TEXTURE FIDELITY BENCHMARK

Michal Haindl[*]*, Miloš Kudělka Jr.*

Institute of Information Theory & Automation
Acadademy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, Prague, 182 08, Czech Republic

ABSTRACT

Automatic texture quality evaluation is important but still unsolved difficult problem. While several generative mathematical texture models were developed, their reliable qualitative evaluation is for now possible only using impractical and expensive visual psycho-physics which hampers their further progress. We present the texture fidelity benchmark created to help the validation of texture fidelity criteria being developed. The benchmark is a web based service (**http://tfa.utia.cas.cz**) designed for performance evaluation, mutual comparison, and ranking of various texture fidelity measures. The benchmark supports rapid verification and development of new fidelity criteria approaches and contains seven color, variable texture quality, series together with their grey-scale counterparts.

Index Terms— texture, benchmark, fidelity criteria, measure,

1. INTRODUCTION

Evaluation of how well various texture models conform with human visual perception is important not only for assessing the similarities between a model output and the original measured texture, but also for optimal settings of model parameters, for fair comparison of distinct models, etc. Few published criteria allow to test selected texture properties such as the texture regularity [1] others claim to test general texture quality [2]. Currently the only reliable, but extremely impractical and expensive option, is to exploit the methods of visual psycho-physics. The psycho-physical methods [3] require a lengthy process of experiment design, tightly controlled laboratory condition, and representative panel of human testing subjects. Such testing obviously cannot be performed on a daily basis. Thus an automatic texture fidelity verification is needed for evaluating the quality of texture-generating algorithms, for database texture retrieval, etc. This problem has not been successfully solved and new measures still emerge. We present the texture fidelity benchmark for validation of texture fidelity measures and the results that were obtained. The point is to show that in most cases neither the general image quality measures do work, nor the texture quality criteria, while performing better, are reliable.

[*]This research was supported by the grant GAČR 14-10911S.

2. IMAGE AND TEXTURE QUALITY MEASURES

We have tested several state-of-the-art image quality measures and several recently published texture criteria that are concisely described in the following text. All of the criteria consider only gray-scale images.

Let us define common naming convention. $\mathbf{x}$ and $\mathbf{y}$ stand for the target and tested image, μ_z and σ_z is a mean value and standard deviation of z, respectively.

2.1. Mean-Squared Error

The mean-squared error (MSE) [6] multispectral criterion is

$$\mathrm{MSE}(\mathbf{x},\mathbf{y}) = \frac{1}{MNd}\sum_{r_1=1}^{M}\sum_{r_2=1}^{N}\sum_{r_3=1}^{d}(\mathbf{x}_\mathbf{r}-\mathbf{y}_\mathbf{r})^2 \ , \quad (1)$$

where $r = \{r_1, r_2, r_3\}$ is a multiindex with the row, column, and spectral indices, M number of rows, N number of columns, d number of spectra. MSE is simple, memory-less, parameter free, and inexpensive to compute, but, it depends strongly on the image scaling and does not measure well human perception.

2.2. Visual Signal-to-Noise-Ratio

The visual signal-to-noise-ratio [4] (VSNR) is a two-stage approach. In the first stage, contrast thresholds for detection of distortions in the presence of natural images are computed via wavelet-based models of visual masking and visual summation in order to determine whether the distortions in the distorted image are visible. The threshold contrast is used in the second step to compute contrast detection thresholds.

2.3. Structural Similarity Index

The structural similarity (SSIM) index [5] is based on an assumption, that structural information about an image can be

978-1-4799-7971-4/14/$31.00

described by a function (usually a simple multiplication) of three terms: luminance l, contrast c and structure s:

$$\text{SSIM}(\mathbf{x},\mathbf{y}) = l(\mathbf{x},\mathbf{y})c(\mathbf{x},\mathbf{y})s(\mathbf{x},\mathbf{y}) = \tag{2}$$
$$= \left(\frac{2\mu_\mathbf{x}\mu_\mathbf{y}+C_1}{\mu_\mathbf{x}^2+\mu_\mathbf{y}^2+C_1}\right)\left(\frac{2\sigma_\mathbf{x}\sigma_\mathbf{y}+C_2}{\sigma_\mathbf{x}^2+\sigma_\mathbf{y}^2+C_2}\right)\left(\frac{\sigma_\mathbf{xy}+C_3}{\sigma_\mathbf{x}\sigma_\mathbf{y}+C_3}\right) \ ,$$

where $\sigma_\mathbf{xy}$ is the sample cross correlation of x and y after removing their means. C_1, C_2, C_3 are small positive constants that stabilize each term.

