

This is the author's final, peer-reviewed manuscript as accepted for publication. The publisher-formatted version may be available through the publisher's web site or your institution's library.

## **A nonparametric-test-based structural similarity measure for digital images**

Haiyan Wang, Diego Maldonado, Sharad Silwal

### **How to cite this manuscript**

If you make reference to this version of the manuscript, use the following information:

Wang, H., Maldonado, D., & Silwal, S. (2011). A nonparametric-test-based structural similarity measure for digital images. Retrieved from <http://krex.ksu.edu>

### **Published Version Information**

**Citation:** Wang, H., Maldonado, D., & Silwal, S. (2011). A nonparametric-test-based structural similarity measure for digital images. *Computational Statistics & Data Analysis*, 55(11), 2925-2936.

**Copyright:** Copyright © 2011 Elsevier B.V. All rights reserved.

**Digital Object Identifier (DOI):** doi:10.1016/j.csda.2011.04.021

**Publisher's Link:** <http://www.sciencedirect.com/science/article/pii/S0167947311001502>

This item was retrieved from the K-State Research Exchange (K-REx), the institutional repository of Kansas State University. K-REx is available at <http://krex.ksu.edu>

# A Nonparametric-Test-Based Structural Similarity Measure for Digital Images

Haiyan Wang<sup>1\*</sup> and Diego Maldonado<sup>2†</sup> and Sharad Silwal<sup>2†</sup>

<sup>1</sup>*Department of Statistics, Kansas State University, Manhattan, KS 66506*

<sup>2</sup>*Department of Mathematics, Kansas State University, Manhattan, KS 66506*

## Abstract:

In image processing, image similarity indices evaluate how much structural information is maintained by a processed image in relation to a reference image. Commonly used measures, such as the mean squared error (MSE) and peak signal to noise ratio (PSNR), ignore the spatial information (e.g. redundancy) contained in natural images, which can lead to an inconsistent similarity evaluation from the human visual perception. Recently, a structural similarity measure (SSIM), that quantifies image fidelity through estimation of local correlations scaled by local brightness and contrast comparisons, was introduced by Wang et al. [2004]. This correlation-based SSIM outperforms MSE in the similarity assessment of natural images. However, as correlation only measures linear dependence, distortions from multiple sources or nonlinear image processing such as nonlinear filtering can cause SSIM to under or overestimate the true structural similarity. In this article, we propose a new similarity measure that replaces the correlation and contrast comparisons of SSIM by a term obtained from a nonparametric test that has superior power to capture general dependence, including linear and nonlinear dependence in the conditional mean regression function as a special case. The new similarity measure applied to images from noise contamination, filtering, and watermarking, provides a more consistent image structural fidelity measure than commonly used measures.

**AMS 2000 subject classifications:** Primary 68U10, 97K80, 62H35; secondary 62G10.

**Keywords and phrases:** Image processing, nonparametric hypothesis testing, image structural similarity, digital image watermarking.

## 1. Introduction

Image similarity indices measure the quality or similarity of an image  $Y$  in relation to a reference image  $X$  and they are of crucial importance in concrete applications. One such application is the use of image similarity indices along with digital image processing. Filtering, compression, transmission, or reproduction of a digital image may result in degradation of its visual quality. An accurate similarity measure can help to decide the parameter settings (e.g. thresholding cut-offs) for optimal results in image processing. Another application appears in the context of content-based image searches or image retrievals. A key step in the design of an image retrieval system is the

---

\*Corresponding author e-mail: [hwang@ksu.edu](mailto:hwang@ksu.edu)

†Research was partially supported by NSF under grant DMS 0901587.

choice of an appropriate image similarity measure. Most traditional and commonly used methods of image search are carried out by adding annotations such as captions, keywords, or descriptions to the images so that retrieval can be performed by means of the annotation words. In order to avoid time-consuming, laborious, and expensive image annotations, whose descriptions may even fail to capture the essence of an image content, there is great interest in content-based image retrieval (CBIR). CBIR aims at retrieving images based on their visual similarity to a user-supplied query image, thus avoiding the use of textual descriptions. In early work on CBIR, similarity measures based on global feature representations, such as color histograms and global shape descriptors, are considered. One problem with all such approaches is the semantic gap (Smeulders et al. [2000]) between low-level content and higher-level concepts. The image domain is too broad and deep for global features to reduce the semantic gap. A major shift has been witnessed in recent years from global feature representations for images to local features and descriptors, such as spatial model features and robust local shape characterizations. Computation of similarity can be performed with feature vectors, region-based signatures, or summarized local features. However, many such methods lack the necessary detail to represent complex image semantics. See Datta et al. [2008] for a review and in-depth discussions.

Most of the literature on image similarity measures is based on the error-sensitivity approach, which summarizes the errors or distortions at each pixel location. Some work uses global or local features of the image and defines dissimilarity based on distortions on the feature vectors. Working with the distortions is consistent with the first stage of the human perception system that focuses on the differences of the images. Mean squared error (MSE) is one of such measures popularly utilized. MSE, which averages squared deviations of all elements of the distortion vector, is used to summarize the dissimilarity between two images. It has the advantages of being easy to compute and providing a convenient Hilbert space structure for mathematical models. Peak signal-to-noise ratio (PSNR) is another popularly used measure of image processing quality. It is proportional to the  $\log_{10}$  ratio of the squared maximum pixel value of the image and MSE. Minkowski distances use  $l_p$  norm to summarize element-wise distortions. As many different distortion vectors can lead to identical MSEs, several distorted images with identical MSE may have very different similarities relative to the original image. PSNR and Minkowski distances share a similar drawback. In addition, all these three measures completely ignore the spatial relationship among the pixels of natural images and treat the distortion at each pixel location equally.

Some recent literature considered region-based image dissimilarity measure to take into account some of the spatial relationships between regions (cf. Wang et al. [2001]; Ko and Byun [2002]). An image is first described by a set of segmented regions and then distances between matching regions are combined into a measure. One of such approaches is the weighted distance (Wang et al. [2001]) that summarizes the difference between two images through a summation of the weighted distance between two sets of feature vectors. Hausdorff distance (HD, Ko and Byun [2002]) is another

region-based measure. The HD is defined as the maximum distance of a set to the nearest point in the other set. That is, each image is first described by a set of vectors and the HD calculates the maximum distance of one set to the closest vector in the other set. As the selection of feature vectors or regions may not fully describe the entire image, other methods have been vigorously pursued by researchers to account for spatial dependence of image pixels. The Kullback-Leibler (K-L) distance has been recently studied in texture retrieval (Do and Vetterli [2002]; Mathiassen et al. [2002]). The K-L distance measures the expected log-likelihood ratio between two distributions with the expectation under one distribution. The unknown distributions need to be estimated to obtain the K-L distance. The estimation is, however, either required to be restricted to a specific parametric family or ignores the spatial dependence among the image data. Wang et al. [2004] proposed a correlation-based structural similarity measure (SSIM) to account for spatial relationship. Among these methods, the SSIM has been the subject of considerable attention in the recent literature on image quality assessment (cf. Wu and Rao [2005]; Wang and Bovik [2006] and the references therein). Extensive experiments have demonstrated consistently better performance of SSIM over MSE (Sheikh et al. [2006]). However, as correlation only captures linear relationship, nonlinear or multiple sources of distortions may significantly limit the performance of SSIM. For example, median filtering is a nonlinear transformation of the original image. Our example in Section 2.1 shows that the SSIM may assign a high score to a median filtered image that has lost the majority of the fine details but assigns a similar or lower score to an image slightly contaminated with Gaussian noise that still contains all the fine details (see Figure 1).

