A Psychophysical Evaluation of Texture Degradation Descriptors

Jiří Filip[1], Pavel Vácha[1], Michal Haindl[1], and Patrick R. Green[2]

[1] Institute of Information Theory and Automation of the ASCR, Czech Republic
[2] School of Life Sciences, Heriot-Watt University, Edinburgh, Scotland

Abstract. Delivering digitally a realistic appearance of materials is one of the most difficult tasks of computer vision. Accurate representation of surface texture can be obtained by means of view- and illumination-dependent textures. However, this kind of appearance representation produces massive datasets so their compression is inevitable. For optimal visual performance of compression methods, their parameters should be tuned to a specific material. We propose a set of statistical descriptors motivated by textural features, and psychophysically evaluate their performance on three subtle artificial degradations of textures appearance. We tested five types of descriptors on five different textures and combination of thirteen surface shapes and two illuminations. We found that descriptors based on a two-dimensional causal auto-regressive model, have the highest correlation with the psychophysical results, and so can be used for automatic detection of subtle changes in rendered textured surfaces in accordance with human vision.

Keywords: texture, degradation, statistical features, BTF, eye-tracking, visual psychophysics.

1 Introduction

Advanced graphics applications such as virtual interior design, cultural heritage ditization, etc. require considerable effort to render the appearance of real-world accurately. When it comes to photo-realistic appearance of materials there is no other way than to use view- and illumination-dependent measurements of real materials. Such measurements can be represented by means of *bidirectional texture functions* (BTF) [1]. Seven-dimensional BTFs represent challenging data due to theirs massive size and thus have high processing and rendering expenses. A number of approaches to BTF compression and modelling have been published in the past as shown in survey [2]. Although BTF generative statistical models exist that are capable to reach huge compression ratios themselves, they can profit from data measurement compression as well, as it can improve their learning and modelling efficiency.

The main disadvantage of most of the compression methods is that they have fixed parameters regardless of the type of sample being compressed. There have been attempts to use data on visual perception for improvement of texture data

compression. Filip et al. [3] applied a psychophysical study to obtain perceptually important subset of view- and illumination-dependent images and thus consequently reduced the amount the data to be processed. On the other hand, Guthe et al. [4] used standard contrast sensitivity in cone response space together with a psychometric difference for improvement of the data compression. Interactions of human gaze fixation with different surface textures have also been analysed [5]. Although these approaches provide pioneering introductions of perceptual methods for improvement of texture compression, they are not suitable for evaluation of subtle visual compression effects.

Contribution of the paper: The main motivation of our research is to find a *computational texture descriptor having responses highly correlated with human vision.* Such a descriptor could then be used for comparison of rendered images resulting from original data and data parameterized by compression methods. Based on the responses from the descriptor the methods could iteratively adapt their parameters to automatically achieve an optimal visual performance. In this paper we test a set of descriptors motivated by standard texture features used in texture retrieval and recognition application. The descriptors we tested are based on a structure similarity, visual difference predictor, local binary patterns, Gabor features, and a causal auto-regressive wide-sense type of Markov random field model. The performance of the descriptors was evaluated by a psychophysical experiment on a group of twelve subjects.

Paper organization: The experimental data are introduced in Section 2, while the tested descriptors are explained in Section 3. Section 4 describes the experimental setup and discusses the results obtained, while Section 5 evaluates the performance of the descriptors with respect to the experimental data. Section 6 summarizes the paper.

2   Test Data Design

To test robustness of the descriptors we designed a set of testing images. Each image features a cube whose three visible faces were rendered using textured materials. We used five different samples (Fig. 1) of view and illumination-dependent textures represented by Bidirectional Texture Functions (BTF) [6] (each sample comprise 81 illumination × 81 view directions, i.e. 6561 texture images of resolution 256 × 256 pixels).

