Perceptibility and acceptability of JPEG 2000 compressed images of various scene types

Authors: Elizabeth Allen, Sophie Triantaphillidou, Ralph Jacobson University of Westminster, London, UK

ABSTRACT

This investigation examines the relationships between image fidelity, acceptability thresholds and scene content for images distorted by lossy compression. *Scene characteristics* of a sample set of images, with a wide range of representative scene content, were quantified, using simple measures (*scene metrics*), which had been previously found to correlate with global scene lightness, global contrast, busyness, and colourfulness. Images were compressed using the lossy JPEG 2000 algorithm to a range of compression ratios, progressively introducing distortion to levels beyond the threshold of detection. Twelve observers took part in a paired comparison experiment to evaluate the perceptibility threshold compression ratios, to identify the level of compression at which the images became visually unacceptable. Perceptibility and acceptability thresholds were significantly correlated for the test image set; both thresholds also correlated with the busyness metric. Images were ranked for the two thresholds and were further grouped, based upon the relationships between perceptibility and acceptability. Scene content and the results from the scene descriptors were examined within the groups to identify and determine the influence of specific common scene characteristics upon both thresholds.

Keywords: Image compression, image quality, image fidelity, acceptability, scene classification

1. INTRODUCTION

An important concern in the evaluation of a lossy compression algorithm is the threshold of perceptibility of artifacts introduced into the image [1, 2], i.e. the fidelity of the compressed image when compared to the original. Psychophysical studies of thresholds and just noticeable differences (JNDs) of distortion, contribute to the development of guidelines around the use of image processes for imaging applications where fidelity is critical (such as forensic and medical imaging [3, 4]), as well as providing reference data against which image quality metrics may be benchmarked. For less specialised imaging applications, image fidelity is not always a requirement and the *acceptability* of image degradations in a given context is also useful. Whereas image fidelity studies involve observer judgements about perceptibility thresholds and just noticeable differences (JNDs), judgements of acceptability are exclusively suprathreshold and are concerned with image quality; in this case distortions are visible, but may or may not be bothersome to the observer.

1.1 Scene dependency in image quality studies

An important challenge to the successful prediction or evaluation of image fidelity and quality is the presence of scene dependencies that may affect the results [5, 6]. Triantaphillidou et al [7] have identified three types of scene dependency affecting psychophysical studies:

Scene dependency resulting from observers' preference criteria relates to the imaging context and the relative specialism of the observer group. For example, the preference criteria of forensic specialists looking at fingerprint images will focus on the specific image attributes important to the extraction of key fingerprint features; in this case, variation in imaging conditions and image attributes may be restricted. In more general applications of imaging, preference criteria will be affected by variation in attributes that make a scene more or less 'pleasing'. As well as varying across individual scenes (and observers), preference varies for broad classes of scenes, meaning that, for example, a slight level of blurring may be more acceptable (or preferable) in a portrait compared to an architectural image [8]. The selection of scenes with varied (but representative) scene content helps to ensure that the results are not skewed by scene dependency as a result

of the content of a single scene. In all cases, images need to be optimised consistently, to a level appropriate to the application (and therefore the observers' preference criteria).

Scene susceptibility due to the visibility of an artifact [7] [9] depends upon the nature of the artifact, and results, effectively, in the perceptibility threshold changing across the image plane. While some image content may accentuate the visibility of an artifact, conversely, some may help to mask it. This is closely related to scene dependency (or susceptibility) of digital processes or image processing algorithms [7]. Where the localisation and nature of the artifacts is determined by the operation and design of the algorithm, their visibility in the image depends upon the interaction of the algorithm with original scene attributes.

These last two aspects of scene dependency can determine the suitability of a process for a particular imaging application. Scene dependency may well be the deciding factor in the adoption of that imaging process, for example, if, the process being evaluated affects an image characteristic that is important to the image quality judgement (for the particular application). Alternatively the process may adversely affect an image characteristic because of the typical scene content or type.

Scene dependency presents a challenge to the development of objective measures of image fidelity and image quality. The most successful metrics tend to be developed for specific imaging contexts, but may not be as useful in predicting image quality if the context is changed. Consideration of scene dependency during the design of an image quality metric could potentially improve the correlation between results obtained from objective metrics and the results from psychophysical studies.

1.2 JPEG 2000, artifacts and image quality

The JPEG 2000 image compression standard was developed in the late 1990s to improve upon JPEG in terms of both flexibility and image quality; also to meet the growing demands in image compression of different digital imaging applications, to produce superior rate-distortion at low bit rates to that of existing standards [10]. JPEG 2000 has been of research interest in a number of specialised sectors of the imaging industries, including forensic imaging, where it has been explored for the compression of images of fingerprints and footprints, and biometric imaging applications [11] [12], as well as in certain fields within medical imaging. Nevertheless, the examination of the performance of JPEG 2000 without restriction to a particular type of image or application is useful in developing an understanding of the influence of scene dependency on image quality.

