Perceptual Similarity Metrics for Retrieval of Natural Textures

Jana Zujovic [#1], Thrasyvoulos N. Pappas [#2], David L. Neuhoff [*3]

[#] *EECS Department, Northwestern University*
2145 Sheridan Road, Evanston, IL 60208, USA
[1] jana.zujovic@eecs.northwestern.edu
[2] pappas@eecs.northwestern.edu

[*] *EECS Department, University of Michigan*
1301 Beal Ave, Ann Arbor, MI 48109, USA
[3] neuhoff@umich.edu

***Abstract*—We investigate perceptual similarity metrics for the content-based retrieval of natural textures. The goal is to find perceptually similar textures that may have significant differences on a point-by-point basis. The evaluation of such metrics typically requires extensive and cumbersome subjective tests. The focus of this paper is on the recovery of textures that are "identical" to the query texture, in the sense that they are pieces of the same texture. This is important in content-based image retrieval (CBIR), where one may want to find images that contain a particular texture, as well as in some near-threshold coding applications. 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. We can thus compare the performance of different metrics on large sets of textures, and derive meaningful statistical results. We evaluate the performance of a recently proposed structural texture similarity metric on grayscale textures, and compare it to that of PSNR, as well as space domain and complex wavelet structural similarity metrics. Experimental results with a database of 748 distinct texture images, indicate that the new metric outperforms the other metrics in the retrieval of identical textures, according to a number of standard statistical measures.**

I. Introduction

Unlike traditional image quality metrics that evaluate the similarity between two images on a point-by-point basis, structural similarity metrics (SSIM) [1] can give high similarity scores even to images with significant pixel-wise differences. The goal is to assess the *perceived* similarity between two images and to allow deviations that do not affect the structure of the image. A number of applications can make use of such metrics, and each application imposes different requirements on metric performance. For example, in image compression it is important to provide a monotonic relationship between measured and perceived distortion, while in image retrieval applications it may be sufficient to distinguish between similar and dissimilar images, while the precise ordering may not be important. In some cases, it is important to have an absolute similarity scale, while in others a relative scale may be adequate. The focus of this paper is on natural textures, and in particular, on the recovery of textures that are "identical" to the query texture, in the sense that they are pieces of the same texture. This is important in content-based image retrieval (CBIR), where one may want to find images that contain a particular texture, as well as in some near-threshold coding applications. The problem of searching for (known) target documents in a database is known in information retrieval community as *known-item search* [2], and specific evaluation measures have been developed to assess the performance of such systems.

The evaluation of image similarity metrics, in general, requires extensive subjective tests, with several human subjects and a large number of image pairs. Depending on the performance requirements, a number of traditional statistical measures can be used for metric evaluation. For example, the Spearman's rank correlation coefficient and Kendall's tau rank correlation coefficient can be used when the relative performance is important [3], while linear regression can be used when absolute performance is needed. In [4], the performance criterion was whether a metric can distinguish between similar and dissimilar pairs, irrespective of the ordering within each group. In such a case, the greater the gap in metric values between similar and dissimilar pairs, the better the metric performance.

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 extend that the texture from which the "identical" pieces are obtained is perceptually uniform. Common measures for this type of retrieval systems 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 precision-recall plots.

In evaluating the similarity of two textures, one has to take into account both the color composition and the spatial texture patterns. In [3] we proposed a new structural texture similarity metric that separates the computation of similarity in terms of grayscale texture and color composition, and then

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00

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 [3]. Thus, for the present study, we focus only on grayscale textures, and compare the performance of the grayscale component of the metric proposed in [3] to that of PSNR, SSIM, complex wavelet SSIM (CW-SSIM) [5]. Experimental results with a database of 748 distinct texture images, extracted from 310 larger texture images, indicate that the new metric outperforms the other metrics in the retrieval of identical textures, according to all of the standard statistical measures we mentioned above.

The remainder of this paper is organized as follows. Section II provides a brief overview of similarity metrics. The structural texture similarity metric is reviewed in Section III. The experimental setup and results are given in Section IV, while the final conclusions are drawn in V.

