-down image formation concept that the local structure of images is the most important aspect of image quality.

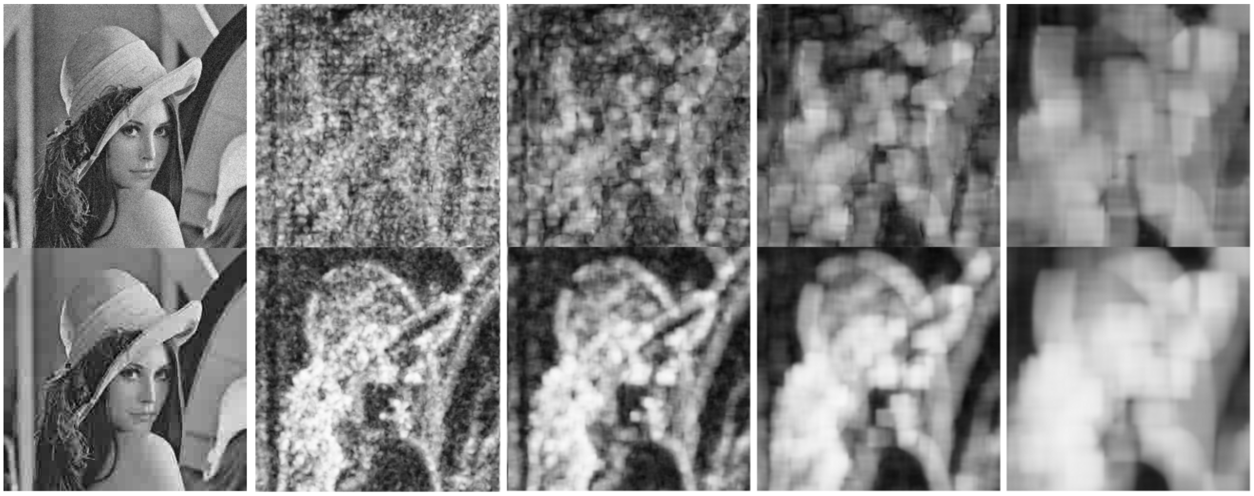

Figure 2. SSIM error maps for different window sizes. The top row shows the Lena image distorted with white Gaussian noise, then error maps for window sizes of 4, 8, 16, and 32. The bottom row shows Lena degraded with DCT quantization and corresponding error maps. In the error maps, darker regions represent higher measured distortion.

3.2. Effect of Window Size

When using SSIM metrics to compare the quality between two images, it is useful to calculate the local distortion between corresponding image patches at many locations. This allows the metric to adapt to the local statistical characteristics at different image locations. The individual quality measurements can then be combined to give a single number that represents the similarity between the images. Applications that need to measure image quality with minimum computation may only want to compute the metric at a few locations within the image.[16] However, for this paper, we measure the similarity with a sliding window at every pixel location, giving a SSIM distortion map. As suggested in Ref. 2, we use a circular Gaussian weighting function on the image patches being compared to smooth the similarity map, and we combine the measurements using a mean operator.

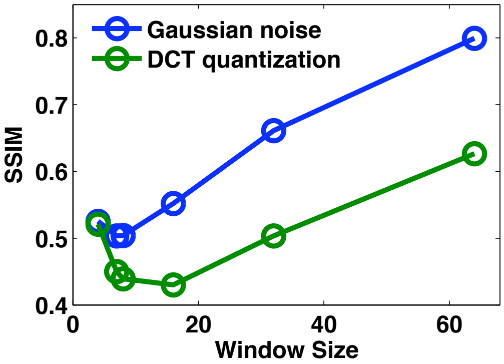

Figure 3. Mean SSIM versus window size for the same data set in Fig. 2.

The choice of the $W \times W$ window size affords a balance between SSIM's ability to adapt to local image statistics versus its ability to accurately compute the statistics within an image patch. A large window allows accurate statistical estimation, at the cost of being less sensitive to fine image distortions. For typical 512×512 images, a window size within the range of 7×7 to 16×16 offers a reasonable operating region.

The effect of window size on SSIM is illustrated in Figures 2 and 3. The first image in the top row of Figure 2 depicts the "Lena" image distorted with white Gaussian noise giving a $PSNR = 28.5dB$ (where $PSNR = 10log_{10}(255^2/MSE)$). The other images in the top row show the SSIM similarity map for a $W \times W$ window of size $W = 4$, $W = 8$, $W = 16$, and $W = 32$, respectively, where the dark regions represent the highest SSIM distortion. White Gaussian noise is most noticeable in the smooth (low frequency) regions of the image such as the shoulder and background. A $W = 4$ window is too small to accurately compute the distortions, giving an almost randomly distributed distortion map. Alternately, a $W = 32$ window is too large to adapt to local statistics, resulting in image quality scores that are too high in most regions. The distortion maps computed with $W = 8$ and $W = 16$ more accurately correspond to human perception.