This investigation explores the perceptibility and acceptability of image distortions introduced as a result of JPEG 2000 in a non-specialised imaging context. The lossy version of JPEG 2000 is designed to be visually lossless above a certain threshold level of compression but at lower bit rates produces various characteristic and localised artifacts (Figure 1).



Figure 1. Characteristic artifacts of JPEG 2000: smudging evident in areas of texture, ringing artifacts at edges

The experiment uses a range of carefully selected scenes, which have been objectively evaluated in terms of their content. The results are considered in terms of the visibility and localisation of the artifacts, as a result of different scene characteristics and thus their effect on the perceived quality and acceptability of the images.

2. METHODOLOGY

The experimental work consisted of three interrelated parts: (i) the quantification of scene attributes of the sample test images using a range of previously tested simple scene analysis tools[7, 13]; (ii) the determination of perceptibility thresholds of distortion of the images when compressed using the JPEG 2000 lossy compression algorithm; (iii) further investigation into the acceptability of the compressed images beyond the perceptibility threshold.

2.1 Image acquisition and processing

The focus during the preparation of the sample image set was to obtain images of optimal quality prior to compression. The images were selected to encompass a range of scene content, allowing investigation into the scene dependency of the algorithm. Different scene types were included in the test set, providing good variation in image attributes, and captured under a range of different lighting conditions typically encountered in consumer photography.

An original sample set of 44 images was captured in raw file format (.cr2) using a Canon EOS 5D mkII D-SLR camera, which has a full frame sensor with resolution of 21Mp (5616 x 3744 pixels), and a Canon EF 24-70mm L II USM lens. Use of the raw file format enabled careful control of the image-processing pipeline prior to image compression and minimized the introduction of further unwanted distortions from other processes.

The images were processed in an sRGB viewing environment, on the same calibrated display that was used later in the psychophysical investigation. The steps in the processing pipeline were based on a typical camera-processing pipeline, but using linear rather than adaptive processing methods.

Using an external raw processing pipeline, the images were optimized scene-by-scene to correct exposure and white balance. Colour noise reduction was applied, the colour space was set to sRGB and the images were down-sampled to the minimum size possible in the raw processor (from 5616 x 3744 pixels to 1536 x 1024 pixels). Further down-sampling was applied(using bicubic interpolation) to optimize for the psychophysical display. The bit depth was reduced from 16-to 8-bits and after a final sharpening stage using an unsharp mask; the images were saved as uncompressed TIFF files.

2.2 Scene analysis and selection

The optimized images were converted to CIELAB for scene analysis. A range of simple image analysis tools were used, [[7]], to evaluate and rank selected scene characteristics in the test images and provide relevant visual scene descriptors (i.e., metrics used to quantify each scene characteristics) [14]. These included:

- first-order statistical measures (median md, variance V, and skewness s); derived from the probability density function (PDF) of the L* channel.
- a busyness metric; a multi-stage segmentation process, applying edge detection with a sobel filter followed by a combination of morphological processes, which evaluated the proportion of detailed areas in the image as a percentage of overall image area; derived from the L* channel.
- chroma variance VC^*_{ab} , the variance of CIELAB C^*_{ab} ; derived from the a* and b* channels.

These measures allowed the images to be broadly classified according to their overall lightness (m, md, s), global contrast (V), spatial content / amount of detail (*busyness metric*, b) and color contrast (VC^*_{ab}) . Histogram skewness was included as a measure of global scene lightness, because as well as correlating with median and mean values, it indicates a predominance of light or dark tones within the image (i.e. the 'key' of the scene).

The original set of 44 images was ranked according to each of the scene descriptors. The mean value for the scene descriptor was first determined. Images were then classified according to whether they fell into the *average* category for the selected scene characteristic, the *greater than* or *less than average* categories (corresponding to more than ± 0.5 standard deviations from the mean). 25 images were selected based upon these results, to ensure that each of the three categories was adequately represented for all five scene descriptors. Examples of images from the three categories are illustrated in Figure 2.



