Beyond Image Quality Comparison

Alexander Bornik[1] and Peter Cech[2] and Andrej Ferko[1,3] and Roland Perko[1]

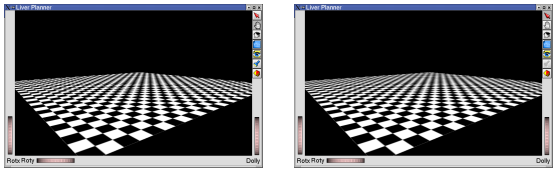
[1] TU Graz, Graphics and Vision, Inffeldgasse 16, A - 8010 Graz, Austria
{bornik, ferko, perko}@icg.tu-graz.ac.at
http://www.icg.tu-graz.ac.at
[2] ETH Zentrum, IFW A44, Haldeneggsteig 4, CH - 8092 Zurich, Switzerland
[3] Comenius University, SK - 842 48 Bratislava, Slovakia

Abstract

We solve the problem of quantitative measuring of image quality, beyond the widely used MSE (mean square error) or even visual comparison. We combine eight feature-based and perceptually-oriented image quality metrics. The need for this comes from virtual archaeology, where various image acquisitions, antialiasing, multiresolution, image reconstruction or texturing methods produce very similar images. The proposed evaluation framework is suitable for any image quality decision-making, being not restricted to virtual archaeology.
In particular, we compare a representative database of visually indistiguishable image pairs from different cameras, anisotropic texture filters and various antialiasing methods. For image registration we propose a modified video processing step. The results support the selection beyond commonly used visual comparison.

1. Introduction

Two commonly used measures of visual quality are MSE or even subjective visual comparison. They cannot give an insight for a qualitative judgement or perceptual significance, especially if the images are very similar. Mathematically, it is possible to estimate the error of a given acquisition, texturing or antialiasing method, but resulting images are perceived by humans. That is why many papers conclude by visual comparison, only. Fortunately, there is a reasonable compromise between the two evaluation extremes - to focus on the image quality measures. If we have two images, which of them is a better "correct image" or "high quality image"?



(a) (b)

Figure 1: *(a) Checkerboard texture (Ripmap). (b) The same texture using Mipmap.*

The paper describes sources of image pairs in Section 2. Section 3 introduces the testing database. Section 4 surveys selected feature-based image quality metrics. Section 5 describes selected perceptual image quality measures. Section 6 is an overview of the implementations and results. Finally, section 7 concludes the paper.

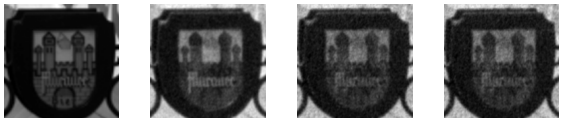
2. Sources of Image Pairs

Anisotropic texture filtering produces images of higher quality for most uses of texturing. P. Heckbert discovered a general theory in the late 80s [9]. Ideally, the pixel projection in texture space can be used to form a line of anisotrophy and the methods vary in sampling strategies along this line. If the line is axially aligned, mipmapping or ripmapping works well. The recently introduced fast Fipmap texture minification [3] offers the solution for any slope of the line of anisotrophy. To the best of our knowledge, there is no widely adopted testing methodology both in the choice of the set of test images and in the methods to evaluate the competence of mipmap, ripmap or another filter. Most papers rely on a subjective visual comparison based on checkerboard images only. The same is observable for antialiasing (see the images below). We see the need to clarify the decision in several areas related to virtual archaeology. Answering this question

seems worthy of obtaining an informed feed-back in any type of image acquisition and synthesis, including photo-realistic rendering, non-photorealistic rendering (NPR) and image-based rendering (IBR). In addition to that, image reconstruction and multiresolution techniques are areas where more precision is needed. All of the areas contribute to a virtual archaeology workflow [5].