II. Background

Traditional quality metrics range from simple MSE and PSNR to more sophisticated metrics that incorporate low-level models of human perception [6] and are typically aimed at near-threshold applications such as perceptually lossless image compression. The idea is to ensure that the original and reconstructed images are perceptually indistinguishable. Traditional metrics evaluate image fidelity on a point-by-point basis. However, for supra-threshold applications, such as CBIR and perceptually lossy compression, we need metrics that can accommodate significant changes as long as the structure of the image is preserved. This was the primary motivation in the development of SSIMs [1], which allow non-structural contrast and intensity changes, and the CWSSIM [5], which also allows small translations, rotations, and scaling changes.

SSIM metrics, whether implemented in the space or wavelet domain, compare two images or image patches (windows) $\mathbf{x}$ and $\mathbf{y}$ by multiplicatively combining a number of terms. Here we assume that the metric is computed in a window of each subband. For the k-th subband, the *luminance* comparison term is

$$l^k(\mathbf{x},\mathbf{y}) = \frac{2\mu_x^k\mu_y^k + C_1}{(\mu_x^k)^2 + (\mu_y^k)^2 + C_1} \tag{1}$$

where μ_x^k and μ_y^k are the means of the two windows; the *contrast* comparison term is

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

where $(\sigma_x^k)^2$ and $(\sigma_y^k)^2$ are the variances of the two windows; and the *structure* term is

$$s^k(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy}^k + C_3}{\sigma_x^k\sigma_y^k + C_3} \tag{3}$$

where σ_{xy}^k is the covariance between the two windows and C_1, C_2, and C_3 are small constants. These terms are then combined to give a composite measure of structural similarity:

$$Q_{\text{ssim}}^k(\mathbf{x},\mathbf{y}) = l^k(\mathbf{x},\mathbf{y})^\alpha c^k(\mathbf{x},\mathbf{y})^\beta s^k(\mathbf{x},\mathbf{y})^\gamma \tag{4}$$

where α, β, and γ are positive weights, typically set to 1. The SSIM is typically evaluated in a small sliding window (e.g., 7×7), and the overall image similarity is obtained as the average over all spatial locations and all subbands.

As we saw above, one of the main thrusts in the SSIM approach is to move away from point-by-point comparisons, and instead, to base the comparisons on region statistics. In an attempt to fully embrace this philosophy, Zhao *et al.* [4] replaced the structure term – which in spite of its name is in fact a point-by-point comparison – with terms that depend on region statistics. They introduced terms that compare the first order correlation coefficients (autocovariance normalized by the variance) in the horizontal $\rho_x^k(0,1)$ and vertical $\rho_x^k(1,0)$ directions as follows:

$$c_{0,1}^k(\mathbf{x},\mathbf{y}) = 1 - 0.5\left(|\rho_x^k(0,1) - \rho_y^k(0,1)|\right)^p \tag{5}$$

The vertical term is defined similarly. Note that these comparison terms take values in the interval $[0,1]$, are symmetrical with respect to $\mathbf{x}$ and $\mathbf{y}$, and have a unique maximum. An additional advantage of eliminating the structure term is that the metric takes only positive values. As in [3], we will assume that $p = 1$.

To compute the overall value of the metric, the two images are decomposed into subbands using a steerable pyramid decomposition, and the statistics are computed, for each orientation and scale, within a small sliding window. The different terms are combined multiplicatively to obtain the similarity coefficient at each location and subband

$$Q_{\text{stsim}}^k(\mathbf{x},\mathbf{y}) = l^k(\mathbf{x},\mathbf{y})^{\frac{1}{4}} c^k(\mathbf{x},\mathbf{y})^{\frac{1}{4}} c_{0,1}^k(\mathbf{x},\mathbf{y})^{\frac{1}{4}} c_{1,0}^k(\mathbf{x},\mathbf{y})^{\frac{1}{4}} \tag{6}$$

Note that the exponents sum to one in order to normalize the metric values, so that metrics with different numbers of terms can be compared. The overall metric value is then calculated, either additively by averaging over all subbands and spatial locations, or multiplicatively by multiplying the coefficients of all subband and then averaging over all spatial locations. We refer to the metrics proposed by Zhao *et al.* in [4] as *structural texture similarity metrics (STSIM).*

III. Review of Structural Texture Similarity

The metric proposed in [3] extends the ideas of [4] by including a broader set of local image statistics. We will refer to this as *STSIM-2*, to distinguish it from the metrics in [4]. The motivation for these comes from the analysis/synthesis literature, and in particular, the work of Portilla and Simoncelli [7], who have shown that a broad class of textures can be synthesized using a set of statistical parameters that characterize the coefficients of a multiscale frequency decomposition.

