Parametric quality assessment of synthesized textures

Darshan Siddalinga Swamy[1], Kellen J. Butler[1], Damon M. Chandler[1], and Sheila S. Hemami[2]

[1] Laboratory of Computational Perception and Image Quality
School of Electrical and Computer Engineering
Oklahoma State University, Stillwater, OK 74078 USA

[2] Visual Communications Lab
School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14850 USA

ABSTRACT

In this paper, we present the results of a study designed to investigate the visual factors which contribute to the perceived quality of synthesized textures. A psychophysical experiment was performed in which subjects rated the quality of textures synthesized from a variety of modern texture-synthesis algorithms. The ratings were given in terms of how well each synthesized texture represented a sample from the same material from which the original texture was obtained. The results revealed that the most detrimental artifact was lack of structural details. Other pronounced artifacts included: (1) misalignment of the texture patterns; (2) blurring introduced in the texture patterns; and (3) repeating the same patch again and again (tiling). Based on these results, we present an analysis of the efficacy of various measureable parameters at predicting the ratings. We show how a linear combination of the parameters from a parametric texture-synthesis algorithm demonstrates better performance at predicting the ratings compared to traditional quality-assessment algorithms.

1. INTRODUCTION

Given an example texture, the goal of texture synthesis is to generate (synthesize) a new sample of texture which, to a human observer, appears to be taken from the same source material. Over the past several decades, a large body of research has focused on developing accurate and efficient texture-synthesis algorithms. Parametric texture-synthesis algorithms operate by first analyzing the input texture to measure various parameters (e.g., low-order statistics), and then synthesizing a new texture sample by molding a random pattern to exhibit the same or similar parameters (e.g., Refs. 1, 2). Other texture-synthesis algorithms are generally considered non-parametric algorithms which operate using a variety of techniques ranging from strategic tiling to model-based approaches (e.g., Refs. 3-6). Some key applications of texture synthesis include its use in texture-synthesis-based image and video coding, computer rendering, and occlusion fill-in. The underlying models used to represent textures have also found use in applications such as texture retrieval and segmentation.

Yet, despite the advancements in texture modeling and synthesis, much less effort has been focused on developing algorithms for assessing the visual quality of a synthesized texture. Given two synthesized versions of an original texture, a human observer can readily determine which of the two synthesized versions better represents the original source. From a computational standpoint, however, this task is extremely challenging. Although a great deal of human vision research has been conducted to investigate the perceptual and neural mechanisms which underlie the visual appearance of texture (see Ref. 7 for a review), it remains unclear how these mechanisms can be applied to quality assessment of synthesized textures. A major factor which complicates this task is that point-by-point comparisons, a common approach used to some extent by all modern quality-assessment algorithms cannot be used to compare the visual similarity of two textures. In the context of texture retrieval, Zujovic and Pappas[8] address quality-assessment of textures by designing a structural similarity index specifically for textures. In Ref. 9, Pappas and Zujovic also demonstrate the utility of their index for texture-synthesis-based image compression.

In this paper, we focus specifically on quality assessment of textures generated by texture-synthesis algorithms. We believe that this specific focus on synthesized textures is an easier and more

D.S.S.: E-mail: darshan.siddalinga_swamy@okstate.edu; K.J.B.: E-mail: kellen.butler@okstate.edu;
D.M.C.: E-mail: damon.chandler@okstate.edu; S.S.H.: E-mail: hemami@ece.cornell.edu

Human Vision and Electronic Imaging XVI, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7865, 78650B · © 2011 SPIE-IS&T
CCC code: 0277-786X/11/$18 · doi: 10.1117/12.876743

feasible task than general texture similarity because the images to be compared (synthesized vs. original) generally exhibit some degree of visual similarity. Here, we present the results of a psychophysical experiment performed to obtain subjective ratings of synthesized-texture quality. Forty-two original textures were chosen from the Brodatz database.[10] The textures were chosen by selecting six textures from each of the seven perceptual clusters previously reported for the Brodatz textures.[11] Six texture-synthesis algorithms were used to generate synthesized versions of each original texture (6x42=252 synthesized textures). We then asked human subjects to rate the quality of each synthesized texture in terms of how well it represented a sample from the same source material from which the original texture was obtained. The ratings were made relative to all other synthesized textures in the database.