In this article, we introduce a new image similarity measure based on hypothesis testing to assess structural information change by evaluating the dependence of local blocks of the error signal on the images being compared. To overcome the disadvantage of SSIM, a nonparametric test with no distributional assumptions will be used to detect general relationship between the error signal and the test images. The rest of the article is organized as follows. Section 2.1 reviews the strengths and limitations of SSIM which has motivated this research. Section 3 describes the details of the new similarity measure and its properties. Section 4 is devoted to some applications and performance comparison with MSE and SSIM followed by a summary at the end. For clarity, all images in the manuscript are given with small size to allow easy organization. The quality of the images can be seen with the zoom in option in the pdf file.

## 2. SSIM: strengths and some limitations

The SSIM was introduced in Wang and Bovik [2002], and formally studied in Wang et al. [2004]. The principle underlying the basic SSIM is that the retention of signal structure should be an important ingredient since the human visual system (HVS) is highly adapted to extracting structural information from visual scenes. Suppose that  $\mathbf{x}$  and  $\mathbf{y}$  are local image blocks taken from the same location of two images that are being compared. The local SSIM index is the product of the similarity measures of three elements of the image blocks: the similarity of the local block brightness

values (luminance), the similarity of the local block contrasts, and the similarity of the local block structures. Namely,

$$SSIM(x,y) = \left( \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \right) \cdot \left( \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \right) \cdot \left( \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \right), \quad (2.1)$$

where  $\mu_x$  and  $\mu_y$  are the local sample means of  $x$  and  $y$ , respectively,  $\sigma_x$  and  $\sigma_y$  are the local sample standard deviations of  $x$  and  $y$ , and  $\sigma_{xy}$  is the sample covariance of  $x$  and  $y$ .  $C_1$ ,  $C_2$ , and  $C_3$  are small positive constants to avoid numerical instability. The final SSIM score is obtained as the average of the local SSIM indices. For further details on SSIM and its variants, see [Wang and Bovik \[2006\]](#).

Hence, by neglecting the constants  $C_1$ ,  $C_2$ , and  $C_3$ , it follows that  $SSIM(X,Y) = 1$  if and only if  $X = Y$ . From the formula in (2.1), it can also be seen that the SSIM index measures the local structural similarity between two images through the sample correlation. As correlation is invariant under location or scale changes, comparison between the local means is considered to penalize location (i.e. luminance) shift, and similarly, comparison between the local sample standard deviations are used to penalize scale changes. It has been shown in [Sheikh et al. \[2006\]](#) through experiments on a wide variety of images and distortion types that the SSIM index gives much more consistent scores than MSE does relative to visual perception.

On the other hand, correlation is a measure of linear relationship. The estimated correlation through samples can change dramatically if there are influential points or if the images have gone through more than one source of distortion. As an example, we illustrate with the images (a) and (d) in [Figure 1](#). The image in [Figure 1](#) (d) was generated by adding noise from a mixture distribution to the original image in (a). The noise imitates two sources of distortion, one of which happens 40% of the time and comes from a t-distribution with 3 degrees of freedom plus a shift of 30 in the mean; the other source happens 60% of the time and is from exponential distribution with a mean of one. A typical scatter plot between the observations in the same local block of size  $11 \times 11$  (default used in SSIM) of images (a) and (d) is shown in the left panel of [Figure 2](#) and the distribution of the sample correlations is depicted with the histogram in the right panel of [Figure 2](#). Without distortion, the data from the same local block of two images are perfectly correlated. The double sources of distortion have a dramatic effect on the sample correlation rendering an average sample correlation of 0.407. This is because the squared sample correlation is the proportion of variations explained by the straight line fit which is small in such cases.

The latest version of SSIM available online actually first applies a locally weighted smoothing ([Wang and Bovik \[2006\]](#)) before estimating the local correlation. Such local smoothing unavoidably makes SSIM lose its sensitivity to edge or boundary loss. Accordingly, SSIM tends to assign overly optimistic similarity scores to blurred images. For example, SSIM gives a score of 0.7137 for image (d) relative to image (a) in [Figure 1](#). Compared to the blurry image (c) in [Figure 1](#), image (d) clearly has more visual structural similarity to the original image, since image (c) has

lost almost all of the fine details present in the original image (see panel (c) of Figure 3 for the error signal). Unfortunately, correlation-based SSIM gives a smaller similarity score for image (d) than for image (c) due to the limitation of using correlation to capture the relationship between two variables. Instead, SSIM judges that images (b) and (c) in Figure 1 have similar structural fidelity to the reference image (a). However, images (b) and (c) have very different visual similarities to the original image. Notice that image (b) is simply the reference image with some Gaussian noise. Such simple noise easily reduces the sample measures of correlations because it reduces the percent of variations explained by the linear relationship.

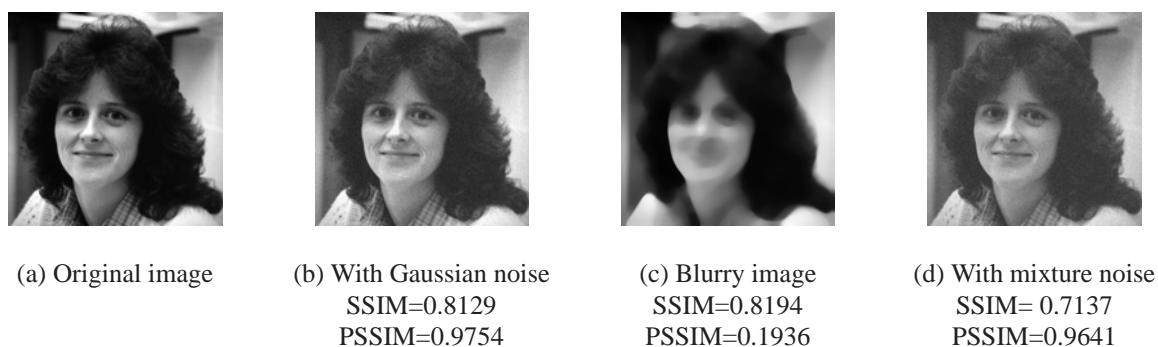


FIG 1. Original image and three versions of distorted images

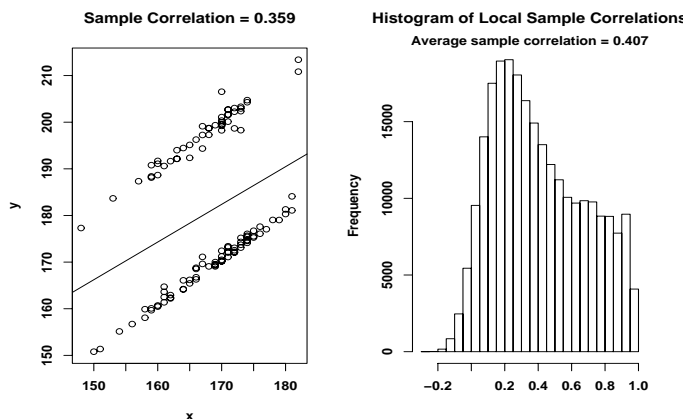


FIG 2. Noise from mixture distribution dramatically affects the sample correlation.

### 3. The proposed $p$ -value-based structural similarity measure (PSSIM)

In order to overcome the correlation-related limitations in the SSIM, we first examine how the human visual system (HVS) processes the information from an image and comes up with a similarity measure. The error-sensitivity principle is commonly used to evaluate the quality of an image compared to a reference image. The idea behind this principle is the decomposition of a distorted image into the sum of the reference image and an error signal. The loss of perceptual quality is

directly related to the visibility of the error signal. As commented in Wang et al. [2004], the error-sensitivity approach simulates the functional properties of early stages of the HVS characterized by both psychophysical and physiological experiments. The point on how poorly MSE performs as a visual similarity measure has been conclusively made by Wang et al. [2004] and Wang and Bovik [2009], where a number of distorted images with the same MSE with respect to a reference image yield inconsistent and even incompatible visual renderings. SSIM is then shown to outperform MSE as an image quality measure. However, as argued above, SSIM is best suited to capture only linear or near linear dependence of the two image signals due to the intrinsic limitations of the correlation as a dependency measure. To address this drawback, we believe it is necessary to take into account nonlinear dependence of the two images being compared both because the distorted image can come from a nonlinear distortion mechanism and because the HVS is a complex and highly nonlinear system as is vigorously argued in Wang et al. [2004].