Figure 2. Examples of scenes from the three classes (less than average, average, greater than average) for the five scene attributes. Images are (left to right, top to bottom), *Lilies, Kids, Afternoon Tea, Cliffs, Lamp, Emporium, Huddle, Players Navy, Flower Garden.*

2.3 Image compression

The processed sRGB TIFF files were compressed as JPEG 2000 files in MATLAB. Default settings were used for the compression parameters (i.e. lossy compression, single quality layer, tile size equal to image size) and the images were compressed to a set of defined compression ratios (CR). CRs for the perceptibility test were 5, 10, 15, 20, 25, 30 and 40; a range found to be adequate for determining thresholds of perceptibility of JPEG 2000 in a previous study [6].An additional pilot test using two observers, was carried out to ensure that the range was suitable under the current experimental conditions. Results indicated that three images ('Afternoon Tea', 'Fred' and 'Bride') had potentially high perceptibility thresholds, and therefore their range was extended up to a CR of 60:1. The sample set consisted of 190 images in total. The acceptability test, being a suprathreshold evaluation, required a larger range of compression. The images were inspected at a range of higher CRs and 70:1 was established as a rate at which most images became unacceptable. The images were therefore further compressed to the following compression ratios: 45, 50, 55, 60, 65, 70,

a total of 150 images. The acceptability test for an individual observer consisted of all images in which the observer had noted a difference during the perceptibility test, in addition to the 150 images at higher compression ratios.

2.4 Psychophysical display and viewing conditions

Perceptibility and acceptability thresholds were evaluated through a two-part paired comparison test. The test interface was developed using MATLAB. The test images were displayed side-by-side, one compressed and the other an uncompressed original. The images were presented in a random sequence and the original and compressed versions were randomized in their presentation on the left or right of the screen. The effective screen size was 518.4mm wide by 324.0 mm; the images took up approximately 45% of the half-screen area on a mid-grey background. Image size was selected to ensure that there would be no interpolation when they were displayed. The viewing distance was fixed at 60cm, giving an angle of subtense of 22.450 degrees of arc (0.392 radians). The time to view the images was unrestricted, the observer controlling when they would move on to the next image using a push button.

The display was an EIZO CG245W, and was calibrated during the period of the test to the sRGB specification [15]. The viewing environment was also calibrated to closely match the sRGB specification, with ambient colour temperature of 5000K and an ambient illuminance of 64 lux.

Twelve experienced observers carried out the perceptibility test, and eleven of them completed the acceptability test. All had normal, or corrected vision.

In the first section of the test, observers were asked to provide a 'yes' or 'no' answer, to the question of whether they could perceive a difference between the two displayed images of the same scene. Observers were given the opportunity to take a break between the perceptibility and acceptability tests and were asked to stop if they felt tired. In the second half of the test, observers were asked whether they found the compressed image acceptable when compared to the original.

2.5 Evaluation of thresholds

The proportions of responses were used to generate estimated psychometric curves against compression ratio for each image, obtained by fitting a logistic function to the data. The psychometric curves directly related the proportion of observers' responses to the original compression ratio and were used to evaluate various points of interest: the point of subjective equality (PSE), or absolute threshold, defined as the 0.5 proportion point; the just-noticeable-difference (JND) (defined by convention as the stimulus increment between the PSE and the 0.75 proportion); and the detection threshold, defined here as the 0.75 proportion point.

The PSE is defined as the statistical point at which observers perceive two images to be equal [16, 17]. Here the PSE corresponded to the compression rate at which 50% of the observers responded 'yes' to perceiving a difference between the two images. Corresponding to this, the absolute acceptability threshold is the point at which 50% of the observers find the images unacceptable (i.e. respond 'no' to the question 'Do you find the image acceptable.').

The compression ratio identified at the 0.75 proportion was the point at which 75% of observers could either perceive a difference between the two images, or found the differences unacceptable; in this study it is this value that is referred to as the threshold of perceptibility/acceptability

2.6 Error estimation and goodness of fit of the psychometric curve

The curve fitting procedure produced a maximum likelihood estimate of the parameters (α , corresponding to threshold at a probability (P) of 0.5 and β , corresponding to the slope) of the psychometric curve for each image, based upon the observer responses across the compression range [17]. A goodness-of-fit test was performed when the estimated curve was generated, based on 1000 simulations of the data [17] and resulting in a ρ -value, the probability that the observed data could be part of the population generated from the estimated model. ρ -values of below 0.05 were deemed to be an unacceptably poor fit. The results of the goodness-of-fit test were used as a means of determining the approach to be used in the error estimation.

To estimate standard error, for a ρ -value of greater than or equal to 0.05, a bootstrapping procedure was again used to generate 400 hypothetical sets of data, based on the parametric description of the observed experimental data (i.e. the estimated curve). A Logistic function was fitted to each set of simulated data to derive the new α and β parameters. Finally the sample standard deviation (i.e. the standard error, equal to population standard deviation divided by number of samples) for each parameter was calculated from their distributions in the simulated sets of data. This *parametric bootstrap* approach used the estimated curve as the starting point and its parameters were the mean values from which the standard deviation was calculated.