We then analyzed the results of the database, both qualitatively and quantitatively, in an attempt to answer the following questions:

1. What types of common texture-synthesis artifacts do subjects find most/least displeasing?
2. To what extent does the impact of an artifact depend on the type (category) of the texture?
3. Can the parameters from a modern texture-synthesis algorithm be used to predict the ratings for textures synthesized from other parametric and non-parametric methods?

This paper is organized as follows: Section 2 presents the details of the psychophysical experiment. Section 3 discusses the experimental results and predictions. General conclusions are provided in Section 4.

2. METHODS

2.1 Apparatus and Subjects

A psychophysical experiment was conducted to rate the performance of six texture synthesis algorithms based on their visual quality. Eight adult subjects including the first author participated in the experiment. All subjects had either normal or corrected-to-normal visual acuity. All subjects were between the ages of 22 to 28 years. Stimuli were displayed on a LACIE 324, 24-inch LCD monitor (1920 × 1200). The display yielded minimum and maximum luminance's of 0.17 cd/m^2 and 266 cd/m^2 respectively and an overall gamma of 2.6. Stimuli were viewed binocularly through natural pupils in a darkened room at a distance of approximately 60 cm.

2.2 Stimuli

Forty two textures *D2, D4, D5, D10, D11, D12, D29, D31, D33, D35, D37, D38, D40, D42, D43, D45, D46, D54, D58, D62, D65, D67, D71, D72, D74, D77, D84, D85, D86, D87, D89, D91, D93, D94, D97, D99, D101, D104, D108, D109, D110 and D111* obtained from the Brodatz database[10] served as original images in the experiment. Six images were chosen for each of the following seven perceptual clusters reported previously for this database.[11]

1) Extended contours (approximately straight)
2) Periodic clusters
3) Extended contours (curved)
4) Rocks
5) Inhomogeneous
6) Stone
7) Grainy

The above digital images were resized to 256×256 pixels from 640×640 pixels using the imresize function in Matlab with bicubic interpolation. The images used were 8-bit grayscale images.

Figure 1 shows seven original textures (one from each cluster) and their synthesized textures. The first column of Figure 1 represents the original textures. All other columns in Figure 1 represent the synthesized textures for these original textures. The texture synthesis algorithms used in the experiments were:

