Quality Assessment of Fractalized NPR Textures: a Perceptual Objective Metric
Pierre Bénard, Joëlle Thollot, François X. Sillion

To cite this version:

HAL Id: inria-00405966
https://inria.hal.science/inria-00405966v3
Submitted on 13 Jul 2012
Quality Assessment of Fractalized NPR Textures: a Perceptual Objective Metric

Pierre Bénard* Joëlle Thollot François Sillion
Grenoble Universities and CNRS – LJK† INRIA

Abstract

Texture fractalization is used in many existing approaches to ensure the temporal coherence of a stylized animation. This paper presents the results of a psychophysical user-study evaluating the relative distortion induced by a fractalization process of typical medium textures. We perform a ranking experiment, assess the agreement among the participants and study the criteria they used. Finally we show that the average co-occurrence error is an efficient quality predictor in this context.

CR Categories: I.4.7 [Image processing and Computer Vision]: Feature Management—Texture

Keywords: texture, perceptual evaluation, non-photorealistic rendering

1 Introduction

Many non-photorealistic rendering approaches aim at depicting 3D scenes with styles that are traditionally produced by 2D media (pigments, strokes,...). The main problem of these methods is the lack of temporal coherence while stylizing dynamic scenes. It results from the conflicting goals of depicting a 3D motion while preserving the 2D characteristics of the medium: in particular, screen-space size and stroke density. Several approaches represent this medium as a texture and partially solve the problem of temporal coherence via a fractalization process that combines many versions of the original texture at different scales through alpha-blending (e.g., [Klein et al. 2000; Cunzi et al. 2003; Breslav et al. 2007; Bénard et al. 2009]).

However, these solutions modify the patterns in the texture: new features and new frequencies may appear, global contrast may be degraded and deformations may occur. As a result the fractalized texture is likely to be visually dissimilar to the original texture targeted by the artist. We believe that the automatic evaluation of this similarity loss can be a valuable tool for comparing existing alpha-blending approaches and may allow the development of new fractalization techniques. This problem is particularly challenging in NPR because assessing the modification of appearance involves multiple factors and perceptual effects.

In this context, we define the texture distortion as the dissimilarity between the original and transformed 2D texture. Our goal is to define a quantitative metric for this distortion. For that, we perform a study in which users are asked to rank pairs of original/distorted textures from the least distorted pair to the most distorted pair. We provide a statistical analysis of the results to derive perceived quality scales for ten classes of NPR media. In this study, we deliberately avoid giving an explanation of the nature of the transformation. We ask the participants themselves to identify the criteria they have used to assess the distortion. In a second step, we study the correlation of these subjective results with several objective metrics that are coming from image quality assessment and texture analysis, and based on global and local image statistics or spectrum analysis. We suggest that the average co-occurrence error (ACE) is a good predictor of the distortion.

This work is a first step toward the quantitative evaluation of the perceived effects of fractalization as we only focus on a bi-dimensional (without perspective projection) and static (a single image) study. This restricted scenario is still of interest though, as it corresponds to slow motions or still frames of an animation.

2 Previous Work

Perceptual experiments take a growing part in computer graphics. Three main indirect methodologies exist for deriving psychophysical scales: rating, paired comparison and ranking. In the following we present an overview of these main approaches.

Rating consists in estimating statistical metrics (root mean square, Pearson correlation and outlier ratio) between the objective model output and the results from viewer subjective rating of the test images. This is the usual setup in image and video quality assessment [Sheikh et al. 2006; Winkler 2005]. The validity and reliability of rating data can only be asserted considering very large number of trials and participants have to be trained before the trial [Kendall 1975].

Paired Comparison In this setup, the whole dataset is presented pair by pair to the user who has to make straightforward forced choices. This approach was used to compare the performances of tone mapping operators [Ledda et al. 2005; Čadík et al. 2008]. Its main limitation is its quadratic complexity. In our case it would mean asking for ninety comparisons, making the experiment tiresome for users.

Ranking uses images organized according to a certain varying degree of quality [Guilford 1954]. Hence, this setup is the least time consuming one while it still provides for relevant data and is more appreciated by participants in general. Stokes et al. [2004] run a series of ranking experiments to define the perceptual importance of different illumination components for efficient high quality global illumination rendering.