2.4. Complex Wavelet - Structural Similarity Index

The Complex wavelet - structural similarity (CW-SSIM) index [7] is basically the SSIM index computed in the complex wavelet domain and it is defined as:

$$\text{CW-SSIM}(\mathbf{c_x},\mathbf{c_y}) = \left(\frac{2\left|\sum_{i=1}^{C} c_{\mathbf{x},i}\cdot c^*_{\mathbf{y},i}\right| + K}{\sum_{i=1}^{C}(|c_{\mathbf{x},i}|^2+|c_{\mathbf{y},i}|^2)+K}\right) \ , \tag{3}$$

where $\mathbf{c_x}$ represents complex wavelet coefficients and $c_{\mathbf{x},i}$ is the i-th coefficient from the image $\mathbf{x}$ (analogically for the image $\mathbf{y}$), z^* stands for complex conjugate, C is the number of complex wavelet coefficients, and K is a small positive stability constant.

2.5. Visual Information Fidelity

Visual information fidelity (VIF) methods [8] explicitly incorporate statistical models of all the components in the communication system interpretation of signal fidelity measurement. VIF is defined as the ratio of the summed mutual information

$$\text{VIF} = \frac{I(C;F|x)}{I(C;E|x)} = \frac{\sum_{i=1}^{N} I(c_i;f_i|x_i)}{\sum_{i=1}^{N} I(c_i;e_i|x_i)} \ , \tag{4}$$
$$I(c_i;e_i|x_i) = \frac{1}{2}\log\frac{|x_i^2 C_U+\sigma_n^2\mathbf{I}|}{|\sigma_n^2\mathbf{I}|} = \frac{1}{2}\sum_{j=1}^{M}\log\left(1+\frac{x_i^2\lambda_j}{\sigma_n^2}\right) \ ,$$
$$I(c_i;f_i|x_i) = \frac{1}{2}\log\frac{|g_i^2 x_i^2 C_U+(\sigma_v^2+\sigma_n^2)\mathbf{I}|}{|(\sigma_v^2+\sigma_n^2)\mathbf{I}|} = \frac{1}{2}\sum_{j=1}^{M}\log\left(1+\frac{g_i^2 x_i^2\lambda_j}{\sigma_v^2\sigma_n^2}\right) \ ,$$

where E and F are models in a wavelet domain for what human visual system (HVS) captures from original and test images, respectively, N is the number of subbands, C_U is a covariance matrix (without considering noise and scale factors) of E, λ_j is the j-th eigenvalue of C_U, x is a realization of an original image, and g is an attenuation factor.

2.6. Structural Texture Similarity Measure

STSIM is an extension of CW-SSIM and has three versions, STSIM-1, STSIM-2 and STSIM-M [2]. STSIM-1 is created from CW-SSIM by replacing the 'structural' term with terms that compare first-order autocorrelations of corresponding subband coefficients $\rho_\mathbf{x}^m(0,1)$ in the horizontal and $\rho_\mathbf{x}^m(1,0)$ in the vertical direction. In the equations for a single subband m, the p is typically set to 1

$$\text{STSIM-1}^m(\mathbf{x},\mathbf{y}) = \tag{5}$$
$$= \left(l_{\mathbf{x},\mathbf{y}}^m\right)^{1/4}\left(c_{\mathbf{x},\mathbf{y}}^m\right)^{1/4}\left(c_{\mathbf{x},\mathbf{y}}^m(0,1)\right)^{1/4}\left(c_{\mathbf{x},\mathbf{y}}^m(1,0)\right)^{1/4} \ ,$$
$$c_{\mathbf{x},\mathbf{y}}^m(0,1) = 1-0.5\left|\rho_\mathbf{x}^m(0,1)-\rho_\mathbf{y}^m(0,1)\right|^p \ ,$$
$$\rho_\mathbf{x}^m(0,1) = \frac{E\left\{[\mathbf{x}^m(i,j)-\mu_x^m]\,[\mathbf{x}^m(i,j+1)-\mu_x^m]^*\right\}}{(\sigma_x^m)^2} \ .$$