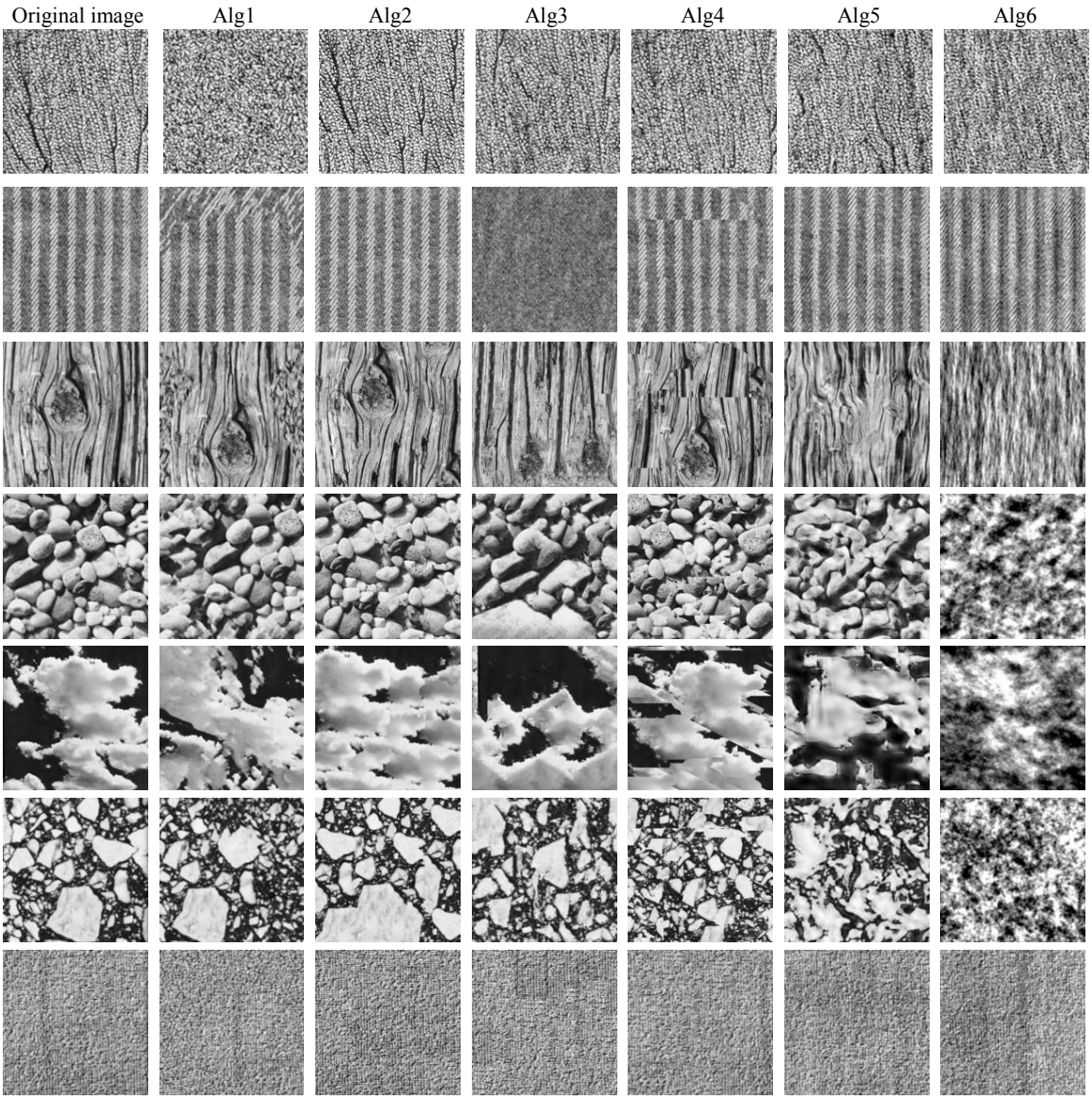

Figure 1. Representative stimuli used in the experiment. *First column*: shows original textures, one from each cluster. *Second column* shows synthesized textures using Alg1. *Third* c*olumn* shows synthesized textures of Alg2. *Fourth* c*olumn* shows synthesized textures of Alg3. *Fifth* c*olumn* shows synthesized textures of Alg4. *Sixth* c*olumn* shows synthesized textures of Alg5. *Seventh* c*olumn* shows synthesized textures of Alg6.

Non-parametric texture synthesis algorithms:

1) Efros and Leung's algorithm[3] (Alg1)
2) Image Quilting algorithm of Efros and Freeman[4] (Alg2)
3) Wei and Levoy algorithm[5] (Alg3)
4) Hertzmann *et al*. algorithm[6] (Alg4)

Parametric texture synthesis algorithms:

5) Portilla and Simoncelli's parametric model[1] (Alg5)
6) Galerne *et al.* algorithm[2] (Alg6)

2.3 Procedures

Subjective ratings were measured by using a continuous rating system in which the six synthesized textures were compared against the original texture. Seven textures were displayed simultaneously on a solid gray background with the original texture fixed on the leftmost position of the screen. The subjects were instructed to place the six synthesized versions to the right of the original such that the horizontal displacement between each synthesized texture and the original was linearly proportional to their subjective assessment of visual quality. Quality was defined as how well each synthesized texture represented a sample from the same source material from which the original texture was obtained. Thus, textures which were placed further to the right from the original image were judged to have lower visual quality relative to the original. The time for each texture was not limited, but the average time for all the subjects for the entire experiment was between 55 and 60 minutes.

When the experiment was completed by all the subjects, an additional experiment was performed by five observers (including the first and third authors) using the same procedures but with simultaneous presentation of 168 textures. This follow-up experiment was necessary to scale the within-texture results to between-texture scores. The raw scores for each subject on each of the original textures were converted to z-scores and these z-scores were averaged over all the subjects. The average z-scores were scaled from within-texture results to between-texture scores, and these results were then normalized to the range [0, 1], where a score of 1 corresponds to the worst synthesized image (DMOS values).

