Journal of Imaging Science and Technology® 51(3): 259–270, 2007.

Image Quality Comparison Between JPEG and JPEG2000. II. Scene Dependency, Scene Analysis, and Classification

S. Triantaphillidou, E. Allen and R. E. Jacobson
Imaging Technology Research Group, University of Westminster, School of Media, Arts and Design, Harrow Campus, Watford Road, Northwick Park, Harrow HA1 3TP, United Kingdom
E-mail: triants@westminster.ac.uk

Abstract. *Image quality assessments have shown that both JPEG and JPEG2000 compression schemes are dependent on scene content. This paper addresses the problem of scene dependency and scene susceptibility in image quality assessments and proposes image analysis as a means to group test scenes, according to basic inherent scene properties that human observers refer to when they judge the quality of images. Experimental work is carried out to investigate the relationship between scene content and the subjective results obtained from experimental work carried out in [E. Allen, S. Triantaphillidou, and R. E. Jacobson, "Image quality comparison between JPEG and JPEG2000. I. Psychophysical Investigation", J. Imaging Sci. Technol. 51, 248 (2007)]. The content of the test images used in this work is analyzed using simple image analysis measures that quantify various image features, such as original scene contrast and global brightness, amount of dominant lines, scene busyness (defined here as a scene/image property indicating the presence or absence of detail), and flat areas within the scene. Preliminary results and conclusions are obtained and suggestions are made to form a basis for further studies on scene dependency and scene classification with respect to image quality measurements. © 2007 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.(2007)51:3(259)]

INTRODUCTION

Image quality, in its strict definition, is concerned ultimately with the subjective impression the image conveys.[1–3] The subjective assessment of image quality is a function of the human visual system and the "quality criteria" of the observers.[4] It is highly dependent on the viewing conditions, the context within which the image is to be used and very importantly *the physical properties of the test stimuli.*

Image measurements relating to quality issues assume that there is a functional relationship between the subjective impression of image quality and some selected physical attributes of the observed stimuli. The main physical attributes, or so-called image quality dimensions, that influence the appearance of image stimuli (see Table I) are assessed either using subjective studies[3,5] or objective measurements[1,2,6–8] to predict the subjective impression of image quality. They are assessed either collectively (i.e., assessments of the overall excellence) or individually, either consciously or unconsciously by the observer when the quality of images or imaging systems is considered. The individual subjective assessment of the image quality attributes has been a subject for discussion, since many have argued that judgements of image attributes are unlikely to be independent from other attributes, while the relationship between them has been studied extensively.[9,10] Table I lists these attributes of image quality, along with the associated perceptual attributes.[11]

In addition to the physical attributes common to all imaging systems, digital images suffer from artifacts that cannot be classified or quantified in a conventional manner. The most common are listed in Table II, along with their origins and the areas within imaging stimuli that are more susceptible to each artifact. One can broadly categorize their origins as follows: Numbers 1–3 are artifacts due to the nature of the digitization process, which involves spatial sampling and quantization, numbers 4–6 are artifacts resulting from digital image processes, such as image compression and sharpening, while 7–9 are types of nonisotropic noise, originating most often from errors in digital printing. Artifacts 10 and 11 are also common to analogue systems, but their impact on images cannot be measured by conventional techniques.

This paper discusses the following. In the next section, a detailed account on scene dependency related to image quality measurements. In the section following that, ideas for a basis for scene classification with respect to image quality measurements. In the section entitled Scene Analysis, a list of simple image analysis tools that can be employed as a starting point for identifying and grouping test images/scenes. Finally, in the section entitled Scene Dependency and Subjective Results, we try to establish trends between original scene content and perceived quality of compressed images with JPEG and JPEG2000 compression schemes.

Table I. Image attributes examined in image quality assessments and associated perceptual attributes (Ref. 10).

Image Attribute	Visual Description
1. Tone	*Macroscopic contrast* or reproduction of *intensity*
2. Color	Differences in *lightness, chrominance* and *hue*
3. Resolution	Discrimination of *fine detail*
4. Sharpness	*Microscopic contrast* or reproduction of *edges*
5. Noise	Spurious information

Received Aug. 15, 2006; accepted for publication Jan. 13, 2007.
1062-3701/2007/51(3)/259/12/$20.00.

Table II. Common digital image artifacts, their sources, and areas within images which are more susceptible to these artifacts. Note: Susceptible areas are defined here as either those affected mostly by the artifact or areas in which the artifact is more evident.

Image Artifact	Cause of Artifact	Susceptible Image Areas
1. Contouring	Poor quantization	Uniform areas, slow varying areas (flat areas)
2. Jaggedness/Pixelization	Insufficient spatial resolution	Slanted edges, slanted lines, high frequency information
3. Aliasing	Sampling	Areas with periodic high frequency information (high frequency lines)
4. Blocking	DCT compression	Areas with high frequency information (busy areas)
5. Smudging/Color bleeding	DWT compression	Areas with high frequency information (busy areas)
6. Ringing or edge echoes	Digital sharpening or DCT compression	Edges, lines
7. Patterning	Dithering	All areas expect from pure blacks and pure whites
8. Streaking	Pixel to pixel nonuniformity in linear arrays (mostly of digital writing devices)	Uniform areas, slow varying areas (flat areas)
9. Banding	Cyclical variations in a property of digital writing devices	Uniform areas, slow varying areas (flat areas)
10. Color misregistration	Optical images for different color channels not geometrically identical	Small amounts: edges, lines, areas with high frequency information. Large amounts: all areas
11. Flare	Stay light in dark areas	Dark areas surrounded by high intensity areas

SCENE DEPENDENCY AND SCENE SUSCEPTIBILITY

In image quality measurements a variety of images representing scenes with different physical properties (or *spatial configurations*[12]) are used for the complete assessment of imaging components and systems. An unsurprising result that stems from these variations in original scene configuration is that in many investigations the subjective quality is found to be scene dependent.[13–16] That is, the results are shown to vary with the content of the images used for the investigations.

In the following paragraphs, the authors attempt to differentiate between three different origins of scene dependency in image quality measurements.

1. Scene dependency resulting from the observer's quality criteria (i.e., observer's preference). It is well known that image classes and scene content exert an influence on quality judgements. Freiser and Biedermann found, as early as 1963, that the sharpness of portraits and landscapes are judged differently,[17] Bartleson and Bray in 1962,[18] and later Hunt et al.,[19] showed that the preference of critical colors, such as green grass, blue sky and flesh colors, is different from the actual colors themselves. Scene characteristics, such as the spatial distribution of subjects[20] and camera to subject distance,[21] have been shown to be important scene dependent parameters in observer's preferences.

An example of the type of scene dependency originating from the observer's preferences is that portraits are often preferred when they are less sharp compared to images with strong lines, such as those of architectural scenes;[22,23] the reason being that "soft focus" renders the skin smoother and thus more pleasant to the viewer, whereas strong lines and edges are usually preferred when they are sharp. This was first reported in 1967 by Biederman,[22] who observed that for very high objective quality portraiture subjective quality decreased as objective quality increased and confirmed what practical photographers had known for years. This is seen in the use of diffusing screens placed over the lens, for example, to soften the image.

In Figure 1, both images have been subjected to the same objective amount of sharpening [top versions, (a) and (c)] and blurring [bottom versions, (b) and (d)]. Most observers would probably judge the blurred version of the portrait to be of a higher quality than the sharpened version, whereas the opposite judgement is more likely for "the door" scene.