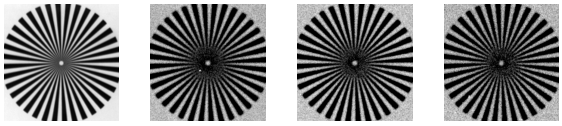
3. Database of Visually Identical Image Pairs

We collected a set of image pairs from three areas: different sensing methods [12], anisotropic texturing [3], and antialiasing [16]. The images of figure 2 respectively 3 were geometrically aligned using an algorithm based on motion estimation [10]. In the future, we plan to extend the database and any contribution is welcome.



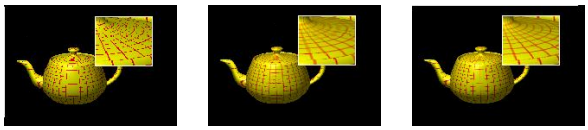
(a) (b) (c) (d)

Figure 2: *Detail of an advertising sign captured with different devices, (a) Philips CCD with* 12*µm pixel size* [13] *(CCD) (b)-(d) analog films scanned with* 5*µm (b) Agfa Scala 200* [1] *(sca) (c) Agfa APX 100* [1] *(apx) (d) Ilford FP4 Plus 125* [11] *(fp4).*



(a) (b) (c) (d)

Figure 3: *Siemens star taken with different sensors for line detection.*



(a) (b) (c)

Figure 4: © *Rosalee Wolfe. Used with permission. (a) No antialiasing. (b) Prefiltering. (c) Supersampling. Images from* ACM SIGGRAPH *page.*

4. Feature-based Image Quality Measures

We compare the quality of images taken by different sensors and investigate three kinds of analog films and one digital camera set. Figure 2 shows one of the test targets. At first sight it is obvious that the image taken by CCD sensor is clearer and contains less noise.

4.1. Noise

Noise is an important criterion for characterizing the image quality and is measured by calculating the entropy in homogenous patches. To get more reliable results, these patches are described by the so-called co-occurrence matrix introduced by Haralick [6]. Common features computed of the co-occurrence matrix are [7]: entropy, energy, maximum probability, contrast, inverse different moment, correlation and homogeneity. Since energy is a homogeneity measure (the larger the value, the more homogeneous the image), and contrast measures the local image variation, these values are considered to be optimal to measure noise.

4.2. Line detection

In this test, a line detector is applied on a Siemens star (see Figure 3) which contains 72 lines. As quality measures, the number of edge elements also called edgels found per line and the average normal distances of all edgels with respect to one line were taken. The results are shown in Table 1. The main conclusion of this test is, that the same results are collected from film-based and CCD-based images, but the results of CCD images are more stable, because more edgels are found for each line.

4.3. Test Patterns

We have captured a wall with several test patterns. Figure 5 shows images of a regular grid and the horizontal profiles. Here, the difference between film-based and CCD-based images is tremendous. Because of the large amount of noise in the film-based images, the grid can hardly be detected.

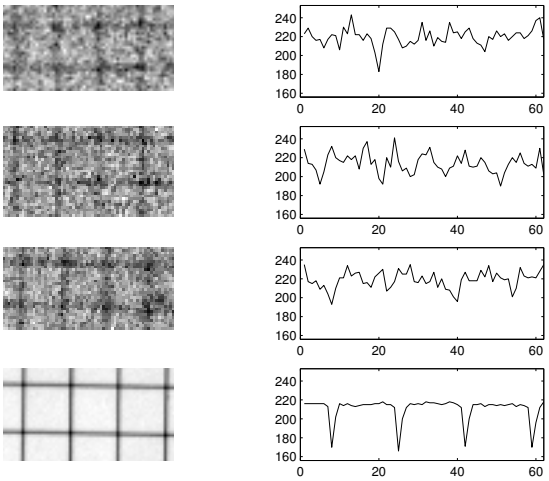


Figure 5: *Regular black grid on white background. left: captured image, right: horizontal profile. From top to down: apx* 15*µm, fp4* 15*µm, sca* 15*µm and CCD.*

	min	max	mean	std
CCD	48	75	62.46	5.943
sca	45	64	52.26	4.976
apx	42	66	52.81	5.545
fp4	36	62	51.99	5.718