In the case of images where the ρ -value <0.05, a *non-parametric bootstrap* was used to evaluate the errors. The experimental data was used instead of the parameters of the curve in the simulations, obtaining hypothetical data based on actual data rather than an 'average' estimated function [17]. The standard deviation for the α and β parameters was again determined from the sampling error across the simulated datasets.

The error estimation procedures did not generate an accurately fitted function ('failed fits') for a few of the images. In these cases, the failed datasets were excluded in the calculations of standard error, as their estimated parameters were deemed an inaccurate fit to the data and would have biased the results.

3. RESULTS

3.1 Scene ranking from objective measures

The images were ranked according to each scene descriptor. The correlations between the individual scene descriptors were evaluated to determine whether there was a predictable relationship between them for the sample image set. They were calculated from the original scenes using Spearman's correlation coefficient[18]: $r = \frac{6\Sigma d^2}{n(n^2-1)}$, where *d*=the difference in rank between the two descriptors for each image, and *n* is the total number of scenes. For 24 degrees of freedom (from n-1), the coefficient has a greater than 95% chance of being significant at a value > 0.406 [19].

The results are shown in Table 1. The coefficient of the correlation between median and skewness (in bold) is the only coefficient that indicates a significant correlation (for this number of images). The results for all images and scene attribute ranks are shown in Table 2.

Scene Measure		Median	Skewness	Variance	Busyness	Chroma Variance
		т	S	V	b	VC* _{ab}
Median	т	1.00				
Skewness	S	0.89	1.00			
Lightness Variance	V	-0.18	-0.32	1.00		
Busyness	b	-0.05	-0.24	0.33	1.00	
Chroma Variance	VC*ab	-0.29	-0.33	0.03	0.30	1.00

Table 1. Spearman Correlation Coefficient values of pairs of measures ranked according to values.

Image	md	Rank <i>md</i>	S	Rank s	% b	Rank b	V	Rank V	VC* _{ab}	Rank VC* _{ab}
accordion	120	6	0.05	9	49.67	11	5760.00	19	144.64	12
afternoon tea	154	18	-0.80	22	2.59	1	3712.70	7	123.24	10
beach goods	128	11	0.17	4	62.80	20	5573.10	17	405.31	24
bride	138	14	-0.40	17	18.23	5	4745.80	14	252.32	20
cliffs	140	16	-0.25	14	61.95	19	2899.60	5	338.50	23
crockery	125	9	-0.10	11	74.24	23	6625.40	21	175.04	15
crown antiques	172	22	-0.48	18	65.23	21	8644.40	24	226.21	17
emporium	60	1	0.54	1	55.49	14	7884.70	23	34.66	1
flags	154	19	-0.73	21	9.60	4	1750.60	1	133.30	11
flower garden	127	10	0.07	7	93.38	25	2863.10	4	270.20	21
formal	129	12	0.13	5	59.11	17	5585.10	18	246.59	18
fred	205	24	-1.42	24	21.61	6	6043.70	20	65.57	2
hive beach	109	3	0.45	3	29.86	8	4743.50	13	118.20	9
huddle	186	23	-1.60	25	23.26	7	1843.60	2	100.49	6
kids	134	13	-0.06	10	6.43	3	2491.90	3	70.98	3
lamp	214	25	-1.36	23	43.50	10	4274.80	10	152.37	13
lilies	110	4	0.11	6	35.22	9	4425.80	12	303.90	22
marle sculpture	158	20	-0.32	15	50.43	13	7565.20	22	79.43	4
pink flowers	124	7	-0.24	13	49.80	12	2977.50	6	250.70	19
players navy	60	2	0.46	2	61.60	18	9484.10	25	443.74	25
pool	153	17	-0.63	19	59.07	16	4984.30	15	208.66	16
seagull	139	15	-0.33	16	72.02	22	4110.10	8	104.93	7
serpent	124	8	-0.19	12	56.68	15	4157.30	9	115.77	8
stones ii	167	21	-0.64	20	84.05	24	5002.30	16	98.19	5
summer	119	5	0.05	8	3.04	2	4300.60	11	157.97	14
average	13.2		-0.32		47.74		4954.05		185.96	
SD	7.2		0.56		24.05		2094.79		107.23	

Table 2. Images ranked according to objective measures: Median (md), Skewness (s), % Busyness (b), Variance (V), Chroma Variance (VC^*_{ab})

3.2 Perceptibility and acceptability thresholds

The results for the thresholds for perceptibility and acceptability for all images are shown in Table 3. The images are presented ranked according to their perceptibility thresholds. The ranks indicate a positive correlation between the perceptibility and acceptability thresholds, as illustrated in Figure 3. This is not unexpected; images are less likely to be judged unacceptable if distortion is less visible. Of more interest are the images where there are differences in their rankings for perceptibility and acceptability, highlighted in Section 4 (grouping of the scenes). Four of the image groups discussed in section 4 are identified as clusters on the graph in Figure 3.

