Subjective Evaluation of Texture Similarity Metrics for Compression Applications

Johannes Ballé
Institut für Nachrichtentechnik
RWTH Aachen University
Aachen, Germany
Email: balle@ient.rwth-aachen.de

***Abstract*—The paper summarizes the results of an experimental subjective evaluation of texture similarity metrics with 25 test subjects. The compared metrics comprise the frequency-weighted log-spectral and Itakura distances, as well as a set of metrics based on an overcomplete Gabor-like filterbank – the STSIM as well as two new ones. The test set consists of 30 synthetic GMRF texture pairs. It turns out that all metrics perform well, but the weighted log-spectral distance tends to outperform the filterbank-based metrics.**

I. Introduction

Traditionally, image and video compression methods rely on pixel fidelity. However, more recent approaches [1], [2], [3] acknowledge that higher compression efficiency can be achieved if this paradigm is overcome. Given that a textured region carries no semantic information to the viewer, a great deal of bit rate can be saved by summarizing its contents in an appropriate manner and synthesizing visually similar texture, rather than relying on traditional methods such as transform coding. However, virtually all of the approaches in the literature are non-parametric and somewhat heuristic. Segmentation algorithms such as [4] are commonly used as the basis for detection of textured regions. Compression is achieved by selecting an appropriate texture sample. Synthesis (reconstruction) is performed by employing non-parametric texture synthesis such as [5] and related methods.

Non-parametric methods are characterized by their ability to cover a wide range of visual texture, as is convincingly demonstrated in [2], [3]. However, the non-parametric nature of the above methods implies that it is difficult to answer questions like: What is the class of texture that a given algorithm is able to detect in natural images, and is it included in the class of texture that is handled by the synthesis algorithm? Is there an equivalent parameter to the quantization step size in transform coders?

Similar questions can be answered precisely and quite naturally considering parametric texture models. A framework for compression of Gaussian Markov random fields (GMRF) within natural images was presented in [1]. In the same paper, it was shown that GMRF texture is fully characterized by the power spectral density of the random field model. Thus, spectral distances can be used to compare two given texture models (i.e., two different sets of parameters) with respect to their visual appearance. An algorithm for simultaneous estimation and quantization of the model parameters, which can be shown to be optimal for the 2D Itakura distance, was derived in earlier work [6]. Similarly, rate–distortion decisions could be optimized for spectral distances in the encoder.

However, before the implementation of a fully optimized system makes sense, it is necessary to evaluate how similarity metrics perform when compared to human judgements. This is the purpose of the present work.

The power spectral density of a GMRF $t(\boldsymbol{x})$ given by its parameters σ and $\mathbf{a}$, a vector of inverse filter coefficients, is equal to:

$$\Phi_t(e^{\jmath 2\pi \boldsymbol{f}}) = \frac{\sigma^2}{\left|1 + A(e^{\jmath 2\pi \boldsymbol{f}})\right|^2} \tag{1}$$

where $A(\boldsymbol{z})$ is the 2D z-transform of the filter corresponding to $\mathbf{a}$ and $e^{(\cdot)}$ denotes element-wise exponentiation. In this work, two classic spectral distances as well as a number of filterbank-based metrics were considered. Given two random field models $s(\boldsymbol{x})$ and $t(\boldsymbol{x})$ and their parameters, $(\sigma_s, \mathbf{a}_s)$ and $(\sigma_t, \mathbf{a}_t)$, respectively:

- We define the log-spectral distance by

$$\mathrm{LSD} = 10\sqrt{\int_{\square} W(\boldsymbol{f}) \left(\log_{10} \frac{\Phi_s(e^{\jmath 2\pi \boldsymbol{f}})}{\Phi_t(e^{\jmath 2\pi \boldsymbol{f}})}\right)^2 \mathrm{d}\boldsymbol{f}} \tag{2}$$

  where the integral is a double integral across the unit square ($\square$) and $W(\boldsymbol{f})$ is an optional weighting function.
- The 2D Itakura distance is defined by

$$\mathrm{Itakura} = \left|\ln \int_{\square} W(\boldsymbol{f}) \frac{\Phi_s(e^{\jmath 2\pi \boldsymbol{f}})}{\Phi_t(e^{\jmath 2\pi \boldsymbol{f}})} \,\mathrm{d}\boldsymbol{f}\right| \tag{3}$$

  Note that the Itakura distance is non-symmetric. Therefore, it is additionally evaluated with reversed arguments (designated 'A' and 'B' in what follows).
- The STSIM (Structural Texture Similarity Index [7]) is defined in the complex wavelet domain. Originally, it is computed on texture images. However, in this work, focus was on comparing texture models (i.e. texture parameter sets). Therefore, global averaging for computation of the STSIM was assumed. Under this condition, the expected value of the STSIM can be computed simply by considering the model PSDs. It could therefore be considered a spectral distance. The steerable pyramid [8] with four different numbers of scales (s) and orientations (o) was used.