(a)

	min	max	mean	std
CCD	0.00037	0.0834	0.0300	0.0293
sca	0.00105	0.1163	0.0381	0.0352
apx	0.00134	0.0806	0.0317	0.0284
fp4	0.00415	0.1187	0.0385	0.0352

(b)

Table 1: *Statistics of (a) number of edgels with respect to one line and (b) average normal distances of edgels according to one line. Minimum, maximum, mean and standard deviation is shown. CCD gives the best results.*

4.4. Stereo matching

A points of interest matcher, using Harris corner detector [8], normalized cross correlation and a least squares method for refining the results to subpixel accuracy, is tested on stereo images pairs. We choose a street-scene for this test, one of the stereo images is shown in Figure 6. Two images taken by the same films/CCD from nearly the same spot (30cm baseline) were matched. 4000 points of interest were searched on a regular grid and the percentage of successfully matched points of interest are 81%, 75%, 70% and 53% for CCD, sca, apx and fp4.
Another quality measure are the normal distances from matched points to their corresponding epipolar lines, whereas the fundamental matrix is calculated via RANSAC algorithm and refinement of results was done using the eight point algorithm [18].

Figure 6: *Street scene taken with CCD sensor.*

5. Perceptual Image Quality Measures

Human perception is influenced by many factors. Most image quality models incorporate only a few of them such as contrast sensitivity or the luminance adaptation. The contrast sensitivity function (and its inverse contrast threshold function) represents the minimum noticeable amount of a change in contrast of the frequency component. The contrast sensitivity depends on the background luminance. The contrast is defined as a ratio between the luminance of stimulus and background luminance. Hence, luminance changes are less noticeable in areas with high luminance. There is also another decrease below 10 cd/m^2. The visibility of a signal can be reduced by the presence of another signal. This phenomenon is called a contrast- or pattern masking. The masking effect is strongest in near spatial, frequency and orientation areas and depends on the type of masking and the masked signal.

• **Summation of Errors** Raw errors might represent a large amount of data. To allow for a good and fast quality overview, the errors have to be summed into a quality map or just one quality number. Summation is done using the well-known Minkowski metrics which include RMSE (relative MSE) measure, a probability summation or a maximum operator.

A summation over frequencies is a preferred first step. In summing over space, the maximum operator is used. For measuring average rather than maximum distortions, the best predictions are given in [2], and [4].

We have chosen three metrics for image quality comparison: MSE, VQM (Visual Quality Metrics) by Xiao [17] and the metrics used in the DCTune algorithm by Watson [15], [14]. All three produce a single number quality measure per image. MSE mean square error can be computed in the spatial domain from the errors in single pixels. There are several metrics defined as a function of MSE, such as RMSE (relative MSE) or SNR (signal to noise ratio).

• **VQM** VQM metrics uses masking properties of the contrast sensitivity function and the luminance adaptation to model human vision. It operates in the DCT (discrete cosine transform) coefficients domain. Luminance masking is incorporated by computation of the local contrast for each DCT block. The inverse of the MPEG quantization matrix was chosen as an approximation. Using this multiplication, the errors are converted and summed using weighted

blending of maximum- and mean errors.

• **DCTune** The DCTune is another algorithm, which works in the DCT coefficients domain. DCTune itself is not a perception quality metrics, but it uses quality metrics internally. DCTune uses spatial frequency threshold approximations introduced by *Ahumada and Peterson* [2].

• **Universal Image Quality Index, UIQI** Unlike the previous three metrics, UIQI index is mathematically defined and does not explicitly utilize human visual system properties like luminance adaptation or contrast masking. Thus (similarly to MSE) it is independent of the observer and the viewing conditions. The UIQI index works in spatial domain, computing correlation, luminance similarity and contrast similarity between the original image and the distorted image. The quality coefficient for the distorted image with respect to the original image is a value ranging from 1 (identical images) to -1 (maximal distortion).

6. Implementation and Results

The new implementation work consists of two independent parts: rendering application and image comparison application. Besides that there are the existing tools for texture reconstruction from multiple views in urban areas.

