Image analysis and compression: Renewed focus on texture

Thrasyvoulos N. Pappas[a], Jana Zujovic[a], David L. Neuhoff[b]

[a]EECS Department, Northwestern University, 2145 Sheridan Rd., Evanston, IL 60201;
[b]EECS Department, University of Michigan, 1301 Beal Ave., Ann Arbor, MI 48109

ABSTRACT

We argue that a key to further advances in the fields of image analysis and compression is a better understanding of texture. We review a number of applications that critically depend on texture analysis, including image and video compression, content-based retrieval, visual to tactile image conversion, and multimodal interfaces. We introduce the idea of "structurally lossless" compression of visual data that allows significant differences between the original and decoded images, which may be perceptible when they are viewed side-by-side, but do not affect the overall quality of the image. We then discuss the development of objective texture similarity metrics, which allow substantial point-by-point deviations between textures that according to human judgment are essentially identical.

Keywords: Image quality, human visual perception, structurally lossless compression, structural texture similarity, dominant colors, image segmentation, semantic analysis.

1. INTRODUCTION

The fields of image analysis and compression have made significant advances during the last two decades, incorporating sophisticated signal processing techniques and models of human perception. One of the keys to further advances is a better understanding of texture. Even though the importance of texture for image quality and semantic analysis is obvious, it is surprising that it has received relatively little attention. For example, image compression techniques have relied on point-by-point comparisons – whether in the original image domain or in a transform domain – that cannot adequately exploit the stochastic nature of texture.[1] Similarly, computer vision has mostly focused on object detection rather than the perception of materials, which is critically dependent on texture.[2] A variety of other fields rely on texture analysis, such as graphics, multimodal interfaces, and we will see, sense substitution. In this paper, we review a number of applications that critically depend on texture analysis, and then focus on the importance of compact texture representations and texture similarity metrics. Our primary focus will be on natural textures.

We first look at image/video compression. In order to achieve high compression ratios while maintaining image quality, image compression algorithms must eliminate all redundant and irrelevant information. The state of the art is dominated by transform-based techniques, including subband and wavelet decompositions. Such techniques work well in smooth regions of the image, where all the energy is concentrated in the low frequency coefficients, but are not very efficient in textured and transition (i.e, containing edges) regions, where there is a lot of energy in the high frequencies. The key to overcoming this problem is better prediction. In fact, better prediction has been the key to the success of video compression algorithms, which use sophisticated motion compensation techniques. This has worked well for transition regions, but the benefits of prediction in textured regions have been limited due to the inability of existing quality metrics to predict perceptual texture similarity.

Another application that critically depends on texture analysis is image classification and retrieval. Most existing approaches[3,4] rely on the extraction of low-level features, such as color, texture, and shape. Ideally such features must be linked to objects in the scene. However, since the detection of objects is quite difficult, an alternative approach is to rely on image segments or fixed-size blocks. Fixed-size blocks are the simplest but may lead to misclassification, as they may not follow object boundaries. The success of this approach thus depends on the availability of semantically meaningful segmentations.[5] A recently proposed approach[6] is based on spatially adaptive color and spatial texture features, and combines an understanding of image characteristics with perceptual models to segment natural scenes into perceptually

Send correspondence to Thrasyvoulos N. Pappas
E-mail: pappas@eecs.northwestern.edu

Visual Information Processing and Communication, edited by Amir Said, Onur G. Guleryuz,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7543, 75430N · © 2010 SPIE-IS&T
CCC code: 0277-786X/10/$18 · doi: 10.1117/12.851682

uniform regions. A third alternative relies on Lempel-Ziv incremental parsing to decompose an image into variable-size rectangular patches, which provide an asymptotically optimal representation for image compression and retrieval applications.[7,8] In all of these cases, the content of the regions – as well as their context within an image – must be analyzed to obtain semantic information. For example, in Refs. 9–13, the extraction of region-wide color and texture features and segment location were used for segment classification. Segment size, boundary shape, and properties of neighboring segments can be used to further improve classification accuracy. Another approach for segment classification is the direct comparison of the segment textures with reference textures. This depends on the existence of texture similarity metrics, which is one of the main focus points of this paper. We will discuss the importance of structural texture similarity metrics in the retrieval of perceptually equivalent textures.

Image segmentation into perceptually uniform regions, combined with texture analysis of the resulting segments is also the basis of a new approach for converting images into tactile patterns. This can help provide greater accessibility to the handicapped (visually impaired) segment of the population. In Ref. 14, Pappas *et al.* proposed a segmentation-based approach for transforming visual to tactile information. The main idea is to map each image segment into a distinct tactile pattern that conveys the same information. The mapping is not obvious because, while for some attributes of visual textures (e.g., directionality, regularity) there are straightforward tactile analogs, for others (e.g., color) there is no obvious correspondence. It is thus necessary to assign an arbitrary mapping from visual concept (e.g., green grass, blue mountains) to tactile pattern. Pappas *et al.* refer to these as *indirect semantic mappings* in contrast to direct mappings, which are also based on semantics. Of course, both can be used, but in either case, texture analysis is key to the mapping. Finally, acoustic signals can be used to enhance or modify the perception of the tactile patterns.

The last observation leads to the importance of multimodal interfaces, whereby the joint perception of visual, acoustic, and tactile textures is considered. Such interfaces are important for the next generation of interactive environments, which will enable telepresence and natural interfaces for communication, commerce, entertainment, education, and medicine, and of course, provide greater accessibility to the handicapped segment of the population (visually impaired or hearing impaired). Initial results in this direction have been reported in Ref. 15. Again, texture analysis and similarity metrics are key here, not only for visual textures, but also for tactile and acoustic, as well as multimodal textures.