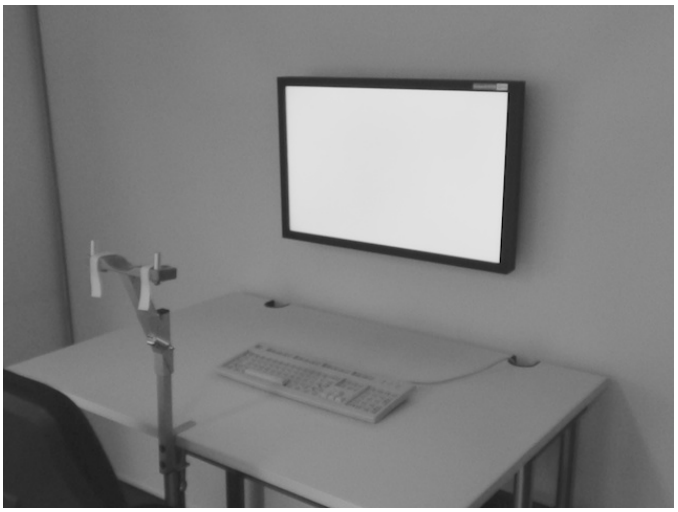


Fig. 1. Experimental Setup

- We define the SSTSIM ("simplified STSIM") by dropping the directional correlation components from the STSIM, thus reducing it to a function of subband power.
- We define the magnitude RMSE ("root mean square error"), likewise, as a function of subband power:

$$\text{Mag. RMSE} = \left(\sum_h \frac{1}{HE_h} \left(\sqrt{P_{s*h}} - \sqrt{P_{t*h}} \right)^2 \right)^{1/2} \tag{4}$$

where P_{s*h} is the power (zeroth-order autocorrelation) of the random field s filtered by the subband filter h, H is the total number of subbands, and E_h is the energy of the subband filter impulse response.

All of the metrics were computed on a 1024×1024 discretization of the PSDs of both texture models and approximating the integrals as sums. The weighting functions included $W(\boldsymbol{f}) \sim 1/|\boldsymbol{f}|^k$, where $|\boldsymbol{f}|$ indicates magnitude of the spatial frequency vector and $k \in \{0, 1, 2\}$.

II. Experimental Setup and Analysis

Subjective scores of texture similarity were acquired in a room with a uniform gray background with dim environmental lighting (Figure 1), using an EIZO SX3031W display set to sRGB color space, gamma 2.1, connected to an Apple Mac Pro 2.66 GHz Quad-Core Intel Xeon. The test subjects were asked to rate the similarity of pairs of texture images on a scale of 1 to 5 using the up and down arrow keys and confirm their selection using the space bar. All keys except the arrow keys and the space bar were removed from the keyboard. A chin rest was used to ensure a controlled viewing distance of approximately 80 cm.

Before being introduced to the experiment, test subjects were checked for visual acuity using the Freiburg Vision Test [9] version 3.7.1 using a viewing distance of 1.7 m. The decimal visual acuity of all test subjects was determined to be at least 1.2, such that the resolution of the display was just below the visual discrimination capabilities of the test subjects, or lower.

A. Prior Distribution of Texture Model Parameters

In earlier work [10] it was observed that the marginal prior distribution of GMRF texture parameters occurring in natural images resembles a double exponential (Laplace) distribution. The texture models were obtained using a Monte Carlo technique. The first texture model of each pair was generated by randomly drawing GMRF parameters from a Laplace distribution. Since textures occurring in natural images do not contain arbitrarily high peaks in their power spectral density functions, a multi-start gradient descent method was used to estimate the global spectral maximum of each generated texture model. New candidate model parameters were randomly drawn until a model was obtained whose power spectral density is bounded by a threshold. To find the second model of each pair, the parameters of the first model were perturbed by a vector of independent Gaussian noise. The second texture model candidates were subjected to the same routine of spectral analysis to avoid high peaks in the second model, as well.