As in [5], [4], STSIM-2 uses the complex-valued steerable filter decomposition of the grayscale component of the two images, which like Gabor filters, is inspired by biological vision and has nice properties, such as translation and rotation invariance. In the following, we use three scales ($N_s = 3$) and four orientations ($N_o = 4$).

In addition to the terms in (6), STSIM-2 uses terms that compare the cross-correlation between subbands. The luminance, contrast and autocorrelation terms in (1), (2), and (5) are calculated on the *raw* subband coefficients, while the cross-correlation statistics are computed on the *magnitudes.*

Note that all the subbands (except the low-frequency band) are *zero-mean* over the *whole* image; however, within small windows, e.g., 7×7, this is not necessarily true. Thus, the average is computed for each sliding window and is used in the variance calculation.

Portilla and Simoncelli [7] base the justification for the use of coefficient correlations within subbands on the fact that the steerable filter decomposition is overcomplete and the existence of periodicities in the textures. They also argue 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 STSIM-2 metric, for each orientation, computes the cross-correlations between the magnitudes of subband coefficients at adjacent scales, and for each scale it computes the cross-correlations between the subband magnitudes of all orientations. Thus, for the 3-scale, 4-orientation decomposition, we have $\binom{4}{2} = 6$ coefficients for each scale, and 2 coefficients for each orientation, for a total of $M = 3 \cdot 6 + 4 \cdot 2 = 26$ new terms. In [3], it was found that the best performance is obtained when all 26 coefficients are used.

The cross-correlations 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

$$\rho_x^{k,l}(0,0) = \frac{E\{(|x_{k,i,j}| - \mu_{x_k})(|x_{l,i,j}| - \mu_{x_l})\}}{\sigma_{x_k}\sigma_{x_l}} \qquad (7)$$

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.

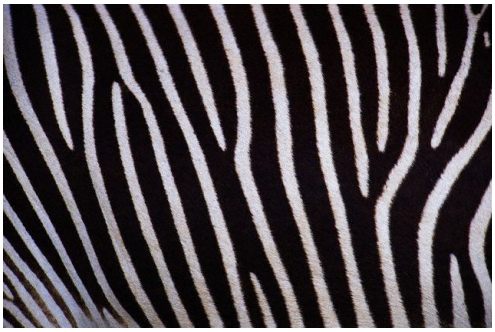
Since the cross-subband correlation coefficients take values in the interval $[-1, 1]$, they can be compared as in (5) to obtain a statistic that describes the similarity between the cross-correlations:

$$c_{0,0}^{k,l}(\mathbf{x}, \mathbf{y}) = 1 - 0.5\left(|\rho_x^{k,l}(0,0) - \rho_y^{k,l}(0,0)|\right)^p \qquad (8)$$

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 $N = N_s \cdot N_o + 2$ subband images, including the highpass and the lowpass subband. (In [3], the lowpass was omitted.) For each of these subbands, we compute the STSIM maps as in (6). We also compute M maps with the new statistics, based on (8). The $N_t = N + M$ matrices are then be combined additively

$$Q_t(\mathbf{x}, \mathbf{y}) = \frac{1}{N_t}\left(\sum_k Q_{\text{stsim}}^k(\mathbf{x}, \mathbf{y}) + \sum_{k,l} c_{0,0}^{k,l}(\mathbf{x}, \mathbf{y})\right) \qquad (9)$$



(a) Original Image

(b) Subimage 1 (c) Subimage 2 (d) Subimage 3

Fig. 1. Extraction of images from an original texture image

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

IV. Experimental Results

A. Database

To construct our database, we downloaded approximately 300 images with perceptually uniform texture from the Corbis website [8]. From each image, we extracted between two and five smaller subimages of size 128×128. The size was small enough to make it possible to extract at at least two subimages from each original original image without significant overlap between the subimages. At the same time, it was large enough to be able to capture several scales of texture. The amount of overlap between the subimages depends on the size of the original image, but in any case, the average mean-squared error (MSE) between two subimages from the same original texture image is quite high (average PSNR is about 15dB). This is illustrated in Figure 1, where we have the original image and three subimages with partial overlap. In similar fashion, we obtained a total of 748 texture images for this experiment. From now on, the term "image" will refer to the subimage, and "original image" to the original, larger image.