The development of texture similarity metrics has been inspired in part by recent research in the area of texture analysis and synthesis. Several authors have presented models for texture analysis/synthesis using multiscale frequency decompositions.[16–23] The most complete results were presented by Portilla and Simoncelli,[23] who developed a elaborate statistical model for texture images that is consistent with human perception. The model is based on the "steerable filter" decomposition[24,25] and captures a very wide class of textures. A related problem is that of generating a texture based on a small sample. Efros and Freeman proposed a technique for stitching together patches from an existing texture to generate a bigger texture.[26] The key to their approach is to ensure continuity between block boundaries. Even though in such cases the similarity is implied by the way the texture is constructed, and verified by perceptual evaluations, it is nevertheless important to have objective similarity metrics for evaluating the quality of the resulting textures."

The development of objective metrics for texture similarity is considerably more challenging than that of traditional image quality metrics because there can be substantial point-by-point deviations between textures that according to human judgment are essentially identical. Such metrics are important not only for image analysis and retrieval applications, but also for image coding applications in which significant changes in the image are permissible, as long as the perceived image quality is unaffected, even though in a side-by-side comparison there may be clearly perceptible differences.

The impetus to develop metrics that deviate from traditional point-by-point fidelity was initiated by the introduction of a broad class of new metrics, the structural similarity metrics (SSIM),[27] which attempt to incorporate "structural" information in image comparisons. Unlike traditional metrics, SSIMs can give high similarity scores even to images with significant pixel-wise differences. The goal is allow deviations that do not affect the structure of the image, which should be what ensures *perceptual* similarity. Wang *et al.* have proposed a number of different metrics, both in the space domain (SSIM)[27] and in the complex wavelet domain (CWSSIM).[28] However, as we will argue below, these metrics still rely on cross-correlations between the two images, and are thus too constrained to capture the perceptual similarity of two textures. In order to overcome such limitations, Zhao *et al.*[29] proposed a structural texture similarity metric (STSIM) that relies entirely on local image statistics, and is thus completely decoupled from point-by-point comparisons. Zujovic *et al.* further developed this idea in Ref. 30, and applied it to the problem of "known-item search"[31] with very encouraging results.[32]

As we saw above, a variety of applications can make use of structural texture similarity metrics. Each application imposes its own requirements on metric performance. Thus, in image compression it is important to ensure a monotonic relationship between measured and perceived distortion, while in image retrieval it may be sufficient to distinguish between similar and dissimilar textures, or to retrieve only textures that are perceptually "identical" to the query texture, as would have been the case if they were different pieces of the same perceptually uniform texture. When relative similarity is important, one should distinguish between the cases when an absolute similarity scale is important, from those where a relative scale may be adequate.

In the remainder of this paper, Section 2 introduces the idea of "structurally lossless" compression. Section 3 discusses content-based retrieval applications of texture similarity. A review of structural texture similarity metrics is presented in Section 4, while Section 5 discusses color similarity metrics.

2. STRUCTURALLY LOSSLESS IMAGE COMPRESSION

Even though storage capacity and transmission bandwidth have been growing, the rapidly increasing amount of visual data and the demand for higher quality and resolution are expected to far outweigh the gains in storage and bandwidth, thus making it imperative to seek higher degrees of compression. On the other hand, the state of the art in image and video compression is quite advanced, reaching the limit of what is possible with conventional approaches, and making it increasingly difficult to further squeeze the bit rate. However, image and video compression is still far from approaching the efficiency of the human brain in storing visual information. Thus, any further gains in compression efficiency will have to come from a better understanding of human perception.

Depending on the application (i.e., quality and bandwidth constraints), image compression ranges from lossless compression, to mathematically lossy but *perceptually lossless* compression, to visually lossy compression. By perceptually lossless, we mean that the compressed image is visually indistinguishable from the original image in a side-by-side comparison. The development of perceptually lossless compression techniques and associated perceptual similarity metrics [33,34] was a significant advance over previous techniques that made only implicit use of the HVS characteristics. They are based on the notion of just noticeable distortion (JND), and typically require extensive subjective tests to establish the associated thresholds and metric parameters.[34]

However, in many applications severe bandwidth limitations dictate the need for further compression. With present methods, this can be done at the expense of severe compression artifacts or a significant reduction in image resolution. For example, Ref. 35 discusses spatial resolution and quantization noise tradeoffs for scalable image compression. Hemami *et al.* conducted systematic studies to quantify perceptual distortion in suprathreshold (visible artifacts) applications. [36–39] However, rather than allowing the encoder to produce the bit rate by introducing distortions that result in significant reductions in visual quality, an alternative is to ask the encoder to produce a reduced bit representation with essentially the same visual quality as the original, by allowing the encoder to make substantial point-by-point changes that do not change the visual quality of the image. Such changes may be perceptible when the original and decoded images are viewed side-by-side, but are not noticeable when the reproduction is viewed by itself. In fact, the quality of the two images should be comparable, so that and it should not be obvious which image is the original.

We refer to this new image compression goal as *structurally lossless compression*, and it is motivated from human perception. In order to achieve high compression efficiency, the human brain makes significant compromises. For example, rather than storing a point-by-point accurate matrix of image intensities, as conventional imaging systems and compression algorithms do, it is well known that the human brain stores images in symbolic form,[40] thus allowing substantial modifications in an image before they are detectable by the eye.

One approach to achieving structurally lossless compression is by allowing significant point-by-point variations in textured areas. This can result in significant compression gains by utilizing spatial and temporal prediction of texture, which would not be allowed by conventional metrics, which typically result in high point-by-point prediction errors for perceptually indistinguishable textures, due to the stochastic nature of texture variations. In addition to the availability of a good texture similarity metrics, the success of this approach depends on texture blending techniques [26] that can eliminate image stitching artifacts. To illustrate the potential success of texture prediction, Fig. 1 shows the image "Baboon," which has a lot of texture, and shows it again with 24% of its pixels replaced by blocks of similar texture. One can see that it is very hard to distinguish the two images, unless they are magnified or are shown in quick temporal succession.