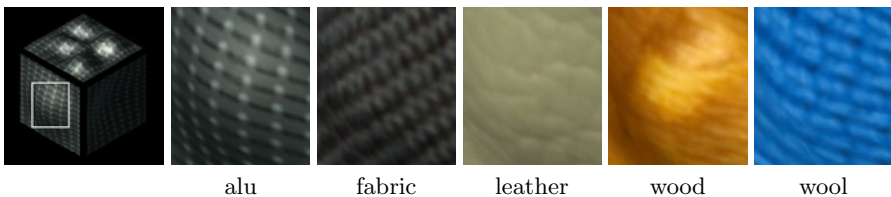alu   fabric   leather   wood   wool

Fig. 1. Examples of five tested material samples shown on a region of one test image

The cube faces were modified in a way to feature different geometry on all three visible faces (top, left, right). To test a range of shapes that occur in the real-world we used different shapes for each cube face: **I**-wide indent, **R**-random bumps, **B**-wide bump, **F**-flat face, **H**-horizontal waves, **V**-vertical waves.

For illumination we used directional light from left and right directions parallel with the upper edge of the cubes. This configuration guaranteed the same illumination of the cubes in all stimuli and similar distribution of light across the top and left/right faces in single cubes. Not all combinations of test cube orientations were used in the experiment as this would result in too high a number of test images. We used only eleven different orientations selected in a way to allow us to compare the most interesting combinations of faces geometry. Additionally, not all orientations were illuminated from both directions as shown in Fig. 2. The figure also shows orientation number (first row) and shapes of left, right, top faces (third row). To simulate possible effects of texture compression we used three filters introducing artificial degradation to the original data modification:

> **A** - illumination/view directions downsampling to 50%
> **B** - spatial filtering (averaging by kernel 3×3)
> **C** - spatial filtering (averaging by kernel 5×5).

1-L	1-R	2-L	2-R	3-L	4-L	5-R	6-L	7-L	8-R	9-R	10-L	11-L

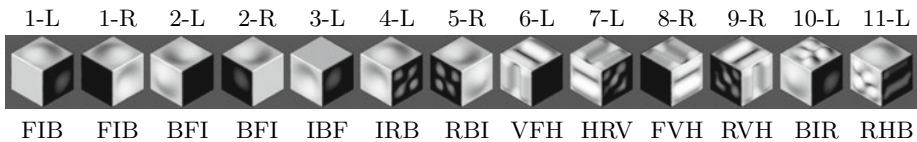| FIB	FIB	BFI	BFI	IBF	IRB	RBI	VFH	HRV	FVH	RVH	BIR	RHB

Fig. 2. Tested combinations of cube orientation and illumination direction

The proposed filters introduce only very subtle differences (Fig. 3) between the original and the modified data and force subjects to perform extensive visual search, which allows us to collect detailed gaze data. Finally, for 13 combinations of cube orientation & illuminations and 5 material samples, we obtained 65 test images for each degradation. These images were used for testing of texture descriptors proposed in the following section and also to generate stimuli in the validation experiment in Section 5.
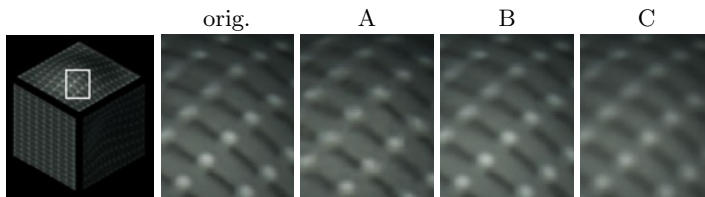
orig.	A	B	C

Fig. 3. Performance of the applied filters on sample *alu*

3   Texture Degradation Descriptors

The 65 images for each degradation were compared with their original counterparts. This means that we always compared images featuring the same sample, cube orientation and illumination direction. The only differences were faint degradation artifacts. Therefore, we do not require the texture descriptors to be view or illumination invariant. The descriptors tested in this paper can be principally divided into those which are translation invariant and those which are not.

3.1   Translation Non-invariant Features

These descriptors are based on perceptually motivated measures of image quality assessment measures computed in pixel-wise manner in a local neighborhood.