STSIM-2 adds cross-band correlation coefficient $\rho_{|\mathbf{x}|}^{m,n}(0,0)$ between subbands m and n

$$\text{STSIM-2}(\mathbf{x},\mathbf{y}) = \tag{6}$$
$$= \frac{\sum_{m=1}^{N_b}\text{STSIM-1}^m(\mathbf{x},\mathbf{y})+\sum_{i=1}^{N_c} c_{\mathbf{x},\mathbf{y}}^{m_i,n_i}(0,0)}{N_b+N_c} \ ,$$
$$c_{\mathbf{x},\mathbf{y}}^{m,n}(0,0) = 1-0.5\left|\rho_{|\mathbf{x}|}^{m,n}(0,0)-\rho_{|\mathbf{y}|}^{m,n}(0,0)\right|^p \ ,$$
$$\rho_{|\mathbf{x}|}^{m,n}(0,0) = \frac{E\left\{\left[\mathbf{x}^m(i,j)-\mu_{|x|}^m\right]\left[\mathbf{x}^n(i,j)-\mu_{|x|}^n\right]\right\}}{\sigma_{|x|}^m\sigma_{|x|}^n} \ ,$$

where N_b is the number of subbands and N_c is the number of possible crossband correlations.

STSIM-M (STSIM-Mahalanobis) chooses another approach. Rather than combining aforementioned terms into a single measure, it uses them to create feature vectors $f_\mathbf{x}$ and $f_\mathbf{y}$ and then calculates the *Mahalanobis distance* between the feature vectors:

$$\text{STSIM-M}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{N_p}\frac{(f_{\mathbf{x},i}-f_{\mathbf{y},i})^2}{\sigma_{fi}^2}} \ , \tag{7}$$

where σ_{fi}^2 is the standard deviation of the i-th feature across all feature vectors in the set. Therefore, to compute the distance between two textures, STSIM-M requires statistics based on the whole set and the results are relative only to the set, which is unfavourable for our cause and therefore the STSIM-M was not included in our tests.

3. BENCHMARK

The development of fidelity evaluation of mathematical texture models or an optimization of their parameters require an automatic validation tool, i.e., a reliable and robust texture quality criterion. To find if some already published criterion

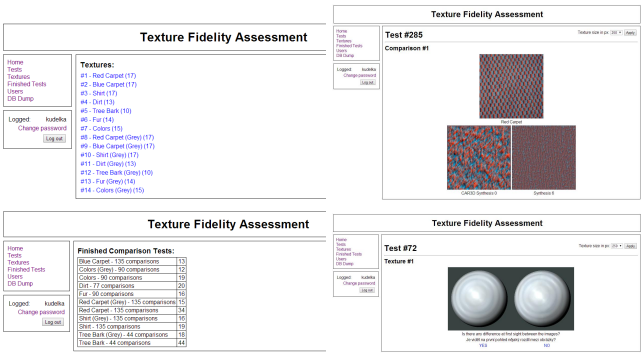

Fig. 1. The benchmark website (http://tfa.utia.cas.cz), home page (upper left), three modifications of the red carpet texture (upper right), partial assessment result (bottom left), and a test screen.

can be used for this purpose, it is necessary to have quality ranked texture series by human observers, so there is a ground truth to compare the criteria with. Although, there are several databases that contain either real world images or textures, no such benchmark with ground true scores or rank exists. Thus we develop the benchmark for viable and reliable way of testing new texture fidelity measures. Texture Benchmark (http://tfa.utia.cas.cz) is a simple website programmed in PHP. The subject can create a user account, log in, navigate through the website (see Fig. 1) to the Test page and begin to evaluate the textures (see Section 3.2).

Fig. 2. Target color benchmark textures.

3.1. Textural Data

We have chosen six natural and one synthetic color texture together with their gray scale versions as the target textures (for the textures see Fig. 2). Synthetic variants of these textures are ranked by benchmark users and the collected data serve to create modelling quality ranks. Textures were mathematically synthesized using various mathematical models and variable quality constraints. The models used were either random field type of models, mainly variants of the auto-regressive Markov random field models [9], or Gaussian mixture models [10].

3.2. Performance Evaluation

The most straightforward way to obtain an ordering is to present human with all possible pairs (see Fig.3 - left), one at a time, and let them choose the one that is more similar to the original. We also measure how long it takes for a user to reach a decision. The ranks are constructed from these collected data. This alternative is not very efficient and testers tend to get a bit frustrated from comparing tenths of image pairs. The main problem is that the number of comparisons grows quadratically with the number of synthesized textures, which already amounts to around a hundred with only a dozen of images.

