# UNDERSTANDING AND SIMPLIFYING THE STRUCTURAL SIMILARITY METRIC

*David M. Rouse and Sheila S. Hemami*

Visual Communications Lab, School of Electrical and Computer Engineering,
Cornell University, Ithaca, NY 14853

## ABSTRACT

The structural similarity (SSIM) metric and its multi-scale extension (MS-SSIM) evaluate visual quality with a modified local measure of spatial correlation consisting of three components: mean, variance, and cross-correlation. This paper investigates how the SSIM components contribute to its quality evaluation of common image artifacts. The predictive performance of the individual components and pairwise component products is assessed using the LIVE image database. After a nonlinear mapping, the product of the variance and cross-correlation components yields nearly identical linear correlation with subjective ratings as the complete SSIM and MS-SSIM computations. A computationally simple alternative to SSIM (c.f. Eq. (6)) that ignores the mean component and sets the local average patch values to 128 exhibits a $1\%$ decrease in linear correlation with subjective ratings to 0.934 from the complete SSIM evaluation with an over $20\%$ reduction in the number of multiplications.

***Index Terms***— quality assessment, human visual system

## 1. INTRODUCTION

Quality assessment (QA) algorithms seek an objective evaluation of image quality consistent with subjective visual quality. These algorithms evaluate a test image $\hat{X}$ with respect to a reference image $X$ to quantify the visual similarity of the test image from the reference image. A challenge for QA algorithms is to generate evaluations consistent with human observer opinions across a variety of image artifacts [1].

The structural similarity (SSIM) [2] metric and its multi-scale extension (MS-SSIM) [3] evaluate visual quality based on the premise that the human visual system (HVS) has evolved to process structural information from natural images, and, hence, a high-quality image is one whose structure closely matches that of the original. To this end, SSIM employs a modified measure of spatial correlation between the pixels of the reference and test images to quantify the degradation of an image's structure. MS-SSIM extends SSIM through a multi-scale evaluation of this modified spatial correlation measure.

SSIM evaluates perceptual quality using three spatially local evaluations: mean, variance, and cross-correlation. Despite its simple mathematical form, SSIM objectively predicts subjective ratings as well as more sophisticated QA algorithms [4, 5]. Furthermore, SSIM's simplicity has intrigued researchers investigating how the HVS evaluates quality [1].

This work investigates how the three SSIM components contribute to its quality evaluation of common image artifacts. A gradient analysis illustrates the value of the SSIM cross-correlation component over the other two components. The performance of individual components and pairwise component products in predicting visual quality is assessed using the LIVE image database [6]. The objective ratings using the product of the variance and cross-correlation components match those of the complete SSIM and MS-SSIM evaluations. A computationally simple alternative to SSIM (c.f. Eq. (6)) that ignores the mean component and sets the local average patch values to 128 exhibits a $1\%$ decrease in linear correlation with subjective ratings to 0.934 from the complete SSIM evaluation with an over $20\%$ reduction in the number of multiplications.

The remainder of this paper has the following organization: Section 2 reviews the SSIM and MS-SSIM metrics. A gradient analysis of the SSIM components is demonstrated in Section 3. The results of individual and combinations of SSIM and MS-SSIM components used to predict subjective ratings of perceptual quality are presented in Section 4. Section 5 analyzes and discusses the results from Section 4. Conclusions are presented in Section 6

## 2. SSIM AND MS-SSIM

SSIM quantifies visual quality with a similarity measure between two patches $x$ and $y$ as the product of three components: mean $m(x,y)$, variance $v(x,y)$, and cross-correlation $r(x,y)$. The two patches, $x$ and $y$, correspond to the same spatial window of the images $X$ and $Y$, respectively. The SSIM value for the patches $x$ and $y$ is given as

$$\begin{aligned}\text{SSIM}(x,y) &= m(x,y)^{\alpha} \times v(x,y)^{\beta} \times c(x,y)^{\gamma} \\ &= \left(\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1}\right)^{\alpha} \times \left(\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}\right)^{\beta} \times \left(\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}\right)^{\gamma} \\ &= m \times v \times r \end{aligned} \quad (1)$$