	I	Perceptibility	Threshold	ls	A	Acceptability '	Threshold	S
	PSE	Threshold	JND P(0.75)-	Rank	PSE	Threshold	JND P(0.75)-	Rank
	P(0.5)	P(0.75)	P0.5)		P(0.5)	P(0.75)	P0.5)	
Afternoon Tea	44.8	54.6	9.8	25	NA	NA	NA	24*
Fred	34.0	47.9	13.9	24	NA	NA	NA	25*
Summer	24.9	32.4	7.5	23	36.5	48.2	11.8	22
Lamp**	24.2	31.1	6.9	22	44.0	55.9	11.9	23
Lilies	23.1	29.7	6.6	21	38.0	48.1	10.1	21
Huddle	19.3	25.5	6.2	20	30.1	36.2	6.1	16
Bride	16.6	24.4	7.7	19	34.5	45.7	11.2	20
Flags	19.2	24.0	4.8	18	28.0	31.5	3.5	11
Emporium	17.9	22.3	4.3	17	34.6	44.6	10.0	19
Pink Flowers	17.5	21.7	4.2	16	25.8	30.7	4.9	10
Kids	14.9	21.6	6.7	15	32.5	40.4	7.9	17
Serpent	15.5	21.6	6.1	14	26.2	35.1	8.9	14
Crockery	17.0	21.6	4.6	13	26.3	33.2	6.9	12
Accordion	14.7	20.4	5.7	12	31.2	41.5	10.3	18
Marle Sculpture	16.2	20.2	4.0	11	26.2	33.7	7.5	13
Pool	14.5	18.2	3.6	10	24.1	36.2	12.0	15
Flower Garden	12.5	17.4	4.8	9	22.1	25.4	3.3	5
Hive Beach	13.7	16.1	2.4	8	20.8	25.6	4.8	6
Formal	12.9	16.0	3.1	7	20.7	24.2	3.5	4
Beach Goods**	12.4	15.7	3.3	6	21.4	27.6	6.2	8
Crown Antiques	11.3	15.6	4.2	5	20.9	29.8	8.9	9
Players Navy	13.8	15.4	1.6	4	22.2	27.6	5.4	7
Seagull	11.6	13.9	2.3	3	18.9	21.7	2.8	2
Stones II**	11.2	13.3	2.0	2	17.5	21.9	4.5	3
Cliffs**	8.0	9.3	1.3	1	12.8	16.7	3.9	1

Table 3. Perceptibility and acceptability thresholds for all images. Ranks are based on the P(0.75) threshold.



Figure 3. Relationship between perceptibility and acceptability threshold compression ratios.

Two scenes with the highest rank in terms of perceptibility threshold (*Afternoon Tea*, and *Fred*, marked with an * in table 3) do not have corresponding values for acceptability thresholds. These images proved to be extremely robust under JPEG 2000 compression. Neither reached the 0.75 proportion point for acceptability at the maximum compression ratio evaluated, meaning that psychometric curves and derived thresholds could not be estimated. The order of their ranking in terms of acceptability is assumed, based upon the proportions of observers who gave 'no' responses at the maximum compression ratio tested (0.4 for *Fred* and 0.7 for *Afternoon Tea*).

Four images (marked with **) are highlighted in Table 3 because they had p-values below the 0.05 threshold point in the perceptibility test, indicating that the estimated curve was of an unacceptably poor fit.

3.3 Correlations between thresholds and scene metrics

Spearman's rank correlation coefficients [18] were calculated to examine the relationships between the perceptibility and acceptability thresholds with each scene descriptor. The results are presented in Table 4.

Table 4. Spearman's correlation coefficients calculated between scene metrics and subjective thresholds.

	Perceptibility	Acceptability
md	0.09	0.09
V	-0.25	0.03
S	0.32	0.27
b	-0.80	-0.73
Vc	-0.28	-0.30

The high negative coefficients corresponding to the busyness metric with both subjective measures are significant. The negative sign of the coefficient indicates that as scene busyness increases, the thresholds of perceptibility and acceptability decrease. This implies scene dependency of the JPEG 2000 algorithm, meaning that it performs less well in images containing lots of detail. Figure 4 illustrates the effects of the algorithm on the most and least busy of the images. The majority of scene measures suggest weak correlation with the subjective thresholds. This implies that scene global lightness, contrast, and color contrast do not play a significant role in JPEG 2000 compression when considered across the entire sample set of images, although correlations may exist within the image groups indicated on Figure 3.