The goal of this article is to introduce a new similarity measure that effectively assesses image structural similarity of two images by capturing general dependence between local blocks of two images. We start with the error signal defined by the distortion of the image compared to the reference image. Denote by  $\mathbf{X} = (X_{ij})_{m \times n}$  and  $\mathbf{Y} = (Y_{ij})_{m \times n}$  the pixel matrices of the original and the distorted images, respectively. The distortion is quantified as  $\mathbf{Z} = (X_{ij} - Y_{ij})_{m \times n}$ . The structural loss of the distorted image is reflected by how much structural similarity the error signal contains compared to the reference image. In Figure 3, the structural loss of the blurry image relative to the reference image happens almost everywhere except for some background in uniform shade.



FIG 3. *The HVS perceives the difference of two images by quantifying the error signal*

We say that an image  $\mathbf{Y}$  has structural loss compared to image  $\mathbf{X}$  if the error signal  $\mathbf{X} - \mathbf{Y}$  contains structural information of  $\mathbf{X}$ . Two images  $\mathbf{X}$  and  $\mathbf{Y}$  have identical structure if and only if  $\mathbf{X}$  has no structural loss compared to  $\mathbf{Y}$  and  $\mathbf{Y}$  has no structural loss compared to  $\mathbf{X}$ . With a pre-defined block size, the structural similarity of the images  $\mathbf{Y}$  and  $\mathbf{X}$ , denoted by  $S(\mathbf{X}, \mathbf{Y})$ , can be defined as the proportion of local blocks where the error signal  $\mathbf{X} - \mathbf{Y}$  does not contain structural information from either image  $\mathbf{X}$  or  $\mathbf{Y}$ . Note that the definition of structural loss is not symmetric, but the structural similarity is symmetric. The error signal in panel (c) contains structural information of both the original image in panel (a) and the blurry one in panel (b) for most of the local blocks.

This provides evidence that the two images have structural difference in those blocks.

As the HVS examines local blocks of both the reference and the distorted images to assess their similarity, we model the dependence of each local block of the distortion on their reference counterparts. Specifically, for images of size  $m \times n$  with  $m \geq 2, n \geq 64$ , we consider local blocks of size  $a \times b$ , where  $a$  is some small value, such as 2, and  $b$  is of reasonable size, such as  $b = 2^6$ . The relatively large  $b$  is to ensure that the local structures of the images are identifiable. The similarity of the two images being compared not only depends on percentages of local blocks from the distortion that resemble the original image profile, but also depends on the luminance similarity at each pixel location. Therefore, we proceed in two steps:

1. For each local block, evaluate if the distortion  $\mathbf{Z}$  is independent of the corresponding block of the reference image  $\mathbf{X}$ . As the distortion could come from any source, such as Gaussian noise, Poisson pulse, median filtering, wavelet filtering (in some cases, the distorted image might even be the result of the reference image going through multiple filters), an effective procedure that works well for most cases would require a distribution-free assumption on the error signal. If the distortion is independent of the original images, then the distortion is simply some additive noise. Otherwise, the distortion contains some structural loss from the reference image.
2. The magnitude of distortion relative to the original image at each pixel location affects visual perception of the similarity in the sense that it affects human perception of local luminance. Therefore, we include comparison of the local luminance in the construction of our new image similarity measure.

For step 1, instead of correlation, we employ a newly-developed test by Wang et al. [2010] for detecting general dependence between two variables. Specifically, we apply the test statistic of Wang et al. [2010] on the local blocks of the error signal and the reference image. Let  $\mathbf{X}^B$  (with elements  $X_{ij}^B$ ) and  $\mathbf{Z}^B$  (with elements  $Z_{ij}^B$ ) be the local blocks of the reference image and the error signal being considered. The null hypothesis states that the conditional distribution of  $Z_{ij}^B$  given  $X_{ij}^B$  is independent of the marginal distribution of  $X_{ij}^B$ . structural deviation from the reference image will be collected in the error signal. Consequently, the tests on some of the local blocks would reject the null hypothesis of independence between the error signal and the reference image. If  $\mathbf{X}$  and  $\mathbf{Y}$  are structurally different images, then  $\mathbf{Z} = \mathbf{X} - \mathbf{Y}$  depends on both  $\mathbf{X}$  and  $\mathbf{Y}$ , and we expect the test to detect significant dependence. The proportion of non-rejections on all local blocks gives a measure of structural fidelity for the testing image relative to the reference image. To describe the test statistic, we first define some notations. Let  $\text{floor}(x)$  be the largest integer not greater than  $x$ . Define  $\hat{F}_{X,i}(x) = b^{-1} \sum_{j=1}^b I(X_{ij}^B \leq x)$ , where  $I$  represents the indicator function. Let  $C_{ic}$  be the set of column indices for  $\mathbf{X}^B$  such that the corresponding  $\mathbf{X}^B$  values in the  $i$ -th row are among the  $k$ -nearest neighbors of (and centered at)  $X_{\Delta}^B$ , where  $\Delta := \text{floor}(c/b), c - \text{floor}(c/b) * b$ ,



for  $c = 1, \dots, ab$ , where  $k$  is a small integer because the inference of the test was developed for a finite number of nearest neighbors. Typical values of  $k = 3, 5, 7$ , or  $9$ , can be used. We recommend to use  $k = 7$  to allow for sufficiently many nearest neighbors to capture spatial dependence and still preserve the local nature. A larger value of  $k$  may be used if a larger value of  $b$  is preferred.

The test statistic is the difference of two quadratic forms multiplied by a standardizing rate

$$\sqrt{ab}(M_{ab} - W_{ab}) = \sqrt{ab}T_B + o_p(1), \quad (3.1)$$

$$\begin{aligned} \text{where } M_{ab} &= ka^{-1}(ab-1)^{-1} \sum_{i=1}^a \sum_{i_1=1}^a \sum_{j_1=1}^b \left[ k^{-1} \sum_{j=1}^b Z_{ij} I \left( b |\widehat{F}_{X,i}(X_{i_1 j_1}^B) - \widehat{F}_{X,i}(X_{ij}^B)| \leq \frac{k-1}{2} \right) \right. \\ &\quad \left. - (abk)^{-1} \sum_{i_2=1}^a \sum_{j_2=1}^b \sum_{j=1}^b Z_{ij} I \left( b |\widehat{F}_{X,i}(X_{i_2 j_2}^B) - \widehat{F}_{X,i}(X_{ij}^B)| \leq \frac{k-1}{2} \right) \right]^2, \\ W_{ab} &= \{a^2 b(k-1)\}^{-1} \sum_{i=1}^a \sum_{i_1=1}^a \sum_{j_1=1}^b \sum_{j=1}^b \left[ Z_{ij} I \left( b |\widehat{F}_{X,i}(X_{i_1 j_1}^B) - \widehat{F}_{X,i}(X_{ij}^B)| \leq \frac{k-1}{2} \right) \right. \\ &\quad \left. - k^{-1} \sum_{j_2=1}^b Z_{ij_2} I \left( b |\widehat{F}_{X,i}(X_{i_1 j_1}^B) - \widehat{F}_{X,i}(X_{ij_2}^B)| \leq \frac{k-1}{2} \right) \right]^2, \\ T_B &= \sum_{i=1}^a \sum_{j \neq j'}^b \frac{(Z_{ij} - E(Z_{ij} | \mathbf{X}^B))(Z_{ij'} - E(Z_{ij'} | \mathbf{X}^B)) K_{ijj'}}{a^2(k-1)b} \text{ and } K_{ijj'} = \sum_{c=1}^{ab} I(j \in C_{ic}) I(j' \in C_{ic}). \end{aligned} \quad (3.2)$$