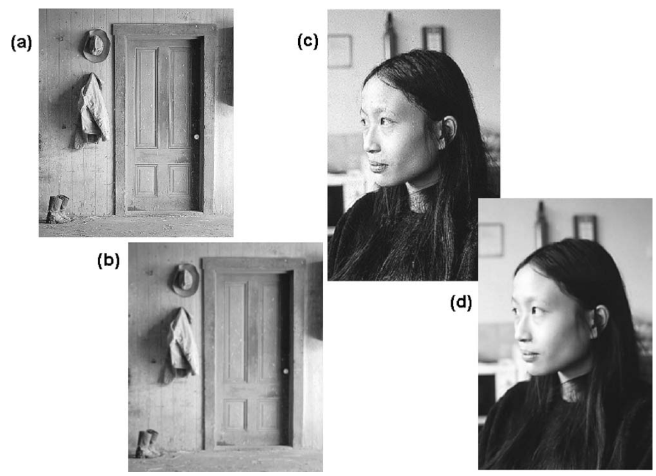

Figure 1. Blurred (top) and sharpen (bottom) versions of two images with different scene content.

Distortion and most objective fidelity and quality measures[1,2,7] will assess images (a) and (c), and (b) and (d) in Fig. 1 in a similar fashion. Subjective tests will probably result in different ratings.

2. Scene dependency due to a visibility of an artifact in some image areas compared to other areas. Keelan[24] addresses this as variations in scene susceptibility. Variations in scene susceptibility occur when the same objective amount of an artifact such as noise, streaking, or banding, for example, is present in images, but it is more or less evident in different types of scenes or different areas with the same scene. Keelan shows that the digital artifact of streaking is more evident in clear-sky image areas (i.e., relatively uniform, light areas) than in image areas of high frequency signal and also in

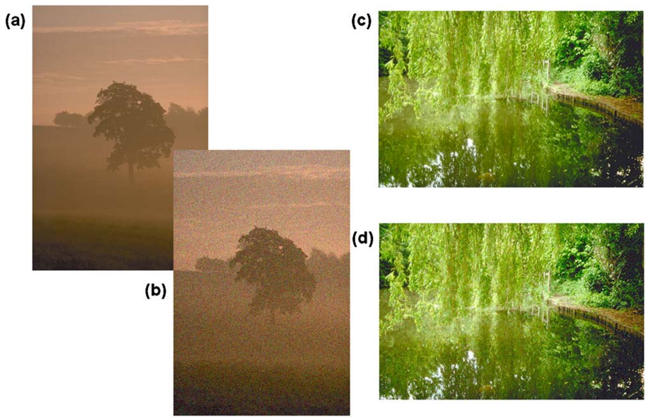

Figure 2. Original (top) and noisy (bottom) versions of two images with different scene content.

extensive dark areas, which visually mask that streaking. Similarly, for a given print granularity, it has been shown that graininess (i.e., subjective measure of photographic granularity) usually decreases with print density[25] and hence dark areas in prints are less visually susceptible to the artifact.

Figure 2 provides a demonstration of scene susceptibility to noise in digital images. Images (b) and (d) are noisy versions of images (a) and (c), respectively. In image (b) most observers would agree that noise reduces significantly the overall quality of the image, since it is largely evident due to low frequency original image content. In image (d) the same amount of uniform noise has been digitally added but, due to the abundance of high frequency information in the original, the noise is hardly visible. Here again, classical distortion and objective quality measures will produce similar results for the two scenes whereas subjective measures will give very different ratings. Device independent image quality models[8] as well as perceived distortion models[7,26] (or fidelity metrics—see overview in Ref. 7) attempt to overcome the problem of scene susceptibility since they utilize weighted information contained in the images themselves rather than the properties of the imaging system. Amongst other drawbacks with perceived distortion models, however, is the fact that they assume that overall perceived distortion is monotonically increasing with perceived error,[27] which is not always in line with subjective impressions of images.

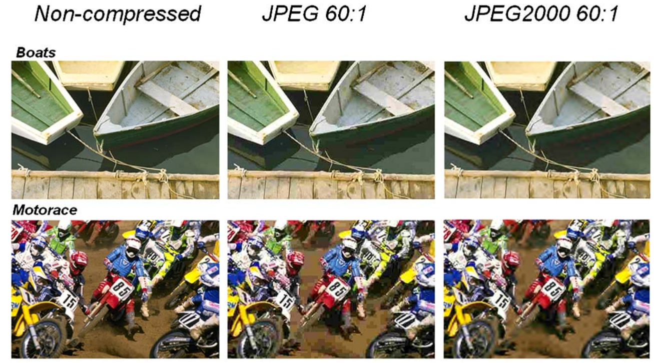

Figure 3. Original and compressed versions of two images with different scene content.

3. Scene dependency (or susceptibility) of digital processes or image processing algorithms, which consequently results in different visual results in different types of scenes. A classical case is that of image compression.[28] Applying the same objective amount of transform based lossy compression (i.e., compression ratio) in two different images, one with mostly high and the other with mostly low frequency information, will discard different quantities of information, since both DCT and DWT compression schemes discard mostly high spatial frequencies. Figure 3 illustrates the results of this effect, where both "*Motorace*" and "*Boats*" images have been compressed at a ratio 60:1. With both JPEG and JPEG 2000 compression schemes, the results of the compression are

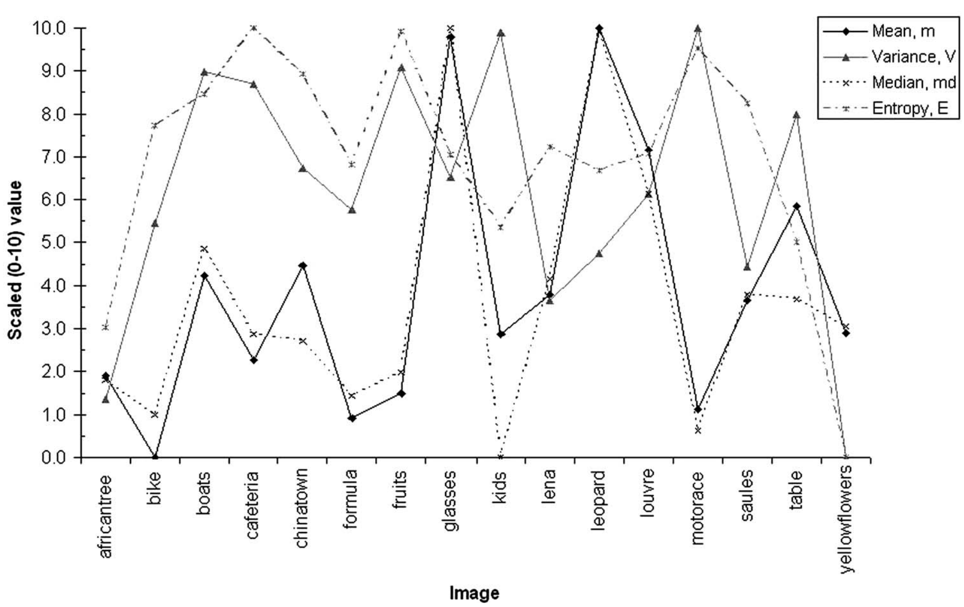

Figure 4. Scaled mean, median, variance, and entropy measures for all 16 original test images.