Figure 4. High and low ranking images in terms of the busyness metric. From left to right: *Fred*, compression ratio 70, perceptibility 54.5, acceptability threshold not reached; *Cliffs*, compression ratio 70, perceptibility 9.3, acceptability threshold 16.7; Close up of *Cliffs* illustrating significant distortion.

4. ANALYSIS OF RESULTS AND DISCUSSION

The results from the psychophysical experiments were examined in terms of the threshold levels, and the relationship between perceptibility and acceptability thresholds The images were grouped in terms of their compression performance. Correlations between scene descriptors within groups were identified.

4.1 Group I: Images with very high thresholds for perceptibility and acceptability

Three images belong to this category: *Lamp, Fred,* and *Afternoon Tea* (figure 5). All have very high acceptability thresholds (neither *Fred* or *Afternoon Tea* reached the acceptability threshold within the experimental CR range). The perceptibility thresholds for *Fred,* and *Afternoon Tea* are higher than the acceptability thresholds for 20 out of the other 23 images. *Afternoon Tea* has the highest perceptibility threshold of all the images.

The psychometric curves for the images indicate a degree of noise in the observers' responses, also confirmed by discussion with the observers after the test. Distortions were difficult to detect and did not tend to affect any of the salient features within the images.

The scene descriptors (Table 5) show similarity in the ranks for median, skewness and busyness, with high ranks for median and skewness, illustrating the predominance of light tones in the images and low ranks for busyness because of the lack of high frequency detail in each of them. Lighter images might result in less visible distortion, due to the reduced contrast in the distorted areas.



Figure 5. Group I images and psychometric curves for perceptibility: Lamp(left), Afternoon Tea (center), Fred (right)

	md	rank	<i>S</i>	rank	b %	rank	V	rank	VC* _{ab}	rank
Afternoon Tea	154	18	-0.80	22	2.59	1	3712.7	7	123.24	10
Fred	205	24	-1.42	24	21.61	6	6043.7	20	65.57	2
Lamp	214	25	-1.36	23	43.50	10	4274.80	10	152.37	13

Table 5. Objective scene descriptors and ranks for Group I images.

4.2 Group II: Images with high thresholds for perceptibility and acceptability

The two images in this category (Figure 6) had high values for both perceptibility and acceptability. The scene descriptors for the group II images indicate similarity in terms of median and skewness (lower than average), busyness (lower than average), and variance (average). The significant features in the images are not busy, and therefore less susceptible to the localized blurring and ringing artifacts introduced by JPEG 2000. In contrast to the images in group I, these images contain relatively large areas of lower than average lightness. The effect on distortions on these areas is

similar to that in the light areas in the images from group I; the distortion contrast is reduced and so they are potentially less visible.



Figure 6. Group II images, with high thresholds for both perceptibility and acceptability Summer(left), Lilies (right)

radie of objective seeme descriptors and ranks for oroup in mages.	Table 6.	Objective	scene descriptors	and ranks for	or Group II images.
--	----------	-----------	-------------------	---------------	---------------------

	md	rank	S	rank	b	rank	V	rank	C Var	rank
Summer	119	5	0.05	8	3.04	2	4300.60	11	157.97	14
Lilies	110	4	0.11	6	35.22	9	4425.80	12	303.90	22

4.3 Group III: Images with high perceptibility thresholds but lower in acceptability threshold rank

These two images (Figure 7) show similarity across all of the scene descriptors. Both images have high median and skewness rankings, low rankings for busyness and variance, and average to low rankings for chroma variance. As for group 1, the dominant light areas of similar colour and tone covering much of the image area do not appear to be very susceptible to visible distortion. At CRs beyond the perceptibility threshold for these images, the important features (the flags in the first image, the blue clothing in the second) are affected by very visible ringing as well as blurring; this may account for the reduction in acceptability once the distortion becomes visible.