Considering its advantages, we use the ranking methodology for our user study. Our work deals with texture transformation for NPR. In this field, perceptual evaluations have been proposed for specific styles, using different methodologies [Santella and DeCarlo 2004]. Among them, Isenberg et al. [2006] conduct an observational study to evaluate how people assess both hand-made stippling illustrations and computer-generated ones. Then, they perform a statistical analysis [Maciejewski et al. 2008] using gray-level co-occurrence matrices (GLCM) to compare the correlation between the different techniques. We take inspiration of this approach to derive an objective distortion metric.

3 Experimental Framework

Our goal is to define some estimate of the quality of NPR renderings based on texture fractalization. For that, we design a ranking experiment and derive an interval scale of perceived distortion to which we correlate an objective metric.
3.1 Stimuli

We choose twenty gray-scale 2D textures sufficiently representative of the main traditional media used in NPR\textsuperscript{1}. To create a sufficient redundancy in the results, we design two sets of ten texture pairs (\(S_1\) and \(S_2\)). For each set, we choose one representative texture per class (pigments on canvas, paint, paper, hatching, cross-hatching, dots, near-regular or irregular patterns, noise and grid).

To construct \(S_1\) and \(S_2\), each of the original textures is transformed using a 2D static version of the fractalization algorithm proposed by Bénard et al.\textsuperscript{[2009].} Here, three scales of the texture are alpha-blended (with the coefficients \(1/3, 1/2\) and \(1/6\)). This number is a minimum to ensure temporal continuity in most dynamic scenes.

Figure 1 shows the twenty pairs – original and transformed version – obtained for the two series.

3.2 Procedure

We realize the ranking experiment via a dynamic web interface to enrich and diversify our panel of subjects. Nevertheless, in order to keep a certain amount of control on the participants’ origin, we have restricted the diffusion of this web site to known people. We are aware that our interface prevents us from having a precise control on the experimental conditions (screen resolution, calibration, viewing angle and ambient lightning). Consequently, we paid special attention to assessing the statistical validity of the resulting data as detailed in Section 4.1. We consider this trade-off admissible with regard to the number of participants and the diversity of their skills in computer graphics (naïve: 58\%, amateur or professional infographists: 8.5\%, researchers: 22.4\% and unknown: 11.1\%).

We realize the ranking experiment via a dynamic web interface to enrich and diversify our panel of subjects. Nevertheless, in order to keep a certain amount of control on the participants’ origin, we have restricted the diffusion of this web site to known people. We are aware that our interface prevents us from having a precise control on the experimental conditions (screen resolution, calibration, viewing angle and ambient lightning). Consequently, we paid special attention to assessing the statistical validity of the resulting data as detailed in Section 4.1. We consider this trade-off admissible with regard to the number of participants and the diversity of their skills in computer graphics (naïve: 58\%, amateur or professional infographists: 8.5\%, researchers: 22.4\% and unknown: 11.1\%).

On each trial, after presenting the general context of this study, the subject is asked to place the pairs in order from the lowest to highest by perceived distortion. This ordering can be partial if texture pairs seem identically distorted. After the ranking of the second set, the subject is asked to specify for each of these last ten pairs the criteria she thought to have used. We provide three basic criteria: contrast, sharpness and scale; but we also allow the user to freely type personal criteria. The two image sets are shown randomly at each user trial to avoid learning effects. The duration of each task (first ranking, second ranking and criteria specification) are recorded.

4 Statistical Analysis

103 people took part in this study. Among them, 45 started with the first image set (\(S_1^1\) then \(S_2^2\)) and 58 with the second (\(S_1^2\) then \(S_2^1\)). The statistical analysis of the resulting two datasets are carried out identically in three steps. After studying the potential learning effect, we propose to derive an interval scale of relative perceived distortion intensity. Finally we run an analysis of the criteria used during the ranking.

4.1 Ranking Duration and Concordance

First, we estimate whether the presentation order of the two image sets has an influence on the final ranking to verify that the subjects are not biased by the first ranking while doing the second. The inset shows the boxplot of the experiments duration for the two series depending on

\(S_1\) and \(S_2\) showdowns. These intervals scales are derived from the z-Scores shown in (b). Texture pairs surrounded by the same dashed line may be considered as perceptually equivalently distorted.
on presentation order. Notice that for each series the distribution of
durations is similar whatever the presentation order. Thus, we can
conclude that the presentation order has no significant influence on
the subjects’ attention. Moreover the mean duration of both series is comparable (211s and 147s respectively).