The bottom row of Figure 2 shows a similar experiment in which the first image is generated using DCT quantization giving a $PSNR = 28.5dB$. Similar to the top row, $W = 8$ and $W = 16$ offer a better balance between estimation accuracy and local adaptation than either extreme value ($W = 4$ and $W = 32$). The dark regions correspond well with the most perceptually annoying compression artifacts: the blockiness on the shoulder, cheeks, and background.

The relationship between SSIM and window size is also shown in Figure 3, which plots the overall quality pooled over the image for each window size. This plot is derived from the same data set that was used in Figure 2. Our experiments indicate that the operating region $7 \leq W \leq 16$ works well for a variety of images and distortions.

4. USING SSIM TO ASSESS VIDEO QUALITY

4.1. Motivation for Exploring Space of Realistic Distortions

The goal of the objective quality metrics that we examine in this paper is to provide a measure of perceptual similarity between a distorted and a reference image (perceptual image fidelity). Of course since the images are intended to be viewed by human observers, the metric predictions should agree with subjective evaluations. Subjective evaluation studies collect opinion scores for a database of distorted images and use statistical analysis to compute a mean opinion score (MOS) for each image. The MOS data can then be compared with the quality metric predictions to validate the effectiveness of the metric (e.g., see Ref. 17). The selection of the database of distorted images is critical for the success of such subjective evaluations, which are quite cumbersome and expensive, as an inadequate selection can lead to inconclusive results. In addition, when designing a metric, an appropriate choice of distorted test images can significantly improve the process of metric design, by providing intuitive insights into ways for improving the metric and eliminating the need for lengthy subjective evaluations after each modification of the metric parameters.

An alternate way of evaluating image quality metrics was proposed in Ref. 18, where the authors suggest a *stimulus synthesis* approach. The idea is to compare two metrics by exploring the image space of metric A while holding metric B constant. This is done via a gradient descent algorithm, producing the “best” and “worst” images in terms of metric A for constant metric B and vice-versa. This allows for efficient evaluation of metrics because the observer only has to look at a few images (the best and worst) to find weaknesses of a metric. A limitation of this approach, however, is that the iterative approach produces distortions that are unlikely to be encountered in compressed video communication systems.

We propose a method of evaluating quality metrics in which we explore the space of realistic distortions that are likely in video compression and communication applications. In our method we hold metric A constant, then examine the results given by metric B with different distortions. This approach provides efficient and valuable intuition for the further improvement of an image quality metric. Section 4.2 explains details of our realistic distortion models, Section 4.3 shows an example distortion space result, and Sections 4.4–4.5 describe specific examples of CWSSIM’s performance for degraded images in the distortion space.

4.2. Coding and Concealment Distortion Models

A variety of distortions can be created in video transmission applications due to source coding or packet loss and concealment. Lossy video compression distorts the video before it is transmitted. If the channel is lossy, the error concealment techniques necessary to reconstruct the video introduce further distortion. In this paper, we propose a set of realistic models for the distortions that are likely in a video transmission system.

We develop models that can be used to simulate video coding and concealment artifacts using a still image. The advantage of this approach is that it allows the study of video distortions in detail without the complexity of evaluating the performance of an entire video compression, transmission, and concealment system. In addition, it provides more flexibility in isolating specific types of distortions (e.g., blocking vs. blurring), which allows us to develop intuition about the effect each distortion type has on a quality metric.

Error concealment is often necessary in applications such as real-time video streaming over a lossy network, where packets are lost or arrive too late to be useful. The most elementary error concealment approach reconstructs lost image data using spatial interpolation, which may result in significant blurring.[19] Improved quality is possible using spatiotemporal approaches that reconstruct by estimating the lost motion vector and then substituting displaced patches of image data from past video frames. Typically, any part of a compressed video stream can be subject to loss, resulting in distortion. For example, loss of motion vectors may lead to spatial block shift distortions, while loss of DC coefficients could create shifts in the mean values of the video blocks affected.