3. RESULTS AND ANALYSIS

3.1 Experimental Results

The experimental results show the performance of parametric and non-parametric algorithms. Textures which preserved the structures of the original texture were rated best by the subjects. The parametric algorithms were not able to regenerate complex structures, for example see Figure 1 third row Alg5 and 6. The DMOS for the parametric algorithms were 0.56 with a standard deviation of 0.24. For the non-parametric algorithms, the average DMOS was 0.25 with a standard deviation of 0.19. These results show that the parametric algorithms were outperformed by the non-parametric algorithms for most of the textures. The parametric algorithms performed well only on the micro-textures[15], i.e., textures with no structures (for example, see Figure 1 second and last row Alg5 and 6).

In addition to the lack of structure, the ratings revealed that subjects disliked textures containing the following artifacts:

1) Too many repetitions of the same patch (tiling); see Figure 1, fifth row, Alg2.
2) Misalignment of the texture patterns; see Figure 1, Alg4.
3) Blurring in the texture patterns, performed to mask the misalignments; see Figure 1, fourth row, Alg1, 3.

The average subjective ratings of textures with these artifacts were 0.452, 0.414 and 0.461 respectively.

Table 1 shows the performance of the algorithms on the seven clusters. Non-parametric algorithms performed well on all seven clusters (see Table 1). Parametric algorithms performed well only on some of the clusters. Parametric algorithms failed on straight and curved extended contours (clusters 1 and 3, respectively) as they were not able to regenerate these shapes. Parametric algorithms performed well on periodic (cluster 2) and grainy (cluster 7) textures, as they were microtextures[15], but not as good as the non-parametric algorithms. Parametric algorithms failed on stone textures (cluster 4) and rocks (cluster 6) which had structures; for example, see Figure 1, fourth and sixth rows, respectively. Parametric algorithms failed on inhomogeneous textures (cluster 5) which also had shapes and structures.

Table 1. Average subjective ratings of the synthesized textures for each of the seven clusters.

Clusters	Average subjective ratings (DMOS)		
	For all six algorithms	Non-parametric algorithms	Parametric algorithms
Cluster 1	0.334	0.251	0.501
Cluster 2	0.324	0.267	0.440
Cluster 3	0.401	0.311	0.582
Cluster 4	0.385	0.245	0.664
Cluster 5	0.384	0.239	0.674
Cluster 6	0.363	0.257	0.575
Cluster 7	0.290	0.225	0.419

3.2 Analysis of Change in Parameters

One way to estimate the quality of a synthesized texture is to compare parameters extracted from the source texture with those extracted from the synthesized texture. We evaluated the following parameters:

1. Portilla and Simoncelli's parameters[1]: The include the following:
    a. Mean, variance, skewness, kurtosis, minimum, and maximum of the pixel values
    b. Autocorrelation of the lowpass and Highpass bands (p_1)
    c. Cross correlation of the orientations (p_2) and scale (p_3)
    d. Cross-scale phase (p_4)
2. Kullback-Leibler divergence[12] (KLD) computed on the Steerable Pyramid coefficients from Ref. 1 (p_5)
3. The first-order entropy of the pixel values[13]

The steerable pyramid used for Portilla and Simoncelli's algorithm and KLD had four scales and six orientations. We modified some of the configurational settings of Portilla and Simoncelli's algorithm for some of the parameters. For auto correlation of the high pass bands, we used a spatial neighborhood of 7×7. We weighted the autocorrelation coefficients for each scale of the high pass bands. Weights used were 0.7 for the first finest scale and 0.1 to the remaining scales. For cross-scale phase, we used a spatial neighborhood of 15×15.

For Portilla and Simoncelli's parameters and for the entropy of the pixels, we computed the absolute difference between each parameter obtained from the original texture and the corresponding parameter obtained from the synthesized texture. For KLD, we computed the KLD of the magnitudes between the steerable pyramid subbands of the original and synthesized textures. The minimum of the KLDs obtained from each subband was assigned as the overall KLD.

Table 2. Correlation of each parameter with the subjective ratings with and without Alg5 obtained from a testing set of one of the trials of cross-validation.

