Visual Significance of Digital Artifacts

K. MacLennan-Brown and R.E. Jacobson
Imaging Technology Research Group, University of Westminster
Middlesex, United Kingdom

Abstract

This paper describes a practical methodology designed to define and compare disparate digital image artefacts the aim of which is to derive a measure of their relative visual impact. The work will recast a methodology for consistently creating artefacts of known value, and from there to create a database of images each with defined Just Noticeable Difference (JND) levels that can be used as a basis for comparing visual impact. A proposed use of the database would be for comparing objective image quality measures against a predetermined set of subjective values for the perceptibility of the artefacts in question. This information can then be used in measures of the acceptability of levels of artefact, and as a basis for comparing and evaluating objective image quality metrics with regard to their subjective correlation. Included in this paper are the preliminary findings of an investigation into the relative visual impact of a cross section of digital artefacts. These artefacts are compared across four different images of real world scenes, both in colour and black & white. Studies of this kind allow research and computational effort to be prioritised to areas of greatest visual significance, thus minimising time and processing power.

Introduction

Artefacts inherent within the digital imaging system can be broadly subdivided into two groups, those directly caused by the discrete nature of the system and those attributable to methods used to limit the first group.

The uniquely digital characteristics of a system that can give rise to artefacts are due to temporal and spatial attributes and the number of recordable levels of each picture element or pixel. The temporal characteristics will not be pursued here, as this study is concerned with static images. It is also assumed that an appropriate system would be used in order to avoid problems such as smear and other motion related artefacts.

The number of recordable brightness levels available at a pixel is called quantization[1] and this directly relates to the likelihood of false contours being detectable within the image. False contours appear as visible transitions or steps in areas of smoothly graduating colour or brightness, giving an appearance similar to the contour lines on a map. The number of quantization levels available can be further attenuated by either the final output device or certain image compression procedures such as JPEG (Joint Photographic Experts Group).[2] It would not be possible by studying the final image to determine where in the imaging chain this artefact appeared.

The spatial attributes of a digital imaging system give rise to several artefacts and methods of reducing these give rise to several more. The artefacts directly attributable to spatial resolution are outlined below:

Pixelation - this is most noticeable on hard edges within the image that do not run orthogonal to the directions of the sensor. The cause of this artefact is that the pixels in the image are now large enough to be individually discernable, due to pixel replication.[3] This is usually first noticed as step-like discontinuities which lead to the other common name for this artefact - jaggies. A similar effect is sometimes seen after JPEG compression when the edges of image blocks become discernible. This artefact is then known as the Gibbs Phenomenon.

Aliasing is the incorrect reproduction of fine detail in the image due to the detail having a greater spatial frequency than the sensor.[4] The effect of this is that fine detail is not accurately captured by the sensor and is reproduced at a lower frequency, giving rise to spurious lines in the image sometimes referred to as the Moiré effect.

Blur, excluding motion blur, has two prime causes. One is low spatial resolution of the sensor; the other is deliberate use of a blur filter to defocus the image on the sensor slightly. This reduces the effect of alias and Moiré.[5] This blur, whether deliberately introduced or not, can be at least partially corrected by the use of various sharpening algorithms.[6] Not all of these are suitable for use in a pictorial application, and even those that are quickly give rise to an artefact known as ringing, fringing or overshoot.

The alternative names for this artefact are descriptive of its appearance in images i.e. a 'ring' or 'fringe' ghosting edges within the image. These false edges are often coloured.

Blur and ringing are good examples of software artefacts induced whilst trying to limit or correct hardware induced artefacts.

Many studies have been carried out on the artefacts inherent or induced within a digital imaging system. Most of these studies, however, were concerned with one artefact in isolation, such as blocking[7,8] or contouring.[9,10] Alternatively, two related artefacts such as aliasing and sharpness[11] have been considered. In this context, the term "related artefacts" refers to artefacts that directly affect one another, i.e. in the above example, as the strength of the anti-aliasing blur filter is increased, so the sharpness of the image decreases and vice versa. There have, however, been few if any attempts to scale disparate artefacts. Disparate artefacts are those that have no interdependency, i.e. the increase of one will not automatically induce the reduction of another.

Ranking disparate artefacts is a task fraught with problems and pitfalls. Many factors combine to influence the visual importance of any given artefact within a scene. Coupled with this contextual problem of assessing the visual importance or impact of an artefact, is also the more fundamental problem that a definitive set of digital artefacts has yet to be compiled.

As a preliminary study, this investigation will suggest a representative set of digital artefacts that, whilst not intended to be exhaustive, would define the master set of artefacts to which all others are sub-classes. Furthermore, it is intended to explore some of the problems that occur in an investigation of this type and suggest some practical measures to deal with them.