We assume a block-based compression technique where the basic units susceptible to distortion are square macroblocks (MB) with $N \times N$ pixels (typically $N = 16$). The first distortion arises from compression, which is typically applied on smaller 8×8 blocks. A number of MBs M (sometimes called a *slice*) are then grouped into a packet. For simplicity, we use a straightforward channel model, whereby a packet is lost with probability P_k. We also include a parameter that controls the grouping of distortions within the MBs that make up a single lost packet.

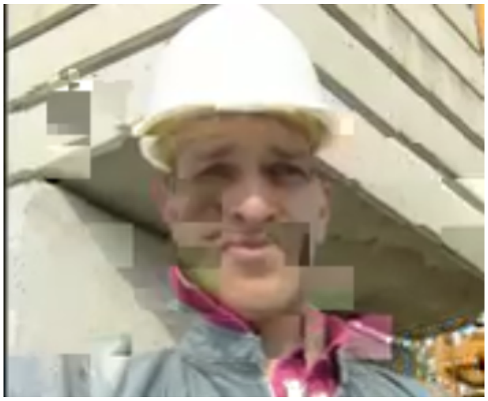

Figure 4. Example frame showing distortions created in a spatial replacement error concealment approach.

Once it is determined that a macroblock is lost, the model applies a distortion consistent with the concealment technique used by the receiver. We consider the following types of distortion: block blur (spatial interpolation), block spatial shift (temporal replacement), and block intensity shift (loss of DC coefficient). To these we also add DCT coefficient quantization, which corresponds to signal compression.

Block blur is modeled as the convolution of the MB image patch $\mathbf{x}$ with a 2-D $(N+1) \times (N+1)$ smoothing filter $\mathbf{f}$:

$$x_{blur}(\mathbf{x}, \mathbf{f}) = x(n_1, n_2) * f(n_1, n_2), \tag{10}$$

where "$*$" is the 2-D convolution operator. This is a model of simple spatial interpolation concealment techniques that may use bilinear interpolation to recreate the lost macroblock data. Of course more sophisticated spatial interpolation techniques exist, but this provides a first order approximation.

Block spatial shift is modeled as a uniform distribution of spatial shifts with a maximum shift of $\pm B$ pixels. The spatially shifted image patch is described as

$$x_{shift}(\mathbf{x}, b) = x(n_1 + b_1, n_2 + b_2), \tag{11}$$

where b_1 and b_2 are independent random values chosen from the uniform distribution described by $f_x(x) = 1/(2B)$ with $-B < x < +B$. This models the effect of temporal replacement concealment techniques that can be used when motion vectors are lost or corrupted, such as motion-compensated temporal replacement.[19]

Block intensity shift is modeled as a uniform distribution of block mean shifts with a maximum shift of $\pm L$ percent with the equation

$$x_{level}(\mathbf{x}, L) = x(n_1, n_2) + L, \tag{12}$$

where L is a random value chosen from the uniform distribution $f_x(x) = 1/(256(0.01L))$ with $-256(0.01L) < x < 256(0.01L)$ for 8-bit grayscale images. This models the distortion that can occur when a DC coefficient is reconstructed with some error.

Finally, we model source coder distortion as the DCT coefficient quantization that results when we apply JPEG compression with a perceptual quantization matrix weighted for a viewing distance of six image heights. This generates 8x8 block distortion across the entire image and serves as a model of what might occur in a communication system where source bit rate is sacrificed in order to achieve improved error resilience.

Many communication systems produce distortions that can be modeled with our four distortions. For example, Figure 4 shows a single frame from a video with spatial translation and intensity shift distortions that occur when using a spatial replacement error concealment approach. This example corresponds to the techniques described in Ref. 20.