The rendering application enables us to render polygons showing well known test textures e.g. the checkerboard texture using different texture filtering methods from exactly the same viewpoint. This involves both, hardware- and software rendering. Hardware rendering is used whenever this is possible, namely for Mipmapping and the anisotropic Ripmap approach provided through OpenGL extensions. Other filtering techniques such as EWA (Elliptic Weighted Average), Summed Area Tables or Fipmap technique could only be implemented using a software renderer. Recently, there are some efforts to use hardware support, but we cannot test the final images.

The image comparison tools we implemented can be used to compare renditions produced be the rendering application, using different image quality measures ranging from the simple RMSE (root mean square error) to more sophisticated methods like VQM or DCTune. We will be able to compare renditions from a real-world checkerboard surfaces acquired using a calibrated digital camera to renditions of a geometrically equivalent artificial scene in the near future. The feature-based image quality metrics were implemented in Matlab.

The implementation of the rendering tool mentioned above is still in progress, so we can only show some preliminary results for hardware-rendered polygons now. Figure 1 shows test images for quality measurements using our images quality measurement framework.

Table 2 lists the results for the images of Figure 1 using RMSE, VQM, DCTune and Universal Image Quality Index in YC_rC_b 4:2:2 space methods. Ripmap image is a reference one, therefore the the values are zero and one.

We have experimented with another set of images, as well. We took them from the *Teaching Texture Mapping Visually* course by R. Wolfe. The original image is improved using mipmapping, supersampling, and by a combination of mipmapping and supersampling. The measurements support the intuitive ordering of images according to the increased perceptual quality as shown in Figure 4.

Discussing the results we have observed primarily, that the eight measures give a spectrum of incommensurable coefficients. Moreover, the perceptually oriented metrics confirm a clear superiority of CCD acquired images. No correlation of perceptual values can be observed with respect to entropy and energy measures. They operate in an independent dimension. We are far from composing the whole spectrum into one number by summing the weighted values. This could reduce the complexity of evaluation. However, it remains to study the competences of mapping of image pairs into the eight-dimensional space of parameters.

7. Conclusion and Future Work

Studying recent anisotropic texture filters and image reconstructions, we noticed that the highest precision improvements might be imperceptible. The detailed study of the image quality metrics (both feature based error metrics [12] and perceptual ones) led us to create an image database using very similar and/or well known images. Our image quality evaluation, combining different approaches, shows both the significant correspondence of results and strong independence of certain quality measures.

Our methodology and testing set of images can be used for measurements of any image pair, even of unknown origin. Our future work is to evaluate real data for selecting the most suitable methods in virtual archaeology workflow, especially for texturing and image reconstruction. However, the new framework contributes in a wide spectrum of applications.

8. Acknowledgements

This work has in part been funded by the European Union under contract No. IST-1999-20273. We wish to thank J. R. Wolfe for antialiasing images courtesy. Furthermore we wish to thank Horst Bischof, Markus Grabner and Franz Leberl for fruitful discussions.

References

1. Agfa-Gevaert AG. Technical Data - Agfa Range of Films, professional, September 1998. 2

	RMSE	VQM	DCTune	UIQI	Entropy	co-occurrence Entropy	co-occurrence Energy	co-occurrence Contrast
Ripmap	0.0	0.0	0.0	1.0	2.21	3.83	0.29	3713
Mipmap	2.439	1.476	2.177	0.995	2.53	4.40	0.26	3195

Table 2: *Quality measures for Mipmap-rendered polygon compared to anisotropic Ripmapping in Figure 1.*

	RMSE	VQM	DCTune	UIQI	Entropy	co-occurrence Entropy	co-occurrence Energy	co-occurrence Contrast
CCD	0.0	0.0	0.0	1.0	4.4455	7.0710	0.012297	93.94
sca	35.236	6.2365	4.4278	0.4114	4.9620	8.2382	0.000768	116.19
apx	29.391	5.7718	3.1568	0.4133	4.9017	8.3326	0.000565	115.57
fp4	29.396	5.8099	3.1829	0.4117	4.9017	8.3482	0.000552	117.26