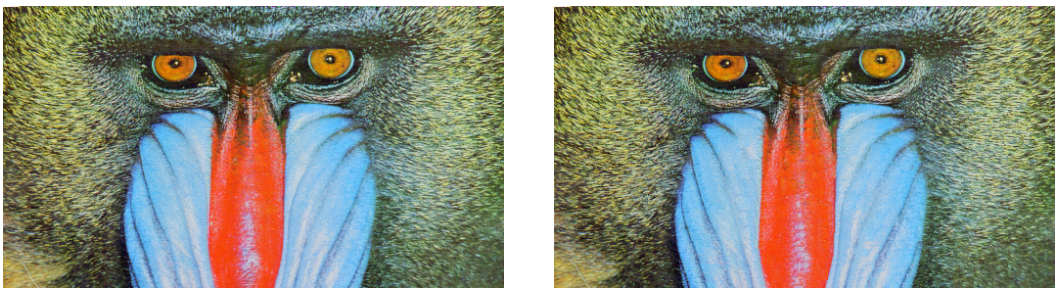


Figure 1. A 288×512 section of original "Baboon" image (left) and texture synthesized version (right) with 24% of the pixels between rows 65 and 288 replaced by pixels from previous blocks.

As we discussed in the introduction, prediction is the key to the success of video compression algorithms, which use sophisticated motion compensation techniques. Indeed, it is primarily improvements in motion-compensated prediction, rather than transform coding of the residual (which has been considerably simplified) that have enabled the substantial gains in compression efficiency offered by the new H264 video compression standard [41] relative to the earlier standards such as H263[42] and MPEG-2.[43,44] The development of texture similarity metrics, in combination with texture blending techniques, may lead to a dramatic increase in spatial and temporal predictions, thus resulting in significant gains in compression efficiency.

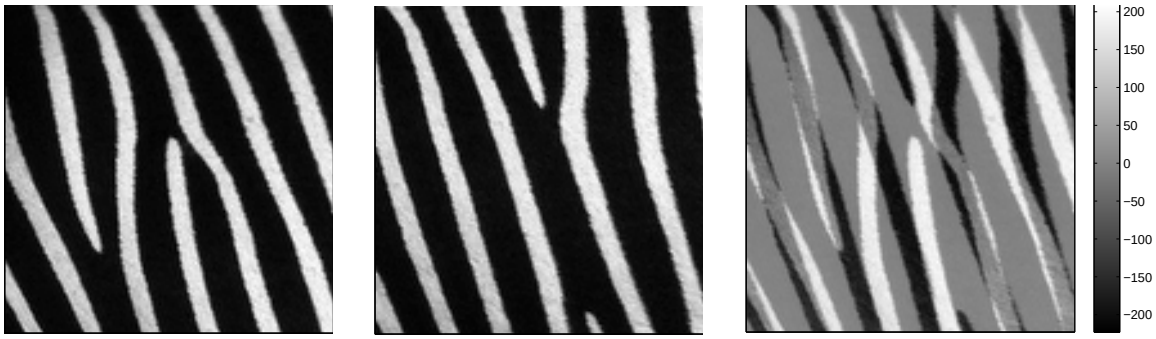
3. CONTENT-BASED RETRIEVAL

As we discussed in the introduction, image classification and retrieval is critically dependent on texture analysis of image segments or patches in order to classify them into semantic categories or to compare them with query textures. Depalov *et al.*,[9–13] proposed region-wide color and texture features and classification techniques for obtaining semantic labels for each segment. An alternative approach for segment classification is the direct comparison of the segment textures with reference textures. This relies on texture similarity metrics, which is the main focus of this paper.

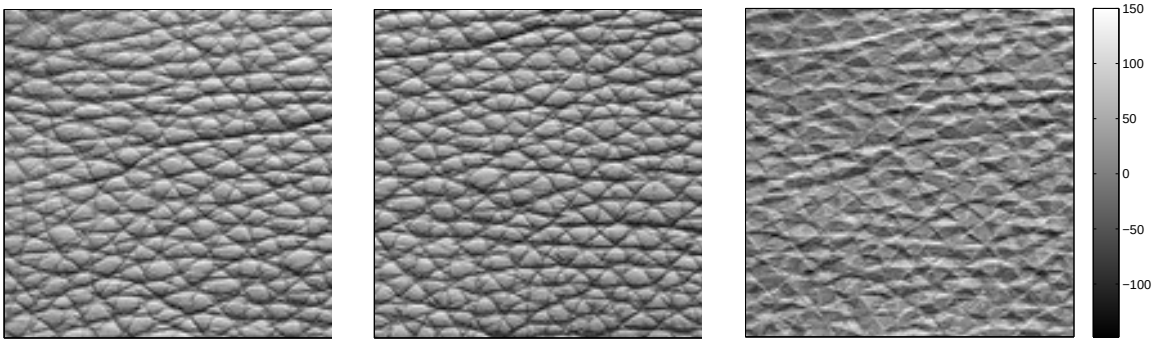
As we mentioned in the introduction, different applications impose different requirements on metric performance. There are several cases to consider in the context of content-based retrieval. For example, it may be important to retrieve all similar images, and to discard the dissimilar ones. In such a case, the precise ordering of the textures is not important, as long as there is a clear distinction between similar and dissimilar textures. Thus, Zhao *et al.* [29] judge a metric by its ability to discriminate between similar and dissimilar texture pairs. Thus, they consider a metric to be good if the values it assigns to similar pairs of textures are consistently above the values it assigns to dissimilar textures. In other words, there is an interval such that metric values above the higher edge of the interval indicate similarity and below its lower edge indicate dissimilarity. The wider the interval the better the metric. Zhao *et al.* [29] showed that STSIM, using the proposed performance metric, outperforms CW-SSIM and a few of its variants. They reported also that SSIM and PSNR performed poorly compared to their proposed metric.