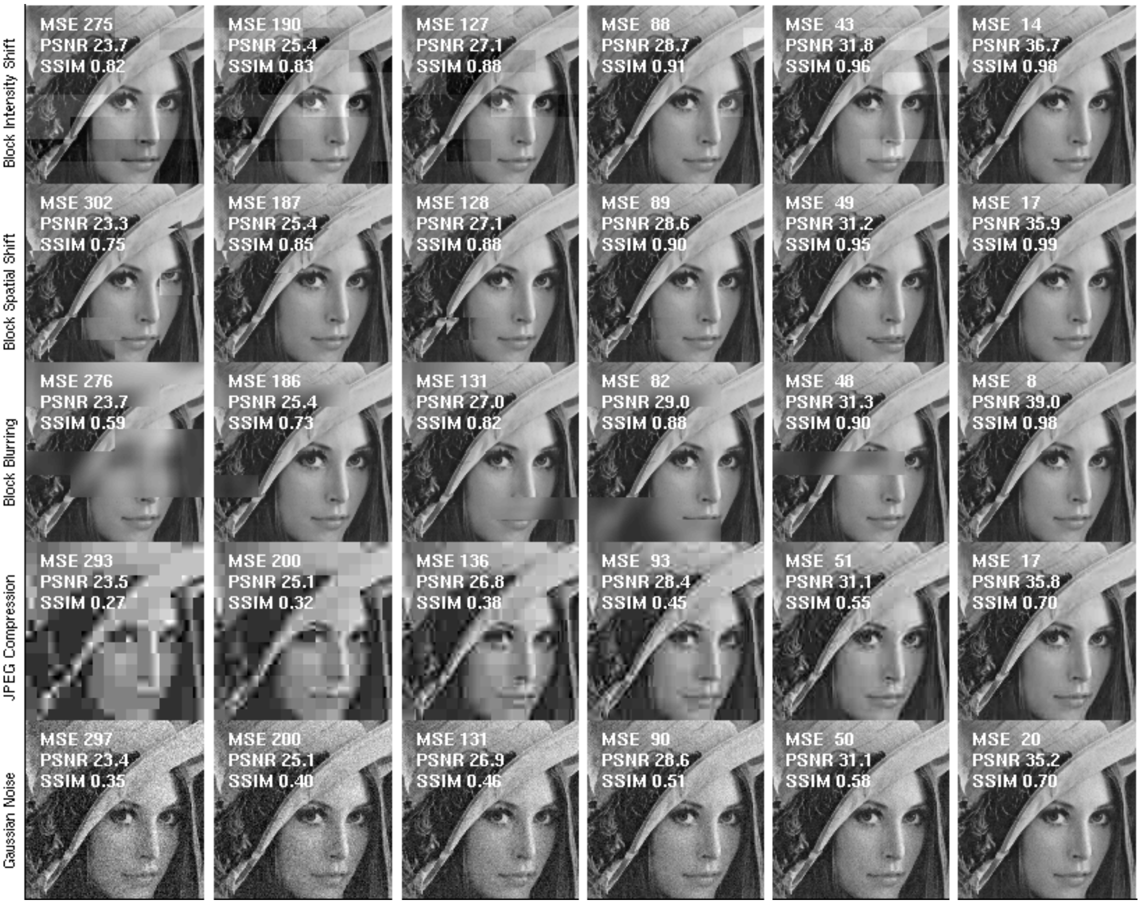
Block Intensity Shift
MSE 275 PSNR 23.7 SSIM 0.82
MSE 190 PSNR 25.4 SSIM 0.83
MSE 127 PSNR 27.1 SSIM 0.88
MSE 88 PSNR 28.7 SSIM 0.91
MSE 43 PSNR 31.8 SSIM 0.96
MSE 14 PSNR 36.7 SSIM 0.98
Block Spatial Shift
MSE 302 PSNR 23.3 SSIM 0.75
MSE 187 PSNR 25.4 SSIM 0.85
MSE 128 PSNR 27.1 SSIM 0.88
MSE 89 PSNR 28.6 SSIM 0.90
MSE 49 PSNR 31.2 SSIM 0.95
MSE 17 PSNR 35.9 SSIM 0.99
Block Blurring
MSE 276 PSNR 23.7 SSIM 0.59
MSE 186 PSNR 25.4 SSIM 0.73
MSE 131 PSNR 27.0 SSIM 0.82
MSE 82 PSNR 29.0 SSIM 0.88
MSE 48 PSNR 31.3 SSIM 0.90
MSE 8 PSNR 39.0 SSIM 0.98
JPEG Compression
MSE 293 PSNR 23.5 SSIM 0.27
MSE 200 PSNR 25.1 SSIM 0.32
MSE 136 PSNR 26.8 SSIM 0.38
MSE 93 PSNR 28.4 SSIM 0.45
MSE 51 PSNR 31.1 SSIM 0.55
MSE 17 PSNR 35.8 SSIM 0.70
Gaussian Noise
MSE 297 PSNR 23.4 SSIM 0.35
MSE 200 PSNR 25.1 SSIM 0.40
MSE 131 PSNR 26.9 SSIM 0.46
MSE 90 PSNR 28.6 SSIM 0.51
MSE 50 PSNR 31.1 SSIM 0.58
MSE 20 PSNR 35.2 SSIM 0.70

Figure 5. Example realistic distortion images where columns have approximately equal mean-squared error. From bottom to top, the lower two rows are baseline images with distortions created by (1) adding white noise and (2) JPEG compression, while the upper three rows result from application of our probabilistic model with (3) block blur, (4) block spatial shift, and (5) block intensity shift distortions. The cropped central 128x128 region of the images is shown, labeled with MSE, PSNR, and CWSSIM results.