B. Experiment setup and Results Evaluation

For the experiment, we used all possible pairs of images, which is slightly less than 280,000 combinations, and calculated the metric values for PSNR, SSIM, CW-SSIM and STSIM-2. This resulted in four 748×748 tables of similarity scores; note all metrics are symmetric. The image labels, which reflect the lineage from an original image, constitutes the ground truth for evaluation of our results.

To evaluate the performance of these metrics in the context of retrieving identical textures, we can look at a number

of statistics. One informative measure of performance is the number of times the first retrieved image has the same lineage as the query, i.e., comes from the same original image. This is commonly referred to as *precision at one*. For PSNR, this happens only in 45 cases, or 6% of the textures. SSIM has a slightly better performance, with 8% success rate or 60 images. CW-SSIM gives considerably better results with 63.6% success rate or 476 cases of correct first match. The precision of the proposed metric is 77.2%, which corresponds to 581 successful queries out of a total of 748.

Another way of assessing the performance of the various metrics is to compute the *mean reciprocal rank (MMR),* i.e., the average value of the inverse rank of the first correctly retrieved image [9]. Again, the PSNR and SSIM metrics perform poorly, with MRRs of 0.1 and 0.11, respectively. The MRR for CW-SSIM is 0.71 and for the STSIM-2 metric it is 0.83.

Note that, since in many cases we extracted more than two patches from the original image, there is more than one correct answer for each query. In such cases, the usual value to report is *mean average precision (MAP)*[10]. The MAP is computed as follows: for each query image *i*, we order the remaining 747 images according to (descending) similarity; then, we calculate the precision for all possible numbers of retrieved documents: the first one, the first two, etc., all the way to 747 images. We then average the precisions of the sets of retrieved images for which the last retrieved image was identical to the query. Finally, we average these values across all images. If we define the number of images that come from the same bigger image as image *i* with n_i (not including image *i*, i.e., n_i is the maximum number of possible correctly retrieved images for query *i*), and if, for the ranked results we define the indicator function as $I_i(j) = 0$ if the j^{th} retrieved image does not come from the same original, and $I_i(j) = 1$ if it does, and if we define $p_i(r)$ the precision of retrieval if we make the cut-off for returning results at the r^{th} retrieved image, average precision is defined as:

$$P_{avg,i} = \frac{1}{n_i} \sum_{r=1}^{747} p_i(r) \cdot I_i(r) \tag{10}$$

with precision at cut-off *r* being

$$p_i(r) = \sum_{j=1}^{r} \frac{I_i(j)}{r}. \tag{11}$$

Therefore, the mean average precision is:

$$\text{MAP} = \underset{i}{\text{mean}}\, P_{avg,i} \tag{12}$$

In this experiment, the MAP for PSNR was 0.095, for SSIM 0.06, for CW-SSIM 0.62 and for the STSIM-2 metric 0.75.

Finally, we generated the precision-recall plots, shown in Figure 2. For the method of generating them, we refer the readers to [11] for a comprehensive explanation.

Given these results, we can conclude that the performance of PSNR and SSIM does not agree with human perception. This should be just as expected because the point-by-point metric

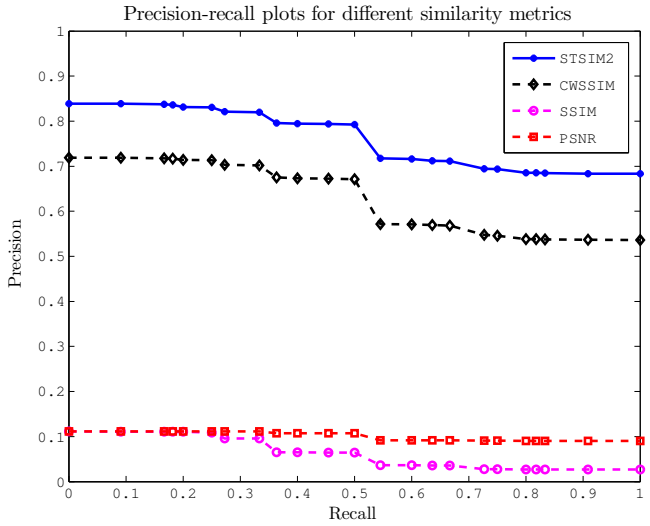


Fig. 2. Precision-recall plots

calculations do not match the statistical nature of the CBIR problem. The CW-SSIM metric, on the other hand, performs fairly well, mainly because it can tolerate small translations, rotations, and scaling changes. However, the STSIM-2 metric outperforms CW-SSIM by a significant margin in all of the reported statistics, and therefore, we can safely claim that it is the best metric for this task.