Figure 7: Group III images: *Huddle (left)* (perceptibility: threshold CR 25.5, rank 20, acceptability: threshold CR 36.2, rank 16), *Flags (left)* (perceptibility: threshold CR 24.1, rank 18, acceptability: threshold CR 31.5, rank 11)

Table 7 Objective scene descriptors and ranks for Group III images.

	md	rank	S	rank	b	rank	V	rank	C Var	rank
Huddle	186	23	-1.60	25	23.26	7	1843.60	2	100.49	6
Flags	154	19	-0.73	21	9.60	4	1750.60	1	133.30	11

4.4 Group IV: Images with low perceptibility and acceptability thresholds and lower acceptability rank

The landscapes in group IV are dominated by natural textures, producing high values for the busyness and chroma variance scene descriptors. The blurring artifact is very noticeable in these scenes, particularly in the foreground areas where the textures are degraded significantly. The low thresholds for these images are unsurprising; the reduction in acceptability threshold rank compared to perceptibility is an indicator that the distortion is bothersome once perceived.



Figure 8: Group IV images: *Flower Garden (left)* (perceptibility threshold CR 17.4, rank 9, acceptability threshold CR 25.4, rank 5), *Formal (left)* (perceptibility threshold CR 16.0, rank 7, threshold CR 24.2 rank 4)

Table 8. Objective scene descriptors and ranks for Group IV images.

	md	rank	<u>s</u>	rank	b	rank	V	rank	C Var	rank
Flower Garden	127	10	0.07	7	93.38	25	2863.10	4	270.20	21
Formal	129	12	0.13	5	59.11	17	5585.10	18	246.59	18

4.5 Group V: Images with low perceptibility thresholds, but an increase in acceptability rank.

These images were very low in terms of perceptibility threshold, but exhibited improved acceptability rankings. The images have high busyness, lightness variance and chroma variance rankings. As well as containing visually important textural features, which proved susceptible to distortion, all three images contain text. The areas of texture are proportionally less than those in Group VI, which may account for their improved acceptability rankings.



Figure 9: Group V images: *Crown Antiques (left)* (perceptibility threshold CR 15.6, rank 5, acceptability threshold CR 29.8, rank 9); *Beach Goods (center)* (perceptibility threshold CR 15.7 rank 4, acceptability threshold CR 27.6 rank 8; *Players Navy (right)* (perceptibility threshold CR 15.4 rank 4, acceptability threshold CR 27.6 rank 7);

Table 8. Objective scene descriptors and ranks for Group V images.

	md	rank	S	rank	b	rank	V	rank	C Var	rank
Beach Goods	128	11	0.167	4	62.803	20	5573.10	17	405.31	24
Players Navy	60	2	0.46	2	61.60	18	9484.10	25	443.74	25
Crown Ant.	172	22	-0.48	18	65.23	21	8644.40	24	226.21	17

4.6 Group VI: Images with very low perceptibility and acceptability thresholds

With the lowest thresholds of all the images, these three images are found to be very busy, with significant proportions of the image areas dominated by texture. The texture was affected by blurring artifacts at very low compression ratio. Because of its visual importance within the images, this can be seen as the main factor influencing the results.



Figure 9: Group VI images: Seagull (left), Stones II (center), Cliff (right)

Table 8. Objective scene descriptors and ranks for Group V images.

	md	rank	S	rank	b	rank	V	rank	C Var	rank
Beach Goods	139	15	-0.33	16	72.02	22	4110.10	8	104.93	7
Players Navy	167	21	-0.64	20	84.05	24	5002.30	16	98.19	5
Crown Ant.	140	16	-0.25	14	61.95	19	2899.60	5	338.50	23

5. CONCLUSIONS

This study explores the relationship between perceptibility and acceptability thresholds of compression across a range of different scenes. The results indicate a significant correlation, for most images, between perceptibility and acceptability thresholds for JPEG 2000.

Although the acceptability context was not clearly defined to observers prior to the experiment, the results for acceptability thresholds across the observers were relatively consistent and the derived psychometric curves fit the observed acceptability data reasonably well, for the majority of images.

Scene characteristics of the test images were evaluated using simple scene descriptors (median, variance, histogram skewness, busyness, chroma variance)[7, 13]. A strong statistical correlation was found between the busyness descriptor and both perceptibility and acceptability thresholds, demonstrating the susceptibility of highly textured scenes to JPEG 2000 distortions as well as the scene dependency of the algorithm, due to the localized nature of the blurring distortions [6, 7]

Other scene characteristics did not correlate with the thresholds consistently across all of the images, but there was good correlation within the image groups, particularly with the descriptors for scene lightness (median and skewness). Images

with high thresholds were found to have low busyness and either higher than average, or lower than average lightness. In these cases, the contrast of the blurring distortions affecting light or dark areas within the images was low and therefore less visible. The reduced contrast also meant that the ringing artifact was not very visible.