Now that we have modeled realistic distortions, we can perform experiments to find out if SSIM quality measurements agree with the intuition that some distortions are more visible than others. In the following experiments, we assume a MB size of $N = 16$ and use the level three subband of the steerable pyramid decomposition averaged over five orientations for CWSSIM.

4.3. The Space of Realistic Distortions: Lena Experiment

The distortion model developed in section 4.2 can be used as a tool for exploring the performance of CWSSIM over a range of distortions. The resulting image data generated by running the distortion model can be viewed as an exploration of the multidimensional image space of realistic distortions in a video communications system. Figure 5 shows some example data generated from the Lena test image with a window size of $W = 8$. The columns have approximately equal MSE with decreasing error from left to right.

One observation that can be gleaned from Figure 5 is that clearly the block blur artifact is the most annoying of the error concealment artifacts. It destroys the image structure within the blurred region and additionally creates obvious block edges. In the fifth column, where the PSNR has a quite high value of 31.3 dB, it is evident that the image with the blur artifacts has much lower perceptual quality than the spatial or intensity shift images.

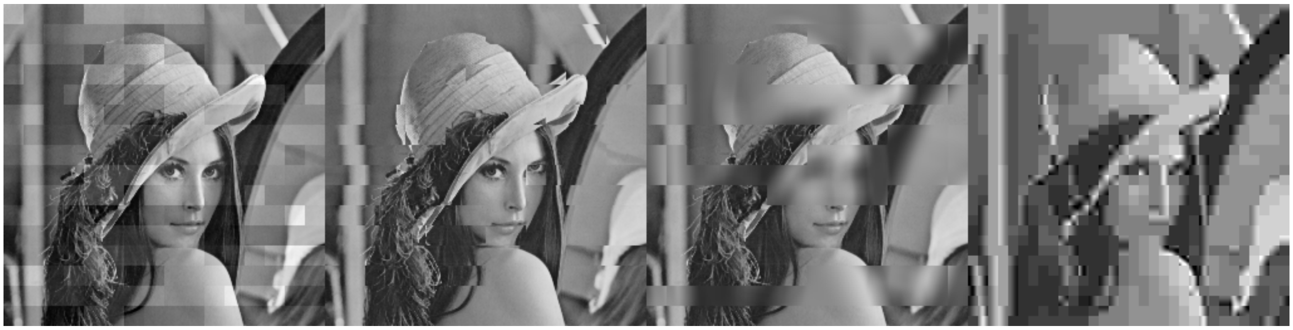

Figure 6. Four degraded images with approximately equal MSE. From left to right, the distortions are: block intensity shifts, block spatial shifts, block blur, and DCT coefficient quantization. These images correspond to the left column of Figure 5.

CWSSIM accurately predicts this difference, giving a value of 0.9 for block blur versus 0.95 and 0.96 for block spatial and intensity shift, respectively.

It is more difficult to pick the least objectionable distortion as we move toward the more highly corrupted images. For example, Figure 6 shows the full-size images that correspond to the left column of Figure 5. Here, one might still argue that the block spatial shift is the least objectionable distortion, but the choice is not as obvious.

4.4. Examples where CWSSIM Accurately Predicts Perceived Quality

One interesting result from the exploration of the space of realistic distortions with the CWSSIM metric is illustrated in Figure 7, which compares the spatial shift and DCT quantization distortions. For equal MSE of 100, CWSSIM gives a much higher quality score (0.78) to the spatial shift image compared to the DCT quantization image (0.30). This result compares very well with informal subjective evaluation of these images, whereby observers quickly see the distortion in the DCT image but typically have to look very closely to even find the problems in the spatially distorted image (e.g., block shifts near her left eye, the base of her left index finger, and by the bottom of her right wrist). CWSSIM accurately predicts that slight spatial translations are much less annoying than excessive DCT quantization source coding.