where $\mu_x$ denotes the mean of $x$, $\sigma_x$ denotes the standard deviation of $x$, $\sigma_{xy}$ is the cross-correlation (inner product) of the mean shifted images $x-\mu_x$ and $y-\mu_y$, and the $C_i$ for $i=1,2,3$ are small positive constants. These constants combat stability issues when either $(\mu_x^2+\mu_y^2)$ or $(\sigma_x^2+\sigma_y^2)$ is

close to zero. The positive exponents $\alpha, \beta$, and $\gamma$ allow adjustments to the respective component's contribution to the overall SSIM value. The original specification for SSIM[1], set $C_3 = \frac{C_2}{2}$ and $\alpha = \beta = \gamma = 1$, which simplifies Eq. (1) to

$$\begin{aligned} \text{SSIM}(x, y) &= \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \times \left(\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \\ &= (m) \times (v \times r). \end{aligned} \tag{2}$$

The overall SSIM image quality index for the images $X$ and $Y$ is computed by averaging the SSIM values computed for small patches of the two images. The SSIM value is computed with $\alpha = \beta = \gamma = 1$ and after downsampling the images $X$ and $Y$ by 2 in both spatial directions [2].

MS-SSIM extends SSIM by computing the variance and cross-correlation components at $K$ image scales, where the $k^{th}$ scale image corresponds to low-pass filtering and subsampling, by a factor of 2 in both spatial directions, the original image $(k-1)$ times. The mean component is only computed at the coarsest scale, $K$. The MS-SSIM index is given by

$$\text{MS-SSIM} = m_K(X, Y)^{\alpha_K} \prod_{k=1}^{K} v_k(X, Y)^{\beta_k} r_k(X, Y)^{\gamma_k}, \tag{3}$$

where $m_k(X, Y)$, $v_k(X, Y)$, and $r_k(X, Y)$ respectively correspond to the mean, variance, and cross-correlation component computed and pooled across patches from scale $k$ with $k = 1$ as the full-resolution image. The exponents $\alpha_K$, $\{\beta_k\}_{k=1}^K$, and $\{\gamma_k\}_{k=1}^K$ vary according to $k$ and adjust the contribution of the components based on experimental results by Wang et al. [3] that examined perceptual image quality across scales for distortions with equal mean-squared error (MSE). The exponents are nonnegative and normalized to sum-to-one across scale (i.e. $\sum_{k=1}^K \beta_k = 1$). The exponents obtained from the experiment by Wang et al. [3] are $\alpha_K = 0.1333$, $\beta_1 = 0.0448$, $\beta_2 = 0.2856$, $\beta_3 = 0.3001$, $\beta_4 = 0.2363$, and $\beta_5 = 0.1333$ with $\beta_k = \gamma_k$ for $k = 1, 2, \cdots, K$.

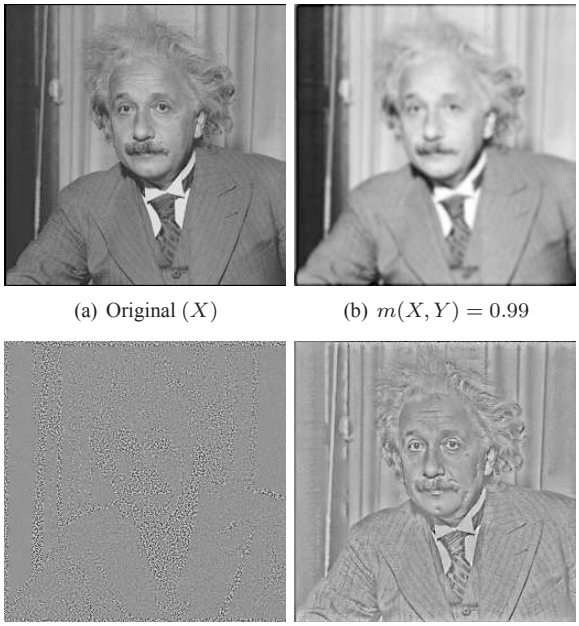
## 3. SSIM COMPONENT GRADIENT ANALYSIS

The SSIM quality metric as given in Eq. (1) combines three components to quantify the visual quality of an image, but it is not immediately obvious how each component evaluates visual quality. A gradient analysis illustrated that for a fixed MSE, the total SSIM quality metric favors an image with increased visual quality [2]. However, a gradient analysis of the individual components of SSIM was not performed.