The left hand side of (3.1) is easy to compute with the pixel data from the images while the right hand side of (3.1) has a clear interpretation that can be seen from (3.2). In fact,  $T_B$  is closely related to the expected conditional local correlation between every pair of pixels in the error signal with correlation induced by their dependence on the reference image  $\mathbf{X}^B$ . The  $K_{ijj'}$  in (3.2) serves as a weight function which connects the error signal locally with the empirical distribution function of  $X_{ij}^B$ . We do not use  $T_B$  because it can not be calculated unless a nonlinear estimation of  $E(Z_{ij'} | \mathbf{X}^B)$  is available and such estimation typically contains bias in addition to further smoothness assumptions. On the other hand,  $\sqrt{ab}(M_N - W_N)$  can be directly obtained from the image data. With Theorem 1 in Wang et al. [2010], it can be seen that this statistic has an asymptotically normal distribution if the marginal cumulative distribution function of  $X_{ij}$  is differentiable and  $b$  is large. The asymptotic mean is zero if the error signal does not contain information of the reference image and the asymptotic variance can be estimated consistently with the  $\widehat{\gamma}_{ab}^2$  below after denoting by  $j_*$  the rank of  $X_{ij}^B$  within the  $i^{th}$  row of the local block:

$$\widehat{\gamma}_{ab}^2 = \sum_{i=1}^a \sum_{j_* < j'_*}^b \left\{ \frac{4 \widehat{\sigma}_i^2(X_{i(j_*)}^B) \widehat{\sigma}_i^2(X_{i(j'_*)}^B)}{ba^3(k-1)^2} \left[ \widehat{V}_{ijj'_*}^2 + \widehat{V}_{ijj'_*} - 2I(j'_* - j_* \leq (k-1)/2) \right] \right\} I(j'_* - j_* \leq k-1),$$

where

$$\widehat{\sigma}_i^2(X_{ij}^B) = \frac{1}{k-1} \left\{ \sum_{l=1}^b Z_{il}^2 I \left[ |\widehat{F}_{X,i}(X_{il}^B) - \widehat{F}_{X,i}(X_{ij}^B)| \leq \frac{k-1}{2b} \right] - \frac{1}{k} \left( \sum_{l=1}^b Z_{il} I \left[ |\widehat{F}_{X,i}(X_{il}^B) - \widehat{F}_{X,i}(X_{ij}^B)| \leq \frac{k-1}{2b} \right] \right)^2 \right\},$$

$$\widehat{V}_{ijj'} = \sum_{i_1, i_1 \neq i}^a \left( \widehat{d}_{i_1 i}(X_{i_1(j_*)}^B) + 1 \right) [k - (j'_* - j_*)] I(j'_* - j_* \leq k-1), \text{ and } \widehat{d}_{i_1 i}(X_{ij}^B) = k^{-1} \sum_{j_4=1}^b I \left( |\widehat{F}_{X,i}(X_{ij}^B) - \widehat{F}_{X,i}(X_{i_1 j_4}^B)| \leq (k-1)/(2b) \right).$$

If  $X^B = Y^B$ , we set the  $p$ -value to be 1.

This test has certain advantages over not only correlation-based approaches including Pearson, Spearman or Kendall's correlations, but also over likelihood-based methods such as linear models, generalized additive models that combine likelihood approach with local smoothing (Wood [2008]), rank-based method for linear models (Terpstra and Mckean [2005]), and copula-based tests of independence (Genest and Rémillard [2004]). We summarize them below and refer to Wang et al. [2010] for details.

- The test is conservative under the null hypothesis of independence but is highly powerful under the alternatives to capture general dependency including nonlinear relationship between  $Z_{ij}^B$  and  $X_{ij}^B$ .
- The response variable (i.e., the error signal in this article) can be continuous or discrete. This allows flexibility and valid inference for all sorts of distortions on the image, including random noise contamination and loss or gain of edge details.
- The variations of the error signal can be different for different rows or columns. Such heteroscedastic variations would typically make the classical regression methods fail due to the violation of the constant variance assumption.
- The whole procedure is distribution-free and resistant to outliers. This is necessary as the distorted image may come from the original image through any kind of filter and a few outliers in the error signal should not lead to serious structural difference in the two images.
- The test is able to detect the dependency of not only the local mean of the error signal on the reference image, but also how the variations of the error signal change with the reference image since both types of changes are under the alternative that the error signal is not independent of the reference image. Particularly, since the variation dependence is already taken into account in the test statistic, it is no longer necessary to include the local contrast of the two images as a component in the construction of the proposed similarity index.

These advantages make this test a better candidate to capture the structural change contained in the error signal. For each local block  $B$ , apply the above test on  $Z^B$  versus  $X^B$  and record the  $p$ -value,  $p_{X,B}$ . This will be used later to assess the structural loss for  $\mathbf{Y}$  compared to  $\mathbf{X}$ . For structural similarity, apply the above test also on  $Z^B$  versus  $Y^B$  and record the  $p$ -value,  $p_{Y,B}$ . The  $p$ -value to be considered for calculating structural similarity is  $p_B = \max\{p_{X,B}, p_{Y,B}\}$ .

We recommend that the horizontal shift between consecutive blocks be no greater than  $\text{floor}(b/2)$  to identify local differences in structure. The vertical shift can take any integer value that is no greater than  $a$ . The default horizontal and vertical shifts are set to be 32 and 2 respectively for  $a = 2$ ,  $b = 64$ . As the tests proceed with local blocks moving across the entire error signal with corresponding blocks of reference and test images, the collection of  $p$ -values  $\{p_B, B = 1, \dots, N\}$  is produced, where  $N$  is the total number of local blocks. Each  $p_B$  reports the sample evidence of

testing whether the corresponding block of the error signal does not contain information about the reference image or the test image. The proportion of  $p_{X,BS}$  greater than a commonly used significance level  $\alpha$  reveals the proportion of local blocks where the test image contains no structural loss compared to the reference image. Similarly, the proportion of  $p_{BS}$  greater than  $\alpha$  estimates the proportion of the local blocks where the two images have similar structure.

The default  $\alpha$  level we recommend to use is  $\alpha = 0.01$ . Under the null hypothesis of independence between the error signal and the reference image, the proportion of false rejections is expected to be no greater than 1%. Correspondingly, the proportion of non-rejections is expected to be at least 99% when the error signal does not contain any structural information of the reference image. Different from the regular hypothesis testing case where the interest is in finding sample evidence to reject  $H_0$ , here we are more interested in the proportion of non-rejections. Consequently, the multiple comparison-adjustment procedure routinely used to control the number of false discoveries, such as Bonferroni correction or false discovery rate control, is not necessary in our current setting. This is because a non-rejection result from a test using  $\alpha = 0.01$  needs to pass a higher threshold than that using the same level with a Bonferroni correction. In the latter case, the threshold is  $0.01/N$ , which gives a less stringent rule to declare that the test image has no structural change compared to the reference image.

For step 2, the two images at each pixel location may look identical except for some shift in their brightness or shade. However, this luminance shift is invariant to the test described in the previous step because the test statistic is invariant under location shifts, as easily seen from (3.2). Therefore, we need to consider the local luminance change separately. SSIM penalizes the similarity measure through the first term in (2.1) using locally estimated means of the two images. As local smoothing induces bias for any finite sample, we use the raw pixel data from each image directly and calculate the average of all the luminance comparisons  $L_{ij}$  to quantify the overall luminance change, where  $L_{ij} = \frac{2X_{ij}Y_{ij}+C}{X_{ij}^2+Y_{ij}^2+C}$  for all  $1 \leq i \leq m, 1 \leq j \leq n$ , and  $C$  is a small positive constant to avoid numerical instability. We take  $C = 0.001$ .

The proposed new similarity measure is the product of the two components described in this section, i.e., the proportion of non-rejections and the average luminance comparison:

$$PSSIM = \frac{\#\{p_B : p_B > \alpha\}}{N} (mn)^{-1} \sum_{i=1}^m \sum_{j=1}^n L_{ij}.$$

SSIM also considers the comparison of local standard deviations as a component because correlation is invariant to both location and scale changes. However, the local standard deviation comparison is not necessary for the new similarity measure because the test already takes into account such comparison.