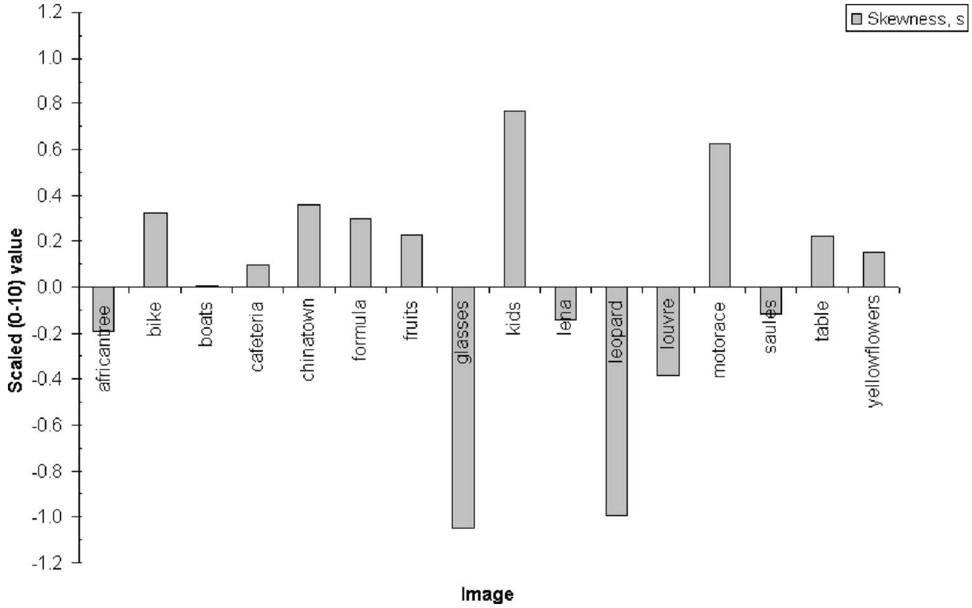

Figure 5. Skewness results for all 16 original test images.

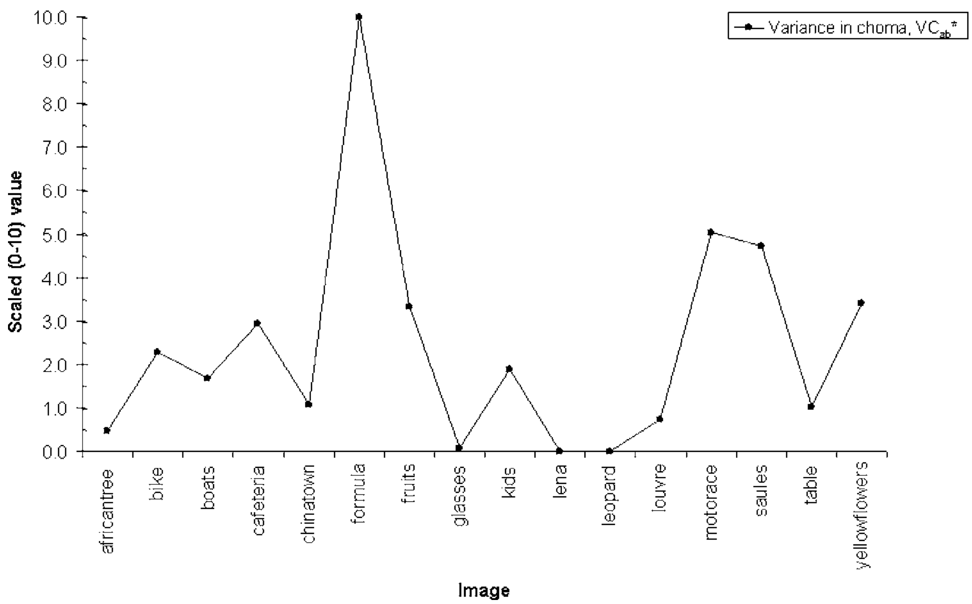

Figure 6. Colorfulness results for all 16 original test images.

more evident in the "*Motorace*" image than in the "*Boats*" image.

Similarly, oversharpening an image with many lines and edges will have more serious visual consequences than oversharpening an image with slow varying information. Essentially, the process of edge sharpening will take place in more pixels in the first image than in the second one. In this last case of scene dependency, both objective and subjective measurements will give different results for the two different scenes.

Independently of the "type" of scene dependency, its main cause (which is the variations in original scene/image properties) as well as the consequences are the same. Scene dependency makes the analysis of results and interlaboratory comparisons problematic. It often biases mean ratings and this is why it is common use to exclude outlying results or the "the odd scenes" in subjective quality measurements. Additionally, the evaluation of objective image quality measures is difficult due to scene dependency in the perceived quality.

There are ways to overcome some of the problems caused by scene dependency. Using a representative set of test stimuli (i.e., well illuminated, all subjects in focus) and excluding atypical stimuli (i.e., such as out of focus subjects, very high or very low contrast scenes) from the set of test stimuli is the most common. Nowadays many experimenters employ the ISO set of test scenes[29] or some commonly used test images. These, however, do not effectively represent the

Table III. Results from all measures applied on the 16 original test scenes and their rank order (from 0 to 16, ranked from lower to higher values) for each measure.

Image	*Mean, m*	*m rank*	*Median, md*	*md rank*	*Variance, V*	*V rank*	*Entopy, E*	*E rank*
African tree	116.06	5	113	5	1296	2	6.77	2
bike	99.62	1	104	3	3002	6	7.50	10
boats	136.34	11	147	13	4474	13	7.62	12
cafeteria	119.33	6	125	8	4353	12	7.86	16
chinatown	138.17	12	123	7	3540	10	7.69	13
formula	107.48	2	109	4	3131	7	7.36	6
fruits	112.57	4	115	6	4514	14	7.84	15
glasses	184.31	15	204	16	3452	9	7.40	7
kids	124.41	7	93	1	4858	15	7.13	4
Lena	132.34	10	139	12	2257	3	7.42	9
leopard	186.19	16	203	15	2715	5	7.34	5
louvre	161.46	14	161	14	3291	8	7.40	8
motorace	109.33	3	100	2	4900	16	7.78	14
saules	131.24	9	135	11	2585	4	7.58	11
table	150.27	13	134	10	4060	11	7.08	3
yellow flowers	124.64	8	127	9	730	1	6.30	1
mean	133.36		133		3323		7.38	
median	127.94		126		3372		7.41	
max	186.19		204		4900		7.86	
min	99.62		93		730		6.30	
stand. dev.	25.78		33		1219		0.41	

Image	*Skewness, s*	*s rank*	*Variance in chroma, VC^*_{ab}*	*VC^*_{ab} rank*	*Busyness, b*	*b rank*	*Log of number of lines, $log_{10}(f)$*	*$log_{10}(f)$ rank*
African tree	−0.190	4	33.24	4	2.38	1	0.00	1
bike	0.323	13	157.15	9	83.65	13	2.40	14
boats	0.003	7	116.57	8	60.29	10	1.54	6
cafeteria	0.094	8	203.33	11	84.60	14	3.96	16
chinatown	0.358	14	74.44	7	65.74	12	2.01	10
formula	0.296	12	691.20	16	32.72	2	2.13	11
fruits	0.226	11	230.49	12	62.20	11	1.08	3
glasses	−1.047	1	5.39	3	44.20	5	1.88	7
kids	0.768	16	130.68	10	40.48	4	1.89	8
Lena	−0.143	5	0.00	1	50.10	8	1.18	4
leopard	−0.994	2	0.00	1	49.30	7	0.00	1
louvre	−0.381	3	50.63	5	52.86	9	1.94	9
motorace	0.627	15	348.89	15	89.25	15	3.17	15
saules	−0.116	6	326.57	14	93.71	16	2.20	12
table	0.223	10	70.44	6	46.25	6	2.27	13
yellow flowers	0.153	9	235.46	13	35.41	3	1.53	5
mean	0.013		167.15		55.82		1.82	
median	0.124		123.63		51.48		1.92	
max	0.768		691.20		93.71		3.96	
min	−1.047		0.00		2.38		0.00	
stand. dev.	0.499		178.61		24.04		1.00	

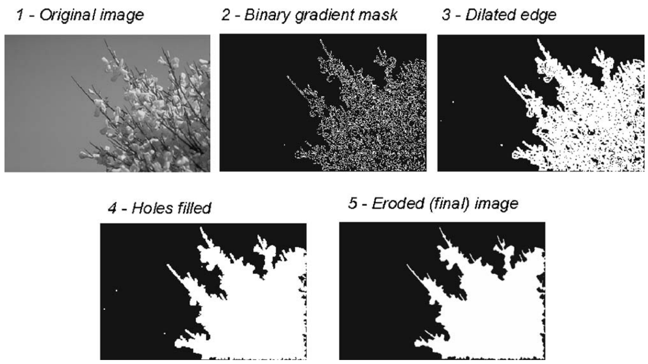

Figure 7. Stages of the image segmentation process used to separate slow varying areas from busier areas of the test image "yellow flowers."