A gradient analysis, inspired by [2], is performed to examine the visual quality evaluation corresponding with the individual components. An original natural image $X$ is selected, and a random image $Y$ is formed whose pixel values are independently and identically drawn from a uniform distribution with mean 128 and standard deviation $1/12$. For example, to optimize according to the mean component of SSIM, $m(X, Y)$, the image $Y$ is updated at iteration $k$ via gradient ascent according to

$$Y \leftarrow Y + \eta(k) \nabla_Y m(X, Y), \tag{4}$$

where $\eta(k)$ is the learning rate at iteration $k$ and $\nabla_Y m(X, Y)$ denotes the gradient of the mean component with respect to $Y$. Here, $m(X, Y)$ denotes the average of the individual patch means $m(x, y)$.



(a) Original $(X)$ (b) $m(X, Y) = 0.99$

(c) $v(X, Y) = 0.99$ (d) $r(X, Y) = 0.98$

**Fig. 1**. Gradient analysis of the individual SSIM components: mean $m(X, Y)$, variance $v(X, Y)$, and cross-correlation $r(X, Y)$. Images (b) – (d) have been rescaled for visibility.

Figure 1 illustrates the effect of maximizing the individual components of SSIM for the natural image *einstein*. At first glance, using the mean component generates an image (Figure 1(b)) that most resembles the original in Figure 1(a) among the three components. However, the maximum for $m(X, Y)$ does not produce a sharp image. The optimization with the SSIM variance component yields a textured image (Figure 1(c)), where the textures occur along the image edges. The variance component optimization does not adequately restrict the possible pixel value configurations to produce an easily recognizable image. The image optimizing the cross-correlation component captures most of the details from the original image. For instance, notice the details in the hair, eyes and mustache in Figure 1(d). Moreover, the fa-

[1] A Gaussian weighting function is used to compute $\mu_x$, $\mu_y$, $\sigma_x$, $\sigma_y$ and $\sigma_{xy}$ [2]. For example, $\mu_x = \sum_{j=1}^n w_j x_j$, where $w_j$ are weights corresponding to a circular-symmetric Gaussian function with $\sum_{j=1}^n w_j = 1$ and $x_j$ denotes the $j^{th}$ pixel in the patch $x$.

cial expression has a more accurate phenomenal appearance in Figure 1(a) with respect to the original than in Figure 1(b), where the expression appears melancholy rather than alert. The SSIM cross-correlation component clearly assesses quality according to the preservation of the reference image edges.

## 4. PREDICTING VISUAL QUALITY WITH SSIM AND MS-SSIM COMPONENTS

The components of SSIM and MS-SSIM are analyzed in terms of the consistency of their objective quality ratings with subjective ratings. The LIVE image database [6] is used to assess the performance of the components. This analysis considers the individual performance of the components and the performance of these components in pairs. That is, the analysis examines the performance of the mean; variance; cross-correlation; mean and variance; mean and cross-correlation; and variance and cross-correlation. Then, the predictive performance of $v \times r$ (c.f. Eq. (2)) is assessed when removing the calculation of the patch means $\mu_x$ and $\mu_y$.

The SSIM components were computed with $\alpha = \beta = \gamma = 1$ and after filtering and downsampling the reference and test images by a factor of 2 in both spatial directions as specified by [2]. The MS-SSIM metric was computed with the exponents as specified in Section 2.

The LIVE image database is a large collection of distorted images for which subjective visual quality ratings have been recorded [6]. The database consists of 29 reference 24-bits/pixel color images and 779 distorted images. Five types of distortions were evaluated: 1) JPEG-2000 (J2K) compression, 2) JPEG (JPG) compression, 3) additive white Gaussian noise (Noise), 4) Gaussian blurring (Blur) , and 5) simulated bitstream errors of a JPEG-2000 compressed bitstream in a fast-fading (FF) channel. Realigned difference mean opinion scores (DMOS) were used for the subjective ratings [7].

The objective ratings were computed from grayscale images generated according to $Y = 0.2989R + 0.5870G + 0.1140B$, where $R$, $G$, and $B$ denote the 8-bit grayscale red, green, and blue image intensities. The nonlinear mapping of the objective ratings $a$ to the subjective ratings $f$ is given as

$$f(a) = \frac{p_1}{1 + \exp(p_2(a - p_3))} + p_4. \quad (5)$$

The parameters $\{p_j\}_{j=1}^{4}$ were fitted to the data via a Nelder-Mead search to minimize the sum-squared error between the nonlinear mapped objective ratings and the subjective ratings. The performance assessment is based on the linear correlation computed between the DMOS and the objective ratings after nonlinear regression.