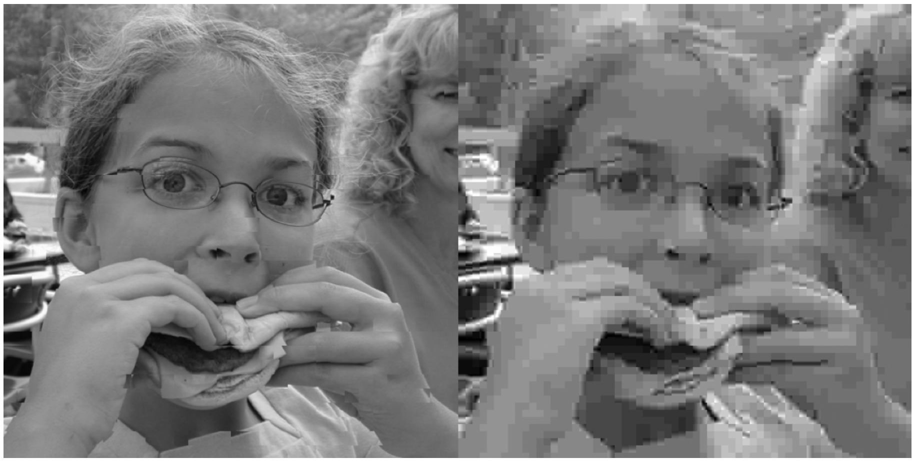

Figure 7. Spatial shift image (left) versus DCT quantization image (right) with equal MSE=100 and CWSSIM of 0.78 and 0.30, respectively. CWSSIM agrees with the subjective observation that the left image has much higher perceptual quality than the right image.

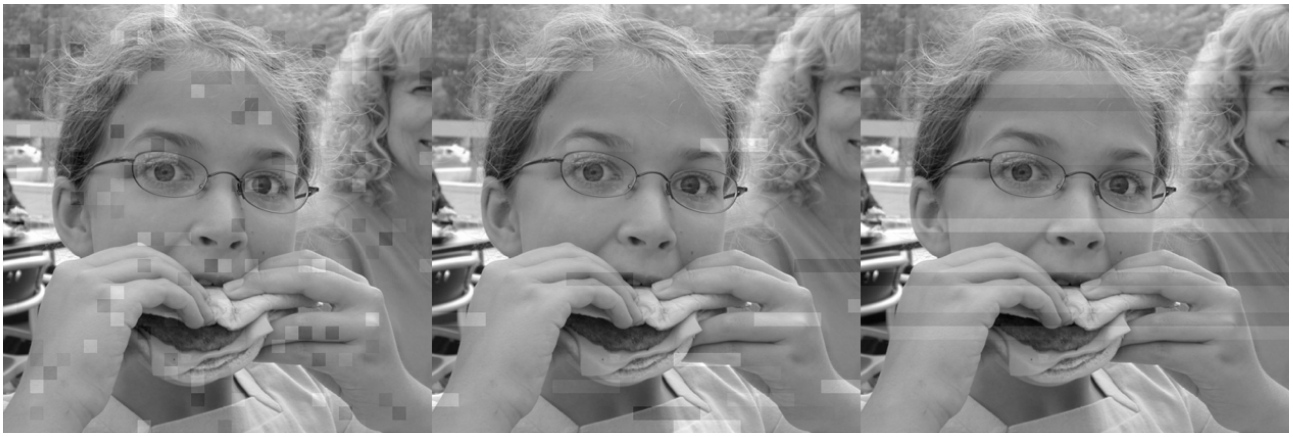

Figure 8. Effect of MB grouping parameter—M blocks per packet—for equal MSE=50. From left to right, $M = 1$, 5, and 32 (entire row) while CWSSIM=0.91, 0.96, and 0.94, respectively. CWSSIM agrees with the subjective observation that the center image has higher perceptual quality than the others.

Another intriguing result of our approach is depicted in Figure 8 where we evaluate the effect of MB grouping into M blocks per packet. For equal MSE of 50, CWSSIM predicts that the images with single block packets ($M = 1$) and packets that contain an entire row of MBs ($M = 32$) have lower quality (CWSSIM=0.91 and 0.94, respectively) than the $M = 5$ image (CWSSIM=0.97). This result coincides with informal perceptual evaluation where the $M = 5$ image is the least annoying.

