SUBJECTIVE AND OBJECTIVE TEXTURE SIMILARITY FOR IMAGE COMPRESSION

Jana Zujovic,[1] *Thrasyvoulos N. Pappas,*[1] *David L. Neuhoff,*[2] *Rene van Egmond,*[3] *Huib de Ridder*[3]

[1]EECS Department, Northwestern University, Evanston, IL 60208
[2]EECS Department, University of Michigan, Ann Arbor, MI 48109
[3]Faculty of Industrial Design Engineering, TU Delft, 2628 CE Delft, The Netherlands

ABSTRACT

We focus on the evaluation of texture similarity metrics for *structurally lossless* or nearly structurally lossless image compression. By structurally lossless we mean that the original and compressed images, while they may have visible differences in a side-by-side comparison, they have similar quality so that one cannot tell which is the original. This is particularly important for textured regions, which can have significant point-by-point differences, even though to the human eye they appear to be the same. As in traditional metrics, texture similarity metrics are expected to provide a monotonic relationship between measured and perceived distortion. To evaluate metric performance according to this criterion, we introduce a systematic approach for generating synthetic texture distortions that model variations that occur in natural textures. Based on such distortions, we conducted subjective experiments with a variety of original texture images and different types and degrees of distortions. Our results indicate that recently proposed structural texture similarity metrics provide the best performance.

Index Terms— perceptual similarity, image quality

1. INTRODUCTION

The conventional problem of image quality evaluation consists of measuring point-by-point distortions between the original and the compressed image. This is true for both the peak-signal-to-noise ratio (PSNR) and *perceptual metrics* that incorporate low-level properties of the human visual system [1]. The goal of the latter is to measure deviations from *perceptually lossless* compression, that is, images cannot be distinguished in a side-by-side comparison at a given display resolution and viewing distance [1]. Instead, our focus is on *structurally lossless* compression [2], whereby the original and compressed images may have visible differences in a side-by-side comparison, but they have similar quality so that one cannot tell which is the original. The goal of the quality metric is then to measure deviations from this standard of performance. This is particularly important for textured regions, which can have significant point-by-point differences, even though to the human eye they appear to be the same.

The point-by-point measurement of image similarity has been shown to not be in accordance with human perception, especially in textured areas [2, 3]. This prevents image (and video) compression algorithms from using spatial (and temporal) prediction for encoding textured regions, because the stochastic nature of textures results in large prediction errors when conventional metrics are used. However, large patches of texture (e.g., of grass, sand, clouds, forest) could be simply replaced with previously encoded patches with indistinguishable characteristics without any significant effect on perceived texture quality. This requires texture similarity metrics that account for the stochastic nature of textures and allow significant point-by-point deviations that do not affect texture appearance [4, 5]. Indeed, the development of such metrics and a better understanding of texture is key to further advances in image compression, as well as other image analysis applications such as computer visions and content-based retrieval [2]. The goal of this paper is to evaluate the performance of texture similarity metrics for the purposes of image compression.

In order to achieve structurally lossless or nearly structurally lossless compression for textured areas, the goal is to have a metric that provides a monotonic relationship between measured and perceived distortion. Such a metric can be used both for quality assessment and as a tool within a compression algorithm. However, as we pointed out in [6], such monotonic metric performance can be achieved only at the high end of the similarity scale, where the structural distortions have either a small effect on perceived quality, or do not affect perceived quality at all. Of course, after a certain point, when the quality is unacceptable, there is no need for monotonicity; the metric should simply give low values. It is also important to have an absolute measure of image similarity, so that consistent image quality can be achieved across different types of image content, both within an image and across different images, as well as across different compression techniques.