The first was **visual difference predictor** (VDP) [7], which simulates low level human perception for known viewing conditions (in our case: display size 37×30 cm, resolution 1280×1024 pixels, observer's distance 0.7 m) and thus is sufficient for our task of perceptually plausible detection of subtle texture degradation artifacts. The VDP provides percentage of pixels that differ with probability p>75% or p>95% from all pixels in the compared images. To ensure consistency with other descriptors, we set the VDP output to $(1-p)$, i.e. giving interval $(0,1)$, where for an output 1 the images are the same.

The **structure similarity index metric** (SSIM) [8] is an empirical measure, which compares in power to VDP. SSIM measures the local structure similarity in a local neighborhood of an $R \times R$ window in an image (we used 11×11 pixels). The basic idea of SSIM is to separate the task of similarity measurement into comparisons of luminance, contrast, and structure. These independent components are then combined into one similarity function. The valid range of SSIM for a single pixel is $[-1, 1]$, with higher values indicating higher similarity. When the local neighborhood is evaluated for each pixel we obtain the SSIM difference of two images as a mean value of SSIM values across all pixels.

3.2   Translation Invariant Features

Markovian features are derived from the multiscale representation assuming a **causal auto-regressive model** (CAR) for each of the K factorisation pyramid levels. The spatial factorization is done using either the Gaussian (GP) or Gaussian-Laplacian (GLP) pyramid. Single model parameters are estimated and the texture features from all pyramid levels are concatenated into a common feature vector.

Let us assume that each multispectral texture is composed of $C = 3$ spectral planes. $Y_r = [Y_{r,1}, \ldots, Y_{r,C}]^T$ is multispectral pixel at location $r = [x, y]$. The spectral planes are either modelled by 3-dimensional (3D) CAR model or by means of a set of C 2-dimensional (2D) CAR models. The CAR representation assumes that the multispectral texture pixel Y_r can be modelled as linear combination of its neighbors:

$$Y_r = \gamma Z_r + \epsilon_r \ , \qquad Z_r = [Y_{r-s}^T : \forall s \in I_r]^T \tag{1}$$

where Z_r is the $C\eta \times 1$ data vector with multiindices r, s, t representing a causal or unilateral neighbourhood, $\gamma = [A_1, \ldots, A_\eta]$ is the $C \times C\eta$ unknown parameter matrix with square submatrices A_s. Some selected contextual index shift set is denoted I_r and $\eta = cardinality(I_r)$. The white noise vector ϵ_r has normal density with zero mean and unknown constant covariance matrix, same for each pixel. Given the known CAR process history, estimation of the parameters $\hat{\gamma}$ can be accomplished using fast, numerically robust and recursive statistics [9]. Five colour invariants were derived from CAR parameter estimates [10]. The texture features are these illumination invariants, which are easily evaluated during the process of estimating CAR parameters. Because the CAR models analyse a texture in some fixed movement direction, additional directions are employed to capture supplementary texture properties. The distance between two feature vectors was computed using L_1, $L_{0.2}$ norms, and by a fuzzy contrast FC_3 [11]. Although the CAR models theoretically assume texture homogenity, they can be still used as statistical descriptors of textured surfaces, and so we expect their ability to detect the degradation artifacts.

The **Gabor features** (GF) [12] are computed from responses of Gabor filters [13], which can be considered as orientation and scale tuneable edge and line detectors. A two dimensional Gabor function $g(r) : \mathbb{R}^2 \to \mathbb{C}$ can be specified as

$$g(r) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi i V x\right] , \tag{2}$$

where σ_x, σ_y, V are filter parameters. The convolution of the Gabor filter and a texture image extracts edges of given frequency and orientation range. The whole filter set was obtained by four dilatations and six rotations of the function $g(r)$, and the filter set is designed so that Fourier transformations of filters cover most of the image spectrum, see [12] for details. The Gabor features [12] are defined as the mean μ_j and the standard deviation σ_j of the magnitude of filter responses computed separately for each spectral plane and concatenated into the feature vector. These feature vectors are compared in the $L_{1\sigma}$ norm [12]. The other tested *Opponent Gabor features* (OGF) [14] are extension to colour textures, which analyses also relations between spectral channels. As our implementation involves FFT the Gabor features were computed only in the square cuts in each cube face.

