Structural Texture Similarity Metrics for Image Analysis and Retrieval

Jana Zujovic, *Member, IEEE,* Thrasyvoulos N. Pappas, *Fellow, IEEE,* and David L. Neuhoff, *Fellow, IEEE*

***Abstract*—We develop new metrics for texture similarity that account for human visual perception and the stochastic nature of textures. The metrics rely entirely on local image statistics and allow substantial point-by-point deviations between textures that according to human judgment are essentially identical. The proposed metrics extend the ideas of structural similarity (SSIM) and are guided by research in texture analysis-synthesis. They are implemented using a steerable filter decomposition and incorporate a concise set of subband statistics, computed globally or in sliding windows. We conduct systematic tests to investigate metric performance in the context of "known-item search," the retrieval of textures that are "identical" to the query texture. This eliminates the need for cumbersome subjective tests, thus enabling comparisons with human performance on a large database. Our experimental results indicate that the proposed metrics outperform PSNR, SSIM and its variations, as well as state-of-the-art texture classification metrics, using standard statistical measures.**

***Index Terms*—natural textures, perceptual quality, statistical models**

I. Introduction

THE development of objective metrics for texture similarity differs from that of traditional image similarity metrics, which are often referred to as quality metrics, because substantial visible point-by-point deviations are possible for textures that according to human judgment are essentially identical. Employing metrics that are insensitive to such deviations is particularly important for natural textures, the stochastic nature of which requires statistical models that incorporate an understanding of human perception. In this paper, we present new *structural texture similarity (STSIM)* metrics for image analysis and content-based retrieval (CBR) applications. We then conduct systematic experiments to evaluate the performance of these metrics and compare to that of existing metrics. For that we focus on a particular CBR application, which facilitates testing on a large texture database, and allows the use of different metric performance statistics, which emphasize different aspects of performance that are relevant for many other image analysis applications.

Traditional similarity metrics evaluate the similarity between two images on a point-by-point basis. Such metrics include mean squared error (MSE) and peak signal-to-noise ratio (PSNR), as well as metrics that make use of explicit low-level models of human perception [1], [2]. The latter are typically implemented in the subband/wavelet domain and are aimed at the threshold of perception, whereby two images, typically an original and a distorted image, are visually indistinguishable. In contrast, our goal is to assess the similarity of two textures, which may have visible point-by-point differences, even though neither one of them appears to be distorted and both could be considered as original images.

The interest in metrics that deviate from point-by-point similarity was stimulated by the introduction of the *structural similarity metrics (SSIM)* [3], a class of metrics that attempt to incorporate "structural" information in image comparisons. Such metrics have been developed in both the space domain (S-SSIM) [3] and the complex wavelet domain (CW-SSIM) [4], and make it possible to assign high similarity scores to pairs of images with significant pixel-wise deviations that do not affect the structure of the image. However, as we discuss below, SSIM metrics still rely on point-by-point cross-correlations between two images or their subbands, and thus retain enough point-by-point sensitivity that they will generally not give high similarity values to textures that are structurally similar. In order to overcome such constraints, Zhao *et al.* [5] proposed a *structural texture similarity metric*, which we will refer to as *STSIM-1*, that relies entirely on local image statistics, and thus completely eliminates point-by-point comparisons; while Zujovic *et al.* [6] included additional statistics to obtain *STSIM-2*. The goal of this paper is to expand on and systematically explore this idea. We present a general framework for STSIMs whose key elements are a multiscale frequency decomposition, a set of subband statistics, formulas for comparing statistics, and pooling to obtain an overall similarity score. An additional goal is to test metric performance on a large database of natural textures.

We develop a number of STSIMs that utilize both intra- and inter-subband correlations, and different ways of comparing statistics. The development of texture similarity metrics has been motivated and guided by recent research in the area of texture analysis and synthesis. Our interest is in texture analysis/synthesis techniques that rely on multiscale frequency

Manuscript received April 24, 2012; revised January 22, 2013; accepted February 13, 2013. Date of publication March 7, 2013. This work was supported in part by the U.S. Department of Energy National Nuclear Security Administration (NNSA) under Grant No. DE-NA0000431. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Erhardt Barth.

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

J. Zujovic was with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL USA. She is now with FutureWei Technologies, Santa Clara, CA 95050 USA (phone: 408-330-4736, e-mail:jana.ehmann@huawei.com).

T. N. Pappas is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: pappas@eecs.northwestern.edu).

D. L. Neuhoff is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail:neuhoff@umich.edu).

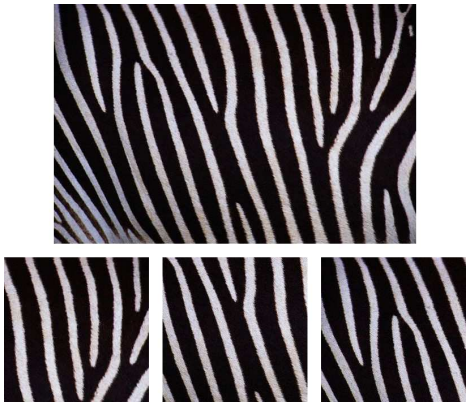


Fig. 1. Constructing a database of identical textures: original and cutouts

decompositions [7]–[11]. The most impressive and complete results were presented by Portilla and Simoncelli [11], who developed a technique based on an elaborate statistical model for texture images that is consistent with human perception. It is based on the steerable filter decomposition [12] and relies on a model with several hundred parameters to capture a very wide class of textures. While in principle a direct comparison of the model parameters can form the basis for a texture similarity metric, our goal is to show that a successful similarity metric can be based on significantly fewer parameters.

A number of applications can make use of STSIMs, and each application imposes different requirements on metric performance and testing procedures. For example, in image compression it is important that the metric exhibit a monotonic relationship between measured and perceived distortion, while in image retrieval applications it may be sufficient for the metric to distinguish between similar and dissimilar images without a need for precise ordering. The focus of this paper is on CBR, and in particular, on the recovery of textures that are *identical* to a query texture, in the sense that they could be patches from a large perceptually uniform texture, as shown in Figure 1. However, these metrics have also been used in image compression [13]. Note that the patches at the bottom of Fig. 1 have visible point-by-point differences, but to a human observer there is no doubt that they are the same texture. The zebra example was chosen to emphasize the point; typical textures are not as coarse as this. Retrieval of identical textures is important in CBR when one may be seeking images that contain a particular texture (material, fabric, pattern, etc.), as well as in some near-threshold coding applications. The problem of searching for a known target image in a database has been extensively studied by the text retrieval community and is referred to as *known-item search* [14]. It has also been addressed by the image processing community for texture retrieval applications [15]–[17].

The evaluation of image similarity metrics, in general, requires extensive subjective tests, with several human subjects and a large number of image pairs. It also requires appropriate statistical measures of performance. Depending on the performance requirements, a number of traditional statistical measures can be used. For example, Spearman's rank correlation coefficient and Kendall's tau rank correlation coefficient can be used when a monotonic relationship between subjective similarity scores and metric values is desired [6], while Pearson's correlation coefficient can be used when a linear relationship is important [18]. In [5], the performance criterion was whether a metric can distinguish between similar and dissimilar pairs, irrespective of the ordering within each group. This idea was further explored in [19], where we argued that the combination of testing procedure and statistical performance measure is critical for obtaining meaningful results.

The advantage of evaluating metric performance in the context of retrieving identical textures is that the ground truth is known, and therefore no subjective tests are required. Of course, the ground truth is known to the extent that the texture from which the identical patches are obtained is perceptually uniform. Another advantage of evaluating a metric in this context is the availability of a number of well-established statistical performance measures, which include *precision at one* (measures in how many cases the first retrieved document is relevant), *mean reciprocal rank* (measures how far away from the first retrieved document is the first relevant one), *mean average precision*, and *receiver operating characteristics.*

In evaluating the similarity of two textures, one has to take into account both the color composition and the spatial texture patterns. In [6] we proposed a new structural texture similarity metric that separates the computation of similarity in terms of grayscale texture and color composition, and then combines them into a single metric. However, our subjective tests indicate that the two attributes are quite separate and that there are considerable inconsistencies in the weights that human subjects give to the two components [6], [19]. Thus, for the present study, we focus only on grayscale textures. We present a general framework for STSIMs that includes the metrics proposed in [5] and [6], as well as a new metric that relies on the Mahalanobis distance between vectors of subband statistics (*STSIM-M*).

Initial experiments with STSIM-2 were performed on a database of 748 natural textures [20]. In this paper, we present experimental results with two databases with a total of 1363 distinct texture images, extracted from 486 larger texture images. Our results indicate that the proposed metrics substantially outperform existing metrics in the retrieval of identical textures, according to all of the standard statistical measures mentioned above, each of which emphasizes different aspects of metric performance.

The paper is organized as follows. Section II reviews grayscale texture similarity metrics, including SSIM metrics. The proposed STSIM metrics are discussed in Section III. Section IV presents the experimental results. Our conclusions are summarized in Section V.

II. Review of Grayscale Similarity Metrics

In this section, we review grayscale image similarity metrics and discuss their applicability to texture images. Such metrics can easily be extended to color by applying the grayscale metric to each of three color components in a trichromatic space, as is sometimes done in compression applications.