Fig. 3. Test variants: comparison of pairs (left), ordering of textures (right)

Faster and simpler alternative is to show all the textures at once and let the subject order them by fidelity in a drag-and-drop manner (see Fig.3 - right). Even though the former approach is quite exhaustive and lengthy for the user, output data can be used not only to create a simple rank, but also to calculate quantitative measure of fidelity by taking into account the number of comparisons where each texture was better ranked than others. The rank calculation can be quite easily weighted considering the time factor. Presented results do not use such 'extended' data, because both measures (see Section 4), i.e., the simple and quantitative rank, give similar results on our data.

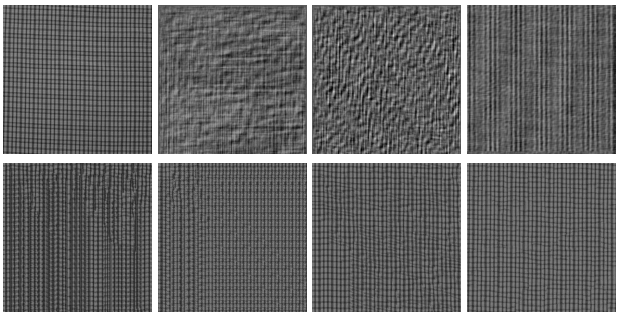

Fig. 4. An example of the texture and several of its syntheses. The top left texture is the original, the remaining textures are synthetic experiments.

4. COMPARATIVE ANALYSIS

To test the measures, we calculated ranks of the criteria from Section 2 and then the Spearman's rank correlation coefficient (using Matlab implementation) between these ranks and the human rank (see Tab. 1). In each row the correlations between both ranks are displayed for all textures. Note, that MSE normally measures the magnitude of the difference between the textures (error). For purposes of uniformity of the results, the rank was reversed, so the meaning of values in the tables is the same for all the measures.

Measure/Img	1	3	5	7
MSE	0.40(0.06)	-0.24(0.83)	0.02(0.49)	-0.17(0.72)
VSNR	0.72(0.00)	0.14(0.29)	-0.01(0.51)	-0.46(0.95)
VIF	0.04(0.43)	0.62(0.01)	-0.07(0.58)	-0.03(0.55)
SSIM	0.29(0.13)	-0.23(0.81)	0.02(0.49)	-0.04(0.56)
CW-SSIM	0.37(0.07)	0.20(0.22)	-0.01(0.51)	0.00(0.50)
STSIM-1-W	0.32(0.11)	0.87(0.00)	0.44(0.10)	0.44(0.06)
STSIM-1-G	0.48(0.03)	0.86(0.00)	0.64(0.03)	0.42(0.07)
STSIM-2-W	0.83(0.00)	0.85(0.00)	0.44(0.10)	-0.03(0.55)
STSIM-2-G	-0.54(0.99)	0.57(0.01)	0.37(0.15)	0.17(0.28)
Measure/Img	8	10	12	14
MSE	0.18(0.24)	-0.08(0.62)	0.13(0.37)	0.50(0.04)
VSNR	0.43(0.05)	0.24(0.17)	0.12(0.38)	0.39(0.09)
VIF	0.10(0.35)	0.72(0.00)	0.10(0.39)	0.41(0.08)
SSIM	0.15(0.29)	-0.05(0.58)	0.08(0.42)	0.39(0.08)
CW-SSIM	0.57(0.01)	0.29(0.13)	-0.08(0.59)	0.20(0.25)
STSIM-1-W	0.49(0.02)	0.89(0.00)	0.61(0.03)	0.71(0.00)
STSIM-1-G	0.61(0.01)	0.89(0.00)	0.78(0.01)	0.66(0.01)
STSIM-2-W	0.64(0.00)	0.93(0.00)	0.58(0.04)	0.28(0.16)
STSIM-2-G	-0.39(0.94)	0.65(0.00)	0.32(0.18)	0.53(0.03)

Table 1. Spearman's rank correlation between the human rank and the criteria results. Textures 1–7 are color images, the remaining are gray-scale.

The variants of STSIM are described as follows: W or G denotes whether the measure was computed on the whole texture (subband) or by the sliding window, where individual results were averaged. The numbers in parentheses are p-values for testing the hypothesis of no correlation against the alternative that there is a positive correlation. In some cases, p-value is zero even if the correlation is lesser than one. It is caused by the Matlab implementation and it only means the p-value is very small and was rounded to zero during the calculation.