4.5. CWSSIM Problem With Local Mean Shifts

It is most straightforward to demonstrate one weakness of single-scale CWSSIM by example. Consider the equal-MSE images in Figure 9. The block intensity distorted image has a CWSSIM of 0.94 while the block spatial shift distorted image has a CWSSIM of 0.84. Most human observers make the opposite choice, giving the spatially shifted image higher quality ratings. Viewing the images at a typical viewing distance of six image heights,

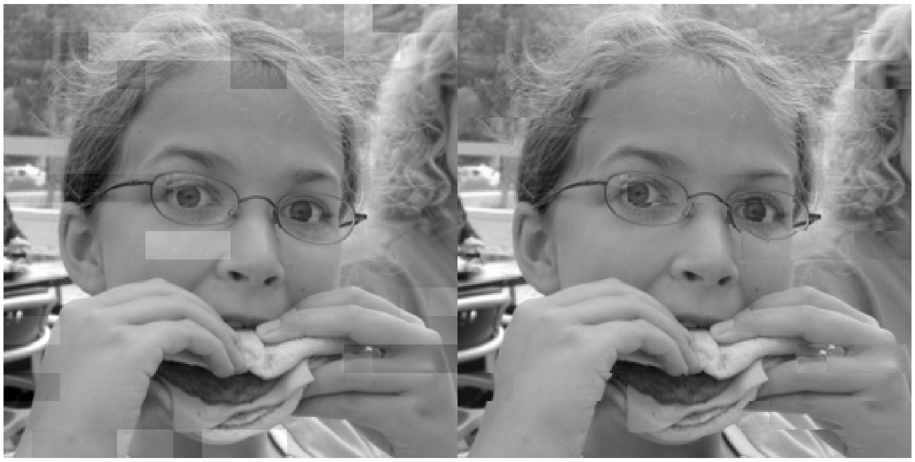

Figure 9. An example where CWSSIM does not agree with the perceptual intuition that the image on the right has higher quality. Two degraded images with equal MSE: block intensity distortions with CWSSIM=0.94 (left) and block spatial shift distortions with CWSSIM=0.84 (right). Our HVS-weighted variant, WCWSSIM, better agrees with perception, giving WCWSSIM=0.87 (left) versus WCWSSIM=0.89 (right).

the artifacts in the intensity distortion image are clearly visible while the artifacts in the spatial shift image are hard to discern.

The main reason single-scale CWSSIM does not agree with perception in this case is that it is not able to account for the low spatial frequency image distortions. These mean shift block distortions can be interpreted as the addition of undesired low frequency structure. Single-scale CWSSIM is looking at a single higher frequency subband, where only the edges of the mean-shift distortions are measured as structural errors. This suggests a natural way to fix the problem, and demonstrates the utility of the proposed exploration of the space of realistic distortions.

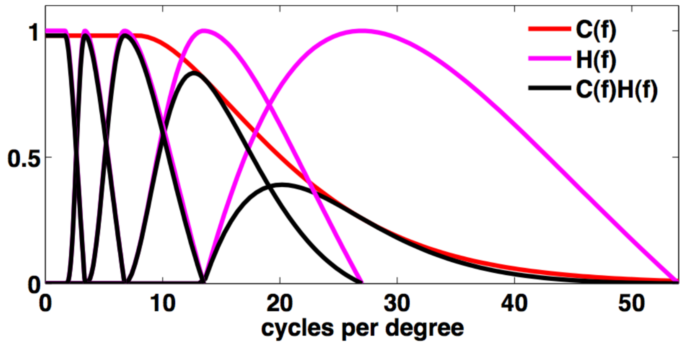

Figure 10. HVS frequency response up to $54\frac{cyc}{deg}$ superimposed on the complex steerable pyramid filter responses.[21] The area under the curves is used to calculate weights for each subband of WCWSSIM, according to (13).

4.6. Weighted Subband Extension of CWSSIM: WCWSSIM

In order to address the weakness described in section 4.5, we propose an extension of CWSSIM that is able to better account for the low frequency distortions evident in Figure 9. We extend CWSSIM to combine the results from multiple wavelet scales using an approach similar to that used in Wang's multi-scale SSIM experiments.[22] However, we derive the subband weights and combine the subbands differently, creating a weighted complex wavelet SSIM implementation (WCWSSIM).

We derive our weights from the Mannos and Sakrinson curve[23] for the HVS contrast sensitivity function (CSF) leveled off at frequencies lower than the peak response, similar to the Daly[24] curve. Figure 10 shows the leveled CSF response superimposed on the complex wavelet steerable filters[21] used in CWSSIM, where the sampling frequency is set to 54 cycles per degree for a six image height viewing distance and a 512 × 512 image size ($\frac{512/2}{\tan^{-1} 1/12} \approx 54$ cycles/degree). The frequency responses of the CSF, $C(f)$, and the wavelet subbands, $H_s(f)$ are combined to find a weight for each subband, s, as follows:

$$W_s = \frac{\int_0^{54} C(f) H_s(f)\, df}{\int_0^{54} H_s(f)\, df}. \tag{13}$$

The resulting weights are: [.254 .254 .25 .18 .061] from low to high spatial frequency. The first weight is applied to the baseband, and the other four to the remaining four levels of the wavelet pyramid. Using these weights to combine the results from multiple subbands, we find that this WCWSSIM version of the metric improves the results of Section 4.5. CWSSIM gave quality ratings of 0.94 and 0.84 for the left and right images in Figure 9, respectively. WCWSSIM improves these results to 0.87 vs 0.89, favoring the block shift image on the right, better agreeing with the higher perceived quality of that image.