Images within groups with low thresholds were also found to correlate across the scene descriptors. Busyness was the biggest influencing factor, but its effect on the thresholds depended upon the visual importance of the busy areas within the image. If the image area contained a large proportion of busy areas, or if important features were very detailed, distortions (particularly the blurring artifact) became both visible and bothersome. The majority of images with low thresholds were low to average in terms of lightness.

Research by Alers et al [20] has shown that image regions are unequally weighted in terms of visual significance by observers in image quality studies. The scene dependency of JPEG 2000 and the localization of its distortions mean that it affects some image areas more than others. The distribution of salient features [21]. (i.e. significant focal points in the image), their area in relation to the overall image area, and their susceptibility to distortion as an influence upon image quality warrants further investigation.

6. REFERENCES

- [1] Klein, S.A., Image quality and image compression: a psychophysicist's viewpoint, in Digital Images and Human Vision, A. Watson, Editor.
- [2] Silverstein, D.A., Farrell, J.E. *The relationship between image fidelity and image quality.* in *IEEE International Conference on Image Processing.* 1996. Lausanne.
- [3] Joong Kim, K., Ho Lee, K., Kang, H., Yeon Kim, S., Hoon Kim, Y., Bohyoung Kim, B., Seo, J., Rafal Mantiuk, R., *Objective index of image fidelity for JPEG2000 compressed body CT images.* Medical Physics 2009. **36** (3218).
- [4] Guarneri, F., Vaccaro, M., Guarneri, C., and Cannavo, T, *JPEG vs. JPEG 2000: benchmarking with dermatological images.* Skin Research and Technology, early view, online version, 2013.
- [5] Corey, G.P., Clayton, M.J. and Cuprey, K.N., *Scene Dependence of Image Quality.* Photographic Science and Engineering, 1982. **27**: p. 9-13.
- [6] Allen, E.T., S., and Jacobson, R.E., *Image Quality of JPEG vs JPEG 2000 Image Compression Schemes, Part 1: Psychophysical Measurements.* Journal of Imaging Science and Technology, 2007. **51**(3): p. 248-258(11).
- [7] Triantaphillidou, S.A., E.; Jacobson, R, Image Quality Comparison Between JPEG and JPEG2000. II. Scene Dependency, Scene Analysis, and Classification. Journal of Imaging Science and Technology, 2007. 51(3): p. 259-270(12).
- [8] Biederman, E.M., Photographic Korrespondent, 1967. **25 and 41**(103).
- [9] Keelan, B., Handbook of Image Quality: Characterization and Prediction, first edition2002: CRC Press.
- [10] Skodras, A.C., C. Ebrahimi, T., , *The JPEG 2000 Still Image Compression Standard*, in *Ieee Signal Processing Magazine*2001. p. 36-58.
- [11] Funk, W., Arnold, M., Busch, C., Munde, A. Evaluation of Image Compression Algorithms for Fingerprint and Face Recognition Systems. in Proceedings of the 2005 IEEE Workshop on Information Assurance and Security. 2005.
- [12] Daugman, J., Downing, C., *EFFECT OF SEVERE IMAGE COMPRESSION ON IRIS RECOGNITION PERFORMANCE.* IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, 2008. **3**(1): p. 52-61.
- [13] Oh K. H., T.S.a.J.R.E. Scene classification with respect to image quality measurements. in Proc. IS&T/SPIE Electronic Imaging 2010: Image Quality & System Performance VII. 2010. San Jose, California, USA.
- [14] Pratt, W.K., *Digital Image Processing*, Interscience, Editor 2007. p. 539.
- [15] Michael Stokes, M.A., Srinivasan Chandrasekar, Ricardo Motta *A Standard Default Colour Space for the Internet- sRGB version 1.10.* 1996.
- [16] Engeldrum, P., in *Psychometric Scaling: A Toolkit for Imaging Systems Development*2000, Imcotek Press.
- [17] Kingdom, F.A.A., and Prins, N., *Psychophysics: A practical Introduction*2010, 32 Jamestown Road, London NW1 7BY, UK: Elsevier Ltd.

- Royal Geographic Society, w.I. Spearman's Rank Correlation Coefficient Excel Guide. [18]
- [19] Zar, J.H., Biostatistical Analysis1984, (Englewood Cliffs, N.J.): Prentice-Hall
- [20] Alers, H., Redi, J., Liu, H., Heynderickx, I., Studying the effect of optimizing image quality in salient regions at the expense of background content. Journal of Electronic Imaging, 2013. **22**(4). Achanta, R., Hemami, S., Estrada F., Süsstrunk, S, Frequency-tuned Salient Region Detection. IEEE
- [21] International Conference on Computer Vision and Pattern Recognition, 2009: p. 1597 - 1604.