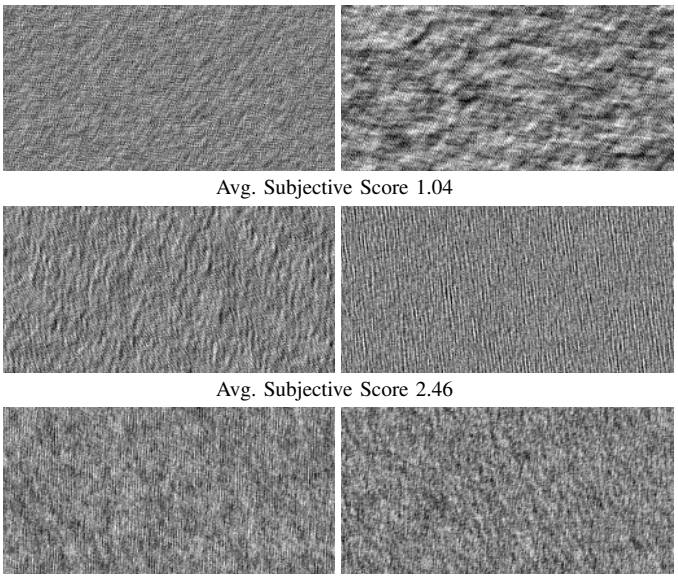
To ensure that the stimuli were approximately equally distributed on the objective and subjective scales, the following procedure was used: After a pair of texture models passed the spectral analysis, the STSIM between the two models was calculated and the pair was classified according to its STSIM value into 30 classes ranging from a STSIM of .85 to an STSIM of 1. The generation of new model pairs was iterated until there was at least one model pair in each class. All except the model pair closest to the central STSIM value of each class were discarded, such that the remaining model pairs encompassed approximately equally spaced STSIM values between .85 and 1. These 30 model pairs were used to generate all double stimuli.

B. Presentation of Stimuli

Each double stimulus consisted of a pair of zero-mean homogeneous texture images of 1024×1024 pixels displayed side-by-side (with a gap in-between) for exactly 6 s. The background was plain gray (corresponding to the zero level of the images). There was no time limit for the rating, and a 2 s pause between the confirmation of each rating and the next double stimulus. Each stimulus was generated by computing the PSD of the corresponding texture model on a discrete grid the same size as the image, taking its square root, multiplying by the DFT of an independent Gaussian noise array (which was randomly generated for each stimulus), and taking the inverse transform. This is effectively the algorithm in [11] applied to Markov fields. To each test subject, 6 runs of double stimuli were presented, where each run consisted of 30 double stimuli. Within each run, each of the 30 model pairs were used exactly once, while their order was randomly permuted for each run.

C. Analysis of Subjective Scores

The experiment was conducted with 25 test subjects. Their responses were analyzed as follows. Firstly, the scores from the first run of double stimuli were discarded for each test subject in order to allow them to adapt to their task without having an effect on the results. Then, the average standard deviation



Avg. Subjective Score 1.04

Avg. Subjective Score 2.46

Avg. Subjective Score 4.19

Fig. 2. Samples of the Texture Model Pairs and Average Subjective Scores

of the remaining responses of each subject was analyzed to check for consistency.

A certain variance of the responses can be explained due to the quantization of the response scale. For example, a test subject may give alternate responses between two scores if the subjective similarity is in-between. In the worst case, this results in a standard deviation of 0.5. Three of the test subjects responded with an average standard deviation above this value. Further analysis also showed that the responses of these test subjects were badly correlated to the responses of all other subjects (Pearson's and Spearman's $\rho < 0.9$). These test subjects were therefore ignored. Another three subjects with equally bad correlation to the rest were also ignored. Analysis showed that their responses were consistent and in principle along the same lines as the others', but either much more conservative or more radical (for example, answering with a score of 1 to more than 60% of all double stimuli).

To counteract singular outliers, the scores were averaged by removing the highest and the lowest score for each test subject and model pair (i.e., removing two of five responses for each subject and model pair) and then taking the arithmetic mean of the remaining values, either per-subject for the subject–subject correlation or across all 19 remaining subjects for the objective–subjective correlation. The mean subject–subject correlation coefficient was computed by taking the arithmetic mean of the correlation coefficients between the ${}^{19(19-1)}/_{2} = 171$ pairs of non-identical test subjects. It was found to be $\rho = 0.930$ (Pearson's), $\rho = 0.924$ (Spearman's), and $\tau = 0.838$ (Kendall's). Examples of textures and their average subjective scores are given in Figure 2.