The **Local Binary Patterns** (LBP$_{P,R}$) [15] are histograms of texture micro patterns, which are thresholded values sampled at each pixel neighbourhood. For each pixel, a circular neighbourhood around the pixel is sampled, P is the number of samples and R is the radius of the circle. Sampled points values are thresholded by a central pixel value and the pattern number is formed as follows:

$$LBP_{P,R} = \sum_{s=0}^{P-1} sgn\left(Y_s - Y_c\right) 2^s, \tag{3}$$

where *sgn* is the signum function, Y_s is a grey value of the sampled pixel, and Y_c is a grey value of the central pixel. Subsequently, the histogram of patterns is computed. Because of thresholding, the features are invariant to any

monotonic grey scale change. The multiresolution analysis is done by growing the circular neighbourhood size. All LBP histograms were normalised to have a unit L_1 norm. The similarity between the histograms is computed using Kullback-Leibler divergence as authors suggested. We have tested combination of $LBP_{8,1}$ and $LBP_{8,3}$ features, and they were computed either on grey-scale images (grey) or on each spectral plane separately (RGB) and concatenated to form the feature vector.

All descriptors compute difference between sets of original images and images obtained for each degradation method, and their responses are averaged across different cube orientations, and illumination directions. It is important to note that the previous textural features are not invariant to texture deformation, which is cased by different shapes. Therefore, the features are always compared between the same surface shapes only.

4   Psychophysical Experiment

We performed a visual search experiment in order to investigate subjects' ability to identify individual introduced visual degradations. We also recorded their gaze fixations in order to analyse relations between their decisions and their fixations statistics.

Experimental Stimuli. For experimental stimuli we have used static images of size 1000×1000 pixels, featuring four cubes, described in Section 2, in individual quadrants (see Fig. 4-middle). We used this layout of stimuli to avoid the central bias in fixations reported in [16], i.e. observers have a tendency to fixate the central area of the screen. In each quadruple, three cubes showed the original data rendering and the remaining one showed a slightly modified rendering. The position of the modified cubes was random. Examples of stimuli are shown in Fig. 4. The edges of the cubes were set to black to mask potentially salient texture seams. The background and the remaining space on the screen were set to dark gray. Fig. 2 shows the 13 conditions of cube orientation and illumination direction that were used. Together with five BTF texture samples, and three different filters, the total number of stimuli was 195 ($13 \times 5 \times 3$).

Participants. Twelve paid observers (three females, nine males) participated in the experiments. All were students or university employees, were less than 35 years of age, and had normal or corrected to normal vision. All were naive with respect to the purpose and design of the experiment.

Experimental Procedure. The participants were shown the 195 stimuli in the same randomly generated order, and asked to identify which of the cubes had a surface texture slightly different from the remaining three cubes. A stimulus was shown until one of four response keys, identifying the different cube, was pressed. There was a pause of one second between stimulus presentations, and participants took on average around 90 minutes to perform the whole experiment, which was split into four sessions. All stimuli were presented on a calibrated 20.1" NEC2090UXi LCD display (60Hz, resolution 1600×1200, color
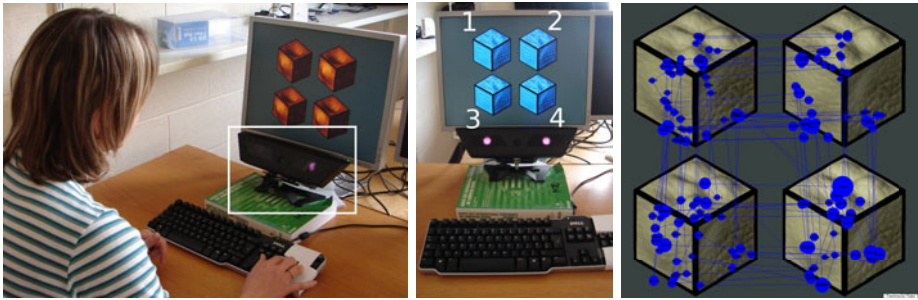
Fig. 4. Setup of the experiment with the eye-tracker highlighted, presentation of stimulus image from subject's view, and a typical gaze fixation pattern