To explore how metric predictions relate to perceived distortion, and in particular, to test metric monotonicity and absolute performance, we need a systematic approach for obtaining different degrees of distortion for a variety of textures and subjective experiments to rate such distortions. However, generating a set images with fine differences in type and level of distortion is a difficult, if not impossible, task in the context

of real applications. We thus chose to generate synthetic distortions that model variations that occur in natural textures. Based on such distortions, we conducted subjective experiments with a variety of original texture images and different types and degrees of distortions. Our results indicate that recently proposed structural texture similarity metrics [4, 5] provide the best performance.

Gide *et al.* [7] have conducted a similar subjective study, where the goal was to build a comprehensive database of texture images with several sets and degrees of distortions, to be used as a benchmark for comparison of different texture quality metrics. While the overall goals are the same, our experiments are aimed at variations that occur naturally in textures, with the goal of exploiting texture self-similarity (in compression applications), rather than distortion induced by traditional coding algorithms such as noise, blur, compression artifacts, shifts due to motion estimation, or synthesis based on a parametric texture model [8].

The remainder of this paper is organized as follows. Section 2 provides a brief overview of similarity metrics. The experimental setup is presented in Section 3, the results in Section 4, while the final conclusions are drawn in 5.

2. SSIM REVIEW

Structural similarity metrics (SSIMs) [9] were developed for supra-threshold applications, such as perceptually lossy compression, where significant changes are allowed that do not affect image structure. When implemented in the image domain, SSIMs allow non-structural contrast and intensity changes, while the complex wavelet version (CWSSIM) [10] also allows *small* translations, rotations, and scaling changes.

SSIM (and CWSSIM) compares two images on a sliding window basis. In each window it computes three terms: the *luminance* term compares the means of pixels (or complex coefficients) in two corresponding windows, the *contrast* term compares their variances, while the *structure* term computes the cross-correlation between the corresponding windows. The three terms are multiplied in each window, and averaged across windows to produce the overal image similarity. For CWSSIM, each subband produces a similarity score; the total similarity is computed as their mean.

For texture similarity, where significant point-by-point differences are possible between two texture images that according to the human eye can be considered as "identical" textures, Zhao *et al.* [4] proposed a new Structural Texture Similarity Metric (STSIM). Their metric removed the structure term from the CWSSIM – which in spite of its name is in fact a point-by-point comparison – and added two new terms that depend only on region statistics. The new terms compare the first order correlation coefficients in the horizontal and vertical direction, so that for each sliding window, they multiply four terms instead of three. Again, the final similarity score is obtained by spatial and frequency pooling.

The metric proposed in [5] extends the ideas of [4] by including a broader set of local image statistics. In addition to the STSIM computation for each subband, STSIM2 also computes terms that compare the cross-correlations between subbands that correspond to adjacent scales for a given orientation, as well as different orientations within the same scale. STSIM2 then combines the results of STSIM and adds the cross-correlation comparison terms to form a final similarity score for two images. We will refer to this as *STSIM2*, to distinguish it from the metric in [4]. In the remainder of this paper, we use a three-scale, four-orientation steerable pyramid decomposition, and sliding windows of size 7×7.

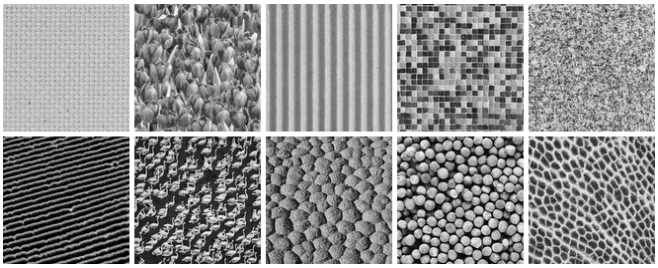


Fig. 1. Original textures

3. EXPERIMENTAL SETUP

Our goal is to determine how successful various similarity metrics are in predicting the perceived degree of distortion in texture images. For our experiments, we chose ten different grayscale texture images, shown in Fig. 1. They range from noise-like to highly structured images, and they exhibit different levels of susceptibility to different types and degrees of distortions. The image resolution is 128×128 pixels.