III. Results

The average subjective scores were compared to the objective metrics using Spearman's ρ and Kendall's τ (Tables I and II). The tables show that all metrics perform well capturing the characteristics of texture important to the human observers

	LSD		Itakura A		Itakura B	
k	ρ	τ	ρ	τ	ρ	τ
0	0.086	0.060	0.213	0.157	0.090	0.078
1	0.689	0.520	0.387	0.285	0.108	0.120
2	**0.957**	**0.833**	**0.788**	**0.626**	**0.730**	**0.575**

TABLE I
Correlation Between Average Subjective Scores and Weighted Spectral Distances

	STSIM		SSTSIM		Mag. RMSE	
s/o	ρ	τ	ρ	τ	ρ	τ
6/6	0.853	0.681	0.874	0.709	0.905	0.755
6/8	0.856	0.686	0.865	0.700	0.897	0.741
8/8	0.938	0.806	**0.940**	**0.815**	**0.913**	0.764
8/10	**0.939**	**0.810**	0.934	0.806	0.910	**0.769**

TABLE II
Correlation Between Average Subjective Scores and Filterbank-Based Metrics

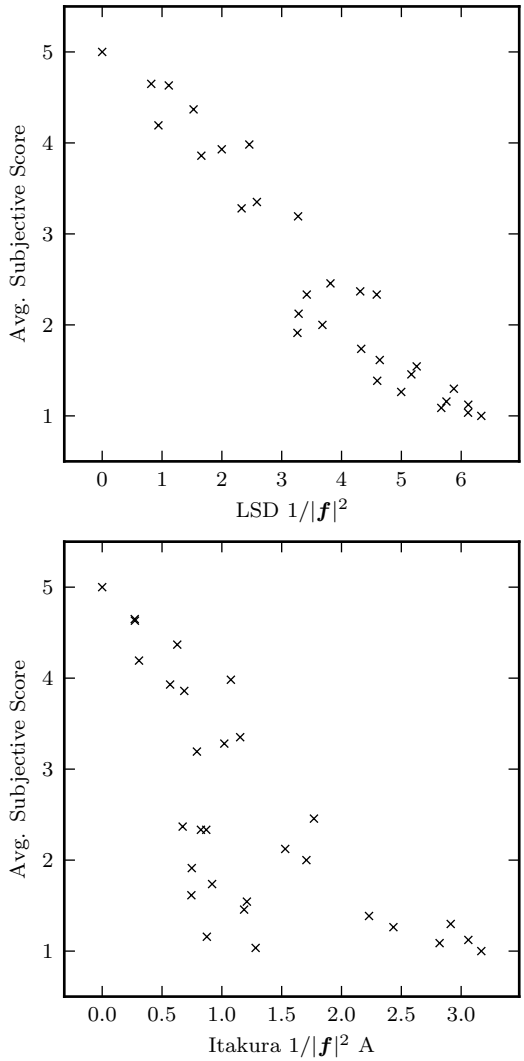


Fig. 3. Scatter Plots for Spectral Distances

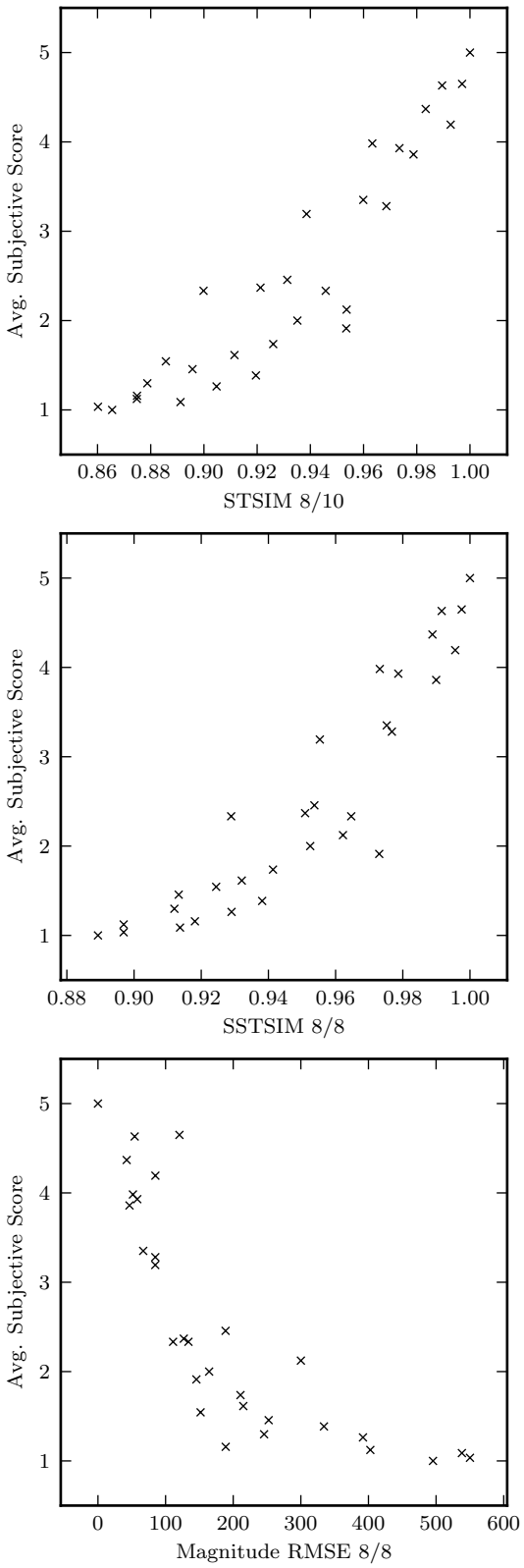


Fig. 4. Scatter Plots for Filterbank-Based Similarity Metrics

when the appropriate weighting or filterbank resolution is used.

This is also evident from Figures 3 and 4, where the average subjective score is plotted against each of the best-performing metrics from the correlation tables (indicated in bold). The $k = 2$ weighting should be expected to correlate well to the subjective scores as this corresponds to the scale-invariant statistics of natural images and the properties of visual cortex cells [12]. This property is also reflected in the steerable filterbank. The primary utility of the Itakura distance is the fact that estimators based on it can be implemented very efficiently in the spatial domain [1]. It appears that this computational benefit comes at the cost of a significantly worse correlation to the subjective scores, even when a weighting is used.

The fact that 22 of 25 test subjects responded very consistently with respect to stimuli that were generated from the same texture model suggests that the randomness introduced into each stimulus had no effect on subjective similarity.

IV. Conclusion

