Proc. Int. Conf. Image Processing (ICIP), (Cairo, Egypt), pp. 2225-2228, Nov. 2009.

STRUCTURAL SIMILARITY METRICS FOR TEXTURE ANALYSIS AND RETRIEVAL

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

[1]EECS Department, Northwestern Univ., Evanston, IL 60208
[2]EECS Department, Univ. of Michigan, Ann Arbor, MI 48109

ABSTRACT

The development of objective texture similarity metrics for image analysis applications differs from that of traditional image quality metrics because substantial point-by-point deviations are possible for textures that according to human judgment are essentially identical. Thus, structural similarity metrics (SSIM) attempt to incorporate "structural" information in image comparisons. The recently proposed structural texture similarity metric (STSIM) relies entirely on local image statistics. We extend this idea further by including a broader set of local image statistics, basing the selection on metric performance as compared to subjective evaluations. We utilize both intra- and inter-subband correlations, and also incorporate information about the color composition of the textures into the similarity metrics. The performance of the proposed metrics is compared to PSNR, SSIM, and STSIM on the basis of subjective evaluations using a carefully selected set of 50 texture pairs.

Index Terms— Steerable filter decomposition, dominant colors, image retrieval, image compression.

1. INTRODUCTION

The development of objective metrics for texture similarity differs from that of traditional image quality metrics because substantial point-by-point deviations are possible for textures that according to human judgment are essentially identical. While the emphasis of this paper is on image analysis and retrieval applications, texture similarity metrics are also important for image coding applications when significant changes in the image are permissible, provided they do not affect the perceived image quality, even though in a side-by-side comparison there may be clearly perceptible differences.

There have been several attempts to develop metrics that deviate from traditional point-by-point fidelity. A broad class of new metrics, the structural similarity metrics (SSIM) [1], attempt to incorporate "structural" information in image comparisons. A number of metrics have been proposed, both in the space domain (SSIM) [1] and the complex wavelet domain (CWSSIM) [2]. In order to overcome some of the limitations of SSIMs when applied to texture analysis applications, Zhao *et al.* [3] proposed a structural texture similarity metric (STSIM) that aims to move further away from point-by-point comparisons by relying only on local image statistics. In this paper, we extend this idea further by including a broader set of local image statistics, basing the selection on metric performance as compared to subjective evaluations. In particular, we utilize the (usually strong) correlations between information contained in different subband decompositions of an image. In addition, we incorporate information about the color composition of the textures into the similarity metrics.

This paper is organized as follows. Section 2 provides a brief overview of similarity metrics. The proposed techniques are described in Section 3. The experimental setup and results are given in Section 4.

2. BACKGROUND

Traditional quality metrics range from simple MSE and PSNR to more sophisticated metrics that incorporate low-level models of human perception [4], and are typically aimed at near-threshold applications such as image compression. The idea is to ensure image fidelity on a point-by-point basis. However, for supra-threshold applications, such as content-based image retrieval (CBIR), 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 in the case of CWSSIM [2], small translations, rotations, and scaling changes as well.

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 the k-th subband. The *luminance* comparison term is defined as

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

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

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

where σ_x^2 and σ_y^2 are the variances of the two windows; and the *structure* term is defined as

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

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

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

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.* [3] 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 $\rho^k_x(0,1)$ and vertical $\rho^k_x(1,0)$ directions as follows:

$$c^k_{0,1}(\mathbf{x},\mathbf{y}) = 1 - 0.5\left(|\rho^k_x(0,1) - \rho^k_y(0,1)|\right)^p \qquad (5)$$

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. Here we assume $p = 1$.

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

$$Q^k_{\text{stsim}}(\mathbf{x},\mathbf{y}) = l^k(\mathbf{x},\mathbf{y})^{\frac{1}{4}} c^k(\mathbf{x},\mathbf{y})^{\frac{1}{4}} c^k_{0,1}(\mathbf{x},\mathbf{y})^{\frac{1}{4}} c^k_{1,0}(\mathbf{x},\mathbf{y})^{\frac{1}{4}} \qquad (6)$$

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.