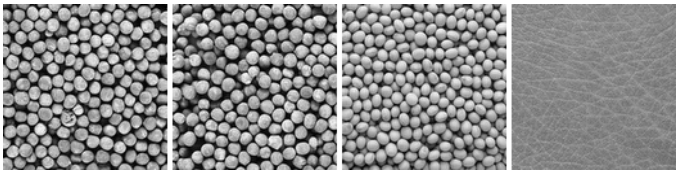


Fig. 2. Illustration of inadequacy of PSNR for texture similarity. Subjective similarity increases left to right, while PSNR indicates the opposite.

However, as we have argued in [6], [21], it is more effective to decouple the grayscale and color composition similarity of an image. So, we restrict our discussion – and this paper – to the grayscale case.

Image similarity metrics can be broadly grouped into two categories: (1) image quality or fidelity metrics that attempt to quantify the (ideally perceptual) difference between an original and a distorted image, and (2) image similarity metrics that compare two images without any judgment about quality. The former are aimed at image compression and the latter at CBR applications. The texture similarity metrics we propose in this paper fall somewhere between these two categories, and are intended for both applications, even though the focus of our experimental results will be on their retrieval abilities. Note that most of the metrics we discuss do not meet the formal definition of a metric, but we will refer to them as metrics anyway.

Another broad categorization of grayscale texture similarity metrics is into statistical and spectral metrics [22], [23]. Spectral analysis (subband decomposition) is essential if a metric is going to emulate human perception, while statistical analysis is necessary for embodying the stochastic nature of textures. It should thus not be surprising that the best metrics combine both attributes.

A. Point-by-point Similarity Metrics

Traditional metrics evaluate image similarity on a point-by-point basis, and range from simple mean squared error (MSE) and peak signal-to-noise ratio (PSNR) to more sophisticated metrics that incorporate low-level models of human perception [1], [2]; we will refer to the latter as *perceptual quality metrics.* Note that even though the former are implemented in the image domain and the latter in the subband domain, in both cases the computation is done on a point-by-point basis. Figure 2 illustrates the failure of point-by-point metrics when evaluating texture similarity. Note that PSNR decreases with increasing texture similarity.

Note also that perceptual quality metrics that are aimed at near-threshold applications, whereby the original and reconstructed images are perceptually indistinguishable, are very sensitive to any image deviations that can be detected by the eye, as for example when comparing the identical texture patches of Fig. 1 and the two textures on the left of Fig. 2.

B. Texture Similarity Metrics

As we mentioned above, image similarity metrics can be grouped into statistical and spectral methods. The statistical methods are based on calculating statistics of the gray levels in the neighborhood of each pixel (co-occurrence matrices, first and second order statistics, random field models, etc.) and then comparing the statistics of one image to those of another, while the spectral methods utilize the Fourier spectrum or a subband decomposition to characterize and compare textures.

We review statistical methods first. One of the best-known methods is based on co-occurrence matrices [24]–[26], which rely on relationships between the gray values of adjacent pixels, typically within a 2×2 neighborhood. However, given the small size of the neighborhood, such methods are not well-suited for computing similarity of textures other than the so-called microtextures [27].

Another approach is to rely on first and second order statistics. Chen *et al.* [28] used the local correlation coefficients for texture segmentation applications. However, as Julesz *et al.* [29], [30] have shown, humans can easily discriminate some textures that have the same second-order statistics. Thus, simple second order statistics of image pixels are not adequate for the evaluation of perceptual texture similarity.

Another class of statistical methods rely on Markov random fields (MRF) to model the distribution of pixel values in a texture [31], [32]. In combination with filtering theory, the MRF models can also be used for texture synthesis [33]. The main drawback of MRF-based approaches is that MRFs can only model a subset of textures.

Ojala *et al.* [16] utilize local binary patterns (LBP) to characterize textures, mainly for retrieval applications. Their method constructs binary patterns that describe the relative value of a pixel to image values in circles of different radii. It then constructs histograms of such patterns for each circle, on the basis of which it computes a log-likelihood statistic that two images come from the same class. This method is very simple yet effective for the task of texture classification. However, as we show in Section IV, it does not provide metric values that are comparable across different texture content.

The main advantage of these statistical approaches is their simplicity and computational efficiency for obtaining the texture features and carrying out comparisons. However, their simplicity is also their main drawback, as is their failure to incorporate models of human perception. Most of these methods have been applied to limited data sets and applications, and are likely to fail in more general problem settings.

The spectral methods provide a better link between pixel image representations and human perception. Initially, spectral methods were based on the Fourier transform, but given that the basis functions for Fourier analysis do not provide efficient localization of texture features [34], they were quickly replaced by wavelet/subband analysis methods, which provide a better tradeoff between spatial and frequency resolution.

Most of the recent spectral techniques extract the energies of different subbands, and use them as features for texture segmentation, classification, and CBR [27], [35]–[38]. One of the most effective classification techniques has been proposed by Do and Vetterli [38]; they use wavelet coefficients as features and show that their distribution can be modeled as a generalized Gaussian density, which requires the estimation of two parameters. They then base the classification on the

Kullback-Leibler distance between two feature vectors.

Some spectral techniques rely on subband decompositions (filter banks) that explicitly model early processing stages of the human visual system (HVS). In addition to different spatial frequency channels, such decompositions are orientation-sensitive, mimicking the orientation selectivity of simple receptive fields in the visual cortex of the higher vertebrates [39]. One example of such decompositions are Gabor filters [40], [41]. Several authors have used features extracted from such decompositions for a variety of applications (e.g., in [8], [35], [36], [42]–[45]). Manjunath and Ma [36] have utilized the mean and the standard deviation of the magnitude of the transform coefficients as features for representing the textures for classification and retrieval applications. Then, a measure of dissimilarity between two texture images is the normalized ℓ^1 distance between their respective two feature vectors.

Some methods for evaluating texture similarity combine the statistical and the spectral approaches. For example, Yang *et al.* [46] combine Gabor features and co-occurrence matrices for CBR applications. One of the MPEG-7 texture descriptors [47], the *homogeneous texture descriptor* also combines spectral and statistical techniques. It consists of the means and variances of the absolute values of the Gabor coefficients. Since these statistics are computed over the entire image, this descriptor is useful in characterizing images that contain homogeneous texture patterns. For non-homogeneous textures, the *edge histogram descriptor* partitions the image into 16 blocks, applies edge detection algorithms and computes local edge histograms for different edge directions. The *texture browsing descriptor*, attempts to capture higher-level perceptual attributes such as regularity, directionality, and coarseness, and is useful for crude classification of textures. These three types of MPEG-7 texture descriptors of MPEG-7 are described in detail in [48]. Ojala *et al.* [16] have shown that the MPEG-7 descriptors are rather limited and provide only crude texture retrieval results. A number of variations of the MPEG-7 techniques have also been developed, e.g., in [49].

Some of the techniques we have reviewed in this section have been shown to be quite effective in evaluating texture similarity in the context of clustering and segmentation tasks. However, there has been very little work towards evaluating their effectiveness in providing texture similarity scores that are consistent across texture content, agree with human judgments of texture similarity, and can be used in different applications. In Section III, we proposed metrics that attempt to achieve these goals, while in Section IV, we present systematic methods for evaluating metric performance.

C. Structural Similarity Metrics

For supra-threshold applications, such as CBR and perceptually lossy compression, there is a need for metrics that can accommodate, i.e., give high similarity scores to, significant (visible) point-by-point differences as long as the overall quality and structure does not change from one image to the other. This was the primary motivation in the development of the SSIMs [3], a class of metrics that attempt to – implicitly – incorporate high-level properties of the HVS. The goal is to allow non-structural contrast and intensity changes, as well as small translations, rotations, and scaling changes, that are detectable but do not affect the perceived quality of an image. The main approach for accomplishing this goal is to compare local image statistics in corresponding sliding windows (for example, 7×7) in the two images and to pool the results of such comparisons. SSIMs can be applied in either the spatial or transform domain. When implemented in the image domain, the SSIM metric is invariant to luminance and contrast changes, but is sensitive to image translation, scaling, and rotation, as shown in [4]. When implemented in the complex wavelet domain, it is tolerant of small spatial shifts up to a few pixels, and consequently also small rotations or zoom [4].

The remainder of this subsection provides a brief review of SSIM in the spatial domain (S-SSIM) [3] and the complex wavelet domain (CW-SSIM) [4]. The main difference between the two implementations is that the former is applied directly to two images, $\mathbf{x} = [x(i,j)]$ and $\mathbf{y} = [y(i,j)]$, whose similarity we wish to assess, while in the latter the images are first decomposed into subbands, $\mathbf{x}^m = [x^m(i,j)]$ and $\mathbf{y}^m = [y^m(i,j)]$, using the complex steerable filter bank [12], and includes an extra subband pooling step. Otherwise, the two implementations are the same.

The SSIM metric can thus be applied to the images $\mathbf{x}$ and $\mathbf{y}$ or the subband images $\mathbf{x}^m$ and $\mathbf{y}^m$. The two cases are differentiated by the presence of m. SSIM fixes a window size and shape (usually square), as well as a set of window positions within the images (typically increments of some sliding stepsize such as the window width). Then for each window position, it performs the following three steps.