Another more commonly used approach is to correlate the metric predictions with subjective assessments. This is of course better suited for visually lossy compression applications, where it is important to quantify the amount of distortion. In this case, it is suitable to use a linear correlation coefficient, since it captures how well the distances between human judgments are preserved in the metric values. If, on the other hand, we are only interested in *ranking* the results and not in the absolute scale, we may use the Spearman's rank or Kendall's tau correlation coefficient that capture how well the results are ordered. In[30] we found that this type of analysis does not yield very meaningful results when the test data includes a lot of dissimilar texture pairs. In such cases, it is difficult even for humans to quantify texture similarity. If quantitative data of this type is necessary, then it makes sense to limit it to comparisons between similar textures, and to ignore variations in metric predictions and human comparisons below a certain threshold. The results of such analysis are reported in the work of Zujovic *et al.*,[30] where it is shown that including a broader set of statistics in the similarity metrics improves the correlation coefficients, both linear and rank ones, even though the overall metric (as well as human) performance is not satisfactory in consistently ordering pairs of texture patches on the basis of their similarity.

As we discussed, another type of search is the "known-item" search,[31] whereby one is interested only in exact matches, that is, samples of the same texture. One advantage of this approach is that the ground truth is known, and therefore



(a) Zebra Z1 (b) Zebra Z2 (c) Z1 − Z2 (PSNR = 7.99dB)



(d) Leather L1 (e) Leather L2 (f) L1 − L2 (PSNR = 14.79dB)

Figure 2. Illustration of inadequacy of PSNR metrics

no subjective tests are required. Of course, this is true to the extent that the texture from which the "identical" pieces are obtained is perceptually uniform. For the known-item search, a number of statistical measures have been developed by the text retrieval community. *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.[45] *Mean average precision*[46] and *precision-recall* plots[47] are also commonly used. Zujovic *et al.*[32] performed extensive experiments for this type of application, using nearly 280,000 pairs of texture images, to assess the performance of their metric to the well known SSIM and CW-SSIM metrics, and demonstrated that it offers considerable advantages.

Finally, in evaluating the similarity of two textures, one has to take into account both the color composition and the spatial texture patterns. In Ref. 30, Zujovic *et al.* proposed separate metrics for grayscale texture similarity and color composition, and then combined them into a single metric. However, their 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.

In the next section, we review structural similarity metrics and their adaptations for texture analysis that were proposed in Refs. 29, 30, and 32 in more detail.

4. REVIEW OF STRUCTURAL SIMILARITY METRICS

The Structural Similarity Metric (SSIM)[27] introduced a new approach for assessing image similarity that focuses on high-level properties of the human visual system (HVS). In contrast to utilizing explicit models of HVS characteristics, as traditional perceptual quality metrics do,[34] it captures the HVS properties implicitly. Multiscale frequency decompositions like Gabor filters can be incorporated, and result in significant performance improvements. The inadequacy of point-by-point metrics, like PSNR, is illustrated in Figure 2.

The idea is to compare statistics of images in a way that would adapt to luminance changes, performs contrast masking and takes into account other high-level HVS characteristics. The images can be analyzed as a whole or on a sliding windows basis, in which case the statistics are compared in corresponding windows. In the remainder of the paper, we will assume that the analysis is done in sliding windows. Also, the analysis can be performed in the image domain or in the transform domain. When complex wavelets are used as transform bases, the Complex Wavelet Structural Similarity Metric (CW-SSIM)[28] is obtained. Like Gabor filters, steerable pyramid filters are inspired by biological visual processing, and have nice properties, such as translation and rotation invariance.[48] In all cases, the idea is to compare the luminance, contrast, and "structure" of image patches (windows), in the image or transform domain, in order to obtain a unique similarity score based on these terms. Luminance is characterized by the mean of intensities within each window, contrast is characterized by the standard deviation, and structure is characterized by the cross-correlation between corresponding windows in the two images being compared.

Let us first introduce notation that will be used throughout the paper:

- x and y are images we want to compare.
- i, j are the spatial indices of the pixel values (or coefficients, in transform domain)
- k, l denote the subband indices, when subband analysis is used.
- For the various variables, the superscript denotes the subband and the subscript denotes the image.
- For images or subbands, the subband index is indicated in the subscript, followed by the spatial indices.
- All variables with names of the form C_n are small constants.
- N is the number of pixels in the window under consideration.
- $E\{\cdot\}$ denotes empirical average.

Assuming we are working in subband domain, the following equations describe necessary computations:

$$\mu_x^k = E\{\mathbf{x}_k\} = \frac{1}{N}\sum_{i,j}\mathbf{x}_{k,i,j} \tag{1}$$

$$\sigma_x^k = \sqrt{E\{(\mathbf{x}_k - \mu_x^k)^2\}} = \sqrt{\frac{1}{N-1}\sum_{i,j}(\mathbf{x}_{k,i,j} - \mu_x^k)^2} \tag{2}$$

$$\sigma_{x,y}^k = E\{(\mathbf{x}_k - \mu_x^k)\cdot(\mathbf{y}_k - \mu_y^k)\} = \frac{1}{N-1}\sum_{i,j}(\mathbf{x}_{k,i,j} - \mu_x^k)\cdot(\mathbf{y}_{k,i,j} - \mu_y^k) \tag{3}$$

The luminance term is defined as:

$$l^k(\mathbf{x},\mathbf{y}) = \frac{2\mu_x^k\mu_y^k + C_0}{(\mu_x^k)^2 + (\mu_y^k)^2 + C_0}, \tag{4}$$

the contrast term is defined as:

$$c^k(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x^k\sigma_y^k + C_1}{(\sigma_x^k)^2 + (\sigma_y^k)^2 + C_1}, \tag{5}$$

and the structure term is defined as:

$$s^k(\mathbf{x},\mathbf{y}) = \frac{\sigma_{x,y}^k + C_2}{\sigma_x^k\sigma_y^k + C_2}. \tag{6}$$

For each position of the sliding window, a similarity value is computed as:

$$Q_{SSIM}^k(\mathbf{x},\mathbf{y}) = l_k(\mathbf{x},\mathbf{y})^\alpha c^k(\mathbf{x},\mathbf{y})^\beta s^k(\mathbf{x},\mathbf{y})^\gamma. \tag{7}$$

Usually, the parameters are set to be $\alpha = \beta = \gamma = 1$ and $C_2 = C_1/2$ to get:

$$Q_{SSIM}^k(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x^k\mu_y^k + C_0)(2\sigma_{x,y}^k + C_1)}{((\mu_x^k)^2 + (\mu_y^k)^2 + C_0)((\sigma_x^k)^2 + (\sigma_y^k)^2 + C_1)}. \tag{8}$$

Typically, SSIM is evaluated in small sliding windows (e.g., 7×7), and the final SSIM metric is computed as the average of $Q^k_{SSIM}(x, y)$ over all spatial locations and all subbands.