In order to quantify the consistency of rankings provided by the
users panel, we compute the Kendall’s coefficient of concordance
(Kendall’s W) [Kendall 1975] over the four results sets $S_{i1}$ and the
$S_i$ obtained by merging the previous series. This non-parametric
measurement is commonly used to evaluate the agreement among raters (from 0 for no-agreement to 1 for full-agreement) without the
need to assume any specific distribution.

As shown in the inset table, the coefficients of concordance of the
merged results sets $S_i$ remain comparable with the un-merged ones. Regarding these statistics the analysis of
merged data seems relevant.

To validate the statistical significance of Kendall’s W we perform a $\chi^2$ test with the null hypothesis $H_0$ that there is no agreement
among the subjects. In the six cases, this hypothesis may be re-
jected at $\alpha = 0.001$ level for 9 degrees of freedom. Hence, we can
conclude that there is some agreement amongst the participants,
that is, the rankings are not effectively random.

4.2 Interval Scale of Relative Perceived Distortion

An ordinal scale can be directly derived by tabulating the raw data
to show how often each pair was placed in each rank position (fre-
quency) and calculating their mean ranks. However, it gives no
quantification of perceived differences between texture pairs: it
doesn’t inform on how much higher one pair is distorted compared
to another. By assuming a normal distribution of the data and us-
ing the Thurstone’s law of comparative judgment [Torgerson 1958],
we can convert the proportion of each pair (frequencies divided by
number of trials) into z-Score (Figure 1b). These z-Scores corre-
spond to relative differences in perceived distortion between texture
pairs on a perceptually-linear scale. For both image sets (Figure 1a),
note that the unstructured textures (noise, pigment, paper) seem to
be more robust to the fructization process. On the contrary, tex-
tures exhibiting more distinctive features are in both cases judged
as the most severely distorted.

In order to go a step further, we perform statistical hypothesis tests
to ensure that the perceived similar distortion intensity is signif-
cant. We use the non-parametric Wilcoxon rank sum test (also
called MannWhitney U test) for assessing whether two independ-
ent samples of observations come from the same distribution. The
$U$ test has both the advantages of working on ordinal data and of
being more robust towards outliers than the Student t-test. The
null hypothesis $H_0$ is that the two considered samples are drawn
from a single population, and therefore that their distribution is
equal. In our case, this test has to be run for each pair of samples:
$(\text{tex}_i,m, \text{tex}_i,n)$ where $1 \leq m \leq n \leq 10$ (i.e. ninety times).

In Figure 1, we frame groups of pairs for which the null hypothesis
cannot be rejected. By looking at the corresponding texture pairs,
we observe that the overall contrast of patterns seems to be the most
significant criterion. In comparison, feature shapes seem to play a
less important role in this grouping.

Note that the interval scales and groups of each image set are not
fully identical. It indicates that the classes of media we defined a
priori only partially fit the classification based on the distortion.
In particular, the two representative textures of the “grid” and “cross-
hatching” classes are judged distorted differently. We think that this
is due to their strong difference in terms of contrast, pattern density or feature shapes.

4.3 Ranking Criteria

We finally analyze the criteria that each subject considered she used
during the ranking. Figure 2c shows the relative frequency at which
the three proposed criteria have been used per series. Overall, they
are quite similar, with a little preference for sharpness and then con-
trast in $S_2$. However when we consider this distribution of criteria
for each texture pair (Figures 2a and 2b), we observe irregular pref-
ences for different criteria. We thus conjecture that the content
of each texture class triggers different criteria when assessing the
similarity between original and transformed textures.

The analysis of the additional criteria proposed by the participants
also shows that the notions of density (15% of these criteria), shape
(10%) and pattern coherence (21%) and to a lesser extent frequency
(4%) and relief (2%) are relevant as well.

Figure 2: Relative frequency at which the criteria have been used:
For each texture pair (a,b) and for each series (c).

5 Correlation with Objective Metrics

In a second step, we review a large range of image metrics and
statistics, with the hope of correlating them with the two previously
derived subjective scales. We use the Pearson product-moment cor-
relation coefficient $r$ and linear regression to evaluate this correla-
tion.