4.7. Conclusions

We have examined the use of structural similarity metrics in suprathreshold video compression/transmission applications, and found that, in general, they agree with subjective evaluations. In order to better explore the space of distortions, we proposed models for typical distortions encountered in these applications. We found that the translation-insensitive complex wavelet SSIM is superior to other SSIM implementations, and proposed a multi-scale HVS weighted extension that accounts for the effects of local mean shift distortions.

REFERENCES

1. T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing, 2nd Edition*, A. C. Bovik, ed., Academic Press, 2005.

2. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing* **13**, pp. 600–612, Apr. 2004.
3. M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing* **70**, pp. 177–200, 1998.
4. T. N. Pappas, T. A. Michel, and R. O. Hinds, "Supra-threshold perceptual image coding," *Proc. Int. Conf. Image Processing (ICIP-96)* **I**, pp. 237–240, 1996.
5. J. Chen and T. N. Pappas, "Perceptual coders and perceptual metrics," in *Human Vision and Electronic Imaging VI*, B. E. Rogowitz and T. N. Pappas, eds., **Proc. SPIE Vol. 4299**, pp. 150–162, (San Jose, CA), Jan. 2001.
6. R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," *Proc. ICASSP-89* **3**, pp. 1945–1948, 1989.
7. P. C. Teo and D. J. Heeger, "DCT quantization matrices visually optimized for individual images," *Human Vision, Visual Proc., and Digital Display IV* **1913**, pp. 202–216, 1993.
8. S. S. Hemami and M. G. Ramos, "Wavelet coefficient quantization to produce equivalent visual distortion in complex stimuli," in *Human Vision and Electronic Imaging V*, B. E. Rogowitz and T. N. Pappas, eds., **Proc. SPIE Vol. 3959**, pp. 200–210, (San Jose, CA), Jan. 2000.
9. M. G. Ramos and S. S. Hemami, "Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis," *J. Opt. Soc. Am. A* , Oct. 2001.
10. D. M. Chandler and S. S. Hemami, "Additivity models for suprathreshold distortion in quantized wavelet-coded images," in *Human Vision and Electronic Imaging VII*, B. E. Rogowitz and T. N. Pappas, eds., **Proc. SPIE Vol. 4662**, pp. 105–118, (San Jose, CA), Jan. 2002.
11. D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *J. Opt. Soc. Am. A* **20**, July 2003.
12. Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, IEEE, ed., *Proc. IEEE* **1**, pp. 1–2, 2002.
13. Z. Wang, A. C. Bovik, and E. P. Simoncelli, "Structural approaches to image quality assessment," in *Handbook of Image and Video Processing, 2nd Edition*, A. C. Bovik, ed., pp. 961–974, Academic Press, 2005.
14. Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, P. Philadelphia, ed., *Proc. IEEE* **II**, pp. 573–576, 2005.
15. E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. H. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Information Theory* **38**, pp. 587–607, Mar. 1992.
16. Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distortion measurement," Jan. 2004.
17. Video Quality Experts Group (VQEG), http://www.its.bldrdoc.gov/vqeg/.
18. Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," in *Human Vision and Electronic Imaging, IX, Proc. SPIE*, **5292**.
19. Y. Wang, J. Osermann, and Y. Zhang, *Video Processing and Communicaitons*, Prentice Hill, Upper Saddle River, New Jersey, 2002.
20. Y. Eisenberg, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Minimizing transmission energy in wireless video communications," in *Proc. Int. Conf. Image Processing (ICIP-01)*, **1**, pp. 958–961, (Thessaloniki, Greece), Oct. 2001.
21. J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. on Image Processing* **12**, pp. 1338–1351, Nov. 2003.
22. Z. Wang, E. P. Simoncelli, and A. C. Bovick, "Multi-scale structural similarity for image quality assessment," *37th IEEE Asilomar Conf. on Signals, Systems and Computers* **37**, 2003.
23. J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. on Info. Theory* **IT-20**, pp. 525–536, July 1974.
24. S. Daly, "Subroutine for the generation of a two dimentional human visual contrast sensitivity function," 1987.