#### 4. Applications and performance evaluation

In this section we give some examples of applications of the PSSIM along with comparisons with MSE and SSIM. The SSIM MATLAB code was downloaded from

<http://www.ece.uwaterloo.ca/~z70wang/research/ssim/ssim.m> and the PSSIM was implemented in R 2.8.1.

#### 4.1. Performance comparisons with commonly used measures on images contaminated with luminance shift and random noises

Here we consider nine  $512 \times 512$  reference images and their distortions to compare our new similarity measure with commonly used measures. These nine images are standard test data in the image processing literature. They are labeled as  $A_{i1}$  for  $i = 1, \dots, 9$ . A batch of ten distorted images were generated for each test image, labeled as  $A_{ij}$ , for  $j = 2, \dots, 11$ . The first five distortions are at low noise level and remaining five are at increased noise level.

##### 1. Low noise level

- D1 **Luminance shift** by 11 for images on scale  $[0, 255]$ . The constant value 11 was added to each entry of  $A_{i1}$  to produce  $A_{i2}$ , for  $i = 1, \dots, 9$ . Since the images have pixel values in the grayscale range of  $[0, 255]$ , values greater than 255 are truncated to 255.  $A_{i2}$  bears almost identical structure as  $A_{i1}$  but is slightly brighter than  $A_{i1}$  unless there are too many truncated pixels for which we have, in effect, luminance shift of less than 11.
- D2 Corruption with additive **Gaussian noise**. After scaling the image to range  $[0, 1]$ , the added noise has mean 0 and variance 0.0018. Common noises in digital camera images such as ‘reset noise’, ‘Johnson noise’, and ‘white noise’ are closely related to Gaussian noise.
- D3 Corruption with **salt and pepper noise** with density 0.006. That is, 0.3 % of the pixels per image were randomly selected to have minimum values and another 0.3 % were selected to have maximum values (i.e., turning entries to 255 or 0 for  $[0, 255]$  scaled images). Since all test images in this section are of size  $512 \times 512$ , the salt and pepper noise affects approximately  $0.006 \times 512^2 = 1572$  pixels. This noise occurs in digital camera images due to errors in signal transmission or memory locations.
- D4 Corruption with **Poisson noise**. If the true pixel in the original image is less than 50, the corresponding pixel in the noisy version is randomly generated from Poisson distribution with mean equal to the true pixel value. When the true pixel value is 50 or more, then its corresponding noisy pixel comes from a Gaussian distribution with both mean and variance equal to the true pixel. This noise is most common in X-rays, MRIs and CT-scans due to the Poisson law of photon-counting in low-light situations in radiography.
- D5 Corruption with multiplicative **speckle noise**. For each  $i = 1, 2, \dots, 9$ , the element of  $A_{i6}$  is obtained by corresponding element of  $A_{i1}$  multiplied by  $(1 + \epsilon)$  where  $A_{i1}$  is scaled to have range  $[0, 1]$  and  $\epsilon$  is uniformly distributed random noise with mean 0 and variance 0.007. This noise is commonly found in active radar, synthetic aperture radar

or ultrasound images and occurs due to the roughness of the surface being of the order of the wavelength that causes interference in the returned waves.

## 2. Increased noise level

D6 Corruption with **Gaussian noise** with mean 0 and variance 0.01 for images on  $[0, 1]$  scale.

D7 Corruption with **Gaussian noise** with mean 0 and variance 0.068 for images on  $[0, 1]$  scale.

D8 Corruption with **salt and pepper noise** with density 0.011.

D9 Corruption with **speckle noise** as in [D5] but the  $\epsilon$  has variance 0.012.

D10 Corruption with **speckle noise**. The  $\epsilon$  has variance 0.12.

All the noises were generated using the built-in MATLAB function “imnoise”. Since the same parameters for the same distortion were used for every test image to create their noisy versions, we only show A11 and its distorted versions in Figure 4. The original images  $A_{i1}$ ,  $i = 2, \dots, 9$  can be seen in Figure 5.

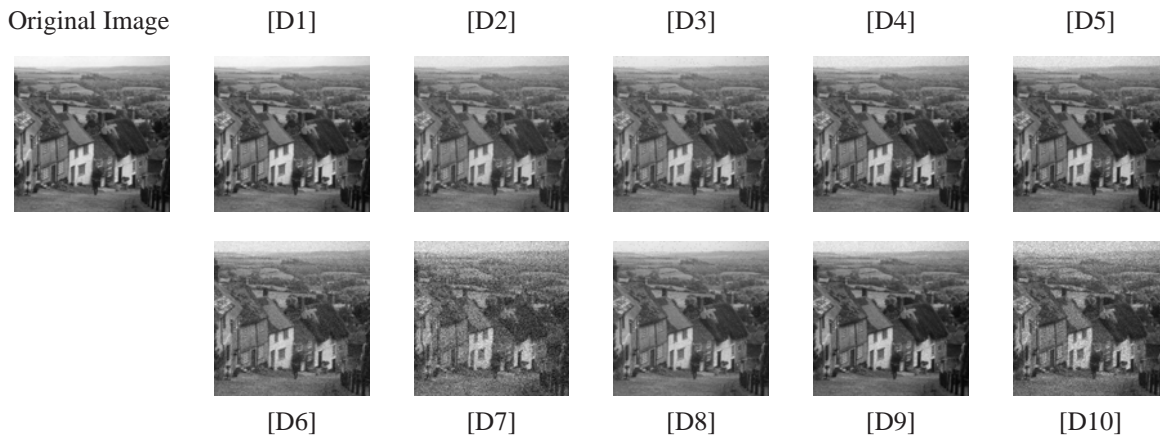


FIG 4. Sample original image and the ten distorted images

For each random noise version, PSSIM performs consistently across all images with the same type of random noise. SSIM, on the other hand, may assign very different similarity scores for different images contaminated with the same level of random noise. Moreover, as the variance of the Gaussian noise changes from 0.0018 in [D2] to 0.01 in [D6] and to 0.068 in [D7], on average, PSSIM decreases from 0.9739 to 0.9281, and then to 0.7755, while SSIM changes dramatically from 0.8822 to 0.6362 and then to 0.3222 for the three noise levels. For the Gaussian noise with variance 0.068 (i.e. [D7]), all SSIM scores are less than 0.39 except for image A21. These scores appear lower than necessary by visually examining the distorted versus the original images. In general, both PSSIM and SSIM have reduced scores with the increased noise level, but SSIM underestimates the structural similarities in such cases.

TABLE 1  
*structural similarity of the distorted versions to their original images*