These tables illustrate the fact, that the state-of-the-art image quality measures do not work almost at all, but some variants of STSIM (mainly STSIM-1-G) show good correlation with the human rank for some textures. Tab. 2 shows scores of STSIM for the texture number 10 (see Fig. 4), which is the best case scenario for STSIM. It is the gray-scale texture, which corresponds with the fact that STSIM does not work with color images.

Measure/Texture	1	2	3	4
STSIM-1-W	0.732	0.725	0.674	0.757
STSIM-1-G	0.871	0.873	0.834	0.901
STSIM-2-W	0.818	0.798	0.808	0.828
STSIM-2-G	0.918	0.892	0.913	0.930
Measure/Texture	5	6	7	8
STSIM-1-W	0.763	0.783	0.722	0.780
STSIM-1-G	0.881	0.902	0.875	0.922
STSIM-2-W	0.802	0.812	0.815	0.835
STSIM-2-G	0.912	0.926	0.930	0.934
Measure/Texture	9	10	11	12
STSIM-1-W	0.785	0.818	0.776	0.799
STSIM-1-G	0.928	0.937	0.924	0.927
STSIM-2-W	0.838	0.838	0.826	0.837
STSIM-2-G	0.898	0.943	0.919	0.899
Measure/Texture	13	14	15	16
STSIM-1-W	0.792	0.838	0.857	0.871
STSIM-1-G	0.921	0.950	0.966	0.958
STSIM-2-W	0.825	0.845	0.845	0.849
STSIM-2-G	0.910	0.940	0.960	0.952

Table 2. STSIM scores for the texture 10 from the Fig. 4. Numbers in the header denote synthetic textures.

As we observed, the criteria results have only very small variance (see Tab. 2). Because the criteria values range between zero and one, this suggests that the quality of all the synthetic textures is very similar, which is undoubtedly wrong. Thus, the question is how reliable these results are with respect to the small range of result values. Tab. 1 also illustrates principally wrong color / texture separation (contrary to some observations, e.g., [2]) - the rank correlation with human observers for color textures is lower (with 0.88 probability) than for their gray-scale variant.

5. CONCLUSIONS

The paper presents the benchmark for validation of texture fidelity criteria. We tested several state-of-the-art image quality measures and also one texture measure in several variants. The results demonstrate that the standard image quality criteria (MSE, VSNR, VIF, SSIM, CW-SSIM) cannot be used for texture quality validation at all. Although, the STSIM texture criterion has significantly higher correlation with human ranking, its results are texture dependent and the criterion has only small variance and therefore this measure is not reliable. The common problem of all tested criteria is that they are only monospectral and cannot use multispectral correlations. As a consequence the rank correlation with human observers for color textures is lower than for their gray-scale variant. This means the problem of texture fidelity assessment is still an open problem and there is a need for more reliable measure.

6. REFERENCES

[1] Wen-Chieh Lin, James Hays, Chenyu Wu, Yanxi Liu, and Vivek Kwatra, "Quantitative evaluation of near regular texture synthesis algorithms," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 1, pp. 427–434, 2006.

[2] Jana Zujovic, Thrasyvoulos N. Pappas, and David L. Neuhoff, "Structural texture similarity metrics for image analysis and retrieval," *Image Processing, IEEE Transactions on*, vol. 22, no. 7, pp. 2545–2558, 2013.

[3] Michal Haindl and Jiri Filip, "Visual Texture: Accurate Material Appearance Measurement, Representation and Modeling," *Advances in Computer Vision and Pattern Recognition*, Springer, 2013.

[4] Damon M. Chandler and Sheila S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *Image Processing, IEEE Transactions on*, vol. 16, no. 9, pp. 2284–2298, 2007.

[5] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[6] Zhou Wang and Alan C. Bovik, "Mean squared error: love it or leave it? a new look at signal fidelity measures," *Signal Processing Magazine, IEEE*, vol. 26, no. 1, pp. 98–117, 2009.

[7] Zhou Wang and Eero P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," *In Acoustics, Speech, and Signal Processing. Proceedings. IEEE International Conference on*, pp. 573–576, 2005.

[8] Hamid R. Sheikh and Alan C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, 2006.

[9] Michal Haindl, "Visual data recognition and modeling based on local markovian models," *Mathematical Methods for Signal and Image Analysis and Representation*, pp. 241–259. Springer, 2012.

[10] Michal Haindl, Jiri Grim, Petr Somol, Pavel Pudil, and Mineichi Kudo, "A Gaussian mixture-based colour texture model," *Pattern Recognition. Proceedings. IEEE 17th International Conference on.* , vol. 3, pp. 177–180, 2004.