Parameter Differences	**Correlation w/ Alg5**	**Correlation w/o Alg5**
Mean	0.08	0.19
Variance	0.36	0.39
Skewness	0.40	0.42
Kurtosis	0.27	0.31
Minimum intensity	0.39	0.42
Maximum intensity	0.28	0.36
Autocorrelation of the Lowpass bands	0.24	0.28
Autocorrelation of the Highpass bands (p_1)	0.70	0.78
Cross correlation of orientations (p_2)	0.59	0.64
Cross correlation of scale (p_2)	0.65	0.73
Cross scale phase (P_4)	0.48	0.48
KLD in frequency (P_5)	0.63	0.72
Entropy	0.36	0.44

To investigate how well each individual parameter (difference) performed in predicting the subjective ratings, nonlinear regression and cross-validation was performed. Cross-validation was used to avoid overfitting; it was performed multiple times on different partitions and the results were averaged over the validations. The 42 textures used in the experiments were randomly split into training and testing sets (containing 21 textures each).

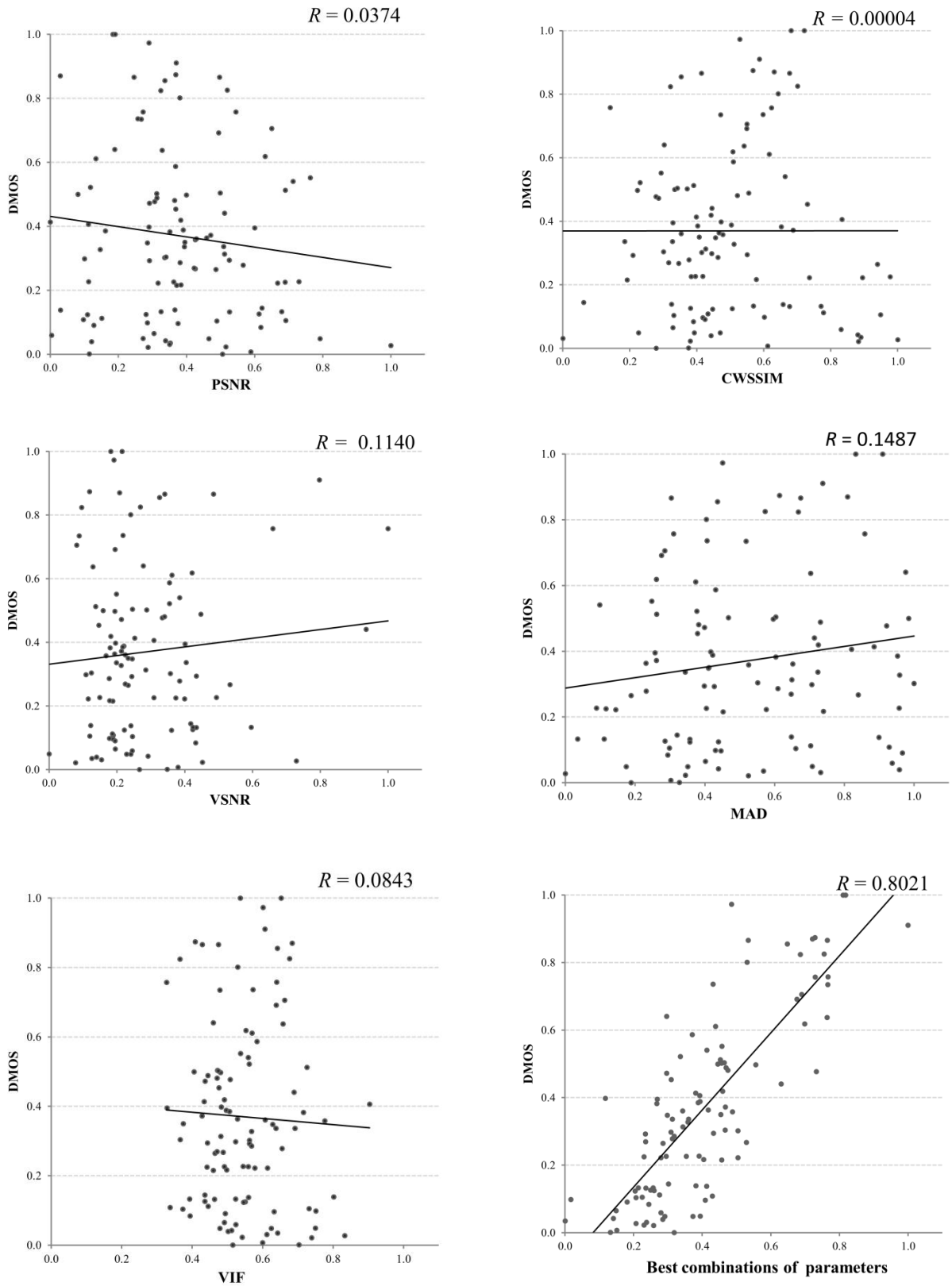

Figure 2. Shows the scatterplots of subjective ratings versus various estimators. A linear curve-fitting was used for the scatterplots and listed correlation coefficients.