reference image	Distortion 1		Distortion 2		Distortion 3		Distortion 4		Distortion 5	
	PSSIM	SSIM	PSSIM	SSIM	PSSIM	SSIM	PSSIM	SSIM	PSSIM	SSIM
A11	0.9921	0.9931	0.9796	0.8980	0.9960	0.9153	0.9852	0.9023	0.9828	0.9124
A21	0.8961	0.9926	0.9763	0.9473	0.9961	0.9528	0.9860	0.9510	0.9871	0.9544
A31	0.9861	0.9930	0.9736	0.9048	0.9962	0.9209	0.9890	0.9007	0.9826	0.9062
A41	0.9332	0.9893	0.9754	0.8129	0.9960	0.8604	0.9836	0.8534	0.9834	0.8825
A51	0.9902	0.9923	0.9775	0.8858	0.9963	0.9177	0.9846	0.8717	0.9815	0.8568
A61	0.9915	0.9928	0.9784	0.8933	0.9960	0.9176	0.9874	0.8966	0.9840	0.9040
A71	0.9874	0.9905	0.9744	0.8578	0.9960	0.8881	0.9852	0.8626	0.9856	0.8751
A81	0.9614	0.9765	0.9418	0.8614	0.9959	0.8978	0.9807	0.8793	0.9825	0.8925
A91	0.9914	0.9924	0.9821	0.8936	0.9960	0.9170	0.9862	0.9022	0.9823	0.9114
reference image	Distortion 6		Distortion 7		Distortion 8		Distortion 9		Distortion 10	
	PSSIM	SSIM	PSSIM	SSIM	PSSIM	SSIM	PSSIM	SSIM	PSSIM	SSIM
A11	0.9407	0.6682	0.7778	0.3143	0.9928	0.8605	0.9829	0.8675	0.9212	0.5365
A21	0.9294	0.7896	0.7723	0.4467	0.9927	0.9179	0.9803	0.9279	0.9156	0.6592
A31	0.9367	0.6843	0.7997	0.3437	0.9928	0.8717	0.9815	0.8574	0.9224	0.5125
A41	0.9200	0.4382	0.7463	0.1826	0.9925	0.6362	0.9777	0.7890	0.9151	0.3756
A51	0.9275	0.6760	0.7777	0.3882	0.9927	0.8483	0.9807	0.8089	0.9193	0.5614
A61	0.9383	0.6768	0.7860	0.3470	0.9926	0.8640	0.9807	0.8576	0.9176	0.5401
A71	0.9181	0.6003	0.7570	0.2963	0.9927	0.8144	0.9818	0.8167	0.9229	0.4983
A81	0.8800	0.6073	0.7246	0.2996	0.9925	0.8207	0.9789	0.8396	0.9229	0.5290
A91	0.9343	0.6191	0.7780	0.3165	0.9929	0.7730	0.9818	0.8458	0.9238	0.4995

TABLE 2  
*Similarity of A11 to the other reference images*

A11 vs	A21	A31	A41	A51	A61	A71	A81	A91
PSSIM	0.0687	0.0894	0.1234	0.1102	0.1150	0.0630	0.0886	0.0779
SSIM	0.0982	0.1666	0.2132	0.1519	0.1449	0.1908	0.1606	0.1541

To see whether the low scores assigned by SSIM are reasonable or not, particularly, a score as low as 0.1826 for A41 compared to its Gaussian noise contaminated version (Distortion 7), we also calculated the SSIM and PSSIM based similarities between different natural images and reported them in Table 2. SSIM assigned the score 0.2132 for A11 versus A41. That is, the original image A11 (Goldhill) in Figure 4 and the original image A41 (lady with black hair) in Figure 1 have more SSIM similarity than A11 versus its Gaussian contaminated image [D7]. Therefore, the assignment of similarity scores by SSIM becomes inconsistent due to the limitation of estimated correlation as a measure of dependence in the presence of noise. We remark that PSSIM assigned lower scores than SSIM for all pairs of different images (see Table 2).

#### 4.2. Performance comparisons on filtered images

Image filtering is a process to modify, enhance, warp, or mutilate an image. Noise removal and edge sharpening are two examples of image filtering. Noise removal can be done via image smoothing with commonly used methods such as Fourier transform, wavelet transform, median filtering, Gaussian smoothing, kernel smoothing, etc. Noise removal is basically equivalent to a low-pass filtering that has a typical problem of blurring fine details or edges. Image sharpening aims at enhancing the line structures or other details in an image. The line structures and edges can be obtained, for example, by applying a difference operator equivalent to a high pass filter on the image.

We generated three blurry versions of each test image to investigate the PSSIM performance on filtered images. The three levels of blurring have SSIM scores in the following ranges: High (0.88-0.92), Medium (0.81-0.85) and Low (0.73-0.77). For  $i = 1, 2, \dots, 9$ , the 3 noisy versions of  $Ai1$  are labeled as  $H_i$ ,  $M_i$  and  $L_i$ , respectively. We used the MATLAB function “spfilt” from the DIPUM package available with the book by Gonzalez et al. [2009] to generate these blurred images. In order to adjust to the desired range of SSIM values, we considered several types of filters (median, arithmetic mean ‘amean’, geometric mean ‘gmean’, contraharmonic mean ‘chmean’) and window sizes ( $2 \times 2$  to  $66 \times 66$ ) for different images. Note that the arithmetic mean is a linear filter and all the other filters considered here are nonlinear filters where each filtered pixel  $Y_{ij}$  is a nonlinear function of the original pixel  $X_{ij}$  and its neighbors. Correspondingly, the difference between the reference image and the filtered image is a nonlinear transformation of the reference image except for the arithmetic mean filter. For example, the geometric mean filter is defined as

$$Y_{ij} = \prod_{(i_1, j_1) \in W_{ij}} X_{i_1 j_1}^{1/(st)} \text{ and Error} = X_{ij} - \prod_{(i_1, j_1) \in W_{ij}} X_{i_1 j_1}^{1/(st)},$$

where  $W_{ij}$  is the local window of size  $s \times t$  at position  $(i, j)$ . This filter is better at preserving edge features than the arithmetic mean filter if the same window size is used. Increasing the window size results in more loss of edge details. See Gonzalez and Woods [2002] for further details on spatial filters and in-depth explanations.

The structural similarity values from PSSIM and SSIM are reported in Table 3. Due to different filters and parameters used, we also present all the original images and their blurred versions in Figure 5. Although PSSIM and SSIM have different numerical ranges, the relative orders of the values are identical among different filtered images, except for one case (A21). However, we believe that some of the SSIM scores assigned to the blurred images are higher than natural. To substantiate this point, we include the D6 Gaussian-noise contaminated images in Figure 5 to serve as a reference in the objective, human-eye based, evaluation of the PSSIM and SSIM scores. From the values in Table 3 we see that, as opposed to PSSIM, SSIM may assign unfair scores to images from different categories. Compared to PSSIM, SSIM is less alert to the loss of edge details for all median and arithmetic mean filtered images. This might not be obvious for images with rich texture (e.g., A11 and A51). However, such loss becomes apparent in images such as L4 and L7. Compared to its original image A41, the eyes, nose, mouth profile, as well as eyebrows, are all missing in L4. On the other hand, these characteristics are clearly shown in the D6 Gaussian-noise-contaminated version. Similarly, the Lena image L7 lost important fine details such as the mouth, the expression in the eyes, and the lines of the decorative fixture on the hat. Relatively, the D6 Gaussian-noise-contaminated version contains a lot more of such details. In both examples, SSIM assigned scores of greater than 0.73 for L4 and L7, but less than 0.61 for the D6 Gaussian noise contaminated version of the images. PSSIM assigned reasonably greater scores for the Gaussian noise contaminated images than their corresponding highly blurred versions L4 and L7.

Additionally, the geometric mean with window size  $2 \times 2$  filtered image H2 preserves more edge features than a median filter with window size  $4 \times 4$  (see Gonzalez and Woods [2002]). But SSIM assigns structural similarity of about 0.91 for them. The proposed PSSIM assigns 0.9613 for H2 relative to its reference image. This set of comparisons clearly illustrate the limitations of using a linear measure (correlation) to quantify nonlinear relationships. The proposed PSSIM has more power to provide an objective evaluation of the structural similarity in such cases.

#### 4.3. Differentiating between original and watermarked images

Digital watermarking is a recent technology for image copyright protection. Spatial-domain and frequency-domain watermarking techniques have been considered by various authors. The Discrete Wavelet Transform (DWT) and the Discrete Cosine Transform (DCT) are two of the frequency-domain transforms in digital image watermarking. DWT has been frequently used due to its spatial localization and multi-resolution characteristics. Let  $\lambda$  denote the wavelet decomposition level and let  $LH_\lambda$  and  $HL_\lambda$  denote the middle-frequency sub-bands and  $HH_\lambda$  denote the high-frequency sub-band at level  $\lambda$ , respectively. A DWT algorithm embeds the watermark in the middle-frequency  $LH_\lambda$  and  $HL_\lambda$  or high-frequency  $HH_\lambda$  sub-bands so that acceptable performance of imperceptibility and robustness can be achieved (Hsieh et al. [2001]; Reddy and Chatterji [2005]; Wang and Lin [2004]).