First, it computes the mean and variance for each image within that window. For example, for $\mathbf{x}^m$, these are

$$\mu_{\mathbf{x}}^m = \mathbb{E}\left\{x^m(i,j)\right\} \tag{1}$$

$$(\sigma_{\mathbf{x}}^m)^2 = \mathbb{E}\left\{\left[x^m(i,j) - \mu_{\mathbf{x}}^m\right]\left[x^m(i,j) - \mu_{\mathbf{x}}^m\right]^*\right\}. \tag{2}$$

where, although the notation does not show it, $\mathbf{x}^m$ refers to the portion of the image within the current window, and where $\mathbb{E}\left\{x^m(i,j)\right\}$ denotes the empirical average of $\mathbf{x}^m$ over spatial locations (i,j) within the window. SSIM also computes the covariance of $\mathbf{x}^m$ and $\mathbf{y}^m$ within corresponding windows:

$$\sigma_{\mathbf{xy}}^m = \mathbb{E}\left\{\left[x^m(i,j) - \mu_{\mathbf{x}}^m\right]\left[y^m(i,j) - \mu_{\mathbf{y}}^m\right]^*\right\}. \tag{3}$$

Second, it compares the corresponding means and variances for the given window position by computing the *luminance* term:

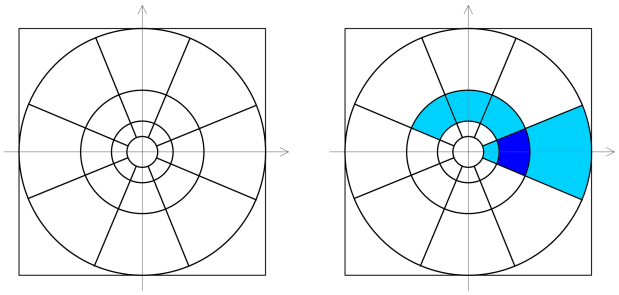
$$l_{\mathbf{x},\mathbf{y}}^m = \frac{2\mu_{\mathbf{x}}^m\mu_{\mathbf{y}}^m + C_0}{(\mu_{\mathbf{x}}^m)^2 + (\mu_{\mathbf{y}}^m)^2 + C_0}, \tag{4}$$

and the *contrast* term:

$$c_{\mathbf{x},\mathbf{y}}^m = \frac{2\sigma_{\mathbf{x}}^m\sigma_{\mathbf{y}}^m + C_1}{(\sigma_{\mathbf{x}}^m)^2 + (\sigma_{\mathbf{y}}^m)^2 + C_1}, \tag{5}$$

where C_0 and C_1 are small positive constants that are included so and that when the statistics are small the term will be close to 1. In addition, the covariance and variances for the window position are used to determine the *structure* term:

$$s_{\mathbf{x},\mathbf{y}}^m = \frac{\sigma_{\mathbf{xy}}^m + C_2}{\sigma_{\mathbf{x}}^m\sigma_{\mathbf{y}}^m + C_2}. \tag{6}$$



(a) (b)

Fig. 3. (a) Steerable filter decomposition. (b) Crossband correlations

which, apart from the small constant C_2, is the cross-correlation coefficient of the two patches.

Finally, it combines these three terms into the similarity value

$$q^m_{\text{SSIM}}(\mathbf{x},\mathbf{y}) = (l^m_{\mathbf{x},\mathbf{y}})^\alpha (c^m_{\mathbf{x},\mathbf{y}})^\beta (s^m_{\mathbf{x},\mathbf{y}})^\gamma, \tag{7}$$

for some choice of positive numbers α, β, and γ, typically all set to 1. Note that CW-SSIM assumes that $\mu^m_{\mathbf{x}} = 0$ for all subbands except the lowpass; it also uses the magnitude of $\sigma^m_{\mathbf{xy}}$ to make sure that all the terms in (7) are real. Note also that all of these terms take values in the interval $[0, 1]$, except for the "structure" term of S-SSIM, which takes values in $[-1, 1]$. The similarity values computed for all window positions are then pooled by averaging to obtain the SSIM value $Q^m_{\text{SSIM}}(\mathbf{x},\mathbf{y})$ over all spatial locations.

In the complex wavelet version of SSIM (CW-SSIM) [4], the images $\mathbf{x}$ and $\mathbf{y}$ are first decomposed into $N_b = N_s \cdot N_o + 2$ subbands using the complex steerable filter bank [12]. Here, N_s denotes the number of scales, N_o the number of orientations, and the "$+2$" accounts for the (innermost) *lowpass* and outermost *highpass* bands, which are not subdivided into different orientations and which have real-valued coefficients, in contrast to the complex coefficients of the $N_s \cdot N_o$ other bands. Figure 3(a) illustrates the passbands of the steerable filter decomposition with $N_s = 3$ scales and $N_o = 4$ orientations. The similarity values are computed for each subband as in (7) and then pooled across subbands, typically by averaging.

Note that SSIM metrics incorporate implicit contrast masking – as opposed to explicit contrast masking in perceptual quality metrics – as the luminance (4) and contrast (5) terms are scaled by the values of the mean and variance, respectively, and are thus weighted by how visible they are. On the other hand, subband noise sensitivities – for a given display resolution and viewing distance – are not implicit but can be easily incorporated into the CW-SSIM metric to obtain a perceptually weighted metric (WCW-SSIM) [50]. Such perceptual weighting is useful for measuring distortions that are dependent on viewing distance, such as white noise and DCT compression [50].

III. Structural Texture Similarity Metrics

As mentioned in Section II, spectral (subband) analysis is needed to model early processing in the HVS, while statistical analysis is necessitated by the stochastic nature of textures. The steerable filter SSIM implementations [4] seem to provide the right ingredients for a perceptual approach to texture similarity. First, steerable filters, like Gabor filters, are inspired by biological visual processing. Second, the most important idea behind the SSIM approach [3] for image quality is the fact that it replaces point-by-point comparisons with comparisons of region statistics. However, the "structure" term of (6), which gives SSIM its name, is actually a point-by-point comparison. This follows from the fact that the cross-correlation between the patches of two images in (3) is computed on a point-by-point basis. Moreover, Reibman and Poole [51] have shown that the image domain SSIM has a direct connection to MSE. This does not hold for CW-SSIM, which is tolerant of small shifts since such perturbations produce consistent phase shifts of the transform coefficients, and thus do not change the relative phase patterns that characterize local image features [4]. However, the amount of shifts the CW-SSIM can tolerate is small and independent of metric parameters. On the other hand, pairs of texture images can have large point-by-point differences and pixel shifts, while still preserving a high degree of similarity.

Thus, in order to fully embrace the SSIM idea of relying on local image statistics, and to develop a metric that can address the peculiarities of the texture similarity problem, we need to completely eliminate point-by-point comparisons by dropping the "structure" term, and to replace it with additional statistics, and comparisons thereof, that reflect the most discriminating texture characteristics. This paper proposes a general framework for STSIM metrics that take the following form:

1) A ***multiscale frequency decomposition:*** Such decompositions can be real or complex. In the following, we will use the three-scale, four-orientation steerable filter decomposition of Figure 3(a) – as in CW-SSIM.
2) A number of ***subband statistics:*** Each statistic corresponds to one image and is computed within one window in that image. Statistics are computed within a subband or across subbands.
3) The window over which the statistics are computed can be ***local*** (sliding window) or ***global*** (the entire image).
4) A means for ***comparing (corresponding) subband statistics***, one from each image whose similarity we wish to assess: The particular formula depends on the range of values that the statistic takes, and yields a nonnegative number that represents the similarity or dissimilarity of the two statistics.
5) Three types of ***pooling*** to obtain an overall (dis)similarity score: One that combines (dis)similarity scores for all statistics that correspond to a given subband, one that pools across subbands, and one that pools across window positions. As we will see, pooling can be done additively or multiplicatively. The order of the pooling can be selected to provide similarity scores for a particular subband or window location.

Note that the "structure" term of SSIM does not fit the above description, because the statistic it computes involves two images, and because it is not a comparison of two statistics.

Moreover, in S-SSIM, it can take negative values.

We now discuss different choices for each of these elements that result in different metric embodiments. Note that all of the metric embodiments we discuss are not scale or rotation invariant. However, if required by an application, they can be modified to account for such invariances, in combination with a scale or orientation detector.

A. STSIM-1

The first structural texture similarity metric was proposed by Zhao *et al.* [5], who replaced the "structure" term of (6) in the CW-SSIM with terms that compare first-order autocorrelations of neighboring subband coefficients in order to provide additional structural and directionality information. We refer to this metric as STSIM-1.

The first-order autocorrelation coefficients can be computed as empirical averages, in the horizontal direction as

$$\rho_{\mathbf{x}}^{m}(0,1)=\frac{\mathbb{E}\left\{[\mathbf{x}^{m}(i,j)-\mu_{x}^{m}]\,[\mathbf{x}^{m}(i,j+1)-\mu_{x}^{m}]^{*}\right\}}{(\sigma_{x}^{m})^{2}} \quad (8)$$

and in the vertical direction as

$$\rho_{\mathbf{x}}^{m}(1,0)=\frac{\mathbb{E}\left\{[\mathbf{x}^{m}(i,j)-\mu_{x}^{m}]\,[\mathbf{x}^{m}(i+1,j)-\mu_{x}^{m}]^{*}\right\}}{(\sigma_{x}^{m})^{2}} \quad (9)$$

Diagonal and anti-diagonal terms could be computed in a similar fashion. However, STSIM-1 did not use them because they did not contribute to any significant improvements in metric performance.

Note that there is no need to consider adding higher order autocorrelations, because this would be equivalent to computing first-order autocorrelations of decimated images. However, this is effectively done when we compute the first-order autocorrelations of the lower frequency subbands, which are lowpass filtered and decimated, which (lowpass filtering) also eliminates aliasing. Thus, by computing first-order autocorrelations on a multi-scale frequency decomposition, we are effectively computing higher-order autocorrelations.