The random splitting process was performed ten times and the results were averaged over the testing sets.

Table 2 shows the Pearson correlation coefficients between the evaluated parameters and subjective ratings using a testing set obtained from one of the trials in cross-validation. Some of the marginal statistics do not correlate well, but other parameters such as cross correlation, autocorrelation of the high pass bands, cross-scale phase, and KLD perform quite well. We also observed that the correlations improve if we do not include the stimuli synthesized using Portilla and Simoncelli's algorithm.

3.3 Combining the Parameters

We next examined whether a weighted combination of some of Portilla and Simoncelli's parameter differences (P_1, P_2, P_3 and P_4, see Table 2) and KLD (P_5) could yield a better prediction of quality than each parameter alone. Equation 1 shows how each of the chosen parameters were combined.

$$Q = \sum_{i=1}^{5} W_i \log_{10} P_i, \tag{1}$$

Here, Q represents the final quality estimate, and W_i represents the weight for the parameter P_i. After performing cross-validation for ten trials, we obtained the following weights: 0.45, 1, 0.1, 0.25 and 0.3 for P_1 through P_5, respectively.

Figure 2 shows scatterplots of subjective ratings plotted against estimates from PSNR, CWSSIM,[14] VSNR[15], VIF,[16] MAD[17], and the above combination on all textures except those from Alg5 (Portilla and Simoncelli). A testing set from one of the trials used in cross-validation was used for the scatterplots. The ratings obtained from each of the quality assessment methods have been normalized between zero and one and the ratings were fitted using a linear function. For the combination of parameters given by Equation (1), cross-validation yielded an average Pearson correlation coefficient of 0.804 and 0.814 for the training and testing sets, respectively.

4. CONCLUSION

Our analysis demonstrates that a parameter-based approach can be successful for quality assessment, despite the fact that parametric texture-synthesis algorithms generally perform worse in texture synthesis than non-parametric methods.

A psychophysical experiment was conducted to obtain visual quality ratings of synthesized textures. Overall, we found that non-parametric algorithms performed better than parametric algorithms. The artifacts found in the synthesized textures most disliked by the subjects were: lack of structure, tiling, misalignments, and blurring artifacts. We also found why the parametric algorithms failed on the some of the seven clusters: The parametric algorithms generally were not able to synthesize complex shapes and structures which subjects tended to look for when gauging visual quality.

In addition, we demonstrated that a texture quality assessment estimator which included KLD in frequency and select parameters from a parametric texture-synthesis algorithm can perform well in predicting the visual quality of the synthesized textures. Traditional image quality assessment algorithms failed on this task.

There are notable cases in which the proposed quality estimator fails to predict the subjective ratings. Figure 3 shows some of these failure cases. The average DMOS values for the synthesized textures in Figure 3 was 0.824 and the average normalized prediction was 0.478. Subjects rated these synthesized textures to be of relatively low quality due to the fact that the synthesized textures lack the objects found in the original textures (gravel, pebbles, and ice cubes). The quality estimator is not able to recognize that these objects are absent.