3. PROPOSED TECHNIQUES

The proposed metrics extend the ideas of [3] by including a broader set of local image statistics. The motivation for this comes from the analysis/synthesis literature, and in particular, the work of Portilla and Simoncelli [5], 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. While texture synthesis requires several hundred parameters, we believe that a lot fewer will be adequate for texture similarity.

As in [2, 3], we use the steerable filter decomposition of the grayscale component of the two images. In the following, we use three scales ($N_s = 3$) and four orientations ($N_o = 4$).

In addition to the terms in (6), we use 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, we need to compute the average for each sliding window and use it in the variance calculation.

Portilla and Simoncelli [5] 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.

In the proposed metric, for each orientation we compute the cross-correlations between the magnitudes of subband coefficients at adjacent scales, and for each scale we compute 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 the Section 4 we will discuss the effect of utilizing all or subsets of these 26 coefficients.

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

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

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 $[-1,1]$, we can compare them as in (5) to obtain a statistic that describes the similarity between the cross-correlations:

$$c^{k,l}_{0,0}(\mathbf{x},\mathbf{y}) = 1 - 0.5\left(|\rho^{k,l}_x(0,0) - \rho^{k,l}_y(0,0)|\right)^p \qquad (8)$$

Note that the $c^{k,l}_{0,0}(\mathbf{x},\mathbf{y})$ 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 $N = N_s \cdot N_o + 1$ subband images (including the highpass but not the lowpass). 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 can then be combined additively

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

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

Color Similarity Metric

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 is what is typically used in image compression applications. An alternative approach that is more effective in image retrieval applications is to use separate metrics for comparing the grayscale textures and the color composition of an image, and then to combine them in order to obtain one number.

The simplest approach for describing and comparing the color composition of images is to use color histograms and simple histogram intersection metrics [6] or a more sophisticated color quadratic distance [7]. However, based on the observation that the human visual system cannot simultaneously perceive a large number of colors, more compact color representations in terms of dominant colors have been employed, e.g., in [8, 9, 10]. In addition, direct histogram comparisons have given way to more sophisticated techniques that account better for the HVS properties. One of the best known techniques is the earth mover's distance (EMD) [11].

In this paper, we adopt the dominant color idea and use a metric that follows the same philosophy as EMD, the Optimal Color Composition Distance (OCCD), which finds the optimal mapping between the color composition features of two segments and computes the average distance between them in the *CIE L*a*b** color space [12]. Since the appearance of an image is best described by the spatial distribution of features, rather than by individual feature vectors [13], we are utilizing a sliding windows approach to assess color similarity, just like we did for texture. In order to account for spatially varying textures, we use the adaptive clustering algorithm [14] to obtain *spatially adaptive* dominant colors. The advantage of this technique is that the averaging is performed only on pixels that belong to the same segment, thus avoiding any blurring along the borders of two segments.

Since the OCCD computes the *distance* between two colors, their *similarity* can be rated as $1 - distance$. 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.

The final step is to linearly combine the texture and color similarity measures Q_t and Q_c with appropriate weights, , w_t and $w_c = 1 - w_t$, as is widely done in the literature

$$Q_{total} = w_t \cdot Q_t + w_c \cdot Q_c \tag{10}$$

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed techniques, we conducted an informal subjective test, whereby people were asked to rate similarity between pairs of images. We used a carefully selected set of 50 texture pairs, obtained from a pool of thirty textures, extracted from the Corbis online database [15]. The size of the texture images was 128×128. Figure 1 shows the seven most similar pairs according to the human judgments, as well as two dissimilar textures.

The subjective test was conducted using Matlab GUI. The number of subjects was ten. Each subject was asked to grade 50 pairs of textures according to their similarity, with the lowest grade being 1 ("completely dissimilar") and the highest grade being 10 ("same texture"). The final score assigned to a pair of textures was the mean value of all the subjective scores. At the beginning of the test, each subject was shown a representative set of textures. The subjects were asked to rank the similarity between two textures without any specific guidelines, e.g., on the relative importance of texture structure, scale, color, etc. Note that since we used a set of natural textures, it was unavoidable that the semantics had an effect on the similarity evaluations.