Experimental Method

The first requirement was to generate a set of images with artefacts to compare both against a reference image and with each other. In the absence of a truly effective image quality metric that was equally valid for all the artefacts under consideration, an alternative approach had to be devised.

It was decided to establish the first JND for each artefact, and then use these perceptually equal images in order to establish whether or not any particular artefact was more or less visually significant.

Four differing real natural scenes were chosen, in order to provide a cross section of the types of images normally encountered. Table 1 gives an overview of these scenes and their attributes, listed with the most significant first. The three colour scenes were chosen to contain areas susceptible to differing artefacts - the scene London with its fine detail was intended to be primarily susceptible to blur whilst the fine shading in the sky of the Landscape should show colour level problems and noise, and finally the Leopard image was intended to be particularly prone to over-sharpening or ringing. The black and white portrait was included to ascertain the veracity of any hierarchy in a monochrome environment. Apart from their susceptibility to certain artefacts, the scenes were also intended to mirror the 'natural' subject matter for an imaging system, as opposed to artificial test targets.

The artefacts induced within the images were created artificially using algorithms applied with the program Matlab.[12] The purpose of using Matlab was twofold: not only did this allow great control over the amount of artefact induced, but also gave artefacts that were generic rather than device-dependent.

The first JND for each artefact was determined according to the following protocol: by a process of trial and error, it was determined which variable for each algorithm created the correct level of control when varying the amount of artefact present in the image. Once this had been established, then a preliminary threshold value was set and a series of five images created around this initial value. This process was repeated for each of the artefacts under consideration. The threshold value was determined using a forced choice pair comparison program that compared the perturbed images with the reference image to which the algorithm had been applied.

Table 1: Image characteristics with greatest first.

Image	Type	Attributes
London	Architecture	Fine detail
		High contrast edges
		Fine shading
Landscape	Landscape	Fine shading
		Fine detail
		High contrast edges
Leopard	Wildlife / Portrait	High contrast edges
		Fine detail
		Flat tone areas
Portrait	Portrait / B&W	Flat tone areas
		Fine detail
		Fine shading

The images were presented on a computer monitor in a darkened room such that the CRT provided the sole illumination. The CRT was degaussed and allowed to warm up for 30 minutes prior to use. Immediately before the tests commenced, the monitor was calibrated using the Nokia Monitor Test program.[13] The surround of the CRT was masked so that extraneous reflections did not enter the surround area of the observers' vision, and the observers were placed approximately 50cm from the screen. The observers were, however, allowed minor adjustments of this distance in order to achieve a comfortable viewing distance. This was to accommodate individual fluctuations in visual acuity, as no eye test was given prior to this test. Allowing for a ± 10cm fluctuation in viewing distance, this gives a 0.01° change in the angle subtended by each screen pixel, which amounts to an approximately 30% latitude in pixel angle to accommodate these potential vision discrepancies.

For each algorithm and image combination, the observers were asked whether or not the perturbed image exhibited more of the artefact in question than the reference image. A total of 15 observers undertook the test with each observer completing the evaluations three times, giving a total of 45 attempts for each image / artefact combination.

The threshold or JND value was set as the algorithm value at which 75% of observers stated that they observed the difference. The choice of 75% as a JND level is based on the work of Gordon,[14] who assumes that if 50% of the observers can actually see a difference, then the remaining 50% will randomly vote for the presence or absence of an artefact, thereby increasing the number who actually state there is a difference to 75%.

Establishing the first JND for each artefact/image combination provided a set of perceptually equal images. This then allowed each set of artefacts to be compared for each image, e.g. London/Blur with London/Ringing and so forth.

In order to compare these images, a computer program was written that would display all possible pairings of the images under consideration and observers were asked which image of the pair presented was preferred. This had the effect of changing the results from the objective "can you see it?" to the subjective "is it objectionable?"

This program also compared each image against itself in order to highlight any bias inherent within the observers or the display device. In order to accomplish this, the images were randomly allocated either the left or right hand side of the screen and the percentage of left or right hand side images that were chosen was then recorded. The results from this bias testing never varied more than ± 4% from the mid point and were not consistent as to direction, so it was concluded that no bias existed. A statistical method of error analysis known as Standard Error of Percentages was applied to the data used to generate the JND values and the errors present were not significant. The results were then processed in order to generate Z-Score rank ordering data.[15]

Results

After placing the artefacts in order according to the Z-Scores for each image, they were then given a score according to their rank with 1 being the most preferred and 6 the least.

These results are given in Table 2, and shown in Figure 1. The final column shows the mean rank value for each artefact. The NPV (No Preference Value) column in Table 2 and shown in Figures 1 and 3 is the average value that could be expected if no preference was attached to any of the artefacts under consideration.