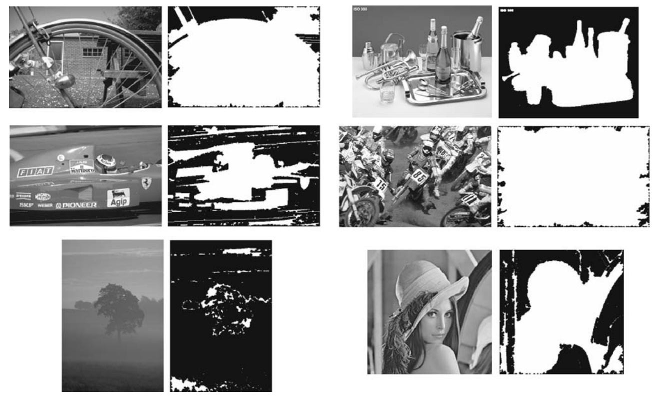

Figure 8. Example of segmented test images with different amount of detail.

range and variety of different scenes that photographers, artists and consumers may wish to record and reproduce with high quality. Furthermore, scenes that deviate in content from the representative set may not be reproduced appropriately, since they are not in accordance with the "average" reproduction derived from image quality results.

Bartleson in his extensive study on psychophysical methods[30] suggested that the test stimuli should be chosen by a clearly defined set of procedures, and proposed five ways for choosing a sample of stimuli: (i) The random independent sample, (ii) the stratified sample, (iii) the contrast sample, (iv) the purposeful sample, and (v) the identical sample. In all cases apart from the randomly chosen samples, some decisions have to be taken from the experimenter on the attributes and content of the test images. A sample set of some kind is generated and the question "How shall we measure the attributes of these?" is asked. Attributes currently are assessed by *inspection* and using the experimenter's intuition and judgment.

SCENE CLASSIFICATION WITH RESPECT TO IMAGE QUALITY MEASUREMENTS

The issue of scene selection and the problem of scene dependency in image quality assessments raise the following questions.[14]

- Are there any quantifiable characteristics/properties/features in scenes that can be used to differentiate one scene from another?
- Can we identify which of these inherent scene characteristics correlate with human criteria used in image quality judgements?
- Can we finally use a weighted set of scene descriptors (or *scene metrics*) to identify, differentiate and group test stimuli a meaningful way that matches the way humans operate when they judge the quality of images?

As early as 1941, Jones et al.[21] suggested that scenes can be classified into relatively homogeneous groups with respect to (a) general illumination characteristics, (b) direc-

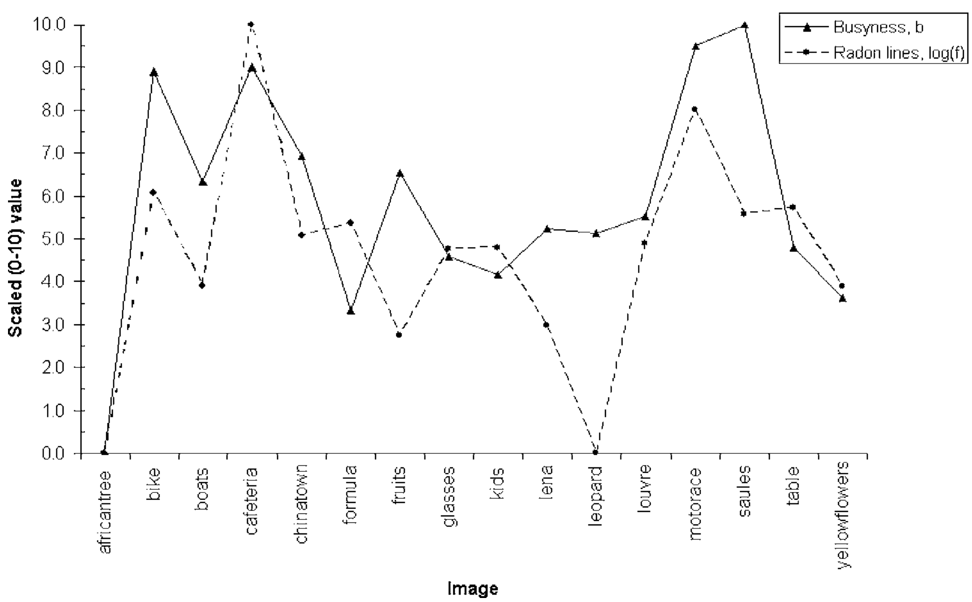

Figure 9. Busyness and log_{10} of the number of lines for all 16 original test images.