The main advantage of SSIM, and when extended to complex wavelet domain, the CW-SSIM, is that it moves away from point-by-point comparisons, attempting to capture structural differences. On the other hand, the structure term, from which the metric takes its name, is actually a *point-by-point* comparison.[29] Therefore, Zhao et al.[29] proposed removing the structure term s in (6), and adding other subband statistics to account for texture characteristics, namely, the first-order autocorrelation coefficients of transform coefficients in the horizontal and vertical directions. Even directional information is implicitly accounted for when the statistics of the different orientation subbands are computed, the argument is that directional information within each subband can be exploited in order to improve the performance of the metric.

The first-order autocorrelation term in the horizontal direction is defined as:

$$\rho^k_x(0,1) = \frac{E\{(x_{k,i,j} - \mu^k_x)(x_{k,i,j+1} - \mu^k_x)\}}{(\sigma^k_x)^2} \tag{9}$$

and in an analogous manner, in the vertical direction:

$$\rho^k_x(1,0) = \frac{E\{(x_{k,i,j} - \mu^k_x)(x_{k,i+1,j} - \mu^k_x)\}}{(\sigma^k_x)^2} \tag{10}$$

The values for the autocorrelation are bounded and lie in the interval $[-1, 1]$. To compare the autocorrelations of a particular subband in the two windows and obtain values in the $[0, 1]$ range, with 1 representing the best possible match, it was found that the following works well:

$$c^k_{0,1}(\mathbf{x}, \mathbf{y}) = 1 - 0.5|\rho^k_x(0,1) - \rho^k_y(0,1)| \tag{11}$$

$$c^k_{1,0}(\mathbf{x}, \mathbf{y}) = 1 - 0.5|\rho^k_x(1,0) - \rho^k_y(1,0)| \tag{12}$$

For each sliding window in each subband, the previously defined l in (4) and c (5) terms are combined with the new ones into the Structural Texture Similarity Metric (STSIM) as:

$$Q^k_{\text{stsim}}(\mathbf{x}, \mathbf{y}) = \left(l^k(\mathbf{x}, \mathbf{y})\right)^{\frac{1}{4}} \left(c^k(\mathbf{x}, \mathbf{y})\right)^{\frac{1}{4}} \left(c^k_{0,1}(\mathbf{x}, \mathbf{y})\right)^{\frac{1}{4}} \left(c^k_{1,0}(\mathbf{x}, \mathbf{y})\right)^{\frac{1}{4}} \tag{13}$$

Zhao *et al.* proposed two approaches for combining the results obtained for each window and each subband. One approach is "additive," whereby the total STSIM is calculated in the same manner as SSIM, taking the mean over spatial locations in each subband, and then taking the mean across frequencies. The other approach is "multiplicative," whereby corresponding STSIM values for each spatial location get multiplied across the subbands, and then the final metric is the spatial mean of these multiplied coefficients.

Zujovic *et al.*[30] extended the ideas of Ref. 29 by including terms that compare cross-correlation terms between subbands (STSIM-2). This was also motivated by the work of Portilla and Simoncelli,[23] who base the justification for the use of coefficient correlations within subbands on the fact that the steerable filter decomposition is overcomplete and on the existence of periodicities in the textures.

They also argued that, while raw coefficients may be uncorrelated, the coefficients magnitudes are *not* statistically independent, and large magnitudes in 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.[48] Therefore, while the luminance, contrast and autocorrelation terms in (4), (5), (11) and (12) are calculated on the *raw* subband coefficients, the cross-correlation statistics are computed on the *magnitudes.*

In STSIM-2, for a fixed orientation, the cross-correlations are computed between the magnitudes of subband coefficients at *adjacent* scales, and for a fixed scale, the cross-correlations are computed between the subband magnitudes of *all possible* orientations.

The covariances between the coefficient magnitudes at subbands k and l are normalized by the variances of the two subbands to obtain the cross-subband correlation coefficient:

$$\rho_x^{k,l}(0,0) = \frac{E\{(|x_{k,i,j}| - \mu_x^k)(|x_{l,i,j}| - \mu_x^l)\}}{\sigma_x^k \sigma_x^l} \tag{14}$$

where $|x_{k,i,j}|$ and $|x_{l,i,j}|$ are the magnitudes of the coefficients of subbands k and l, respectively, and μ_x^k and μ_x^l are the corresponding means of the magnitudes in the window. The expected value is an empirical average over the window.

Since the cross-subband correlation coefficients take values in the interval $[-1, 1]$, they are compared as in (11) to obtain a statistic that describes the similarity between the cross-correlations:

$$c_{0,0}^{k,l}(\mathbf{x}, \mathbf{y}) = 1 - 0.5\left(|\rho_x^{k,l}(0,0) - \rho_y^{k,l}(0,0)|\right)^p \tag{15}$$

Typically, $p = 1$. Note that the $c_{0,0}^{k,l}(\mathbf{x}, \mathbf{y})$ values are in the interval $[0, 1]$, just like the STSIM terms.

For a steerable pyramid with N_s scales and N_o orientations, we have a total of $N_i = N_s \cdot N_o + 1$ subband images (including the highpass but not the lowpass). For each of these subbands, the STSIM maps are computed as in (13). We also have M maps with the new statistics, based on (15). The $N_t = N_i + M$ matrices are then combined additively

$$Q_t(\mathbf{x}, \mathbf{y}) = \frac{1}{N_t}\left(\sum_k Q_{\mathrm{stsim}}^k(\mathbf{x}, \mathbf{y}) + \sum_{k,l} c_{0,0}^{k,l}(\mathbf{x}, \mathbf{y})\right) \tag{16}$$

or multiplicatively to obtain a single similarity matrix. Finally, spatial summation over the matrix values gives a single value for the similarity metric.