### 4.1. Prediction with Individual Components and Pairwise Products of Components

The nonlinear mapping of Eq. (5) was fitted using the objective evaluations for the entire set of distorted images (ALL) for each component and component pair tested. Table 1 summarizes the linear correlation coefficients of the SSIM and MS-SSIM metrics, their individual components, and the pairwise products of the components after nonlinear regression.

**Table 1**. Linear correlation coefficients between DMOS [7] and the individual and pairwise SSIM-based metric component values after nonlinear regression for each artifact type in LIVE database [6]. Refer to Section 4 for artifact acronyms. Within each SSIM-based metric, the rows are ordered by the linear correlation coefficient for the entire set (ALL).

| *Metric* | *Artifact Type* | | | | | |
|---|---|---|---|---|---|---|
| *Component* | ALL | J2K | JPG | Noise | Blur | FF |
| SSIM | **.937** | **.966** | **.979** | **.907** | **.947** | **.948** |
| $v \times r$ | **.937** | **.966** | **.979** | **.908** | **.947** | **.948** |
| $r$ | .932 | .960 | .968 | .925 | .945 | .946 |
| $m \times r$ | .932 | .960 | .968 | .924 | .946 | .946 |
| $m \times v$ | .883 | .948 | .942 | .863 | .906 | .929 |
| $v$ | .880 | .948 | .940 | .861 | .903 | .929 |
| $m$ | .834 | .874 | .928 | .837 | .860 | .691 |
| MS-SSIM | **.934** | **.967** | **.981** | **.905** | **.952** | **.919** |
| $v \times r$ | **.934** | **.967** | **.981** | **.905** | **.952** | **.919** |
| $r$ | .930 | .965 | .975 | .925 | .952 | .916 |
| $m \times r$ | .930 | .965 | .975 | .925 | .952 | .916 |
| $v$ | .881 | .944 | .948 | .865 | .907 | .918 |
| $m \times v$ | .660 | .766 | .803 | .849 | .809 | .717 |
| $m$ | .284 | .588 | .318 | .560 | .405 | .435 |

Individually, the SSIM cross-correlation component predicts subjective evaluations the best among the individual components and nearly as well as the corresponding complete SSIM definition across the six artifact types. The SSIM and MS-SSIM mean component ($m$) exhibits poor correlation with the subjective ratings across most of the artifact types with the exception of the Gaussian noise (Noise) type. The SSIM and MS-SSIM variance component ($v$) correlate well with subjective ratings for each artifact type, but overall demonstrate poorer performance than the cross-correlation component ($r$).

Among the pairwise combinations of the SSIM components, the product of the variance and cross-correlation components ($v \times r$) performs nearly identically to the corresponding complete metric definition that uses all three components. The product of the mean and variance components ($m \times v$) predict subjective ratings well, but it is evident that the incorporation of the cross-correlation component significantly improves the objective quality evaluation. Even the product of the mean and cross-correlation components ($m \times r$) predicts subjective ratings well across the six artifact types.

### 4.2. Prediction without Computing $\mu_x$ or $\mu_y$ for SSIM

The predictive performance of the mean component with the LIVE image database casts doubt on its relevance in an objec-

tive quality assessment for typical image artifacts[2]. However, removing the mean component $m$ from the SSIM index does not significantly reduce the computational complexity, since the variance and cross-correlation components use the terms from $m$: $\mu_x$, $\mu_y$.

Removing or fixing the values of $\mu_x$ and $\mu_y$ produces significant computational savings. When $\mu_x$ and $\mu_y$ are computed for two patches $x$ and $y$ of $n$ pixels, the computation of $v \times r$ over $n$ pixels requires $8n + 8$ multiplications. However, if $\mu_x$ and $\mu_y$ are fixed or set to zero, the computation of $v \times r$ reduces to $6n + 8$ multiplications. For a patch of size $n = 11$, this leads to a reduction of more than $20\%$ in the number of multiplications.