Note also that in contrast to the variances, which are unbounded and nonnegative, the correlation coefficients are bounded and their values lie in the unit circle of the complex plane. Thus, the statistic comparison terms cannot take the form of (4) and (5). Hence, new terms were suggested in [5]:

$$c_{\mathbf{x},\mathbf{y}}^{m}(0,1)=1-0.5|\rho_{\mathbf{x}}^{m}(0,1)-\rho_{\mathbf{y}}^{m}(0,1)|^{p} \quad (10)$$

$$c_{\mathbf{x},\mathbf{y}}^{m}(1,0)=1-0.5|\rho_{\mathbf{x}}^{m}(1,0)-\rho_{\mathbf{y}}^{m}(1,0)|^{p}. \quad (11)$$

We will refer to these as *correlation terms.* Typically, $p=1$.

Note that the means, variances, and autocorrelations are calculated on the *raw*, complex subband coefficients. Since the subband decomposition (apart from the lowpass subband) does not include the origin of the frequency plane, the subbands will ordinarily have *zero-mean* over the *entire* image; however, within small windows, e.g., of size 7×7, this does not have to be true; thus, the means $\mu_{\mathbf{x}}^{m}$ have to be computed in each sliding window, and used in the variance calculations.

For each window, the similarity scores corresponding to the four statistics are combined into one score for each subband and window location:

$$q_{\text{STSIM-1}}^{m}(\mathbf{x},\mathbf{y})=(l_{\mathbf{x},\mathbf{y}}^{m})^{\frac{1}{4}}(c_{\mathbf{x},\mathbf{y}}^{m})^{\frac{1}{4}}(c_{\mathbf{x},\mathbf{y}}^{m}(0,1))^{\frac{1}{4}}(c_{\mathbf{x},\mathbf{y}}^{m}(1,0))^{\frac{1}{4}} \quad (12)$$

Note that the exponents were selected to sum to 1 in order to normalize the metric values so that metrics with different numbers of terms are comparable [5]. The overall metric value is obtained by pooling over all subbands and spatial locations.

For spatial pooling, Zhao *et al.* [5] considered two approaches. In the "additive" approach, the metric values are averaged across all subbands. In the "multiplicative" approach, the metric values are multiplied across the subbands. In both cases, the final metric is calculated as the spatial average over all the sliding window locations.

In [5], the STSIM-1 was shown to outperform SSIM and CW-SSIM, in the sense that it provides texture similarities that are closer to human judgments.

B. Selection of Subband Statistics

In the remainder of this section, we develop metrics that extend the ideas of [5] by including a broader set of image statistics. The motivation comes from the work of Portilla and Simoncelli on texture analysis/synthesis [11], who have shown that a broad class of textures can be synthesized using a set of statistics that characterize the coefficients of a multiscale frequency decomposition (steerable filters). Based on extensive experimentation, they claim that the set of statistics they proposed are necessary and sufficient. Now, if a set of statistics is good for texture synthesis, then these statistics should also be suitable as features for texture comparisons. However, while texture synthesis requires several hundred parameters, we believe that many fewer will suffice for texture similarity.

Among the various statistics that Portilla and Simoncelli proposed, the proposed metrics adopt the mean and variance of the original SSIM metrics, the correlations coefficients of the STSIM-1 metric, and add *crossband* correlations (between subbands). The argument for adding crossband correlations lies in the fact that the image representation by steerable filter decomposition is overcomplete, and thus, the subband coefficients are correlated. More importantly, we compute the crossband-correlation statistics on the *magnitudes* of the coefficients. The raw complex coefficients may in fact be uncorrelated, since phase information can lead to cancellations. As shown by Simoncelli [52], the magnitudes of the wavelet coefficients are *not* statistically independent and large magnitudes in subbands of natural images tend to occur at the same spatial locations in subbands at adjacent scales and orientations. The intuitive explanation may be that the "visual" features of natural images do give rise to large local neighborhood spatial correlations, as well as large scale and orientation correlations [11].

The crossband-correlation coefficient between subbands m and n (excluding the lowpass and highpass bands) is computed as:

$$\rho_{|\mathbf{x}|}^{m,n}(0,0)=\frac{\mathbb{E}\left\{\left[|\mathbf{x}^{m}(i,j)|-\mu_{|\mathbf{x}|}^{m}\right]\left[|\mathbf{x}^{n}(i,j)|-\mu_{|\mathbf{x}|}^{n}\right]\right\}}{\sigma_{|\mathbf{x}|}^{m}\sigma_{|\mathbf{x}|}^{n}} \quad (13)$$

Among all subband combinations, we have decided to include the correlations between subbands at adjacent scales for a given orientation and between all orientations for a given scale; an example is shown in Figure 3(b). This is in agreement with the findings of Hubel and Wiesel [53] that spatially close simple cells in the primary visual cortex exhibit amplification of the responses of cells whose preferred orientations are similar. Note that for computing crossband correlations, it is important that the subbands have the same sampling rates (number of coefficients); this can be achieved if in the steerable filter decomposition we just filter without subsampling. However, all other statistics can be computed using subsampled coefficients, provided they are normalized for pooling.

The total number of crossband correlations is equal to:

$$N_c = N_s \cdot \binom{N_o}{2} + N_o \cdot (N_s - 1),$$

where the first term comes from the correlations across all possible orientation combinations for a given scale and the second term comes from the correlations of adjacent scales for a given orientation. (Note that each subband in the first and last scales has only one adjacent subband.) Thus, if we use a steerable filter decomposition with $N_s = 3$ scales and $N_o = 4$ orientations, as shown in Figure 3(b), there are 26 new subband statistics. Overall, the proposed STSIM metrics incorporate the following statistics, which are computed over the complex subband coefficients or their magnitudes. For each of the N_b subbands we compute:

- mean value $|\mu_{\mathbf{x}}^m|$ (to make it real) or $\mu_{|\mathbf{x}|}^m$,
- variance $(\sigma_{\mathbf{x}}^m)^2$ or $(\sigma_{|\mathbf{x}|}^m)^2$,
- horizontal autocorrelation $\rho_{\mathbf{x}}^m(0,1)$ or $\rho_{|\mathbf{x}|}^m(0,1)$,
- vertical autocorrelation $\rho_{\mathbf{x}}^m(1,0)$ or $\rho_{|\mathbf{x}|}^m(1,0)$,

and for each of the N_c pairs of subbands we compute

- crossband correlation $\rho_{|\mathbf{x}|}^{m,n}(0,0)$.

for a total of $N_p = 4 \cdot N_b + N_c = 82$ statistics.

C. Local Versus Global Processing

In SSIM, CW-SSIM, and STSIM-1 the processing is done on a sliding window basis. This is essential when comparing two images for compression and image quality applications, where we want to ignore point-by-point differences, but want to make sure that local variations on the scale of the window size are penalized by the metric. Note that the window size determines the texture scale that is relevant to our problem. Thus, if the window is large enough to include several repetitions of the basic pattern of the texture, e.g., several peas, then the peas are treated as a texture; otherwise, the metric will focus on the surface texture of the individual peas. On the other hand, when the goal is overall similarity of two texture patches, then the assumption is that they constitute uniform (homogeneous) textures and the global window produces more robust statistics, unaffected by local variations. Thus, in the following, we will consider both global and local metric implementations (for all metrics except the SSIM, for which the global implementation does not provide much information [3], [50]).

D. Complex Versus Real Steerable Filter Decomposition

The complex steerable filters decompose the real image $\mathbf{x}$ into complex subbands $\mathbf{x}^m$. The real and the imaginary parts of such subbands are not independent of each other, in fact, the imaginary part is the Hilbert transform of the real part, that is, they are in quadrature. Quadrature filters are used for envelope detection and for local feature extraction in images. By applying filters in quadrature, we are able to capture the local phase information, which is consistent with receptive field properties of neurons in mammalian primary visual cortex [54].

However, Aach *et al.* [55] have shown that the spectral energy signatures from the subbands obtained with quadrature filters are linearly related to the energies obtained by the "texture energy transform," which performs local variance estimation on the image filtered with the in-phase filter. This is true when we perform the calculations over the windows that are the same size as the filter support. Thus, the same performance is expected when using either complex or real steerable pyramids when a global window is applied. For a local window, which may be different than the filter support, the conclusions from [55] no longer hold and the complex transform is favorable, given its invariance to small rotations, translations and scaling changes, as shown by Wang *et al.* [4].

E. Comparing Subband Statistics and Pooling – STSIM-2

We are now ready to define STSIM-2, a metric that incorporates the statistics we defined in Section III-B. The metric will use the mean value $|\mu_{\mathbf{x}}^m|$, variance $(\sigma_{\mathbf{x}}^m)^2$, and autocorrelations $\rho_{\mathbf{x}}^m(0,1)$ and $\rho_{\mathbf{x}}^m(1,0)$, computed on the complex subband coefficients, and the crossband correlation $\rho_{|\mathbf{x}|}^{m,n}(0,0)$, defined on the magnitudes. If we adopt the SSIM approach for comparing image statistics, all we need to do is add a term for comparing the crossband-correlation coefficients to the STSIM-1 metric. Like the STSIM-1 comparison terms in (10) and (11), this term should take into account the range of the statistic values and should also produce a number in the interval $[0,1]$:

$$c_{\mathbf{x},\mathbf{y}}^{m,n}(0,0) = 1 - 0.5|\rho_{|\mathbf{x}|}^{m,n}(0,0) - \rho_{|\mathbf{y}|}^{m,n}(0,0)|^p \qquad (14)$$

Again, typically, $p = 1$.