As we discussed in the introduction, we chose to generate synthetic distortions that model variations that occur in natural textures. Since these include variations in position, orientation, and color [8], we implemented the following types of distortion: (1) *random rotation* of small local patches, (2) *random shifts* of small local patches, and (3) *image warping*, whereby the images are distorted according to random deviations of the control points of an underlying mesh. The severity of each type of distortion can be easily manipulated by varying the distortion parameters (probabilistic distribution of rotations, shifts, and mesh points). For the rotation angles, shifts, and the control points of the warping, we assumed a uniform distribution of the corresponding variables, the range of which determines the degree of distortion. Each of the ten textures was distorted with three distortion algorithms, with three degrees of severity for each distortion. Examples of the distorted images corresponding to three different originals, are given in Fig. 2. From left to right, we have three rotation-distorted images, three shift-distorted images, and three warped images. For each type of distortion, the severity is increasing from left to right.

In our experiments, we used 11×11 pixel patches, while the warping meshes were 5×5. This is because the smaller meshes result in artifacts of similar scale as those of the larger

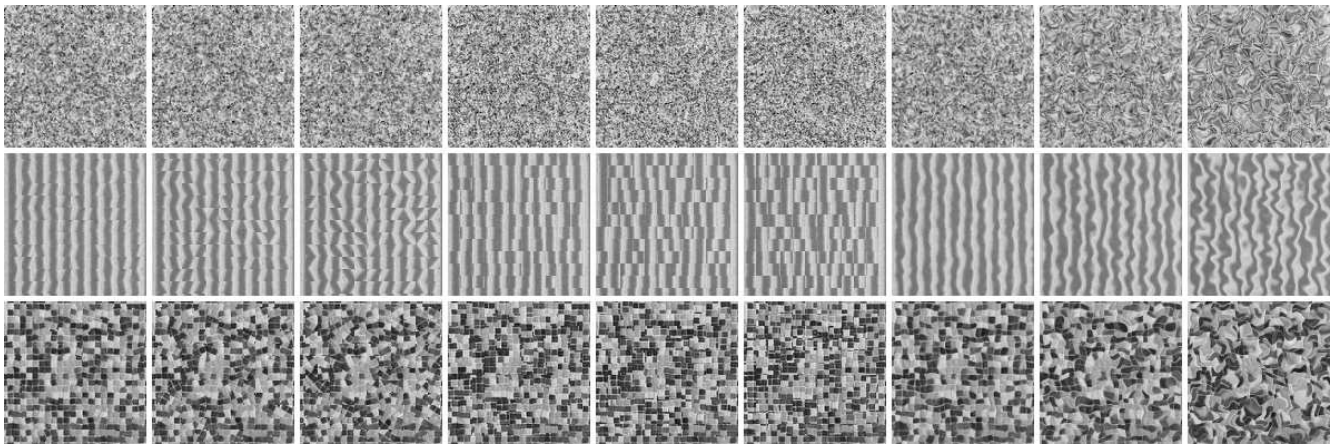
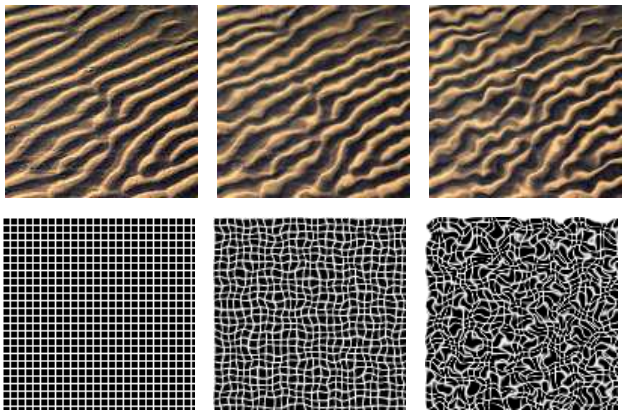