temperature 6500K, gamma 2.2, luminance $120 \, \mathrm{cd/m^2}$). The experiment was performed in a dark room. Participants viewed the screen at a distance of 0.7m, so that each sphere in a pair subtended approximately 9° of visual angle. Subjects' gaze data was recorded during the experiment using a Tobii x50 infrared-based binocular eye-tracking device as shown in Fig. 4. The device was calibrated for each subject individually and provided the locations and durations of fixations at a rate of 50 samples/s. The shortest fixation duration to be recorded was set to 100 ms.

Results – Responses accuracy. On average, the subjects were able to find the modified cube in 67% of the stimuli, which was surprisingly high in relation to the chance level 25%, given the subtle changes introduced by filters used (see Fig. 3). Informal interviews after the experiment revealed that the subjects were certain in less than 50% of stimuli and for the rest they believed that they were only guessing the right answer. The obtained rates suggest that in the difficult cases they often successfully relied on low level visual perception. The responses accuracy of individual filters is shown in Fig. 5-a and reveals that modifications introduced by the filter **A** are the hardest to spot while the smoothing by filter **C** is the most apparent; this was expected, since smoothing effect is uniform and generally more apparent that the slight illumination and view direction dependent changes in reflectance caused by reduction of directions reduction (filter **A**). While success rates across textures were quite similar for smoothing filters **B** and **C**, their values for filter **A** varied much more.

Results – Fixations. Twelve subjects performed 62 916 fixations longer than 100 ms. Average fixation duration was 242 ms. Each stimulus was on average fixated for 11 s by means of 26 fixations. Figures Fig. 5-b,c,d show subjects' gaze fixation statistics as (b) average number of fixations per stimuli, (c) average time spent fixating stimuli , and (d) average fixation time. The first two statistics are highly inversely correlated with subjects' response accuracies Fig. 5-a, with correlation coefficients $R_{(b)} = -0.904$ and $R_{(c)} = -0.930$, respectively. The figures also reveal apparent differences between the tested samples. For samples *leather* and *wood*, the subjects were less successful in identification of the modified cube,
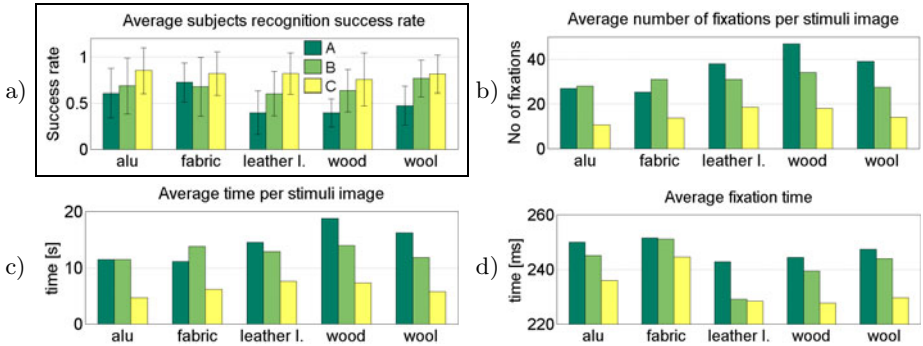
Fig. 5. Subjects' average (a) recognition success rate, (b) number of fixations per stimuli, (c) time spend on stimuli, (d) fixation duration for individual degradations and tested samples. Error-bars represent twice the standard error across subjects, different cube orientations and illuminations.

they fixated the stimuli for longer, and made significantly more fixations, which were shorter those on the other materials. We suspect that a lower local texture contrast in these samples makes detection of degradation artifacts more difficult.