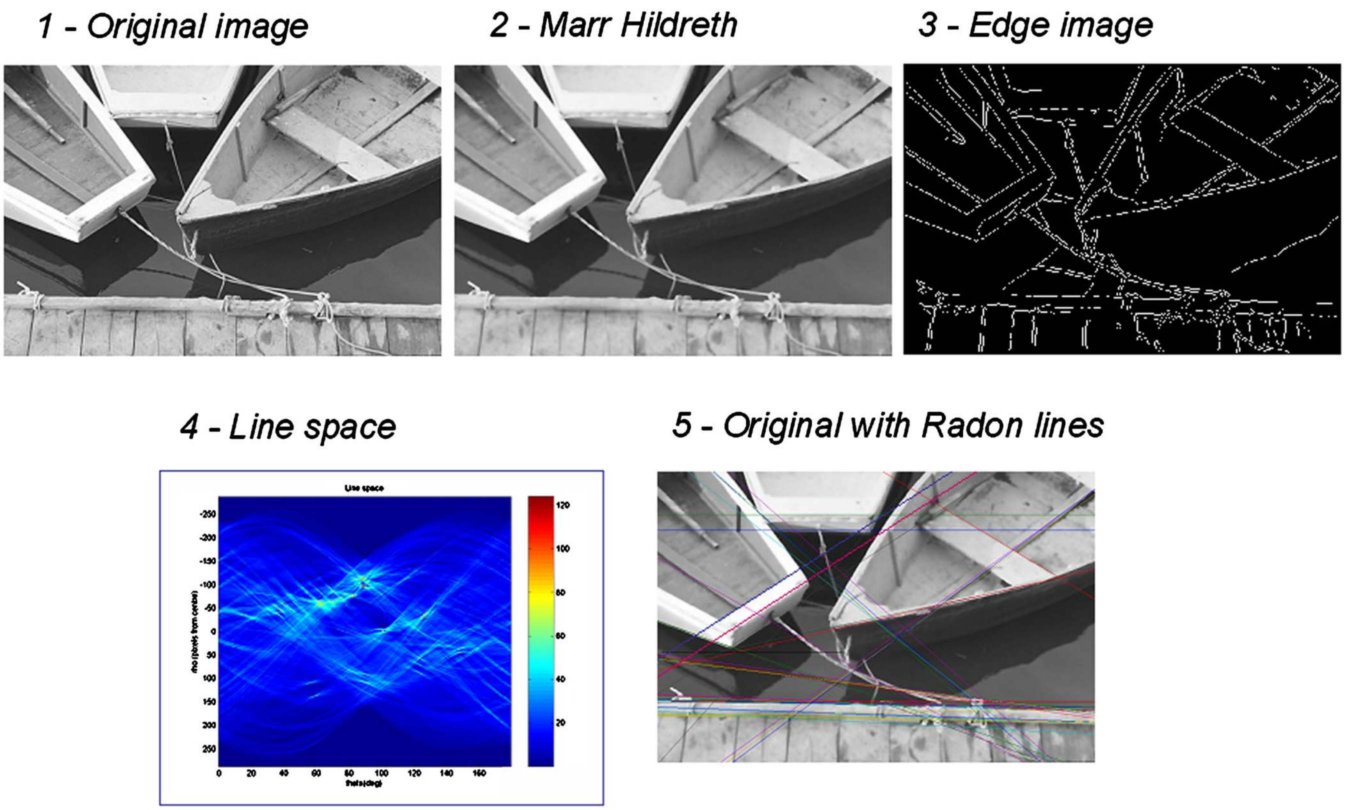

Figure 10. Stages of the line extraction process, used to identify strong lines and edges in the test image *boats*.

tional viewing aspect, and (c) spatial distribution of scene elements and local illumination conditions. This was one of the very first efforts to categorize exterior scenes *by inspection*. Today, with the use of digital imagery we can quantify many aspects of a digitized scene or image. Also, we can be in a position to understand the influence that different scene properties may have on our perception of those scenes, as well as on image quality metrics.

One way to identify scene features that have a significant role in image quality assessments is to look closely at *the visual description of the different image quality attributes* (Table I) as well as *the susceptible image areas in various digital image artifacts* (Table II). By doing this one can observe that a relationship between low-level visual features, such as

- illumination and contrast
- hue and chrominance and color contrast
- lines and edges (hard and soft)
- slow varying areas (low frequency information)
- busy areas with high amount of detail/texture (high frequency information) and human judgments on image quality exists.

Quantification of these features, at global and local (region-of-interest) levels can help identify, differentiate and group test stimuli, along with the quantification of other scene characteristics such as

- camera to subject distance
- spatial distribution of subjects.

Table IV. Spearman correlation coefficients, *r*, of pairs of scaled measures. Negative correlations indicate anticorrelated data. Unsuccessful correlations are indicated with a strikethrough.

	Median, md	*Variance V*	*Entropy E*	*Skewness s*	*Busyness b*	*Lines* $log_{10}(f)$
Mean, m	**0.93**	~~0.05~~	~~0.07~~	**–0.80**	~~0.18~~	~~0.32~~
Median, md		~~0.17~~	~~0.00~~	**–0.94**	~~0.11~~	~~0.34~~
Variance, V			**0.70**	~~0.37~~	~~0.46~~	~~0.51~~
Entropy, E				~~0.07~~	**0.73**	~~0.44~~
Skewness, s					~~0.22~~	~~0.46~~
Busyness, b						**0.66**

Table V. Collective scene metric values and their relative rank order (from 1 to 16–in parenthesis) from *Group 1* of subjective results presented in Ref. 34. The bold values indicate similarity in measures.

Scene Metric	Lena	Glasses	Leopard
m	132.34 (10)	**184.31 (15)**	**186.19 (16)**
md	139 (12)	**204 (16)**	**203 (15)**
s	**−0.143(5)**	**−1.047(1)**	**−0.994(2)**
V	**2257 (3)**	3452 (9)	**2715 (5)**
E	**7.42 (9)**	**7.39 (7)**	7.34 (5)
b	**50.10 (8)**	**44.20 (5)**	**49.30(7)**
log(*f*)	**1.18 (4)**	**1.88 (7)**	0 (1)
VC^*	**0.00 (1)**	**5.39 (2)**	**0.00 (1)**

Table VII. Collective scene metric values and their relative rank order (from 1 to 16—in parenthesis) from *Group 3* of subjective results presented in Ref 34.

Scene Metric	Louvre	Fruits	Bike
m	161.46 (14)	112.57 (4)	99.62 (1)
md	161 (14)	115 (6)	104 (3)
s	−0.381 (3)	0.226 (11)	0.323 (13)
V	3291 (8)	4514 (14)	3002 (11)
E	7.40 (8)	7.84 (15)	7.50 (10)
b	52.86 (9)	62.20 (11)	83.65 (13)
log(*f*)	1.94 (9)	1.08 (3)	2.40 (14)
VC^*	50.63 (5)	230.49 (12)	157.15 (9)

Low-level visual descriptors have been shown to relate to semantic scene/image categories.[31–33] There is a need for reliable and workable low-level feature extraction algorithms that correlate with human assessment of the quantities that are measured.

SCENE ANALYSIS

Experimental work, based on previous work by two of the authors,[13] was carried out to investigate the relationship between scene content and results from subjective assessments on image compression, published in Ref. 34. The objective was to investigate a variety of simple image analysis tools/processes and use them to quantify some of the scene features listed in the preceding section. At this stage only *global image content* was investigated. By applying these processes to the set of original (noncompressed) images used in the subjective evaluation of JPEG and JPEG2000 (Appendix A) we try to establish (a) how global spatial, luminance and chromatic content may affect image compression and (b) whether the selected image analysis tools are reliable measures of the selected scene features.

Image analysis was applied mostly to the CIELAB L^* channel (the L^* channel of each image was converted to an 8-bit gray scale image for analysis), to derive measures relating to image tones and spatial frequency content. The image's C_{ab}^* values were calculated from the CIE a^*, b^* channels (ranged from −127 to 128; this range covers largely the CIELAB a^*, b^* limits of sRGB 8-bit images) and were used to derive a measure relating to image colorfulness.

Statistical Measures

The following first order statistical measures[35] were derived from the probability density functions (PDF) of the L^* channel in MATLAB[36] environment:

- mean value, *m*: relates to the average global intensity in the scene,
- median value, *md*: also relates to the average global intensity in the scene,
- skewness, *s*: a descriptor of the imbalance of the PDF, indicator of average intensity, low or high key scenes,
- variance, *V*: relates to the global scene contrast,
- entropy, *E*: relates to the information content, amount of detail and random changes in the scene; it is computed as

$$E = -\sum_{i=0}^{k-1} P(i)\log_2 P(i), \tag{1}$$

where $P(i)$ is the probability density of the *i*th gray level and *k* the total number of gray levels.

Table VI. Collective scene metric values and their relative rank order (from 1 to 16–in parenthesis) from *Group 2* of subjective results presented in Ref. 34. The bold values indicate similarity in measures.

Scene Metric	Formula	Cafeteria	Motorace	Kids
m	**107.48 (2)**	**119.33 (6)**	**109 (3)**	**124.41 (7)**
md	**109 (4)**	**125 (8)**	**100 (2)**	93 (1)
s	0.296 (12)	0.094 (8)	**0.627 (15)**	**0.768 (16)**
V	3131 (7)	**4353 (12)**	**4900 (16)**	**4858 (15)**
E	7.35 (6)	**7.86 (16)**	**7.78 (14)**	7.12 (4)
b	32.72 (2)	**84.60 (14)**	**89.25 (15)**	40.48 (4)
log(*f*)	2.13 (11)	**3.96 (16)**	**3.17 (15)**	1.89 (8)
VC^*	**691.20 (16)**	203.33 (11)	**348.89 (15)**	130.68 (10)

Table VIII. Collective scene metric values and their relative rank order (from 1 to 16—in parenthesis) from *Group 4* of subjective results presented in Ref 34.