Fig. 2. Examples of distorted texture images: rotation (columns 1–3), translation (columns 4–6), warping (columns 7–9),



(a) Original (b) Low distortion (c) High distortion

Fig. 3. Warping distortion with underlying meshes

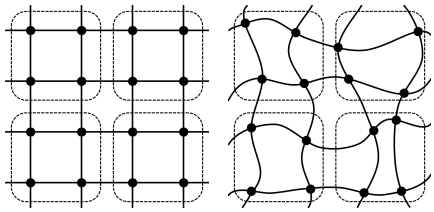


Fig. 4. Grid points in the original and distorted meshes

patches, as shown in Fig. 3. The reason for this is illustrated in Fig. 4, which shows the control points of the underlying mesh. Note that each disjoint set of four points on the grid can be seen as the control points of one rectangle (dashed lines), with space between rectangles for smooth transitions, effectively producing deformations on rectangles of size 10×10.

For each of the ten original images, the subjects were asked to rank the distorted images from *best* to *worst*, compared to the original image. They were not allowed to give the same ranking to two distorted images. As a result, every user provided rankings between 1 and 9 for the nine distorted images corresponding to each original. This, however, produces data that can only be processed for a given original, not across originals. In order to determine metric performance across different originals, the subjects were also asked to rank the worst distortions for each of the original textures.

4. EXPERIMENTAL RESULTS

Eleven subjects participated in this experiment that compared the PSNR, SSIM, CWSSIM, STSIM, and STSIM2 metrics.

4.1. Analyzing the ranking results per original image

Analyzing the ranking data can be done in a number of ways. In all cases, for each original image, we extract a 1-D vector that describes the subjective similarity between the original and the nine distorted images. This vector is compared to the values of each similarity metric. To measure the goodness of fit, we use the Pearson's correlation coefficient, which evaluates absolute metric performance, and the Spearman rank correlation coefficient, which describes how well a metric ranks the distorted images compared to the subjective rankings. The reported values for Pearson's and Spearman's ρ are the average correlation coefficients taken over all the originals.

The simplest approach is to find the mean ranking for each distorted image and use that as its "subjective" position with respect to the original. This is perhaps one of the oldest techniques, proposed in 1770 by Jean-Charles de Borda, and today usually known as Borda's rule. He called this method "election by order of merit," i.e., the cumulative preference given to a candidate is its final score.

One popular way to analyze this type of data is to use Thurstonian scaling [11]. It is applied on the *preference* matrix P, where $P(i, j)$ denotes how many times image i was preferred to image j, i.e., how many times image i was ranked as closer to the original than image j. After pooling all the results, the preference matrix is scaled to represent percentages ("image i was preferred to image j in p percent of cases") and percentages are converted into z-scores. This can produce singular values when we have perfect agreement among raters, so an alternative has been proposed by Krus *et al.* [12], which avoids such undesirable behavior.

Algorithm	Borda's rule	Thurstonian scale	MDS
PSNR	0.72	0.72	0.72
SSIM	0.74	0.74	0.74
CWSSIM	0.84	0.84	0.83
STSIM	**0.88**	**0.88**	**0.87**
STSIM2	**0.88**	**0.88**	**0.87**

Table 1. Pearson's ρ for different analysis methods

Algorithm	Borda's rule	Thurstonian scale	MDS
PSNR	0.67	0.66	0.67
SSIM	0.72	0.71	0.72
CWSSIM	0.81	0.81	0.81
STSIM	0.85	0.85	0.85
STSIM2	**0.86**	**0.86**	**0.86**

Table 2. Spearman's ρ for different analysis methods

Yet another way to analyze the data is to treat the ranks as distances between images. For example, the image that was ranked as number 1 and the image that was ranked as number 5 would be assigned distance of 4. After aggregating all the data from all the users, we can then perform multidimensional scaling (MDS) [13,14] to extract the perceptual dimensions embedded in the data. For each methods, we report the Pearson's ρ in Table 1, and the Spearman's ρ in Table 2.