Several spectral distances have been compared to filterbank-based metrics such as the STSIM. It turns out that, considering GMRF texture, the weighted LSD tends to outperform STSIM, while the latter also performs reasonably well. Additionally, the study provides evidence supporting the earlier theoretical result that the information about the visual appearance of GMRF texture can be comprehensively summarized by the PSD [1]. Future research will include optimization of existing compression methods with respect to the results of this evaluation.

References

[1] J. Ballé, A. Stojanovic, and J.-R. Ohm, "Models for static and dynamic texture synthesis in image and video compression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1353–1365, Nov. 2011.

[2] M. Bosch, F. Zhu, and E. J. Delp, "Segmentation based video compression using texture and motion models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1366–1377, Nov. 2011.

[3] F. Zhang and D. R. Bull, "A parametric framework for video compression using region-based texture models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1378–1392, Nov. 2011.

[4] R. J. O'Callaghan and D. R. Bull, "Combined morphological-spectral unsupervised image segmentation," *IEEE Transactions on Image Processing*, vol. 14, no. 1, pp. 49–62, Jan. 2005.

[5] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proc. of IEEE International Conference on Computer Vision ICCV*, Sep. 1999, pp. 1033–1038.

[6] J. Ballé and M. Wien, "A quantization scheme for modeling and coding of noisy texture in natural images," in *Proc. of IASTED Conference on Signal and Image Processing SIP '09*. Honolulu, HI, USA: ACTA Press, Calgary, Aug. 2009.

[7] X. Zhao, M. G. Reyes, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for retrieval applications," in *Proc. of IEEE International Conference on Image Processing ICIP '08*, San Diego, CA, USA, Oct. 2008, pp. 1196—1199.

[8] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[9] M. Bach, "The Freiburg Visual Acuity Test – automatic measurement of visual acuity," *Optometry & Vision Science*, vol. 73, no. 1, pp. 49–53, Jan. 1996.

[10] C. Feldmann and J. Ballé, "Improved entropy coding for component-based image coding," in *Proc. of IEEE International Conference on Image Processing ICIP '11*, Bruxelles, Belgium, Sep. 2011.

[11] B. Galerne, Y. Gousseau, and J. M. Morel, "Random phase textures: Theory and synthesis," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 257–267, Jan. 2011.

[12] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America A*, vol. 4, no. 12, pp. 2379–2394, Dec. 1987.