Table 3: *Quality measures for test targets in Figure 2.*

	RMSE	VQM	DCTune	UIQI	Entropy	co-occurrence Entropy	co-occurrence Energy	co-occurrence Contrast
CCD	0.0	0.0	0.0	1.0	4.8612	7.9895	0.001819	424.18
sca	49.806	9.0277	2.9879	0.5118	4.4064	7.5960	0.059547	333.08
apx	40.968	8.3467	3.2370	0.5625	4.3913	7.6003	0.063221	418.84
fp4	51.930	0.0829	3.3187	0.5113	4.3775	7.6068	0.064357	393.40

Table 4: *Quality measures for Siemens stars in Figure 3.*

	RMSE	VQM	DCTune	UIQI	Entropy	co-occurrence Entropy	co-occurrence Energy	co-occurrence Contrast
Prefiltering	0.0	0.0	0.0	1.0	2.3245	3.3475	0.4456	132.74
Supersampling	2.0568	1.3240	2.2245	0.95241	2.3257	3.3396	0.4456	136.49
No antialiasing	4.0712	4.8191	7.2473	0.82634	2.3326	3.2695	0.4340	263.31

Table 5: *Quality measures for teapot images in Figure 4.*

2. A. Ahumada and H. Peterson. Luminance-model-based DCT quantization for color image compression. *Human Vision, Visual Processing, and Digital Display*, 3:365–374, 1992. 3, 4
3. A. Bornik and A. Ferko. Texture minification using quad-trees and fipmaps. In *Eurographics 2002 - Shorts Presentations*, September 2002. 1, 2
4. M. Eckert and A. Bradley. Perceptual quality metrics applied to still image compression. *Signal Processing*, 70:177–200, 1998. 3
5. J. Cosmas et al. 3D MURALE: A multimedia system for archaeology. In *Proceedings of the International Symposium on Virtual Reality, Archaeology and Cultural Heritage 2001*, November 2001. 2
6. Robert M. Haralick. Statistical and structural approaches to texture. *Proceedings IEEE*, 67(5):786–803, 1979. 2
7. Robert M. Haralick and Linda G. Shapiro. *Computer and Robot Vision*, volume 1. Addison-Wesley Publishing Company, 1992. 2
8. C. Harris and M. Stephens. A combined corner and edge detector. *Proceedings 4th Alvey Visual Conference*, 1988. 3
9. P. S. Heckbert. Fundamentals of texture mapping and image warping. Master's thesis, University of California, 1989. 1
10. David J. Heeger. Notes on motion estimation. Psych 267/CS 348D/EE 365, Department of Psychology, Stanford University, Stanford, CA 94305, 1996. 2
11. Ilford. Fact Sheet FP4 Plus, March 2002. 2
12. Roland Perko and Michael Gruber. Comparison of quality and information content of digital and film-based images. *Photogrammetric Computer Vision - ISPRS Commission III Symposium*, XXXIV(3B):206–209, September 2002. 2, 4
13. Philips Semiconductors. Data Sheet FTF3020-M Full Frame CCD Image Sensor, November 1999. 2
14. A. Watson. DCT quantization matrices visually optimized for individual images. In *Human Vision, Visual-Processing, and Digital Display*, volume 4, pages 1913–1914, 1993. 3
15. A. Watson. DCTune: A technique for visual optimization of DCT quantization matrices for individual images. In *Society for Information Display: Digest of Technical Papers XXIV.*, pages 946–949, 1993. 3
16. R. J. Wolfe. Teaching texture mapping visually. SIGGRAPH Education Slide Set. http://www.siggraph.org/education/materials/HyperGraph/mapping/r_wolfe/r_wolfe_mapping_1.htm, 1997. 2
17. F. Xiao. DCT-based video quality evaluation. Technical report, Stanford University, 2000. 3
18. Zhengyou Zhang. Determining the epipolar geometry and its uncertainty: A review. Research report, INRIA, Sophia-Antipolis, France, July 1996. 3