4.2. Analyzing the ranking results across originals

In the final step of the test, each user had to rate ten images they previously labeled as the "worst" ones, from "best" to "worst." Thus, every user gives a ranking of a subset of all possible pairs of images across different originals. Given that there are 9^{10} possible different subsets, it is clear that the data gathered in this manner produces very sparse matrices.

Methods for comparing the similarity or preference matrices with incomplete data do exist. However, in this case, when analyzed with non-metric multidimensional scaling techniques [15], the results are very unreliable, due to the high sparsity of the formed preference matrix.

An alternative is to use a metric that estimates the agreement between the subjective scores and a metric's scores. One possible test to perform is Kendall coefficient of agreement [16] (or Kendall's W), which is designed to measure inter-rater agreement. To analyze the performance of a metric, we can treat it as yet another rater and then compute the joint agreement between the metric values and the subjective scores. The higher the overall coefficient of agreement, the better the metric represents the subjective data. Values for this test are reported in Table 3.

5. CONCLUSION

We have presented a novel approach for generating synthetic texture distortions that model variations that occur in natural textures. We conducted subjective tests to learn how the human subjects perceive the severity of such distortions, and to determine which texture similarity algorithm correlates the best with the subjective judgements. The STSIM2 algorithm always outperforms, or is tied with, the other algorithms, which shows its usefulness for compression applications.Future research will include a broader range of image deformations, as well as more thorough subjective testing and the incorporation of the metric into a compression algorithm.

PSNR	SSIM	CWSSIM	STSIM	STSIM2
0.53	0.60	0.66	0.67	**0.68**

Table 3. Kendall's coefficient of agreement

6. REFERENCES

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

[2] T.N. Pappas, *et al.*, "Image analysis and compression: Renewed focus on texture," *Vis. Inf. Proc. Comm.*, Jan. 2010, Proc. SPIE vol. 7543, pp. 75430N–1–12.

[3] P. Ndjiki-Nya, *et al.*, "Generic and robust video coding with texture analysis and synthesis," *ICME-07*, pp. 1447–1450.

[4] X. Zhao, *et al.*, "Structural texture similarity metrics for retrieval applications," *ICIP-08*, San Diego, CA, Oct. 2008, pp. 1196–1199.

[5] J. Zujovic, *et al.*, "Structural similarity metrics for texture analysis and retrieval," *ICIP-09*, Cairo, Egypt, Nov. 2009, pp. 2225–2228.

[6] J. Zujovic, *et al.*, "A new subjective procedure for evaluation and development of texture similarity metrics," *Proc. 10th IVMSP Wksp*, June 2011, pp. 123–128.

[7] M.S. Gide, L.J. Karam, "On the assessment of the quality of textures in visual media," *Conf. Inf. Sciences Sys. (CISS), 2010*, pp. 1–5.

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

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

[10] Z. Wang, E.P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," *ICASSP-05*, Philadelphia, PA, 2005, v. II, pp. 573–576.

[11] L.L. Thurstone, "A law of comparative judgment.," *Psychological Review*, v. 34, no. 4, pp. 273, 1927.

[12] D.J. Krus, P.H. Krus, "Normal scaling of the unidimensional dominance matrices: the domain referenced model," *Educ. Psych. Meas.*, v. 37, pp. 189, 1977.

[13] W.S. Torgerson, *Theory and methods of scaling*, Wiley, New York, NY, 1958.

[14] J.B. Kruskal, M. Wish, *Multidimensional scaling*, Sage Publications, Beverly Hills, CA, 1977.

[15] J.B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, v. 29, no. 1, pp. 1–27, 1964.

[16] M.G. Kendall, B.B. Smith, "On the method of paired comparisons," *Biometrika*, v. 31, pp. 324–345, 1940.