We have tested our most recently proposed metric (STSIM-2) in two of the aforementioned settings: comparing to subjective tests and in the known-item search scenario. For the first case, our results were reported in Ref. 30; here, we briefly discuss the Spearman rank correlation coefficients. The Spearman's rank correlation coefficient for PSNR was 0.283, while SSIM performed considerably better, with 0.515; the CW-SSIM and STSIM metrics achieved even better results with 0.579 and 0.598, respectively. STSIM-2 had the best overall performance with 0.659. This was obtained with weights $w_t = 0.6$ and $w_c = 0.4$. However, as found by Markov *et al.*,[49] these optimal weights vary for different datasets and cannot be universally chosen.

To give intuitive meaning to these correlation coefficients, we also tested how humans performed against humans: for each of the human subjects, we removed their judgments from the pool, and computed the mean grades of the remaining subjects; then, we conducted the Spearman rank correlation tests, and, to be fair, we recalculated the correlation coefficients between STSIM-2, and the same mean grades of the remaining subjects. The mean value of the correlations of human judgments against one human was 0.794, as compared to the value of 0.661 for STSIM-2. As mentioned earlier, this subjective test has shown that humans are inconsistent with grading similarity on the given dataset, and we would need to redesign the tests to achieve better benchmark subjective ratings.

For the known-item search, comprehensive results are reported in Ref. 32. The conclusion was that STSIM-2 performs best, with 77.2% success in retrieving the correct document as the first returned result (precision at one), whereas for PSNR this happens only for 6% of the images. SSIM has a slightly better performance, with 8% success rate, while CW-SSIM gives considerably better results with 63.6%. Mean average precision results also show a clear advantage for STSIM-2 (MAP = 0.75) over PSNR (0.095) ad SSIM (0.06), and a definite improvement over CW-SSIM, whose MAP is equal to 0.62.

5. COLOR SIMILARITY METRICS

Color is perhaps the most expressive of all the visual features and has been extensively studied in the image retrieval research during the last decade.[50] Determining color similarity or quality of color images is application-dependent, and different objectives call for different approaches. In image compression applications, a straightforward approach for extending an image quality metric to color is to apply the grayscale metric to each of three color components in a trichromatic space. This approach is suitable for lossless and visually lossy compression algorithms; for perceptually and structurally lossless coding applications, the methods and metrics must be adjusted for the way humans perceive color, which may be different from methods applicable in grayscale image analysis. An alternative approach that may be more effective in image retrieval applications and perceptually and structurally lossless coding, is to use separate metrics for comparing the

grayscale textures and the color composition of an image, and then to use them separately or to combine them in order to obtain one number. We will adopt this approach, since it is arguably more appropriate for our target applications.

The simplest approach for describing and comparing the color composition of images is to use color histograms and simple histogram intersection metrics[51] or a more sophisticated color quadratic distance. [52] However, as shown in Refs. 50, 53, the human visual system cannot simultaneously perceive a large number of colors. People, in fact, see only a few prominent colors in a given image, which are typically referred to as *dominant* colors. Thus, the color descriptors have moved away from direct histogram acquisitions to more compact color descriptors, as in Refs. 6, 54, 55.

In addition, comparing the dominant colors of two images has to be trusted to more sophisticated techniques that account better for the HVS properties. One of the best known techniques is the earth mover's distance (EMD). [56] EMD is based on the minimal cost that must be paid to transform one color distribution into the other; informally speaking, EMD measures how much work needs to be applied to move earth distributed in piles p_x so that it turns into the piles p_y.

An approach that follows the same philosophy as EMD is Optimal Color Composition Distance (OCCD) developed by Mojsilovic *et al*.[53] In this case, the color composition descriptors are the extracted dominant colors and their respective percentages. OCCD is an approximation of EMD in the sense that it quantizes color percentages into *units*, thus transforming the linear optimization problem of EMD into the weighted graph matching problem, which is solvable by deterministic algorithms. The chosen color space is *CIELAB* (or *L*a*b**), since it exhibits approximate perceptual uniformity, in the sense that the Euclidean distance separating two similar colors is proportional to their visual difference. [57]

Since the appearance of an image is best described by the spatial distribution of features, rather than by individual feature vectors,[58] we are utilizing a sliding windows approach to assess color similarity, just like we did for texture. Given the fact that, as we stated earlier, people don't perceive a lot of different colors at the same time, and they perform local averaging (as opposed to noticing all the detailed variations of colors), the first step in extracting the color composition is image filtering. For this, we choose the adaptive clustering algorithm (ACA), [59] which accounts for spatial variations in the textures to segment the image into areas of uniform color. ACA can be used to perform local spatial averaging (using a small window, typically 7×7) within regions while preserving region boundaries. This is important because blurring across boundaries can create new colors that are not present in the image.

Since the OCCD computes the *distance* between two colors, their *similarity* can be rated as $1 - \mathrm{distance}$. Thus, as a similarity measure, we use the map $Q_c(\mathbf{x}, \mathbf{y}) = 1 - OCCD$. The mean of $Q_c(\mathbf{x}, \mathbf{y})$ map is taken as the color similarity measurement Q_c. The color texture similarity is determined on a sliding windows basis, thus producing a color similarity map, similar to those obtained for the grayscale texture.

As widely done in the literature,[60,61] we can then linearly combine the texture (Q_t) and color (Q_c) similarity measures with appropriate weights, , w_t and $w_c = 1 - w_t$, to obtain a final similarity metric:

$$Q_{total} = w_t \cdot Q_t + w_c \cdot Q_c \tag{17}$$

However, as we discussed above and as reported in Refs. 30 and 49, it is difficult to find the right weights for such a combined metric, as humans are not consistent in how they combine color and spatial texture information. Thus, in many applications it is best to keep the two metrics separate.