Scene Metric	Yellow Flowers	Saules
m	124.64 (8)	131.24 (9)
md	127 (9)	135 (11)
s	0.153 (9)	−0.116 (6)
V	730 (1)	2585 (4)
E	6.30 (1)	7.58 (11)
b	35.41 (3)	93.71 (16)
log(*f*)	1.53 (5)	2.20 (12)
VC^*	235.46 (13)	326.57 (14)

Table IX. Collective scene metric values and their relative rank order (from 1 to 16—in parenthesis) from *Group 5* of subjective results presented in Ref. 34. The bold values indicate similarity in the measures.

Scene Metric	Boats	Chinatown	Table
m	**136.34 (11)**	**138.17 (12)**	**150.27 (13)**
md	**147 (13)**	123 (7)	**134 (10)**
s	**0.003 (7)**	0.358 (14)	**0.223 (10)**
V	**4474 (13)**	**3540 (7)**	**4060 (11)**
E	**7.62 (12)**	**7.687 (13)**	7.08 (3)
b	**60.29 (10)**	**67.74 (12)**	46.25 (6)
log(*f*)	1.54 (6)	**2.01 (10)**	**2.27 (13)**
VC^*	**116.57 (8)**	**74.44 (7)**	**70.44 (6)**

The variance of the CIELAB C_{ab}^* of the images, VC_{ab}^*, was used as a measure of color contrast or colorfulness. This has been shown to correlate successfully with the perceived image colorfulness.[37]

In Figure 4, *m*, *md*, *V*, and *E* values are plotted for all 16 original test images. They are scaled between 0 and 10 for presentation and comparison purposes. Figure 5 shows separately image skewness, *s*, and Figure 6 the image variance in chroma, VC_{ab}^* (0–10 scaled values).

Results from all statistical operations along with their relative ranked order—ranked from the lower to the higher metric value—are listed in Table III. Note that a negative *s* indicates a PDF imbalance toward high lightness, a positive *s* toward low lightness and *s* close to zero no PDF imbalance.

Image Segmentation

In this paper *busyness* is defined as a scene/image property indicating the presence or absence of detail. A simple image segmentation technique was applied in MATLAB, to separate the slow varying areas from busier areas within the images. This involved.[38]

- the calculation of the gradient image of the CIELAB L^* channel, by applying the Sobel edge detector in both horizontal and vertical orientations, and using a very low threshold of 0.04,
- the dilation of the binary image to amplify the detail, using flat linear structuring elements,
- the use of a flood filling operation to fill the holes in the dilated image,
- the erosion of the binary image to get rid of spurious noise.

Figure 7 illustrates the different stages of this process for one image and Figure 8 six original images and their segmented version.

The common threshold for the calculation of the gradient image was derived empirically, after careful observations and is considered appropriate for the segmentation process of sRGB images, displayed at the given resolution. The value *b*, expressed as a percentage, was finally used as a metric of scene busyness. It was calculated from the ratio of the number of white pixels (that indicated the busy image areas) to the total number of pixels.

Figure 9 illustrates graphically scaled results obtained from this segmentation process and Table III lists the *b* values, along with their rank order.

Line Detection

In an effort to extract and quantify hard edges and strong lines, we used the Marr-Hildreth edge detection[39] and the Radon transformation in MATLAB.[38] The zero crossings were determined using the Canny edge detector, with low and high thresholds set to 0.05 and 0.15, respectively. In the line space (see Figure 10), noise was filtered using a 5×5 Wiener kernel and the smoothened peaks (shown as bright spots in Fig. 10) that corresponded to individual lines, *f*, were identified at a certain peak and above (peak $>=70$).

The problem with Radon transformation is that it produces more than one peak in the accumulation (line) space per line in the image space, since recorded lines and edges are usually noisy. The procedure worked relatively well for most of the images, especially those with an average number of strong lines or with no strong lines at all. However, it was difficult to assess its performance in very busy scenes, were the number of lines found was overexaggerated.

By observing the results, we found that the $\log_{10}$ of the number of lines [$\log_{10}(f)$] resulting from the above line extracting operation was a relatively successful measure of busyness *and* amount of lines within the scenes. Figure 9 illustrates graphically scaled results obtained from this process and column 9 in Table III lists the actual $\log_{10}(f)$ values, along with their rank order.

Correlations

The Spearman correlation coefficient[40] was used to relate any given pair of scaled measures (i.e., scene metric values) derived earlier. The correlation coefficients are presented in Table IV. It can be seen that mean, *m*, and median, *md*, in L^* plane, and the PDF skewness, *s*, all correlate successfully; they all provide information on image global intensities. Entropy, *E*, correlates rather successfully with variance, *V*, in the L^* plane and with the metric for busyness, *b*.

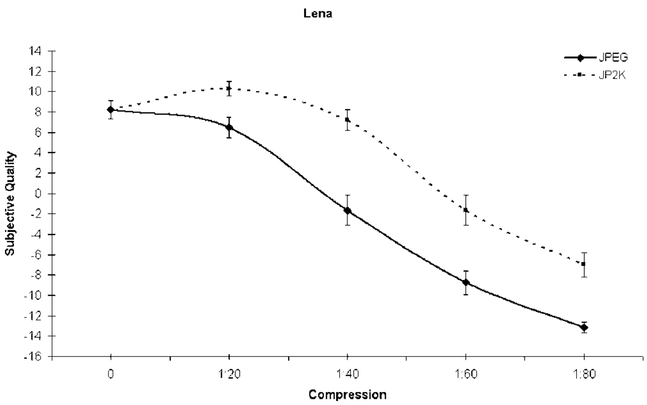

Figure 11. Subjective results from test image "*Lena*" (as presented in Ref. 34).

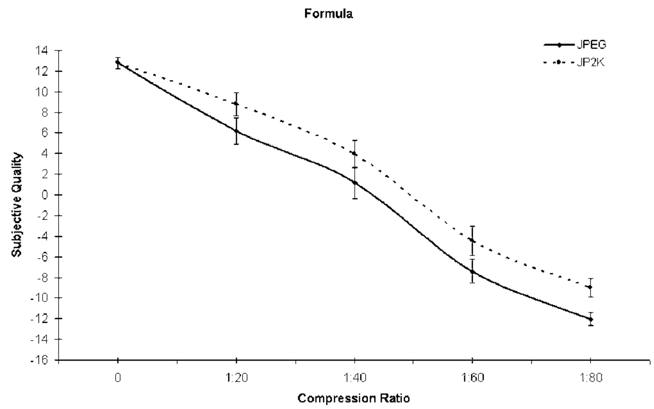

Figure 12. Subjective results from test image "*formula*" (as presented in Ref. 34).

SCENE DEPENDENCY AND SUBJECTIVE RESULTS

In an effort to establish a number of preliminary relationships between original scene content and perceived quality in image compression, the scene metric values produced in the preceding section were examined against groups of scenes shown to produce common subjective rating in Ref. 34.

To discuss the proximity of the scene metric values between the different scenes, and therefore the similarity in their characteristics, we binned these values into four ranges, ordered by relative distance from the average. The median value of each scene metric was used as the average value for that metric. Values found to be within one sigma from the median were considered "average" and were split into "average to low," if below the median and "average to high," if above the median. Another two bins included "extreme" values, i.e. values that were more than one standard deviation away from the median, comprising the "very low" category if more than one sigma below the median, and "very high" if more than one sigma above the median. Further, in Tables V–IX the values in bold belong to a common bin for a specific scene metric (although it is not indicated to which one of the 4 bins) and therefore indicate similarity in the relevant scene characteristic.