The computation of $v \times r$ with $\mu_x = \mu_y = 128$ (c.f. Eq. (6)) predicts subjective quality ratings very well across all distortion types. Table 2 summarizes the linear correlation coefficients for $v \times r$ when the values $\mu_x$ and $\mu_y$ are fixed to 128. For comparison, the linear correlation of $v \times r$ from Table 1 is included. Moreover, the performance for $\mu_x = \mu_y = 128$ is very similar to the complete SSIM computation.

## 5. ANALYSIS AND DISCUSSION

The gradient analysis of the SSIM components along with the results in Section 4 emphasizes the significance of the cross-correlation component when assessing perceptual quality. Human evaluations of perceptual quality demonstrate a preference for images that preserve image edge information across image scales [8]. This finding is consistent with the principle of global precedence, which contends that the HVS processes a visual scene in a global-to-local order [9]. The MS-SSIM cross-correlation component explicitly evaluates the pixel values across image scales, which provides a measure of how well the edges of two images match. For both SSIM and MS-SSIM, the image that maximizes the cross-correlation component with respect to a reference image possess identical edge information.

A simple analysis explains the prediction performance of $v \times r$ when the local average pixel values are set to 128 (c.f. Table 2). Let $\mu$ denote a fixed mean offset subtracted from an image before computing the product of the SSIM variance and cross-correlation components. In terms of the SSIM definitions of $\mu_x$, $\mu_y$, $\sigma_x^2$, $\sigma_y^2$, and $\sigma_{xy}$, the product of the modified variance and cross-correlation components for a fixed mean offset $\mu$ is given as

$$\hat{v}(x,y) \times \hat{r}(x,y) = \frac{2\sigma_{xy} + C + AB}{\sigma_x^2 + \sigma_y^2 + C + A^2 + B^2}, \qquad (6)$$

where $A = \mu_x - \mu$ and $B = \mu_y - \mu$. Eq. (6) is very similar to the $v \times r$ component of Eq. (2). The additional constant $AB$ in the numerator only shifts the objective rating, and the additional constant $A^2 + B^2$ in the denominator rescales the objective rating. Using the minimum MSE estimate of the mean pixel value, $\mu = 128$, ensures that on average other values of $\mu$ will demonstrate poorer predictive performance. Objective quality evaluation with Eq. (6) does not significantly alter the linear correlation between the DMOS and the objective ratings as demonstrated by the results in Table 2.

[2]The LIVE database contains image artifacts representative of typical imaging applications, where there is limited variation to the luminance.

**Table 2**. Linear correlation coefficients between DMOS [7] and $v \times r$ for fixed $\mu_x = \mu_y = \mu$ after nonlinear regression for each artifact type in LIVE image database [6].

| | *Artifact Type* | | | | | |
|---|---|---|---|---|---|---|
| *Metric* | ALL | J2K | JPG | Noise | Blur | FF |
| $v \times r$ | .937 | .966 | .979 | .908 | .947 | .948 |
| $\mu = 128$ | .925 | .936 | .965 | .898 | .917 | .927 |

## 6. CONCLUSIONS

This work investigates how the SSIM components (mean, variance, and cross-correlation) contribute to its quality evaluation of common image artifacts. The objective ratings using the product of the variance and cross-correlation components match those of the complete SSIM and MS-SSIM evaluations. A computationally simple alternative to SSIM (c.f. Eq. (6)) that ignores the mean component and sets the local average patch values to 128 exhibits a $1\%$ decrease in linear correlation with subjective ratings to 0.934 from the complete SSIM evaluation with an over $20\%$ reduction in the number of multiplications.

### 7. REFERENCES


[1] A. C. Brooks and T. N. Pappas, "Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions," in *Proc. SPIE: HVEI XI*, B. E. Rogowitz, T. N. Pappas, and S. J. Daly, Eds., San Jose, CA, Jan. 2006.

[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 Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. of the 37th IEEE Asilomar Conf. on Sig., Sys. and Comp.*, Pacific Grove, CA, Nov. 2003.

[4] 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.

[5] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[6] H. R. Sheikh, Z. Wang, L. Cormack, and A. Bovik. Live image quality assessment database release 2. http://live.ece.utexas.edu/research/quality.

[7] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3441–3452, Nov. 2006.

[8] 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. Amer. A*, vol. 20, no. 7, Jul. 2003.

[9] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive Psychology*, vol. 9, pp. 353–383, 1977.