REFERENCES

[1] Brooks, A. C., Zhao, X., and Pappas, T. N., "Structural similarity quality metrics in a coding context: Exploring the space of realistic distortions," *IEEE Trans. Image Process.* **17**, 1261–1273 (Aug. 2008).

[2] Adelson, E. H., "On seeing stuff: The perception of materials by humans and machines," in [*Human Vision and Electronic Imaging VI*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **4299**, 1–12 (Jan. 2001).

[3] Rui, Y., Huang, T. S., and Chang, S.-F., "Image retrieval: Current techniques, promising directions and open issues," *J. Visual Communication and Image Representation* **10**, 39–62 (Mar. 1999).

[4] Smeulders, W. M., Worring, M., Santini, S., Gupta, A., and Jain, R., "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1349–1379 (Dec. 2000).

[5] Forsyth, D. A. and Ponce, J., [*Computer Vision - A Modern Approach*], Prentice-Hall (2002).

[6] Chen, J., Pappas, T. N., Mojsilovic, A., and Rogowitz, B. E., "Adaptive perceptual color-texture image segmentation," *IEEE Trans. Image Processing* **14**, 1524–1536 (Oct. 2005).

[7] Bae, S. and Juang, B.-H., "Incremental parsing for latent semantic indexing of images," in [*Proc. Int. Conf. Image Processing (ICIP-08)*], 925–928 (Oct. 2008).

[8] Bae, S. and Juang, B.-H., "Parsed and fixed block representations of visual information for image retrieval," in [*Human Vision and Electronic Imaging XIV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **7240**, 724017–1–12 (Jan. 2009).

[9] Depalov, D., Pappas, T. N., Li, D., and Gandhi, B., "Perceptually based techniques for semantic image classification and retrieval," in [*Human Vision and Electronic Imaging XI*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., **Proc. SPIE Vol. 6057**, 6057OZ–1–6057OZ–10 (Jan. 2006).

[10] Depalov, D., Pappas, T. N., Li, D., and Gandhi, B., "A perceptual approach for semantic image retrieval," in [*Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-06)*], **II**, 417–420 (May 2006).

[11] Depalov, D., Pappas, T. N., Li, D., and Gandhi, B., "Perceptual feature selection for semantic image classification," in [*Proc. Int. Conf. Image Processing (ICIP-06)*], 2921–2924 (Oct. 2006).

[12] Depalov, D. and Pappas, T. N., "Analysis of segment statistics for semantic classification of natural images," in [*Human Vision and Electronic Imaging XII*], Rogowitz, B. E., Pappas, T. N., and Daly, S. J., eds., **Proc. SPIE Vol. 6492**, 6492OD–1–6492OD–11 (Jan. 29 – Feb. 1 2007).

[13] Pappas, T. N., Chen, J., and Depalov, D., "Perceptually based techniques for image segmentation and semantic classification," *IEEE Commun. Mag.* **45**, 44–51 (Jan. 2007).

[14] Pappas, T. N., Tartter, V., Seward, A. G., Genzer, B., Gourgey, K., and Kretzschmar, I., "Preceptual dimensions for a dinamic tactile display," in [*Human Vision and Electronic Imaging XIV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **7240**, 72400K–1–12 (Jan. 2009).

[15] van Egmond, R., Lemmens, P., Pappas, T. N., and de Ridder, H., "Roughness in sound and vision," in [*Human Vision and Electronic Imaging XIV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **7240**, 72400B–1–12 (Jan. 2009).

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

[17] Porat, M. and Zeevi, Y. Y., "Localized texture processing in vision: Analysis and synthesis in gaborian space," *IEEE Trans. Biomed. Eng.* **36**(1), 115–129 (1989).

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

[19] Heeger, D. J. and Bergen, J. R., "Pyramid-based texture analysis/synthesis," in [*Proc. Int. Conf. Image Processing (ICIP-95), vol. III*], 648–651 (Oct. 1995).

[20] Portilla, J., Navarro, R., Nestares, O., and Tabernero, A., "Texture synthesis-by-analysis based on a multiscale early-vision model," *Optical Engineering* **35**(8), 2403–2417 (1996).

[21] Zhu, S., Wu, Y. N., and Mumford, D., "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling," in [*IEEE Conf. Computer Vision and pattern Recognition*], 693–696 (1996).

[22] De Bonet, J. S. and Viola, P. A., "A non-parametric multi-scale statistical model for natural images," *Adv. in Neural Info. Processing Systems* **9** (1997).

[23] Portilla, J. and Simoncelli, E. P., "A parametric texture model based on joint statictics of complex wavelet coefficients," *Int. J. Computer Vision* **40**, 49–71 (Oct. 2000).

[24] Simoncelli, E. P., Freeman, W. T., Adelson, E. H., and Heeger, D. J., "Shiftable multi-scale transforms," *IEEE Trans. Inform. Theory* **38**, 587–607 (Mar. 1992).

[25] Simoncelli, E. P. and Freeman, W. T., "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in [*Proc. ICIP-95, vol. III*], 444–447 (Oct. 1995).

[26] Efros, A. A. and Freeman, W. T., "Image quilting for texture synthesis and transfer," in [*Proc. 28th Intl. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH-01)*], 341–346 (Aug. 2001).

[27] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.* **13**, 600–612 (Apr. 2004).

[28] Wang, Z. and Simoncelli, E. P., "Translation insensitive image similarity in complex wavelet domain," in [*IEEE Int. Conference on Acoustics, Speech, and Signal Processing*], **II**, 573–576 (2005).

[29] Zhao, X., Reyes, M. G., Pappas, T. N., and Neuhoff, D. L., "Structural texture similarity metrics for retrieval applications," in [*Proc. Int. Conf. Image Processing (ICIP-08)*], 1196–1199 (Oct. 2008).