Table V tabulates the collective scene metric values from the *Group 1* of subjective curves that includes the scenes "Lena," "glasses," and "leopard."[34] In this group of images JPEG2000 outperformed, in most cases, JPEG at all compression rates. Little or no loss in perceived quality is noticed until 40:1 compression with JPEG2000, as demonstrated for one test image, "Lena," in Figure 11. The most prominent collective features in these scenes are

- Very low variation in chroma, VC^*—these are the three out of the sixteen scenes with no, or very low chrominance information—and
- High global intensity, indicated by average to high mean, m, and median, md, and very low PDF skewness, s.

Otherwise, the scenes are of average busyness and entropy.

Table VI tabulates the collective scene metric values from the *Group 2* of subjective curves, including the scenes "formula," "cafeteria," "motorace," and "kids."[34] In this group, again, JPEG2000 outperformed JPEG at all compression rates, but the difference in the perceived quality between the two compression schemes is less distinct than in Group 1, while the uncompressed image (i.e., the TIFF version) maintains always better quality than any compressed versions. An example is illustrated in Figure 12.

Common scene features in this group include

- Very high variation in chroma, VC^*, or very high variance V, or both.
- Two of the scenes in this group are amongst the busiest scenes in the entire set, with high E, b, and $\log_{10}(f)$.
- Global intensity is average-to-low or very low.

Also, all scenes in this group include lettering (see Appendix A).

In *Groups 3–5* there are no distinct common scene characteristics[34] (see Tables VII–IX respectively). It is perhaps interesting to notice that none of the images of *Group 5*, "*boats*," "*Chinatown*" and "*table*" (where there is virtually no subjective difference in perceived quality between the two different compression schemes at all levels of compression—see example in Figure 13) have "extreme" features (Table IX). The scenes present

- average or average to high global intensity,
- average to high variance,
- average or average-to-high busyness, entropy, and number of lines,
- average to low variation in chroma.

Finally, in one scene, the "*African tree*," JPEG clearly outperformed JPEG2000 over the entire compression range (i.e., up to 80:1 compression rates), while up to 40:1 JPEG compression, the compressed versions of this scene were judged of a higher quality than the uncompressed TIFF,[34] as demonstrated in Figure 14. Table X indicates that "*African tree*" is a scene with

- very low variance, V (i.e., low contrast) and

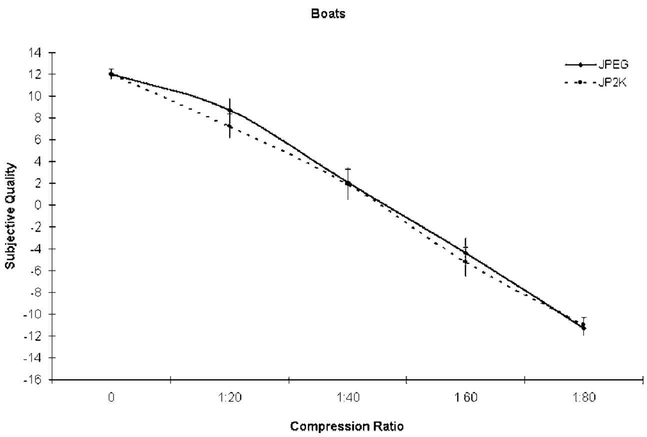

Figure 13. Subjective results from test image "*boats*" (as presented in Ref. 34).

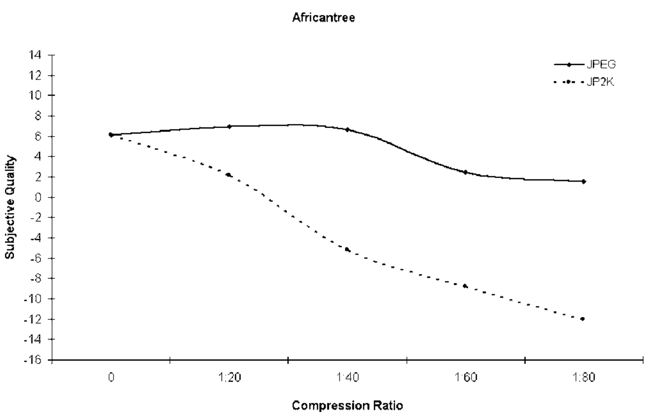

Figure 14. Subjective results from test image "*African tree*" (as presented in Ref. 34).

- average to low global intensity and chromatic information.
- This is the least busy scene in the entire set, with very low entropy, E, the lowest busyness, b, and zero number of lines, $\log_{10}(f)$.

Overall, the general observations drawn from the quantification of scene features match fairly successfully visual observations of the scenes provided in Ref. 34, which also provides explanations on how these features affect the perceptibility of related artifacts. However, the small number of test-scenes and the unaccounted for Human Visual System and display scene metrics presented here do not allow for concrete conclusions that relate scene content and image compression. The following overall trends can be identified from this study: JPEG2000 clearly performs better than JPEG in high global intensity, low chroma scenes, with average contrast and average busyness. JPEG2000 performance is not as good, but is still better than JPEG in very busy scenes, of high contrast and/or high color contrast with average global intensity. JPEG, on the other hand, far outperforms JPEG2000 (and in some cases the TIFF original at relatively low compression rates) in a very low contrast, rather low global intensity and low chroma test scene, a scene of distinctly low busyness.

Table X. Collective scene metric values and their relative rank order (from 1 to 16—in parenthesis) from the scene "*African tree*."

Image Metric	African Tree
m	116.06 (5)
md	113 (5)
s	−0.190 (4)
V	1296 (2)
E	6.77 (2)
b	2.38 (1)
$log(f)$	0 (1)
VC^*	33.24 (4)

DISCUSSION

We have described in detail the long-standing issue of scene dependency related to image quality assessments. The identification, selection and classification of scenes/images are unresolved challenges, not only within the imaging science community and to-date there has not been a universal way for scene selection and classification. Many image classifiers are far too complicated and insufficiently comprehensive, and are often employed mostly by the teams who introduce them, whereas others that are commercialized are sold very expensively. It is also important to notice that, often different image classifiers serve different purposes.

In this paper, we propose image analysis for tackling the problem of scene selection and classification with respect to image quality measurements. To start with, our approach is simple but the preliminary work carried out here indicates that there is potential to extract quantitatively features in images using simple image analysis tools, such as image statistics, image segmentation, edge detection etc. in order to develop selected image feature metrics for each of these properties.

The successful metrics can be combined and weighted appropriately to produce a means for selecting and grouping test scenes/images in image quality assessments. In the third section, we proposed an initial list of image features that play a significant role when the quality of images is judged.

Further work on the topic should include identification of successful metrics that best describe the selected image properties/features. This can be succeeded by seeking correlations between human assessments and derived metrics values for each selected property (e.g., human judgments of scene business, colorfulness, amount of strong/weak lines, etc., can be put against respective metric values to identify best correlations). The successful image metrics should be probably weighted according to location and visual importance.[41]

It is also important that an appropriate space is used to describe the digital image values. In this work we use the CIELAB space, which is limited, since it does not take into account color appearance issues[42] that are present in complex scenes. Using a color appearance model, such as

CIECAM02,[43] to describe the image in appearance coordinates is a better option.

Once the image is transformed to color appearance coordinates, preprocessing of the digital image values prior to image analysis should accommodate for the human visual system[44] and for the display medium. This task can be modular in a similar fashion to the prefiltering that takes place in the iCAM image quality framework, introduced by Johnson and Fairchild.[8] It needs to include processes such as spatial prefiltering to accommodate for the human contrast sensitivity function, local contrast detection, and spatial localization as well as the modulation transfer function of the display/print medium.