Note that since the crossband correlation comparison terms involve two subbands, it does not make sense to multiply them with the other STSIM-1 terms in (12). We thus need a separate term. For a given window, the overall STSIM-2 metric can then be obtained as a sum of two terms: one that combines the STSIM-1 values over all subbands, and one that combines all the crossband correlations.

$$q_{\text{STSIM-2}}(\mathbf{x},\mathbf{y}) = \frac{\sum_{m=1}^{N_b} q_{\text{STSIM-1}}^m(\mathbf{x},\mathbf{y}) + \sum_{i=1}^{N_c} c_{\mathbf{x},\mathbf{y}}^{m_i,n_i}(0,0)}{N_b + N_c}. \qquad (15)$$

When the metric is applied on a sliding window basis, spatial pooling is needed to obtain an overall metric value $Q_{\text{STSIM-2}}(\mathbf{x},\mathbf{y})$. As we saw above, spatial pooling can be done before or after the summation in (15).

F. Comparing Subband Statistics and Pooling – STSIM-M

Another approach for comparing image statistics, that is better suited for comparing entire images or relatively large image patches, is by forming a feature vector that contains all the statistics we identified in Section III-B over all subbands, and computing the distance between the feature vectors. We found that it is most effective when the statistics are computed on the ***magnitudes*** of the subband coefficients.

One of the advantages of this approach is that we can add different weights for different statistics, depending on the application and database characteristics. For example, we could put a lower weight on statistics with large variance across the database, thus de-emphasizing differences that are expected to be large and paying more attention to differences that are not commonly occurring. This can be accomplished by computing the *Mahalanobis distance* [56] between the feature vectors, which if we assume that the different features are mutually uncorrelated, is a weighted Euclidean distance with weights inversely proportional to the variance of each feature. We refer to the resulting metric as STSIM-M, where "M" stands for Mahalanobis. Note that as a distance this is a dissimilarity metric that takes values between 0 and ∞.

The feature vector for image or image patch $\mathbf{x}$ has a total of $N_p = 4 \cdot N_b + N_c$ terms and can be written as:

$$F_{\mathbf{x}} = [f_{\mathbf{x},1}, f_{\mathbf{x},2}, \ldots, f_{\mathbf{x},N_p}]$$

The STSIM-M metric for $\mathbf{x}$ and $\mathbf{y}$, is then given by the Mahalanobis distance between their feature vectors $F_{\mathbf{x}}$ and $F_{\mathbf{y}}$:

$$Q_{\text{STSIM-M}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{N_p} \frac{(f_{\mathbf{x},i} - f_{\mathbf{y},i})^2}{\sigma_{f_i}^2}}. \qquad (16)$$

where σ_{f_i} the standard deviation of the i^{th} feature across all feature vectors in the database. Thus, unlike the other SSIM and STSIM metrics, computation of the distance between two texture images using STSIM-M requires statistics based on the entire database.

IV. Experiments

As we discussed in the introduction, one of our goals was to conduct systematic experiments over a large image database that will enable testing different aspects of metric performance. We have chosen to test metrics in the context of retrieving identical textures (known-item search), which as we argued, essentially eliminates the need for subjective experiments, thus enabling comparisons with human performance on a large database. While this seems to restrict testing to a very specific problem, we will argue that the conclusions transcend the particular application and have important consequences for other image analysis applications, including compression.

As we pointed out in Section III, the metrics we proposed in this paper are not scale or rotation invariant. Accordingly, in the experimental results, textures with different scales and orientations will be considered as dissimilar.

A. Database Construction

For our experiments, we collected a large number of color texture images. The images were carefully selected to (a) meet some basic assumptions about texture signals, and (b) facilitate the construction of groups of identical textures.

To address the first point, we need a definition of texture. The precise definition of texture is not widely agreed on in the literature. However, several authors (e.g., Portilla and Simoncelli [11]) define texture as *an image that is spatially homogeneous and that typically contains repeated structures, often with some random variation (e.g., random positions, size, orientations or colors).* The textures we collected had to meet the requirement of spatial homogeneity and repetitiveness; the latter we defined as at least five repetitions, horizontally or vertically, of a basic structuring element. We also made sure that there is a wide variety of textures and a wide range of similarities between pairs of different textures.

To address the second point, we collected images of what we considered to be perceptually uniform textures, from which we cut smaller patches of identical textures – each of which met the basic texture assumptions. The group of patches originating from the same larger texture are considered to be identical textures, and thus considered *relevant* to each other in a statistical sense.

Our subjective experiments were conducted on two different texture databases, obtained from the *Corbis* [57] and *CUReT* databases [58], [59], respectively.

To construct the first database, we downloaded around 1000 color images from the *Corbis* website [57]. All of the textures were photographic, mostly of natural or man-made objects and scenes. No synthetic textures were included. The resolution varied from 170×128 to 640×640 pixels. Roughly 300 of those were discarded, as they did not represent perceptually uniform textures. Of the remaining 700 images, we selected 425 for the known-item-search experiments. To obtain groups of identical textures, each of the 425 images were cut into a number of 128×128 patches. Depending on the size of the original image, the extracted images had different degrees of spatial overlap, but we made sure that there were substantial point-by-point differences, such as those shown in Figure 1. The idea was to minimize overlap while maintaining texture homogeneity. In some cases – when the original image was large enough – we downsampled the image, typically by a factor of two, in order to meet the repetitiveness requirement. A minimum of two and a maximum of twelve patches were obtained from each original texture. Overall, we obtained 1180 texture patches originating from 425 original texture images.

The second database was constructed in similar fashion using 61 images from the *CUReT* database [58], [59], which contains images of real-world textures taken at different viewing and illumination directions. We selected images from lighting and viewing condition 122 [59]. From each of the 61 images, we cut out three 128×128 patches at random positions, making sure that the entire patch overlapped the texture portion of the image. The total number of test images was thus 183. The advantage of the *CUReT* database is that the textures were carefully chosen and photographed under

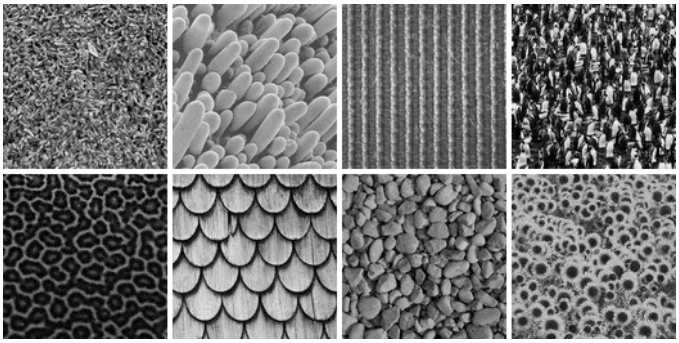


Fig. 4. Samples from the *Corbis* database

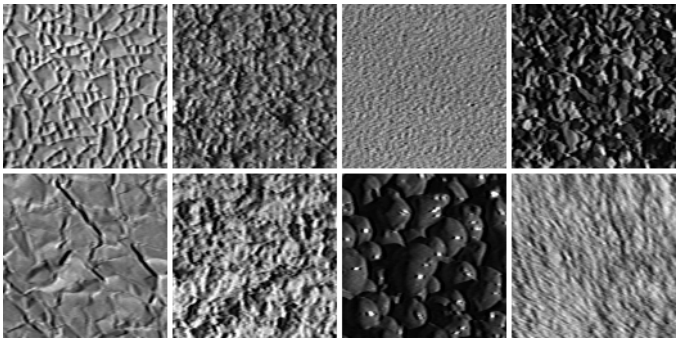


Fig. 5. Samples from the *CUReT* database

controlled conditions. More importantly for our experiments, all textures are more or less perceptually uniform. On the other hand, the variety of materials is limited. Since our primary interest is on the variety of textures rather than the detailed effects of viewing conditions, the *Corbis* database is better suited to the goals of this paper.

Figures 4 and 5 show examples of images from the two databases. In both cases, we used the grayscale component of the images. From now on, we will refer to the selected textures from the two databases as the *Corbis* and *CUReT* databases.

B. Performance Based on Information Retrieval Statistics

We treat the known-item search experiment as a retrieval task: an image is queried and the similarity scores between the query and the rest of the database are ordered according to decreasing similarity. The first retrieved document is the image with highest similarity to the query; the second retrieved document is the one with the second-highest similarity, etc.

One informative measure of performance is the number of times the first retrieved image is *relevant*, i.e., it comes from the same original image and has the same label as the query. This is commonly referred to as *precision at one*. Another way of assessing metric performance is to compute the *mean reciprocal rank (MMR)*, i.e., the average value of the inverse rank of the first relevant retrieved image [60]. This measure tells us, on average, how far down the list the first relevant image is.

When there is more than one relevant image for a given query, as is the case for many of the entries of our database, the usual value to report is *mean average precision (MAP)* [61]. The MAP is calculated as follows: for each query and positive integer n less than or equal to the size of the database, we compute the fraction of the n highest ranked images that are relevant (precision), and then average these fractions over all values of n for which the nth highest ranked image was actually relevant, to obtain the MAP for that query. Finally, we average these values across all images.