Some authors have considered combined DWT and DCT watermarking (Nikolaidis and Pitas



TABLE 3

The structural similarity of blurry images relative to the reference images assessed by MSE, SSIM, and the proposed PSSIM. The blurry images were obtained through spatial filtering with the `spfilt` command in the MATLAB DIPUM package. The type and size of each spatial filtering are given in the table. The columns labeled as D6, corresponding to Distortion 6 in Table 1, are listed for reference.

Reference	image	A11				A21			
filter	Blurred	H1	M1	L1	D6	H2	M2	L2	D6
	type	median	median	chmean		gmean	gmean	gmean	
	size	(4,4)	(7,7)	(8,8)		(2,2)	(5,5)	(6,6)	
measures	MSE	81.70	115.07	202.37		185.06	115.07	202.38	
	SSIM	.9120	.8246	.7358	.6682	.9195	.8246	.7358	.7896
	PSSIM	.8779	.7451	.5956	.9407	.9613	.8256	.8060	.9294
Reference	image	A31				A41			
filter	blurred	H3	M3	L3	D6	H4	M4	L4	D6
	type	median	median	median		median	median	median	
	size	(4,4)	(6,6)	(8,8)		(14,14)	(32,32)	(66,66)	
measures	MSE	81.70	115.07	202.38		48.34	157.24	438.11	
	SSIM	.8955	.8246	.7358	.6843	.9034	.8194	.7380	.4382
	PSSIM	.8337	.7648	.6948	.9367	.4067	.1936	.1042	.9200
Reference	image	A51				A61			
filter	blurred	H5	M5	L5	D6	H6	M6	L6	D6
	type	amean	amean	gmean		median	median	median	
	size	(5,5)	(7,7)	(9,9)		(3,3)	(5,5)	(10,10)	
measures	MSE	163.92	259.94	389.74		189.52	307.42	352.69	
	SSIM	.9123	.8372	.7583	.6760	.9178	.8111	.7311	.6768
	PSSIM	.8779	.6144	.5327	.9275	.6898	.6411	.5895	.9383
Reference	image	A71				A81			
filter	blurred	H7	M7	L7	D6	H8	M8	L8	D6
	type	median	median	median		amean	amean	amean	
	size	(6,6)	(11,11)	(18,18)		(7,7)	(11,11)	(16,16)	
measures	MSE	85.10	131.10	242.13		117.80	215.49	347.32	
	SSIM	.9136	.8388	.7486	.6003	.9132	.8309	.7345	.6073
	PSSIM	.7253	.5414	.3554	.9181	.6097	.4156	.2655	.8800
Reference	image	A91							
filter	blurred	H9,	M9	L9	D6				
	type	median	gmean	gmean					
	size	(4,4)	(6,6)	(9,9)					
measures	MSE	99.33	166.60	234.68					
	SSIM	.9166	.8344	.7316	.6191				
	PSSIM	.8919	.6900	.5406	.9343				

[2003]; Al-Haj [2007]) with the hope that the combined transforms could compensate for the drawbacks of each other. In the DWT-DCT watermarking algorithm by Al-Haj [2007], watermarking is done by altering the wavelet coefficients of the 2nd level ( $\lambda = 2$ ) DWT sub-bands  $HL_2$  or  $HH_2$ , followed by the application of the DCT transform on the selected sub-bands. The watermarked image is produced by an application of inverse DCT and inverse DWT transform on the DWT-DCT transformed image including the modified sub-bands.

A  $512 \times 512$  ‘Lena’ image was used in Al-Haj [2007] as the original cover host image, and a  $256 \times 256$  grey-scale image of the expression ‘copyright’ was embedded in the host image with DWT only and combined DWT-DCT algorithms on  $HL_2$  or  $HH_2$  sub-bands to produce watermarked images (see top panel of Figure 6). In Al-Haj [2007], the combined DWT-DCT water-

marked image has higher peak signal to noise ratio (PSNR) than the DWT watermarked image.

As pointed out by Wang and Bovik [2009],  $PSNR = 10 \log_{10} L^2 / MSE$ , where  $L$  is the dynamic range of allowable image pixel intensities. The PSNR is useful if images having different dynamic ranges are being compared, otherwise it contains no more information than the mean squared error MSE. Therefore, the drawbacks of MSE as a measure of image structural difference also apply to PSNR. We refer to Wang and Bovik [2009]; Wang et al. [2004] for a detailed illustration of such drawbacks. Since DWT-DCT and DWT are two different image processing methods that would possibly modify different locations of an image, PSNR may not give an objective evaluation of the structural fidelity due to the drawback mentioned in the introduction. Here we evaluate the structural similarity of the aforementioned DWT only and DWT-DCT watermarked images relative to the original host image. The error signals are presented in the second row of Figure 6. They contain both the watermarking information and any structural difference between the watermarked image and the original host image caused by transformation and inverse transformation of DWT and DCT.

Watermarking with combined DWT and DCT may produce smoother images than DWT only. However, as shown in the error signals, the double transformation by combined DWT and DCT also leads to more significant structural difference seen in the (maybe enhanced) edge and line details compared to the image produced with DWT only. Correspondingly, the proposed PSSIM identifies such structural difference and reports that the watermarked images through DWT-DCT have less than 60% of the pixels maintain identical structure as the original host image. On the other hand, SSIM gives a very opposite evaluation yielding that DWT-DCT watermarked images have more structural fidelity to the original image than the DWT watermarked images. This is clearly not supported by the amount of visual structural information contained in the error signals resembling the host image. We remark that the previous comment applies to structural gain or loss for the processed images instead of visual quality. Should smoother images with enhanced edge details be preferred by a user, the DWT-DCT watermarked image could be the better choice.

### Summary

In this work, we proposed a new digital image structural similarity measure based on nonparametric hypothesis testing of independence between the error signal and the two images being compared. The new similarity measure quantifies the structural fidelity of an image to its reference image through the percentage of local windows where the error signal does not contain information about either image scaled by the average of the luminance comparisons. Whether the error signal contains information from the images being compared or not was determined through a nonparametric test. This test is powerful enough to efficiently capture general dependence including nonlinear relationships. We compared the performance of the newly proposed similarity measure PSSIM with commonly used measures with images from three types of image processing techniques: noise contamination, filtering, and watermarking. As a structural similarity measure, the

proposed PSSIM offers an improved performance over SSIM and MSE. Regarding PSSIM's limitations, we mention that the major potential application domain of PSSIM would be the evaluation of image denoising methods. In the context of CBIR, PSSIM as presented is well-suited only when the target images are registered (i.e., they are overlaid in the same coordinate system) which excludes the case of image transformations such as dilations and rotations. In order for PSSIM to handle images involving those transformations, a pre-processing step of registering images would be necessary. Such a pre-processing step is beyond the scope of this work.

### Acknowledgements

The authors would like to thank the anonymous referees for their constructive comments that helped improve the presentation and clarity of the article.

### References

- Al-Haj, A. (2007). Combined DWT-DCT digital image watermarking. *Journal of Computer Science*, 3:740–746.
- Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40:Article 5.
- Do, M. N. and Vetterli, M. (2002). Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Trans. Image Process*, 11:146–158.
- Genest, C. and Rémillard, B. (2004). Tests of independence and randomness based on the empirical copula process. *Test*, 13:335–369.
- Gonzalez, R. and Woods, R. (2002). *Digital Image Processing*. 2nd. Ed. Prentice-Hall.
- Gonzalez, R., Woods, R., and Eddins, S. (2009). *Digital Image Processing Using Matlab*. Gatesmark Publishing.
- Hsieh, M., Tseng, D., and Huang, Y. (2001). Hiding digital watermarks using multiresolution wavelet transform. *IEEE Trans. on Industrial Electronics*, 48:875–882.
- Ko, B. and Byun, H. (2002). Integrated region-based image retrieval using regions spatial relationships. *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*.
- Mathiassen, J. R., Skavhaug, A., and Bo, K. (2002). Texture similarity measure using Kullback-Leibler divergence between gamma distributions. *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Nikolaidis, A. and Pitas, I. (2003). Asymptotically optimal detection for additive watermarking in the DCT and DWT domains. *IEEE Trans. Image Processing*, 2:563–571.
- Reddy, A. and Chatterji, B. (2005). A new wavelet based logo-watermarking scheme. *Pattern Recognition Letters*, 26:1019–1027.
- Sheikh, H., Sabir, M., and Bovik, A. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Processing*, 15(11):3449–3451.
- Smeulders, A., Worring, M., Gupta, S. S. A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell*, 22:1349–1380.