As can be seen from Table 2, the rank order scores for each artefact were broadly similar across all four images. It was therefore decided to test this correlation using Spearman's rank correlation[16] shown in Eq.1.

Table 2. Rank order.

	Lndn.	Land.	Leop.	Portr.	NPV	Mean Rank
Blur	2	1	1	3	3.5	1.75
Gibbs	3	4	3	2	3.5	3
Levels	5	3	6	5	3.5	4.75
Noise	4	5	4	4	3.5	4.25
Resolution	1	2	2	1	3.5	1.5
Ringing	6	6	5	6	3.5	5.75

$$r_s = 1 - \frac{6\sum x^2}{n(n^2-1)} \qquad (1)$$

where r_s is Spearman's Coefficient, x is the difference in rank for each artefact and n is the number of items in the sample.

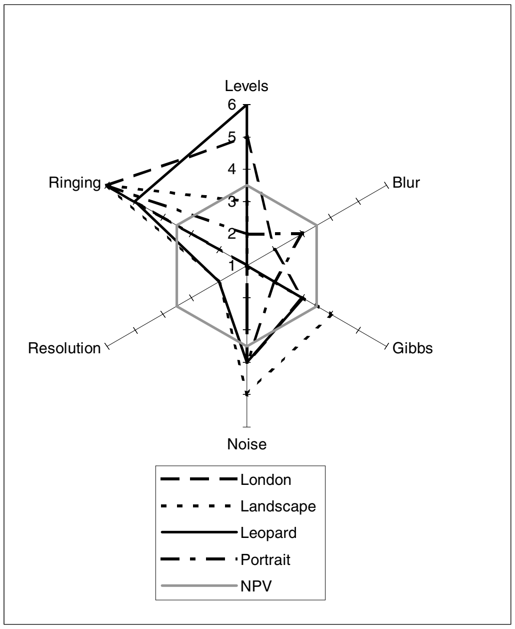

Figure 1. Rank order for the acceptability of artefacts; the differing line styles denote the four images used for this experiment. NPV is the value associated with a no-preference scenario.

This correlation was applied to each of the possible image pairs, the data being presented in Table 3. Though only two values, those for London/Leopard and London/Portrait actually fall within the 95% confidence limit for a sample of this size, the other correlation coefficients are sufficiently large to show that the values across the image range are in agreement.

Table 3: Spearman rank correlation coefficients.

Image Pairing	R_s
London/Landscape	0.771
London/Leopard	0.886
London/Portrait	0.943
Landscape/Leopard	0.657
Landscape/Portrait	0.600
Leopard/Portrait	0.771

As the Spearman test showed agreement across the images, it was decided to sum the rank scores for each artefact and image combination in order to provide an overall value for each artefact, and these values are tabulated in Table 4. As an assessment of the errors inherent within these figures, a method involving the deviation between the actual rank position and the average rank position was used; Eq.2, and the mean rank scores with these error bars are shown in Figure 2.

Table 4: Artefacts ranked in order of preference.

Artefact	Mean Rank
Resolution	1.5
Blur	1.75
Gibbs	3
Noise	4.25
Levels	4.75
Ringing	5.75

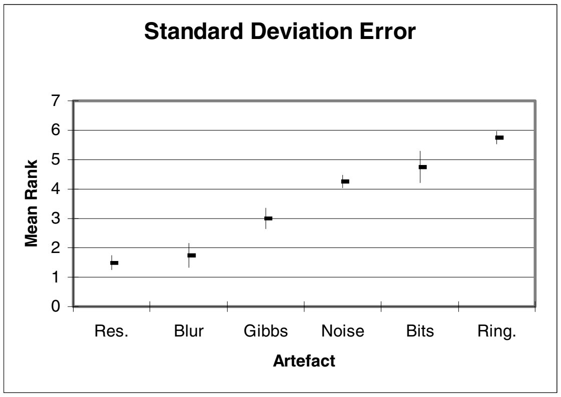

Figure 2. Rank order of artefacts, showing the error generated by the σ of their respective mean rank.

$$\frac{\sqrt{(x_i - y)^2 \ldots (x_n - y)^2}}{n} \qquad (2)$$

Figure 3 shows the mean rank scores for each artefact and image combination, and also where the no-preference value would fall.

Discussion

Having established a set of objective first JND levels for the image/artefact/display combination outlined above, a series of rank orders for the subjective significance of the artefacts as relating to each image was then calculated. Though not consistent across all four images (Figure 1), a definite trend is seen to emerge as to a hierarchy of visual significance for the artefacts under consideration (Figure 2). The effects of scene dependency that give rise to the inconsistencies seen in Figure 2 are not as pronounced as would have been expected, though they are not so pronounced as to affect the hierarchy seen to emerge in Figure 2.

