Influence of affective image content on subjective quality assessment

Ian van der Linde1,2* and Rachel M. Doe3

1Department of Computing and Technology, Anglia Ruskin University, East Road, Cambridge CB1 1PT, UK
2Vision & Eye Research Unit (VERU), Postgraduate Medical Institute, Anglia Ruskin University, East Road, Cambridge CB1 1PT, UK
*Corresponding author: i.v.d.linde@anglia.ac.uk

Received June 4, 2012; revised July 26, 2012; accepted August 6, 2012; posted August 7, 2012 (Doc. ID 169787); published August 27, 2012

Image quality assessment (IQA) enables distortions introduced into an image (e.g., through lossy compression or broadcast) to be measured and evaluated for severity. It is unclear to what degree affective image content may influence this process. In this study, participants (n = 25) were found to be unable to disentangle affective image content from objective image quality in a standard IQA procedure (single stimulus numerical categorical scale). We propose that this issue is worthy of consideration, particularly in single stimulus IQA techniques, in which a small number of handpicked images, not necessarily representative of the gamut of affect seen in true broadcasting and unrated for affective content, serve as stimuli. © 2012 Optical Society of America

1. INTRODUCTION

The desire for images and video of high visual quality that can be compactly stored and efficiently transmitted has fuelled the development of lossy compression algorithms that selectively jettison information purported to be of low perceptual importance to human observers. To evaluate the effectiveness of such algorithms, along with algorithms used in applications such as image restoration, enhancement, and watermarking, and to evaluate transmission, display and broadcast systems, both objective and subjective image quality assessment (IQA) and video quality assessment (VQA) procedures are used. For a review of modern techniques for images and video, see [1,2], respectively. Subjective quality assessment methods require that human participants rate individual images or video sequences in a presentation series, and are consequently both expensive and time consuming to administer and cannot be applied in real time to make automatic output-contingent adjustments [3,4]. Two broad approaches to objective quality assessment exist: those that measure signal fidelity (typically relative to a reference image) by statistical means and those that attempt to measure perceptually relevant image properties, sometimes referred to as perceptual visual quality metrics (PVQMs) [3]. Fidelity-based objective quality assessment methods typically rely upon relatively unsophisticated numerical metrics, such as mean absolute error (MAE), mean square error (MSE), peak signal to noise ratio (PSNR), and linear correlation coefficient (LOC), among others [5], and are fast and easy to compute [6], but since they often correlate poorly with human responses [4,7] they are of limited utility where the ultimate receiver is the human visual system (HVS). More specifically, not all numerically equivalent image degradations are equally noticeable [3], and not all image regions enjoy equal attention [8]. Conversely, PVQMs may employ a model of, or derive inspiration from, the sensory computations performed in the early HVS [9,10] or utilize psychophysically derived knowledge of visual performance. Considerable research effort has been invested in the development of PVQMs that accurately reflect the responses that would be elicited by typical human observers [11]. Prominent examples include the universal quality index [12], the noise quality measure [13], the structural similarity index [14], and the visual information fidelity metric [4]. One recent study takes the novel step of focusing automated image quality assessment at those regions most likely to be fixated by human observers, leading to an increased correlation with human ratings [15]. However, despite efforts to develop automated objective quality assessment algorithms that accurately predict human ratings, subjective image quality assessment is still the only truly reliable approach [16], perhaps because simplified HVS performance models do not adequately capture the subtle and multifaceted rating criteria that human observers actually employ.

Image databases, such as LIVE [17], IVC [18], CSIQ [19], TID [20,21], MIC/Toyama [22], and WIQ [23], are in frequent use in IQA research; crucially, they support the comparison of results across laboratories. Table 1 shows the size and constitution of these databases. Each provides a number of lossless reference images (not always exclusive to each database; see below), along with a set of degraded versions of each image, typically distorted to different degrees with a range of different distortion methods (e.g., representing the effect of network packet loss or the artifacts introduced by extant compression algorithms). Many databases incorporate baseline subjective (human rating) data that may be compared with automated IQA techniques.