5   Perceptual Evaluation and Discussion

In this section we evaluate performance of the proposed descriptors by comparison with subjects' responses obtained from the psychophysical experiment. The evaluation was based on computation of correlation coefficient $R_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$, where X, Y are compared data vectors, i.e. subjects responses and descriptor responses, and μ and σ are their means and variances.

The overall comparison of descriptors is shown in Tab. 1. From the results we observe low performance of SSIM and VDP descriptors. This can be caused 1) by their translation non-invariance, so that they give high responses even to a slight, perceptually insignificant, planar shift of texture caused by the filter **A** (this is most apparent for sample *alu* Fig. 3, and 2) by their lower sensitivity to the very subtle degradations that were tested. The Tab. 1 shows also the approximate speed of computation of differences between two textures, and the sizes of feature vectors ℓ_{FV}. We observe that although the CAR 3D has a slightly shorter feature vector than its 2D variant, it does not achieve the same performance. The table also shows for the CAR model comparison of different feature vector distances. While for 2D CAR the best performance was achieved for $L_{0.2}$ norm, for 3D CAR the best results were for L_1 norm. A high correlation with the psychophysical results was achieved by descriptors based on CAR model and LBP features. Fig. 6 shows performance of the best combination of parameters for each type of descriptor, i.e. (a) SSIM, (b) VDP (p>75%), (c) 2D CAR (GP 1), (d) 2D CAR (GLP 2), (e) Gabor (GF RGB), and (f) LBP (RGB). Generally, the best results were obtained for 2D CAR model without any pyramid (GL 1), where

Table 1. Correlation of the tested degradation descriptors with data obtained from the psychophysical experiment. The best variant of each descriptor type is highlighted.

SSIM (speed: $\sim$ 2 s)	R (11×11) **0.125**	**VDP** (speed: $\sim$ 10 s)	R(p>75%) **0.107**	R(p>95%) 0.097

CAR (speed: $\sim$ 4 s)

pyramid type	levels	model dimens.	ℓ_{FV}	FC$_3$	R L$_{0.2}$	L$_1$
GP	1	2D	195	0.777	**0.787**	0.677
		3D	177	0.550	0.542	0.581
	2	2D	390	0.710	0.752	0.644
		3D	354	0.517	0.552	0.573
GLP	2	2D	390	0.654	**0.714**	0.638
		3D	354	0.360	0.362	0.573
	3	2D	585	0.648	0.677	0.620
		3D	531	0.422	0.439	0.475

LBP (speed: $\sim$ 1 s)

data	ℓ_{FV}	R
grey	512	0.610
RGB	1536	**0.712**

Gabor (speed: $\sim$ 8 s)

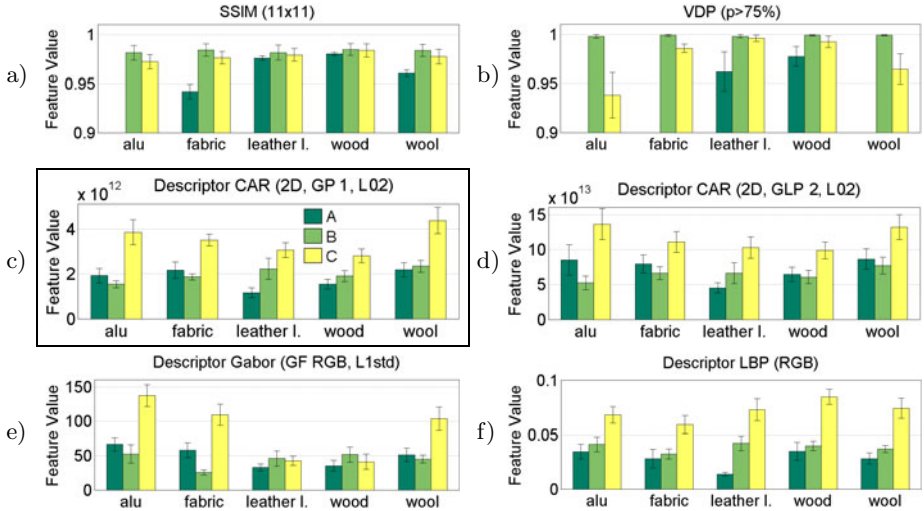| data	method	ℓ_{FV}	R
grey	GF	48	0.569
RGB	GF	144	**0.578**
	OGF	252	0.322