The above tasks are labor intensive but they are comprehensive and inclusive. Once individual procedures for image data calibration are set and metrics for individual scene properties are selected and combined, they can be standardized so that they provide a universal mean for scene classification with respect to image quality measurements.

(Appendix available as Supplemental Material on the IS&T website, www.imaging.org)

REFERENCES

[1] R. E. Jacobson, "An evaluation of image quality metrics", J. Photogr. Sci. **43**, 7 (1995).

[2] E. J. Farrell, "Image quality evaluation", in *Colour Imaging: Vision and Technology*, edited by L. W. MacDonald and M. R. Luo (John Wiley and Sons, London, UK, 1999), Chap. 15, p. 285.

[3] P. G. Engeldrum, *Psychometric Scaling* (Imcotek Press, USA, 2000).

[4] S. Triantaphillidou, Ph.D. thesis, University of Westminster, London, 2001.

[5] C. L. Bartleson, *Optical Radiation Measurements, Volume 5, Visual Measurements Part II* (Academic Press, New York, 1984), Chaps. 7–9, p. 367.

[6] B. W. Keelan, *Handbook of Image Quality* (Marcel Dekker Inc., New York, 2002).

[7] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation", in *Handbook of Image and Video Processing*, edited by A. Bovik, (Elsevier, Amsterdam, 2005), Chap. 8.2, p. 939.

[8] M. D. Fairchild and G. M. Johnson, "The iCAM framework for image appearance, image differences, and image quality", J. Electron. Imaging **13**, 126 (2004).

[9] C. J. Bartleson, "The combined influence of sharpness and graininess on the quality of colour prints", J. Photogr. Sci. **30**, 33 (1982).

[10] G. C. Higgins and N. R. Wolfe, "The relation of definition to sharpness and resolving power", J. Opt. Soc. Am. **45**, 121 (1955).

[11] A. M. Ford, Ph.D. thesis, University of Westminster, London, 1997.

[12] P. G. Engeldrum, Ref. 4, p. 25.

[13] R. E. Jacobson and S. Triantaphillidou, "Metric approaches to image quality", in *Colour Image Science: Exploiting Digital Media*, edited by L. W. MacDonald and M. R. Luo (John Wiley and Sons, London, UK, 2002), Chap. 18, p. 371.

[14] A. M. Ford, "Determination of compressed image quality", in *Colour Imaging: Vision and Technology*, edited by L. W. MacDonald and M. R. Luo (John Wiley and Sons, London, UK, 1999), Chap. 16, p. 315.

[15] S. Triantaphillidou, R. E. Jacobson, and A. M. Ford, *Proc. ICPS-98, The International Conference of Imaging Science*, Vol. 2 (KVCV, Leuwer, Belgium, 1998) p. 204.

[16] M. Klima, P. Pata, K. Fliegel, and P. Hanzlik, *Proc. IEEE, CCST'05 on Security Technology* (IEEE Press, Piscataway, NJ, 2005) p. 19.

[17] H. Frieser and K. Biedermann, "Experiments on image quality in relation to the modulation transfer function and graininess of photographs", Photograph. Sci. Eng. **7**, 28 (1963).

[18] C. J. Bartleson and C. P. Bray, "On the preferred reproduction of flesh, blue-sky and green-grass colors", Photograph. Sci. Eng. **6**, 9 (1962).

[19] R. W. G. Hunt, I. T. Pitt, and L. M. Winter, "The preferred reproduction of blue sky, green grass and Caucasian skin in colour photography", J. Photogr. Sci. **22**, 144 (1974).

[20] L. A. Jones and H. R. Condit, "The Brightness scale of exterior scenes and the computation of correct photographic exposure", J. Opt. Soc. Am. **31**, 651 (1941).

[21] G. P. Corey, M. J. Clayton, and K. N. Cuprey, "Scene dependence of image quality", Photograph. Sci. Eng. **27**, 9–13 (1982).

[22] E. M. Biederman, Photogr. Korresp. **103**, 25, 41 (1967).

[23] E. M. Granger and K. N. Cuprey, "An optical merit function (SQF), which correlates with subjective image judgments", Photograph. Sci. Eng. **16**, 221 (1972).

[24] B. W. Keelan, *Handbook of Image Quality* (Marcel Dekker Inc., New York, 2002), Chap. 10.

[25] R. W. G. Hunt, *The Reproduction of Colour*, 6th ed. (John Wiley and Sons, New York, 2004), p. 310.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", IEEE Trans. Image Process. **13**, 600 (2004).

[27] Z. Wang, A. C. Bovik, and L. Lu, *Proc: IEEE, ICASSP'02* (IEEE Press, Piscataway, NJ, 2002) p. 3313.

[28] R. Ansari, C. Guillemot, and N. Memon, "Lossy image compression: JPEG and JPEG2000 standards", in *Handbook of Image and Video Processing*, edited by A. Bovik (Elsevier, Amsterdam, 2005), Chap. 5.5, p. 709.

[29] ISO 12640:1997 Graphic Technology: Prepress digital data exchange—CMYK standard color image data (CMYK/SCID), www.iso.org

[30] C. L. Bartleson, *Optical Radiation Measurements, Volume 5, Visual Measurements Part II* (Academic Press, New York, 1984), Chap. 8, p. 443.

[31] R. Schettini, C. Brambilla, A. Valsasna, and M. De Ponti, *Proc. IS&T/SID 9th Color Imaging Conference* (IS&T, Springfield, VA, 2001) p. 35.

[32] A. Mojsilovic and B. E. Rogowitz, "Semantic metric for image library exploration", IEEE Trans. Multimedia **6**, 828 (2004).

[33] D. Depalov, T. Pappas, D. Li, and B. Gandhi, *Proc. IEEE, ICASSP'06*, Vol. 2 (IEEE Press, Piscataway, NJ, 2006) p. 417.

[34] E. Allen, S. Triantaphillidou, and R. E. Jacobson, "Image quality comparison between JPEG and JPEG2000 I-Psychophysical investigation", J. Imaging Sci. Technol. **51**, 248 (2007).

[35] W. K. Pratt, *Digital Image Processing*, 2nd ed. (John Wiley and Sons, New York, 1991).

[36] MATLAB The Language of Technical Computing, V. 6.1., Mathworks, Cambridge, UK.

[37] S. Triantaphillidou, conference presentation, *Digital Futures 2004* (2004), www.isg-rps.org

[38] S. Triantaphillidou, 2006, MATLAB code provided at: http://www.wmin.ac.uk/ITRG/sophie/mfiles_Sanalysis

[39] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing using MATLAB* (Pearson, Prentice Hall, Englewood Cliffs, NJ, 2004).

[40] S. S. Cohen, *Practical Statistics* (Edward Arnold, London, 1988), p. 81.

[41] R. Schettini, "Indexing and retrieval in colour image databases", in *Colour Image Science: Exploiting Digital Media*, edited by L. W. MacDonald and M. R. Luo (John Wiley and Sons, London, UK, 2002), Chap. 10, p. 183.

[42] M. D. Fairchild, *Color Appearance Models*, 2nd ed. (John Wiley and Sons, New York, 2004).

[43] N. Moroney, M. D. Fairchild, R. W. G. Hunt, C. Li, R. M. Luo, and T. Newmann, *The CIECAM02 Color Appearance Model*, *Proc. IS&T/SID 10th Color Imaging Conference* (IS&T, Springfield, VA, 2004) p. 2.

[44] L. K. Cormack, "Computational models of early vision", in *Handbook of Image and Video Processing*, edited by A. Bovik (Elsevier, Amsterdam 2005), Chap. 4.1, p. 325.