- Terpstra, J. T. and Mckean, J. (2005). Rank-based analyses of linear models using R. *Journal of Statistical Software*, 14:<http://www.jstatsoft.org/>.
- Wang, H., Tolos, S., and Wang, S. (2010). A distribution-free nonparametric test to detect dependence between a response variable and covariate in presence of heteroscedastic treatment effects. *To appear in Canadian Journal of Statistics*.
- Wang, J., Li, J., and Wiederhold, G. (2001). Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23:947–963.
- Wang, S. and Lin, Y. (2004). Wavelet tree quantization for copyright protection watermarking. *IEEE Trans. Image Processing*, 13:154–164.
- Wang, Z. and Bovik, A. (2002). A universal image quality index. *IEEE Signal Processing Lett.*, 9(3):81–84.
- Wang, Z. and Bovik, A. (2006). *Modern Image Quality Assessment*. Morgan & Claypool Publishers.
- Wang, Z. and Bovik, A. (2009). Mean squared error, love it or leave it. *IEEE Signal Processing Magazine*, 26(1):98–117.
- Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612.
- Wood, S. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *J. Roy. Statist. Soc. Ser. B*, 70:495–518.
- Wu, H. and Rao, K. (2005). *Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications)*. CRC Press.

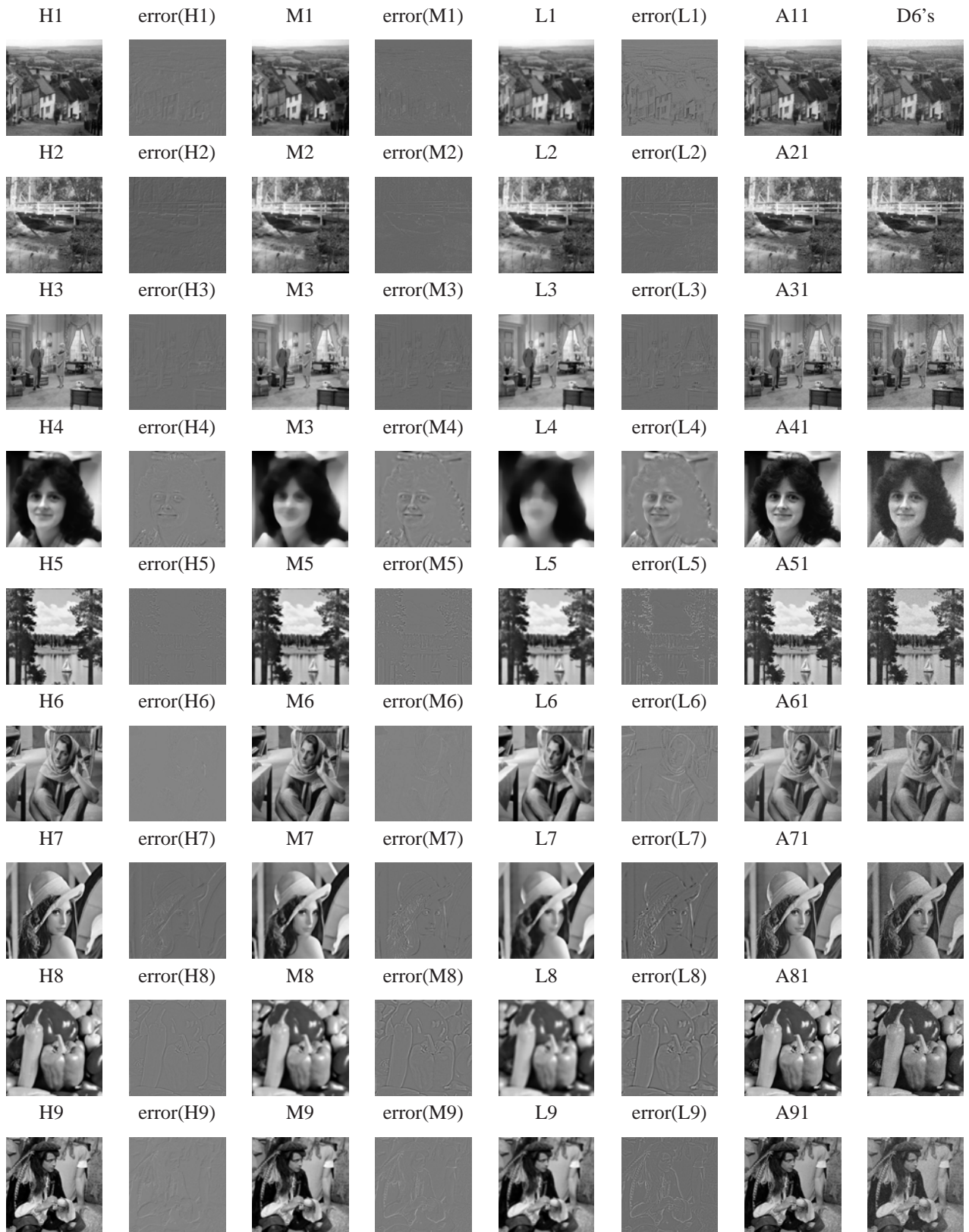


FIG 5. Blurred images with error signals, their original image, and Distortion 6 as an additional reference

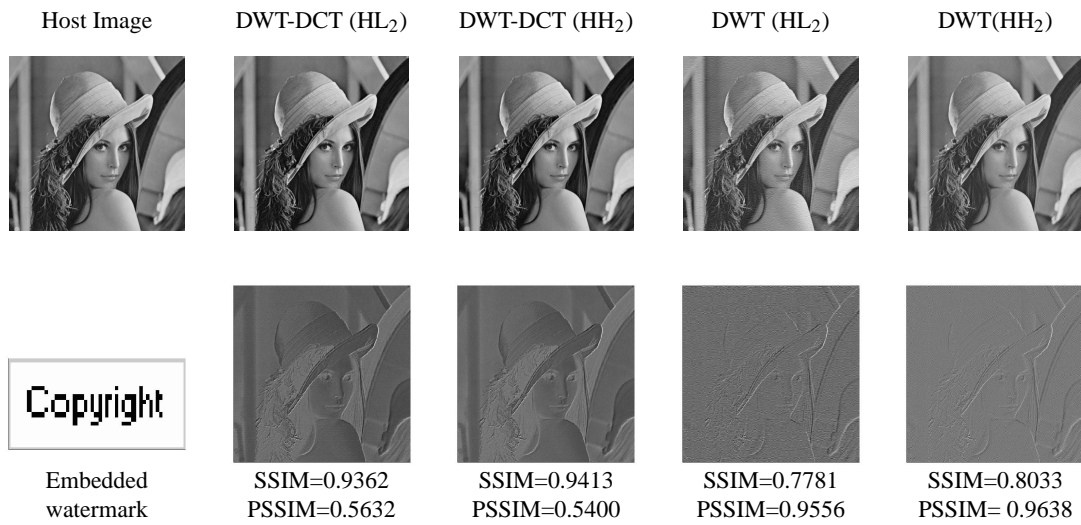


FIG 6. 'Lena' host image and four versions of watermarked images with grey image of the word 'Copyright'. The second row shows the added watermark and error signals of the watermarked images compared to the host image.