Fig. 6. Best performance of the tested descriptors (a) SSIM, (b) VDP (p>75%), (c) 2D CAR (GP 1), (d) 2D CAR (GLP 2), (e) Gabor (GF RGB), (f) LBP (RGB)

the difference of the feature vectors was evaluated using $L_{0.2}$ norm. Additionally, the CAR model enable to adjust pyramid type and size with regards to the type and intensity of degradation. Although the LBP features (f) are fast and have also quite high correlation with the human judgments, their responses clearly do not follow the trend of values across the samples present in Fig. 5-a. We tested also other variants of LBP features such as $LBP_{24,3+8,1}^{riu2}$ and $LBP_{16,2}^{u2}$, however their descriptive abilities were clearly worse than of those shown in Tab. 1.

6   Conclusions

Our results show that the statistical texture descriptors based on the causal auto-regressive model have the best performance in detection of subtle texture differences with respect to human judgments obtained in a psychophysical study. We conclude that these descriptors are the best, out of the tested features, for the automatic prediction of subtle perceptual differences in rendered view- and illumination-dependent surface textures in accordance with human perception. This highly demanded property can be used as automatic feedback for optimization the visual performance of texture compression and rendering methods.

References

1. Dana, K., van Ginneken, B., Nayar, S., Koenderink, J.: Reflectance and texture of real-world surfaces. ACM Transactions on Graphics 18(1), 1–34 (1999)
2. Filip, J., Haindl, M.: Bidirectional texture function modeling: A state of the art survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(11), 1921–1940 (2009)
3. Filip, J., Chantler, M., Green, P., Haindl, M.: A psychophysically validated metric for bidirectional texture data reduction. ACM Transactions on Graphics 27(5), 138 (2008)
4. Guthe, M., Müller, G., Schneider, M., Klein, R.: BTF-CIELab: A perceptual difference measure for quality assessment and compression of BTFs. Comput. Graph. Forum 28(1), 101–113 (2009)
5. Filip, J., Chantler, M., Haindl, M.: On uniform resampling and gaze analysis of bidirectional texture functions. ACM Transactions on Applied Perception 6(3), 15 (2009)
6. Database BTF, Bonn (2003), `http://btf.cs.uni-bonn.de`
7. Mantiuk, R., Myszkowski, K., Seidel, H.P.: Visible difference predictor for high dynamic range images. In: IEEE International Conference on Systems, Man and Cybernetics, October 2004, vol. 3, pp. 2763–2769. IEEE, Los Alamitos (2004)
8. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)
9. Haindl, M., Šimberová, S.: A Multispectral Image Line Reconstruction Method. In: Theory & Applications of Image Analysis, pp. 306–315. World Scientific Publishing Co., Singapore (1992)
10. Vacha, P., Haindl, M.: Image retrieval measures based on illumination invariant textural MRF features. In: CIVR, pp. 448–454. ACM, New York (2007)
11. Santini, S., Jain, R.: Similarity measures. IEEE Trans. Pattern Anal. Mach. Intell. 21(9), 871–883 (1999)
12. Ma, W.Y., Manjunath, B.S.: Texture features and learning similarity, pp. 425–430. IEEE, Los Alamitos (1996)

13. Bovik, A.: Analysis of multichannel narrow-band filters for image texture segmentation. IEEE Trans. on Signal Processing 39(9), 2025–2043 (1991)
14. Jain, A., Healey, G.: A multiscale representation including opponent colour features for texture recognition. IEEE Transactions on Image Processing 7(1), 125–128 (1998)
15. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 971–987 (2002)
16. Tatler, B.W.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision 7(14), 1–17 (2007)