[30] Zujovic, J., Pappas, T. N., and Neuhoff, D. L., "Structural similarity metrics for texture analysis and retrieval," in [*Proc. Int. Conf. Image Processing*], (Nov. 2009). To appear.
[31] Meadow, C. T., Boyce, B. R., Kraft, D. H., and Barry, C., [*Text information retrieval systems*], Emerald Group Publishing (2007).
[32] Zujovic, J., Pappas, T. N., and Neuhoff, D. L., "Perceptual similarity metrics for retrieval of natural textures," in [*Proc. IEEE Workshop on Multimedia Signal Processing*], (Oct. 2009).
[33] Eckert, M. P. and Bradley, A. P., "Perceptual quality metrics applied to still image compression," *Signal Processing* **70**, 177–200 (1998).
[34] Pappas, T. N., Safranek, R. J., and Chen, J., "Perceptual criteria for image quality evaluation," in [*Handbook of Image and Video Processing*], Bovik, A. C., ed., 939–959, Academic Press, second ed. (2005).
[35] Bae, S., Pappas, T. N., and Juang, B.-H., "Spatial resolution and quantization noise tradeoffs for scalable image compression," in [*Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-06)*], **II**, 945–948 (May 2006).
[36] Hemami, S. S. and Ramos, M. G., "Wavelet coefficient quantization to produce equivalent visual distortion in complex stimuli," in [*Human Vision and Electronic Imaging V*], Rogowitz, B. E. and Pappas, T. N., eds., **Proc. SPIE Vol. 3959**, 200–210 (Jan. 2000).
[37] Ramos, M. G. and Hemami, S. S., "Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis," *J. Opt. Soc. Am. A* **18**, 2385–2397 (Oct. 2001).
[38] Chandler, D. M. and Hemami, S. S., "Additivity models for suprathreshold distortion in quantized wavelet-coded images," in [*Human Vision and Electronic Imaging VII*], Rogowitz, B. E. and Pappas, T. N., eds., **Proc. SPIE Vol. 4662**, 105–118 (Jan. 2002).
[39] Chandler, D. M. and Hemami, S. S., "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *J. Opt. Soc. Am. A* **20**, 1164–1180 (July 2003).
[40] B.Pelz, J., Canosa, R., Babcock, J., and Barber, J., "Visual perception in familiar, complex tasks," in [*Proc. Int. Conf. Image Processing (ICIP-01)*], **2**, 12–15 (Oct. 2001).
[41] Wiegand, T., Sullivan, G. J., Bjontegaard, G., and Luthra, A., "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.* **13**, 560–576 (July 2003).
[42] ITU-T Recommendation H.263, "Video coding for low bit rate communication," (Feb. 1998).
[43] ISO/IEC DIS 13818-2:2000, "Information technology – generic coding of moving pictures and associated audio information: Video," (2000).
[44] Haskell, B. G., Puri, A., and Netravali, A. N., [*Digital Video: An Introduction to MPEG-2*], Kluwer Academic (1996).
[45] Voorhees, E. M., "The trec-8 question answering track report," in [*In Proceedings of TREC-8*], 77–82 (1999).
[46] Voorhees, E. M., "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information Processing & Management* **36**, 697–716 (September 2000).
[47] Manning, C. D., Raghavan, P., and Schtze, H., [*Introduction to Information Retrieval*], Cambridge University Press, New York, NY, USA (2008).
[48] Portilla, J. and Simoncelli, E. P., "Texture modeling and synthesis using joint statistics of complex wavelet coefficients," in [*IEEE Workshop on Statistical and Computational Theories of Vision, Fort Collins*], (1999).
[49] Markov, I., Vassilieva, N., and A. Yaremchuk, A., "Image retrieval: Optimal weights for color and texture features combining based on query object," *Proceedings of RCDL* , 195–200 (2007).
[50] Manjunath, B. S., Ohm, J.-R., Vasudevan, V. V., and Yamada, A., "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.* **11**, 703–715 (June 2001).
[51] Swain, M. and Ballard, D., "Color indexing," *Int. J. Computer Vision* **7**(1), 11–32 (1991).
[52] S.Sawhney, H. and Hafner, J. L., "Efficient color histogram indexing," in [*Proc. ICIP-94*], **II**, 66–70 (Nov. 1994).
[53] Mojsilović, A., Hu, J., and Soljanin, E., "Extraction of perceptually important colors and similarity measurement for image matching, retrieval, and analysis," *IEEE Trans. Image Processing* **11**, 1238–1248 (Nov. 2002).
[54] Ma, W. Y., Deng, Y., and Manjunath, B. S., "Tools for texture/color based search of images," in [*Human Vision and Electronic Imaging II*], Rogowitz, B. E. and Pappas, T. N., eds., **Proc. SPIE, Vol. 3016**, 496–507 (Feb. 1997).
[55] Mojsilović, A., Kovačević, J., Hu, J., Safranek, R. J., and Ganapathy, S. K., "Matching and retrieval based on the vocabulary and grammar of color patterns," *IEEE Trans. Image Processing* **1**, 38–54 (Jan. 2000).
[56] Rubner, Y., Tomasi, C., and Guibas, L. J., "The earth mover's distance as a metric for image retrieval," *Int. Journal of Computer Vision* **40**(2), 99–121 (2000).

[57] Kasson, J. M. and Plouffe, W., "An analysis of selected computer interchange color spaces," *ACM Transactions on Graphics* **11**(4), 373–405 (1992).
[58] Rubner, Y., Puzicha, J., Tomasi, C., and Buhmann, J. M., "Empirical evaluation of dissimilarity measures for color and texture," *Computer Vision and Image Understanding* **84**, 25–43 (Oct. 2001).
[59] Pappas, T. N., "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Signal Processing* **SP-40**, 901–914 (Apr. 1992).
[60] Guerin-Dugue, A., Ayache, S., and Berrut, C., "Image retrieval: A first step for a human centered approach," *Proceedings 2003 Joint Conference of the Fourth Int. Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia.* **1**, 21–25 (Dec 2003).
[61] Markov, I. and Vassilieva, N., "Image retrieval: Color and texture combining based on query-image," in [*Image and Signal Processing*], *Lecture Notes in Computer Science* **5099**, 430–438, Springer (2008).