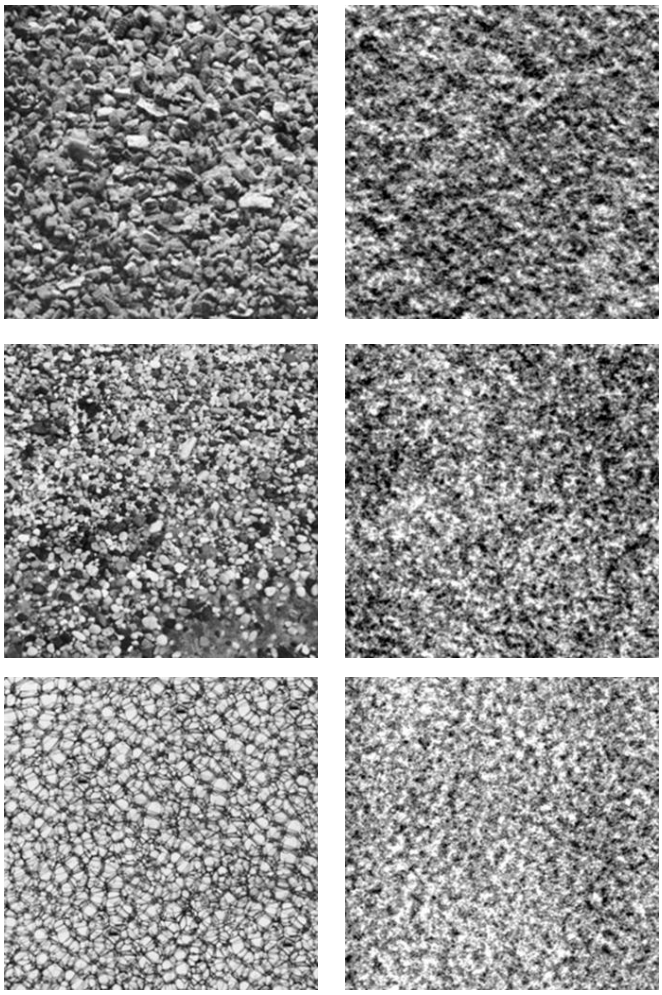

Figure 3. Shows three failure cases in the prediction. *First column* represents the original textures. *Second column* shows the synthesized textures for the originals respectively.

5. REFERENCES

1. J. Portilla and E.P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. Int'l Journal of Computer Vision, vol.40 (1), pages 49–71, December 2000. http://www.cns.nyu.edu/~eero/texture/
2. B. Galerne, Y. Gousseau and J. M. Morel, Random Phase Textures: Theory and Synthesis preprint CMLA N°2009-24, 2009. http://www.ipol.im/pub/algo/ggm_random_phase_texture_synthesis/
3. A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In IEEE International Conference on Computer Vision, pages 1033–1038, Corfu, Greece, September 1999. http://www.rubinsteyn.com/comp_photo/texture/
4. A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. Proceedings of SIGGRAPH 2001, pages 341–346, August 2001. http://www.cs.princeton.edu/~cdecoro/imagequilting/
5. Wei, L.-Y., and Levoy, M. 2000. Fast texture synthesis using tree structured vector quantization. Proceedings of SIGGRAPH 2000 (July), 479–488. http://www.its.caltech.edu/~matthewf/Links.html

6. Image analogies A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, D. Salesin SIGGRAPH 2001 Conference Proceedings. http://mrl.nyu.edu/projects/image-analogies/lf/
7. M. S. Landy and N. Graham, "Visual Perception of Texture," in The Visual Neurosciences, MIT Press, pages 1106-1118, 2004.
8. J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Perceptual similarity metrics for retrieval of natural textures," in Proc. IEEE Workshop on Multimedia Signal Processing, 2009
9. T. N. Pappas, J. Zujovic, and D. L. Neuhoff, "Image analysis and compression: Renewed focus on texture," in *Visual Information Processing and Communication*, Proc. SPIE, Vol. 7543, (San Jose, CA), Jan. 2010.
10. P. Brodatz, Textures: A Photographic Album for Artists and Designers, Dover, New York, 1996.
11. B. Balas. Attentive texture similarity as a categorization task: Comparing texture synthesis models Proceedings of pattern recognition society 2008, pages 972-982, August 2007
12. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second Edition ISBN: 0-471-24195-4.
13. R. C. Gonzalez, R.E. Woods, S.L. Eddins, *Digital Image Processing Using MATLAB*, New Jersey, Prentice Hall, 2003, Chapter 11
14. M. P. Sampat, Z. Wang, S. Gupta, A.C. Bovik and M.K. Markey, "Complex wavelet structural similarity: a new image similarity index" IEEE Transactions on Image Processing, Vol: 18 No: 11, October 2009 Page(s): 2385-2401
15. D. M. Chandler and S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," IEEE Transactions on Image Processing, 16, 2007.
16. H. R. Sheikh. and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol.15, no.2, pp. 430- 444, Feb. 2006.
17. E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," Journal of Electronic Imaging, 19 (1), March 2010.
18. MIT Vision and Modeling Group. Vision Texture, [Online]. Available: http://vismod.www.medis.mit.edu