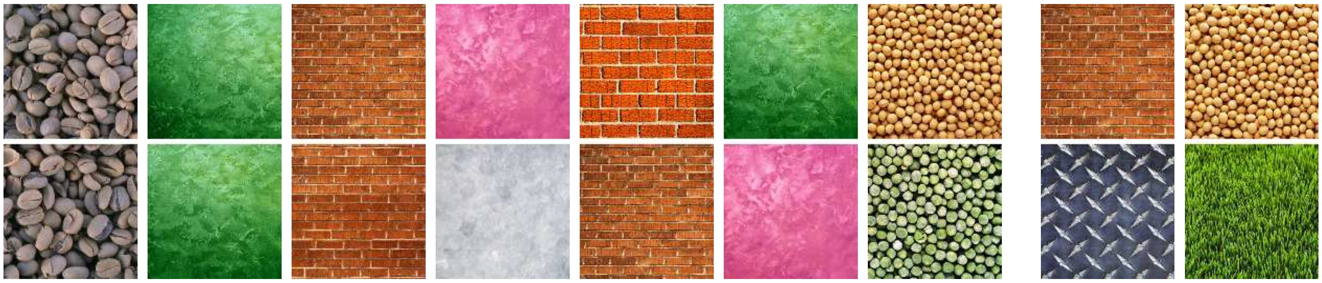
As we discussed, in our experiments we used a 3-scale, 4-orientation decomposition. The sliding window for all the metrics was 7×7. In order to determine the best set of parameters for the proposed metric, that is, the best combination of texture maps in (9) and the optimal values for the weights w_t and w_c, we tried different combinations and selected the one that results in the best correlation with human judgments.

Evaluating performance of similarity evaluation systems is difficult. Since we are using a relatively small set of pairs, we are mostly interested in how the metric rankings of texture similarity compare to the subjective rankings. Thus, instead of Pearson's, we used the Spearman rank correlation coefficient, "a non-parametric measure of correlation – that is, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about their frequency distribution" [16].

In addition to selecting parameters for the proposed metric, we compared its performance to other metrics. As might be expected, the Spearman correlation coefficient for PSNR was 0.283. The SSIM [1] (based on the Matlab implementation downloaded from [17]) performed considerably better; we used the same varying weights, and the best result was 0.515. The CW-SSIM and STSIM metrics performed better with 0.579 and 0.598, repsectively. The best performance of the proposed metric, and best overall, was 0.659. This was obtained for the additive approach including all the terms in (9) and with weights $w_t = 0.6$ and $w_c = 0.4$ in (10). We also tried another well-established test, the Kendall tau rank correlation coefficient, and the results were similar.

To distinguish the effect of including the new terms (9) vs. including color information only, it should be noted that the Spearman rank correlation coefficient for the proposed metric without using color ($w_t = 1$ and $w_c = 0$) was 0.638, well above the best performance without the new terms (0.598). This shows that the crucial improvement was accomplished by adding new cross-correlation terms.

Since raw correlation numbers like $\rho = 0.65$ are not very descriptive when one does not know what to expect, we performed the following calculation. For each of the subjects, we removed their judgments from the pool, and computed the mean grades of the remaining subjects. Then, we conducted



(a) 9.7 (b) 9.6 (c) 9.3 (d) 8.1 (3) 7.4 (f) 7.4 (g) 7.2 (h) 2.5 (i) 2.2

Fig. 1. Selected texture pairs used in the experiments and average human scores (a-g) similar, (h-i) dissimilar.

the Spearman rank correlation tests, to find out how well a human performs against other humans; to be fair, we recomputed the correlation coefficients between our best performing method, and the same mean grades of the remaining subjects. In other words, we compared the performance of our metric with the performance of each human subject versus that 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 the proposed metric. For practical purposes, a metric is considered to be useful if the Spearman rank correlation coefficient exceeds 0.8; this result shows that, in future, we need to design the subjective tests more carefully and have more reliable benchmark human judgements.