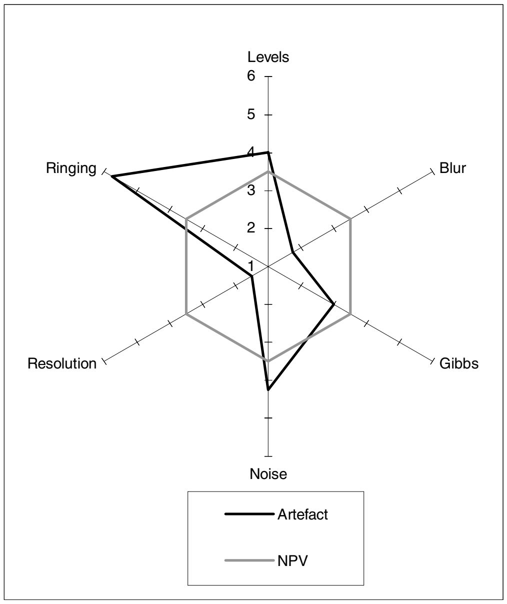

Figure 3. The overall mean rank score for each artefact with six as the worst.

Though the findings of this study are by no means conclusive, early results are significant and a hierarchy can be seen to exist. Further work is required to extend both the number and skill base of the observers and the variety of images under consideration. This further work would establish whether the hierarchy observed is consistent for all skill levels of observer, or if trained observers are ‘conditioned’ to be susceptible to certain artefacts. Increasing the number of images under consideration would allow further research into the scene dependency of artefacts.

Conclusions

- Definite trend of visual significance apparent.
- Scene dependence not as pronounced as first thought.
- Effect of observer skill level needs investigation.
- Further work required to investigate/quantify scene content.

Acknowledgement

K. MacLennan-Brown acknowledges the Engineering and Physical Sciences Research Council for supporting this work with a CASE award.

References

1. Milan Sonka, Vaclav Hlavac and Roger Boyle, *Image Processing, Analysis and Machine Vision,* International Thomson Computer Press, London, 1993, pg.28
2. Kenneth R. Castleman, *Digital Image Processing*, Prentice-Hall, International, London, 1996, pg.441
3. Rafael C. Gonzalez, Richard E. Woods, *Digital Image Processing*, Addison – Wesley Publishing, England, 1992, p.34.
4. Milan Sonka, Vaclav Hlavac and Roger Boyle, *Image Processing, Analysis and Machine Vision,* International Thomson Computer Press, London, 1993, pp.24-27
5. Kenneth R. Castleman, *Digital Image Processing*, Prentice-Hall, International, London, 1996, p.262
6. Rafael C. Gonzalez, Richard E. Woods, *Digital Image Processing,* Addison – Wesley Publishing, England, 1992, pp.195-201
7. Shanika A. Karunesekera, Nick G. Kingsbury, *IEEE Transactions on Image Processing*, 4, 6, 713 (June 1995).
8. Francois-Xavier Coudoux, Marc Georges Gazalet, Patrick Corlay, Jean-Michel Rouvaen, *Journal of Visual Communication and Image Representation*, Vol. **8**, No. 4, 327 (1997).
9. Cathleen M. Daniels and Douglas W. Christoffel, The Perceived Image Quality of Reduced Color Depth Images, *Proc. PICS*, pg. 196. (1998).
10. M. R. Pointer, G. G. Attridge, *Color Research and Application,* **23,** 1,112 (1998).
11. Michael A. Kriss, Tradeoff Between Aliasing Artifacts and Sharpness in Assessing Image Quality, *Proc. PICS* pg. 247. (1998).
12. *Matlab*, The Mathworks Inc.,
13. Nokia Monitor Test V1.0a, Nokia Monitors, P.O. Box 37, S-164, 93 Kista, Sweden.
14. I.E. Gordon, *Theories of Visual Perception*, John Wiley and Sons, Chichester, 1989, p.24.
15. Woodlief Thomas Jr., Editor, *SPSE Handbook of Photographic Science and Engineering*, John Wiley and Sons, London, 1973, p.1115.
16. S.S. Cohen, *Practical Statistics*, Edward Arnold, London, 1988, p.81.

Biography

Ken MacLennan-Brown received an honours BSc. Degree in Photographic and Electronic Imaging Science from the University of Westminster in 1996. Now a visiting lecturer in image science at the University of Westminster Ken has just submitted his PhD thesis for examination in the field of evaluating relative visual response to digital artefacts. Kens research interests lie in the fields of subjective image quality and colour science.