5.1 Image Quality Metrics

We first examine eleven well-known objective quality assessment
metrics (implemented by Matthew Gaubatz in the MeTrix MuX
Matlab\textsuperscript{\textregistered} package\footnote{Available at: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/}): the peak signal-to-noise ratio (SNR), the
signal-to-noise ratio, the structural similarity (SSIM) index and
multi-scale SSIM index, the visual signal-to-noise ratio, the vi-
sual information fidelity (VIF) and pixel-based VIF, the informa-
tion fidelity criterion, the universal quality index, the noise quality
measure and the weighted SNR.

\begin{table}[ht]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
\textbf{series} & \textbf{contrast} & \textbf{sharpness} & \textbf{scale} & \textbf{other} & \textbf{empty} \\
\hline
$S_1$ & 21.96\% & 26.86\% & 24.83\% & 17.40\% & 8.95\% \\
$S_2$ & 28.27\% & 35.21\% & 21.72\% & 7.12\% & 7.07\% \\
\hline
\end{tabular}
\end{table}

\textsuperscript{2}Empirically verified using a Q-Q plot and a Shapiro-Wilk test: the null hypothesis of normally distributed data cannot be rejected.
None of these metrics shows a significant correlation with the subjective interval scales. This conclusion was predictable as these metrics have been designed to assess the quality of images suffering limited amount of distortion (noise, blur, compressions artifacts, etc). In comparison, the fractalization process may strongly modify the appearance of the distorted texture.

5.2 Global and Local Image Statistics

Because the three criteria – contrast, sharpness and scale – have to be considered simultaneously, we cannot expect global image statistics to give significant results, especially when one considers that our textures are not “natural images”. Our experiments on histograms, power spectra and distribution of contrast measurements [Balboa and Grzywacz 1993] are, as expected, inconclusive.

Taking inspiration from the field of texture analysis [Tuceryan and Jain 1993], we choose a statistical method – the gray level co-occurrence (GLC) model – which estimates local image properties related to second-order statistics (like variance, standard deviation and correlation). Moreover, psychophysical experiments have shown that the GLC model matches certain levels of human perception [Julesz et al. 1976] and it has been successfully used and perceptually validated for texture synthesis [Copeland et al. 2001]. Finally, this model might be related to the density and pattern coherence criteria freely proposed by the subjects.

![Figure 3: Linear regression of ACE against z-Scores for S1 (left) and S2 (right).](image)

We measure the distortion between original and distorted textures by computing their GLC matrices and then the distance between these two sets of matrices with the average co-occurrence error (ACE) of Copeland et al. This error metric is highly correlated with the perceptual interval scales for both series. We obtain the maximum absolute Pearson’s correlations of 0.953 for S1 and 0.836 for S2 (with, respectively, a p-value < 0.0001 and 0.003 for 8 degrees of freedom) considering the GLC matrices for all displacements up to $T_{NX} = T_{NY} = 4$ pixels with a $G = 32$ gray levels quantization (with the notation of Copeland et al.).

Figure 3 shows the corresponding linear regression of the ACE against z-Scores for both series ($r^2 = 0.9075$ and 0.6992 respectively, with the same p-values as for Pearson’s correlations). This high correlation for both image sets confirms that the ranking differences observed in Section 4.2 are coherent and that an a priori classification is not a suitable predictor of the distortion.

6 Conclusions

In this work, we proposed the average co-occurrence error as a meaningful quality assessment metric for fractalized NPR textures. We validated the relevance of this predictor by showing its strong correlation with the results of a user-based ranking experiment.

We plan to investigate potentially better suited texture and vision descriptors to derive an improved objective quality metric. Image retrieval approaches, based on extracted texture features, seem a promising field of inspiration. Longer term future work will consider the dynamic version of the fractalization process. In this case, the methodology we developed for the current study will have also to consider the trade-off between temporal continuity and texture dissimilarity to the original medium.

Acknowledgments

We would like to thank all the participants of this user study. We especially thank Jean-Dominique Gascon, Olivier Martin, Fabrice Neyret, Pierre-Edouard Landes, Pascal Barla, Alexandra Orzan and the anonymous reviewers for their valuable comments and suggestions.

References