C. Points of Failure and Future Research

It is interesting to comment on the cases in which the proposed metric failed. One of the most common cases in which the metric assigns a high similarity value to two images that come from different original textures is when the two textures are quite similar except for the difference in mean gray level. That's because the gray level doesn't count much in the proposed metric, and the differences in the pattern outweigh the differences in gray level. An example is shown in Figure 3. Note that the image on the right has the right gray level, but the orientation of the lines seems to be more at odds with the query image than that of the pattern in the middle. Indeed, we consider there to be considerable perceptual similarity between the query and best match in this example, so this type of failure is understandable from a human perceptual point of view. If mean gray value is important, the problem can be remedied by introducing a new term that compares the local average grayscale values directly in the image domain. This term could be easily combined with the transform-based terms to produce the desired result. However, the intention of the original SSIM metrics, and the reason for (1), was to deemphasize differences in average luminance.

Another group of failures comes from images with strong texture elements, such as the tiled walls shown in Figure 4, but with different colors. This could be explained by the large variation of colors in the true match, thus bringing the more uniformly colored texture as a closer match. Related to this is the example shown in Figure 5, where it appears that the

two textures are similar enough to override differences due to spatial shifts in the identical texture samples.

A large part of failures come from the fact that images in our database have different scales. This can be seen in Figure 6, where the query image is a texture at a large scale and the retrieved image at a smaller scale. 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.

Finally, the failure shown in Figure 7 is due to the fact that the metric cannot adequately discriminate texture orientations, while the failure in Figure 8 is due to the metrics inability to identify periodic textures.

V. Conclusions

We examined the performance of a perceptual structural similarity metric (STSIM-2) for the retrieval of images of natural textures. The metric relies on subband decomposition of grayscale images, and the local statistics computed on small sliding windows. Our focus was on the recovery of textures that are "identical" to the query texture, in the sense that they are pieces of the same texture. This eliminates the need for cumbersome subjective tests. We used a large database of 748 distinct texture images, and performed pair-wise calculations for all possible pairs. We compared the performance of the STSIM-2 metric to that of PSNR, SSIM and CW-SSIM on all texture pairs, using standard statistical measures (precision at one, mean reciprocal rank, mean average precision, and precision-recall plots). We have shown that the STSIM-2 metric outperforms the other metrics in this experiment. The remaining drawbacks of the metric include problems with the mean gray value of the textures, texture periodicities, and scale. These will be addressed in our future research.

Fig. 3. Failed example: query image (left), first best match (middle), first "identical" match (right)

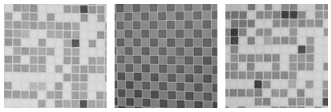


Fig. 4. Failed example: query image (left), first best match (middle), first "identical" match (right)

Fig. 5. Failed example: query image (left), first best match (middle), first "identical" match (right)

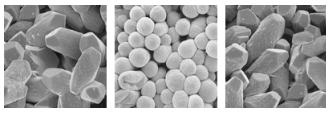


Fig. 6. Failed example: query image (left), first best match (middle), first "identical" match (right)

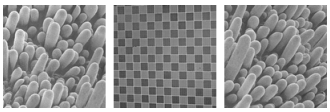


Fig. 7. Failed example: query image (left), first best match (middle), first "identical" match (right)

Fig. 8. Failed example: query image (left), first best match (middle), first "identical" match (right)

References

[1] 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. [Online]. Available: http://www.uta.edu/faculty/zhouwang/publications/ssim.html

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

[3] 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, submitted.

[4] 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-08)*, San Diego, CA, Oct. 2008, pp. 1196–1199.

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

[6] 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.

[7] 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.

[8] "Corbis stock photography," http://www.fotosearch.com/corbis/.

[9] E. M. Voorhees, "The trec-8 question answering track report," in *In Proceedings of TREC-8*, 1999, pp. 77–82.

[10] ——, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information Processing & Management*, vol. 36, no. 5, pp. 697–716, September 2000.

[11] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.