	Corbis Database			*CUReT* Database		
Metric	P@1	MRR	MAP	P@1	MRR	MAP
PSNR	0.04	0.07	0.06	0.11	0.17	0.17
S-SSIM	0.09	0.11	0.06	0.06	0.11	0.10
CW-SSIM	0.39	0.46	0.40	0.69	0.77	0.72
CW-SSIM global	0.27	0.36	0.28	0.31	0.45	0.35
STSIM-1	0.74	0.80	0.72	0.81	0.85	0.80
STSIM-1 global	0.86	0.90	0.81	0.93	0.94	0.90
STSIM-2	0.74	0.80	0.74	0.81	0.86	0.81
STSIM-2 global	0.93	0.95	0.89	**0.97**	**0.97**	**0.95**
STSIM-M	**0.96**	**0.97**	**0.92**	0.96	**0.97**	**0.95**
Gabor features	0.92	0.94	0.88	0.96	0.96	**0.95**
Wavelet features	0.84	0.89	0.80	0.92	0.95	0.93
LBP	0.90	0.92	0.86	0.93	0.94	0.89

TABLE I
INFORMATION RETRIEVAL STATISTICS

In our experiments, we compared the following metrics:

- PSNR
- S-SSIM with 7×7 local window
- CW-SSIM with 7×7 local window
- CW-SSIM over the entire image (global)
- STSIM-1 with 7×7 local window
- STSIM-1 over the entire image (global)
- STSIM-2 with 7×7 local window
- STSIM-2 over the entire image (global)
- STSIM-M over the entire image (global)
- Normalized ℓ^1 distance on Gabor features [36]
- Kullback-Leibler distance on wavelet features [38]
- Local Binary Patterns (LBP) [62]

The implementation of the texture similarity algorithms of Manjunath and Ma [36] and of Do and Vetterli [38] were downloaded from the respective authors' websites. For simplicity, we will refer to them in tables and plots as *Gabor features* [36] and *Wavelet features* [38]. The implementation of the LBP method [62] was downloaded from the authors' website and uses the $LBP^{riu2}_{8,1} + LBP^{riu2}_{24,3}$ combination of features. Additionally, to avoid $\log 0$ terms causing the LBP metric to produce undefined values, any such term was replaced by $\log 10^{-8}$.

The results are summarized in Table I for the two databases. The highest value for each statistic is highlighted. Even though the databases are quite different, the results are qualitatively the same. Based on these results, and according to all three statistics, the global STSIM-M and STSIM-2 metrics outperform all other metrics. Note that including the extra statistics results in a substantial gain over STSIM-1. Note how poor is the performance of the point-by-point metrics (PSNR and S-SSIM). Another observation is that, with the exception of CW-SSIM, the global methods have a significantly higher performance than the local, sliding window-based ones. This can be explained by the fact that we are comparing more or less homogeneous texture images and it is in our interest

Metric	p-values
STSIM-1 local & STSIM-2 local	0.692
STSIM-1 global & Wavelet features	0.098
STSIM-2 global & Gabor features	0.269
LBP & Gabor features	0.061

TABLE II
COCHRANE'S Q TEST P-VALUES > 0.01 (*Corbis*)

to capture their global, overall image statistics, rather than comparing the images on a window-by-window basis. The small sliding windows may in fact not include enough of the texture image to capture its statistical regularities. This is particularly true for higher scale (coarser) textures, for which the image in a small window may not qualify as a texture. Thus, an implicit assumption is that the smallest window over which the texture statistics are computed qualifies as a texture, as we defined it earlier in this section. When this is true, the STSIM metrics are very tolerant of non-structural deformations, but when it is violated, then the performance of the metric deteriorates. To avoid such cases, the scale of the textures can either be known *a priori* or the application of STSIMs can be coupled with with a texture scale detector, so that the metric can be chosen adaptively.

C. Statistical Significance Tests (For Corbis Database)

To test whether the differences in performance based on the information retrieval statistics are due to chance, we performed standard statistical tests with significance level $\alpha = 0.01$.

1) Precision at One: Since there are only two possible outcomes for each query – the first retrieved image is relevant or not – we performed the Cochrane's Q test [63] to determine the significance of the results. The test was applied to each pair of metrics, and found that all differences are statistically significant except the ones listed in Table II, for which the p-values are greater than $\alpha = 0.01$. Thus, the global STSIM-2 and STSIM-M metrics significantly outperform all other metrics, based on precision at one.

2) Mean Reciprocal Rank and Mean Average Precision: Since the MRR statistic is ordinal and MAP is non-Gaussian, we performed the Friedman test [64], [65] followed by the Tukey-Kramer Honestly Significant Difference test [66] to determine the significance of the results. The results are represented by the plots of Figs. 6 and 7, which show the mean performance ranks for each metric and the confidence intervals, for $\alpha = 0.01$. When the confidence intervals of two particular metrics overlap, the difference in their performance scores is not considered statistically significant. These statistical tests confirm that, based on these retrieval statistics, the superior performance of STSIM-2 and STSIM-M is not by chance due to the limited size of the database.

D. Performance Based on Receiver Operating Characteristic

Another approach for comparing metric performance is to treat the known-item search problem as a binary classification problem, where the task is to determine whether two images are identical textures (null hypothesis) or not (alternate hypothesis). The test variable is the similarity value that a metric

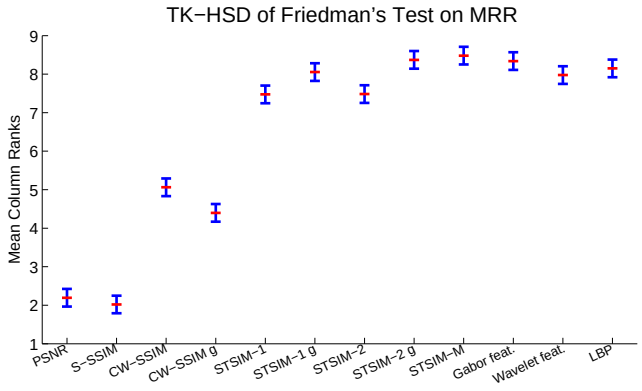


Fig. 6. Friedman's test on mean reciprocal rank values (*Corbis*)

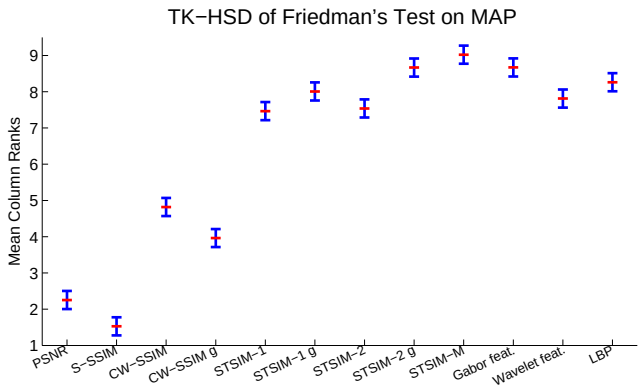


Fig. 7. Friedman's test on mean average precision values (*Corbis*)

produces for a pair of images. The system should then decide whether the two images correspond to identical textures by comparing the probability of the given similarity value under each of the hypotheses. The probability density functions for the two hypotheses are modeled by the histograms of metric values, one for the pairs of identical textures and one for the pairs of non-identical textures. Figure 8 (top) shows an example of well-separated distributions, that correspond to the STSIM-2 metric over the *Corbis* database. Note that the distribution for identical textures is peaky, which tells us that the metric provides similar values for similar textures irrespective of content. This is a much stronger indicator of metric performance than the retrieval statistics of Section IV-B because it establishes that there is an absolute threshold for metric values above which textures can be considered identical. On the other hand, the distribution for non-identical textures is much broader, which is expected given the variety of textures in the database. Figure 8 (bottom) shows an example of distributions with a lot of overlap; these correspond to PSNR over the *Corbis* database. Note that in addition to the overlap, the distribution for identical textures is fairly broad.

Given the probability density functions, we can compare metric performances by plotting the receiver operating characteristic (ROC) curve for each metric. The ROC curve plots the true positives rate (TPR) as a function of the false positives rate (FPR). The ROC curves obtained when the different metrics were tested on the *Corbis* database are plotted in Figure 9.

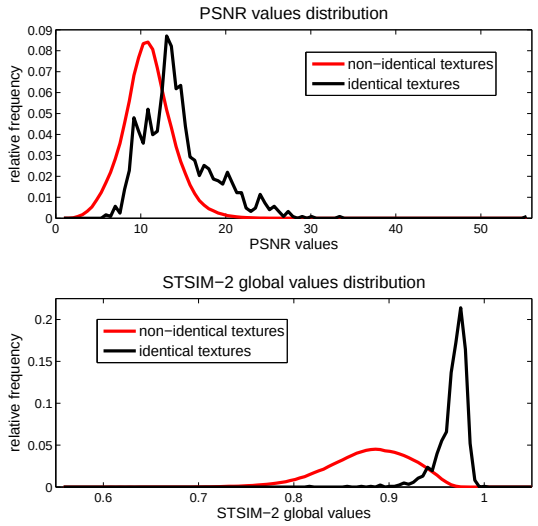


Fig. 8. Probability density functions for identical texture detection (*Corbis*)

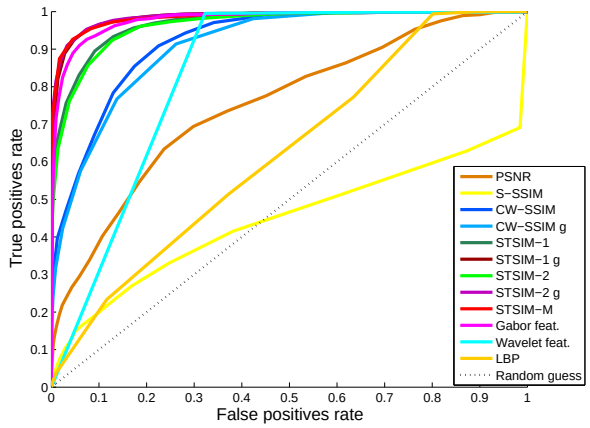


Fig. 9. ROC curves

The area under the curves can be used as a measure of performance. Ideally, the area is equal to 1. The areas under the curves are given in Table III. Again, note that the global STSIM metrics outperform all other metrics, and that the global metrics outperform the local ones. These results are consistent with the results based on the information retrieval statistics, with one notable exception. The LBP algorithm has very poor performance. This is because while it has relatively good classification performance, it does not result in consistent similarity values, that is, it cannot be used to determine an absolute threshold for metric performance.