As another indication of metric performance, Figure 2 shows a scatter plot of the subjective tests versus the metric values for the CW-SSIM, STSIM, and the proposed metric (all variables are standardized, i.e., centered around 0 and scaled by their standard deviations). The slope of the mean-square fit, which is in fact the Pearson's correlation coefficient, is an indication of performance. Note that, while the proposed technique is closest to the ideal slope of one, it is still a long way from ideal performance. Looking at the seven most similar pairs as judged by the subjects, we get an idea of the difficulty of the task. One possibility for improvement may be to decouple the grayscale from the color judgment. Another issue to be addressed more carefully is texture scale.

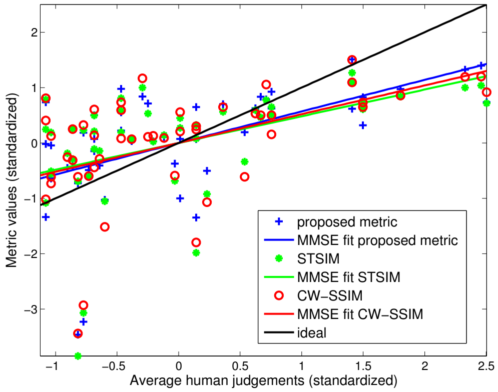


Fig. 2. Scatter plot of metric values vs. average human scores.

5. REFERENCES

[1] Z. Wang, et al., "Image quality assessment: From error visibility to structural similarity," *IEEE Tr. Image Proc.*, vol. 13, pp. 600–612, Apr. 2004.

[2] Z. Wang, E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," *Proc. ICASSP*, 2005, vol. II, p. 573.

[3] X. Zhao, et al., "Structural texture similarity metrics for retrieval applications," *Proc. ICIP*, 2008, p. 1196.

[4] T. N. Pappas, et al., "Perceptual criteria for image quality evaluation," in *Handbook Image and Video Proc.*, A. C. Bovik, Ed., pp. 939–959. Academic Press, 2005.

[5] J. Portilla, E. P. Simoncelli, "A parametric texture model based on joint statictics of complex wavelet coefficients," *Int. J. Comp. Vis.*, vol. 40, pp. 49–71, 2000.

[6] M. Swain, D. Ballard, "Color indexing," *Int. J. Comp. Vis.*, vol. 7, pp. 11–32, 1991.

[7] H. S.Sawhney, J. L. Hafner, "Efficient color histogram indexing," *Proc. ICIP*, 1994, vol. II, p. 66.

[8] W. Y. Ma, et al., "Tools for texture/color based search of images," in *Human Vision and Electronic Imaging II*, 1997, Proc. SPIE, Vol. 3016, pp. 496–507.

[9] A. Mojsilović, et al., "Matching and retrieval based on the vocabulary and grammar of color patterns," *IEEE Tr. Image Proc.*, vol. 1, pp. 38–54, Jan. 2000.

[10] J. Chen, et al., "Adaptive perceptual color-texture image segmentation," *IEEE Tr. Image Proc.*, vol. 14, pp. 1524–1536, Oct. 2005.

[11] Y. Rubner, et al., "The earth mover's distance as a metric for image retrieval," *Int. J. Comp. Vision*, vol. 40, pp. 99–121, 2000.

[12] A. Mojsilović, et al., "Extraction of perceptually important colors and similarity measurement for image matching, retrieval, and analysis," *IEEE Tr. Image Proc.*, vol. 11, pp. 1238–1248, Nov. 2002.

[13] Y. Rubner, et al., "Empirical evaluation of dissimilarity measures for color and texture," *Comp. Vision Im. Underst.*, vol. 84, pp. 25–43, Oct. 2001.

[14] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Tr. Signal Proc.*, vol. 40, pp. 901–914, Apr. 1992.

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

[16] "Wikipedia," http://www.wikipedia.org.

[17] H. R. Sheikh, et al., "Image and video quality assessment research at LIVE."