E. Points of Failure and Future Research

The results presented so far are quite good. However, a close study of the cases where the STSIM metrics fail to retrieve the correct image are quite revealing, as they can point to both weaknesses of the proposed metrics and strengths of the proposed approaches to texture similarity. Figure 10 shows different failure examples. It shows the query image, the best match and the first correct match. The most benign type of

Metric	Area	Metric	Area
PSNR	0.753	STSIM-2	0.963
S-SSIM	0.446	STSIM-2 global	**0.986**
CW-SSIM	0.921	STSIM-M	0.985
CW-SSIM global	0.910	Gabor features	0.979
STSIM-1	0.967	Wavelet features	0.836
STSIM-1 global	0.985	LBP	0.625

TABLE III
AREA UNDER THE ROC CURVE FOR *Corbis* DATABASE

failure is when the metric (STSIM-M) retrieves a texture that is quite similar to the query, like the one shown in Example A of Figure 10. Another type of failure is shown in Example B, where the retrieved image has similar statistics to the query, with the only difference being that one is quasi periodic and the other is more random. This is a common type of failure and difficult for the proposed metrics to handle. In Example C, the images have more or less the same underlying texture except for weak edges that are too sparse to be captured by high-frequency subbands and have too little contrast to be captured by the low-frequency subbands. This type of failure is also difficult for the proposed metrics to handle. In Examples D and E, the differences are more substantial, but it does not help that the orientations of the textures of the identical pairs are not well matched. Finally, some failures come from the fact that images in our database have different scales. This can be seen in Example F, where the retrieved image is a texture at a larger scale than the query image. Note that our metric weights similarity equally across scales. In general, the metric in its current form has difficulties handling textures of larger scales. There are a number of possibilities for improvement, e.g., by explicitly detecting the scale of each image.

V. CONCLUSIONS

We developed structural texture similarity metrics, which account for human visual perception and the stochastic nature of textures. The metrics allow substantial point-by-point deviations between textures that according to human judgment are essentially identical. They are based on a steerable filter decomposition and rely on a concise set of subband statistics, computed globally or in sliding windows. We investigated the performance of a progression of metrics (STSIM-1, STSIM-2, STSIM-M) in the context of known-item search, the retrieval of textures that are identical to the query texture. This eliminates the need for cumbersome subjective tests, thus enabling comparisons with human performance on a large database. We compared the performance of the STSIMs to PSNR, SSIM,CW-SSIM, as well as state-of-the-art texture classification metrics in the literature, using standard statistical measures. We have shown that global metrics perform best for texture patch retrieval, and that the STSIM-2 and STSIM-M metrics outperform all other metrics.

ACKNOWLEDGEMENT

This work was supported in part by the U.S. Department of Energy National Nuclear Security Administration (NNSA) under Grant No. DE-NA0000431. Any opinions, findings, and

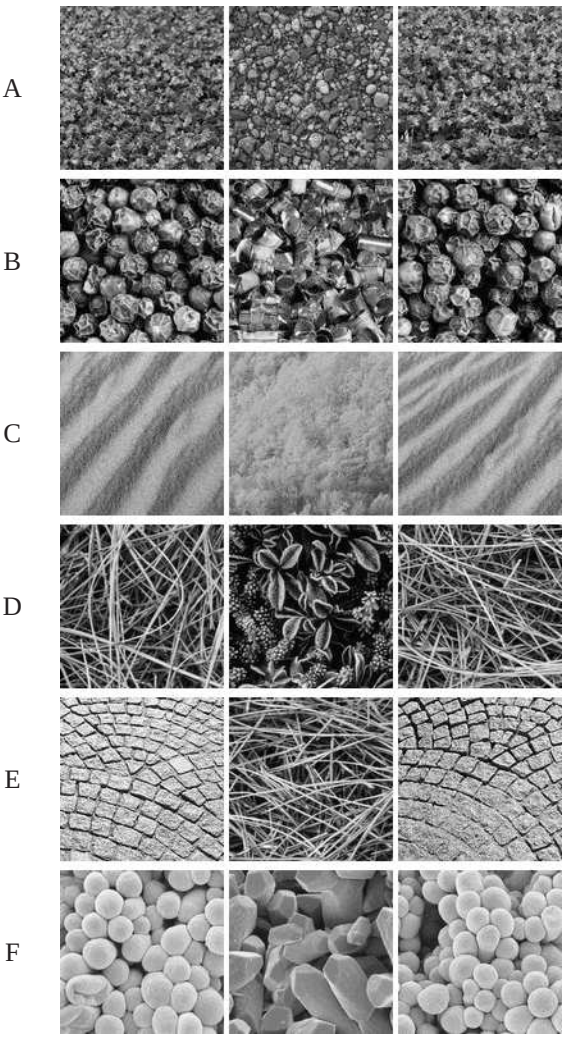


Fig. 10. Query image (left), best match (middle), first identical match (right)

conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NNSA.

REFERENCES

[1] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing*, vol. 70, pp. 177–200, 1998.

[2] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, 2nd ed., A. C. Bovik, Ed. Academic Press, 2005, pp. 939–959.

[3] 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.

[4] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. II, Philadelphia, PA, 2005, pp. 573–576.

[5] X. Zhao, M. G. Reyes, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for retrieval applications," in *Proc. Int. Conf. Image Processing (ICIP)*, San Diego, CA, Oct. 2008, pp. 1196–1199.

[6] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Structural similarity metrics for texture analysis and retrieval," in *Proc. Int. Conf. Image Processing*, Cairo, Egypt, Nov. 2009, pp. 2225–2228.

[7] D. Cano and T. H. Minh, "Texture synthesis using hierarchical linear transforms," *Signal Processing*, vol. 15, pp. 131–148, 1988.

[8] M. Porat and Y. Y. Zeevi, "Localized texture processing in vision: Analysis and synthesis in Gaborian space," *IEEE Trans. Biomed. Eng.*, vol. 36, no. 1, pp. 115–129, 1989.

[9] K. Popat and R. W. Picard, "Novel cluster-based probability model for texture synthesis, classification, and compression," in *Proc. SPIE Visual Communications '93*, Cambridge, MA, 1993.

[10] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *Proc. Int. Conf. Image Processing (ICIP), vol. III*, Washington, DC, Oct. 1995, pp. 648–651.

[11] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statictics of complex wavelet coefficients," *Int. J. Computer Vision*, vol. 40, no. 1, pp. 49–71, Oct. 2000.

[12] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[13] G. Jin, Y. Zhai, T. N. Pappas, and D. L. Neuhoff, "Matched-texture coding for structurally lossless compression," in *Proc. Int. Conf. Image Processing (ICIP)*, Orlando, FL, Oct. 2012, accepted.

[14] C. T. Meadow, B. R. Boyce, D. H. Kraft, and C. Barry, *Text information retrieval systems*. Emerald Group Publishing, 2007.

[15] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Trans. Image Process.*, vol. 11, no. 2, pp. 146–158, Feb. 2002.

[16] T. Ojala, T. Menp, J. Viertola, J. Kyllnen, and M. Pietikinen, "Empirical evaluation of MPEG-7 texture descriptors with a large-scale experiment," in *Proc. 2^{nd} Int. Wksp. Texture Anal. Synthesis*, 2002, pp. 99–102.

[17] Z. He, X. You, and Y. Yuan, "Texture image retrieval based on non-tensor product wavelet filter banks," *Signal Processing*, vol. 89, no. 8, pp. 1501–1510, 2009.

[18] J. Zujovic, T. N. Pappas, D. L. Neuhoff, R. van Egmond, and H. de Ridder, "Subjective and objective texture similarity for image compression," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 1369–1372.

[19] ——, "A new subjective procedure for evaluation and development of texture similarity metrics," in *Proc. IEEE 10th IVMSP Wksp.: Perception and Visual Signal Analysis*, Ithaca, New York, Jun. 2011, pp. 123–128.

[20] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Perceptual similarity metrics for retrieval of natural textures," in *Proc. IEEE Wksp. Multimedia Signal Proc.*, Rio de Janeiro, Brazil, Oct. 2009.

[21] J. Zujovic, "Perceptual texture similarity metrics," Ph.D. dissertation, Northwestern Univ., Evanston, IL, Aug. 2011.

[22] M. Tuceryan and A. K. Jain, "Texture analysis," in *Handbook Pattern Recognition and Computer Vision*, C. H. Chen, L. F. Pau, and P. S. P. Wang, Eds. Singapore: World Scientific Publishing, 1993, ch. 2, pp. 235–276.

[23] H. Z. Long, W. K. Leow, and F. K. Chua, "Perceptual texture space for content-based image retrieval," in *Proc. Int. Conf. Multimedia Modeling (MMM)*, Nagano, Japan, Nov. 2000, pp. 167–180.

[24] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, no. 6, pp. 610–621, 1973.

[25] S. K. Saha, A. K. Das, and B. Chanda, "CBIR using perception based texture and colour measures," in *Proc. 17th Int. Conf. Pattern Recognition (ICPR)*, vol. 2, 2004, pp. 985–988.

[26] T. Mita, T. Kaneko, B. Stenger, and O. Hori, "Discriminative feature co-occurrence selection for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1257–1269, 2008.

[27] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Trans. Image Process.*, vol. 4, no. 11, pp. 1549–1560, Nov. 1995.

[28] P. Chen and T. Pavlidis, "Segmentation by texture using correlation," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 1, pp. 64–69, 1983.

[29] B. Julész, E. Gilbert, and J. Victor, "Visual discrimination of textures with identical third-order statistics," *Biological Cybernetics*, vol. 31, no. 3, pp. 137–140, 1978.

[30] B. Julesz, "A theory of preattentive texture discrimination based on first-order statistics of textons," *Biological Cybernetics*, vol. 41, no. 2, pp. 131–138, 1981.

[31] R. L. Kashyap and A. Khotanzad, "A model-based method for rotation invariant texture classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-8, no. 4, pp. 472–481, July 1986.

[32] B. S. Manjunath and R. Chellappa, "Unsupervised texture segmentation using markov random field models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 5, pp. 478–482, 1991.

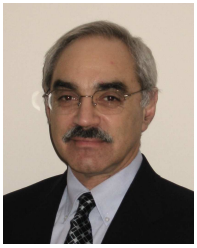
[33] S. C. Zhu, Y. Wu, and D. Mumford, "Filters, random fields and maximum entropy (frame) – towards a unified theory for texture modeling," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 107–126, 1998.

[34] P. Ndjiki-Nya, D. Bull, and T. Wiegand, "Perception-oriented video coding based on texture analysis and synthesis," in *Proc. Int. Conf. Image Processing (ICIP))*, Nov. 2009, pp. 2273–2276.
[35] M. Clark, A. C. Bovik, and W. S. Geisler, "Texture segmentation using a class of narrowband filters," in *Proc. ICASSP*, Apr. 1987, pp. 571–574.
[36] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.
[37] V. Wouwer, G. Scheunders, P. Livens, and S. van Dyck, "Wavelet correlation signatures for color texture characterization," *Pattern Recognition*, vol. 32, pp. 443–451, 1999.
[38] M. N. Do and M. Vetterli, "Texture similarity measurement using Kullback-Leibler distance on wavelet subbands," in *Proc. Int. Conf. Image Proc.*, vol. 3, Vancouver, BC, Canada, Sep. 2000, pp. 730–733.
[39] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, pp. 106–154, Jan. 1962.
[40] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Optical Soc. America A*, vol. 2, pp. 1160–1169, 1985.
[41] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *J. Neurophysiology*, vol. 58, no. 6, pp. 1233–1258, Dec. 1987.
[42] D. Dunn and W. E. Higgins, "Optimal Gabor filters for texture segmentation," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 947–964, Jul. 1995.
[43] J. Ilonen, J. K. Kamarainen, P. Paalanen, M. Hamouz, J. Kittler, and H. Kalviainen, "Image feature localization by multiple hypothesis testing of Gabor features," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 311–325, 2008.
[44] S. Arivazhagan, L. Ganesan, and S. Priyal, "Texture classification using Gabor wavelets based rotation invariant features," *Pattern Recognition Letters*, vol. 27, no. 16, pp. 1976–1982, December 2006.
[45] J. Mathiassen, A. Skavhaug, and K. B, "Texture similarity measure using Kullback-Leibler divergence between gamma distributions," *Computer Vision ECCV 2002*, pp. 19–49, 2002.
[46] G. Yang and Y. Xiao, "A robust similarity measure method in CBIR system," in *Proc. Congr. Image Signal Proc.*, vol. 2, 2008, pp. 662–666.
[47] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.
[48] B. S. Manjunath, P. Salembier, and T. Sikora, *Texture Descriptors*. John Wiley and Sons, 2002, ch. 14, pp. 213–228.
[49] Y. Lu, Q. Zhao, J. Kong, C. Tang, and Y. Li, "A two-stage region-based image retrieval approach using combined color and texture features," in *AI 2006: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006, vol. 4304/2006, pp. 1010–1014.
[50] A. C. Brooks, X. Zhao, and T. N. Pappas, "Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1261–1273, Aug. 2008.
[51] A. Reibman and D. Poole, "Characterizing packet-loss impairments in compressed video," in *Proc. Int. Conf. Image Proc. (ICIP)*, vol. 5, San Antonio, TX, Sep. 2007, pp. 77–80.
[52] E. P. Simoncelli, "Statistical models for images: compression, restoration and synthesis," *Conf. Record Thirty-First Asilomar Conf. Signals, Sys., Computers*, vol. 1, pp. 673–678, Nov. 1997.
[53] D. Hubel and T. Wiesel, "Ferrier lecture: Functional architecture of macaque monkey visual cortex," *Proc. Royal Society of London. Series B, Biological Sciences*, vol. 198, no. 1130, pp. 1–59, 1977.
[54] K. Foster, J. Gaska, S. Marcelja, and D. Pollen, "Phase relationships between adjacent simple cells in the feline visual cortex," *J Physiol (London)*, vol. 345, p. 22p, 1983.
[55] T. Aach, A. Kaup, and R. Mester, "On texture analysis: Local energy transforms versus quadrature filters," *Signal processing*, vol. 45, no. 2, pp. 173–181, 1995.
[56] P. C. Mahalanobis, "On the generalized distance in statistics," in *Proc. of the National Institute of Science,India*, vol. 2, 1936, pp. 49–55.
[57] "Corbis stock photography." [Online]. Available: www.corbis.com
[58] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real-world surfaces," *ACM Trans. Graphics*, vol. 18, no. 1, pp. 1–34, Jan. 1999.
[59] "CUReT: Columbia-Utrecht Refelctance and Texture Database." [Online]. Available: www1.cs.columbia.edu/CAVE/software/curet/
[60] E. M. Voorhees, "The trec-8 question answering track report," in *In Proc. of TREC-8*, 1999, pp. 77–82.
[61] ——, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information Processing & Management*, vol. 36, no. 5, pp. 697–716, Sep. 2000.
[62] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 971–987, 2002.
[63] W. G. Cochran, "The combination of estimates from different experiments," *Biometrics*, vol. 10, no. 1, pp. 101–129, 1954.
[64] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.
[65] J. D. Gibbons, *Nonparametric statistics: An Introduction*, ser. Quantitative Applications in Social Sciences 90. London: Sage Publications, 1993.
[66] J. W. Tukey, "Quick and dirty methods in statistics," in *Part II: Simple Analyses for Standard Designs. Quality Control Conference Papers*, 1951, pp. 189–197.

Jana Zujovic (M'09) received the Diploma in electrical engineering from the University of Belgrade in 2006, and the M.S. and Ph.D. degrees in electrical engineering and computer science from Northwestern University, Evanston, IL, in 2008 and 2011, respectively. From 2011 until 2013, she was working as a postdoctoral fellow at Northwestern University. Currently she is employed as a senior research engineer at FutureWei Technologies, Santa Clara, CA. Her research interests include image and video analysis, image quality and similarity, content-based retrieval and pattern recognition.

Thrasyvoulos N. Pappas (M'87, SM'95, F'06) received the S.B., S.M., and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, in 1979, 1982, and 1987, respectively. From 1987 until 1999, he was a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ. In 1999, he joined the Department of Electrical and Computer Engineering (now EECS) at Northwestern University. His research interests are in image and video quality and compression, image and video analysis, content-based retrieval, perceptual models for multimedia processing, model-based halftoning, and tactile and multimodal interfaces.

Dr. Pappas is a Fellow of the IEEE and SPIE. He has served as an elected member of the Board of Governors of the Signal Processing Society of IEEE (2004-07), editor-in-chief of the IEEE Transactions on Image Processing (2010-12), chair of the IEEE Image and Multidimensional Signal Processing Technical Committee (2002-03), and technical program co-chair of ICIP-01 and ICIP-09. He has also served as co-chair of the 2005 SPIE/IS&T Electronic Imaging Symposium, and since 1997 he has been co-chair of the SPIE/IS&T Conference on Human Vision and Electronic Imaging. In addition, Dr. Pappas has served on the editorial boards of the IEEE Transactions on Image Processing, the IEEE Signal Processing Magazine, the IS&T/SPIE Journal of Electronic Imaging, and the Foundations and Trends in Signal Processing.



David L. Neuhoff received the B.S.E. from Cornell and the M.S. and Ph.D. in Electrical Engineering from Stanford. Since graduation he has been a faculty member at the University of Michigan, where he is now the Joseph E. and Anne P. Rowe Professor of Electrical Engineering. From 1984 to 1989 he was an Associate Chair of the EECS Department, and since September 2008 he is again serving in this capacity. He spent two sabbaticals at Bell Laboratories, Murray Hill, NJ, and one at Northwestern University. His research and teaching interests are in communications, information theory, and signal processing, especially data compression, quantization, image coding, image similarity metrics, source-channel coding, halftoning, sensor networks, and Markov random fields. He is a Fellow of the IEEE. He co-chaired the 1986 IEEE International Symposium on Information Theory, was technical co-chair for the 2012 IEEE SSP workshop, has served as an associate editor for the IEEE Transactions on Information Theory, has served on the Board of Governors and as president of the IEEE Information Theory Society.