The question of how the images featuring in standard image databases were selected is rarely subject to scrutiny. All databases listed above contain images at a conspicuously low spatial resolution; an important justification for this is that several core applications (such as IQA and the study of HVS properties) require that images are presented to observers...
on a monitor without algorithm-dependent resampling, i.e., are shown at their native resolution on a display device that has a corresponding spatial resolution, imposing a limit on the maximum useful pixel density. Another common characteristic of all databases is the apparently arbitrary (or at least scantily documented) protocol for the selection of images. For CSIQ [19], not much is said about the selection of the 30 original images, except that they are divided into five categories: animals, landscape, people, plants, and urban. The same can be said about the images selected in the IVC database [18], although many of these are standard test images in widespread use by the image processing community. The images in the LIVE database [17] were selected to promote diverse image content, and most originate from the Kodak Lossless True Color Image Suite [24].

Specifically, images include pictures of faces, people, animals, close-up shots, wide-angle shots, natural scenes, man-made objects, images with distinct foreground/background configurations, and also images without any specific object of interest. Almost all images featuring in the MICT/Toyama database originate from the Kodak suite, and all bar three also feature in the LIVE database. Reference images used in TID2008 are obtained by cropping images from the Kodak Lossless True Color Image Suite, although again, the selection procedure is largely undocumented. Since almost all common databases contain a significant proportion of images taken from the Kodak Lossless True Color Image Suite, it is reasonable to suggest that these pictures have borne a significant influence on the IQA field. Conversely, it is both intuitive and supported by empirical evidence to suggest that reliable results may be obtained by using image databases that are large and diverse [25].

In addition to low spatial resolution, the application of a relatively modest range of distortion types, and the tendency to derive distorted test stimuli from a small number of source images, it is also likely that the content of each of the images featuring in standard image databases will affect quality assessment performance. Although some researchers have acknowledged the impact of image content in image and video quality assessment, “content” is typically interpreted as the visual/statistical composition of an image, viz., the magnitude of motion, range of textures, degree of homogeneity, whether predominated by natural or man-made objects, etc. [26], rather than semantic or affective properties of the scene. Though it has been recognized that human raters’ emotional state may affect their subjective judgment [3], and objective methods to evaluate aesthetic quality have recently been proposed [27], little work has been done to quantify the impact of image-evoked aesthetic content and emotion on subjective quality assessment [28]. It is likely that a portion of the variance in image and video quality assessment may be attributed not to the purely mechanical properties of the pixels featuring in the presented scene, such as the presence of noticeable artifacts, but the degree of image “appreciability,” something that is inherently difficult to quantify and even more difficult to compute automatically. A study that begins to address this issue found that content “desirability” (rather than affect) positively biases quality assessment judgments for video (VQA) and that content enjoyment can reduce the importance of video quality [29]. In this article, we examine the potential impact of affective content on a typical image quality assessment procedure. Image affect may be controlled experimentally by the use of a second class of image database, in which constituent images are rated for the degree to which they are likely to elicit an emotional reaction [typically along two dimensions: valence (the degree of pleasantness) and arousal (the intensity of response or “urgency”)]. Image affect can be obtained by capturing behavioral ratings from human observers but has also been measured by unconscious physiological responses, such as elevated electrodermal activity and heart rate, specific patterns of facial electromyographic (EMG) activity [30], and scalp-recorded event-related potentials [31], and has been correlated with activity at specific neural loci using functional magnetic resonance imaging [32]. Affective image content is also known to influence the way we examine images, changing eye movement parameters such as fixation durations and the degree to which fixations are clustered [33]. Interestingly, the execution of repetitive eye movements is also thought to reduce the impact of affective imagery and autobiographical memories [34]. The most commonly known affective image database is IAPS [35], though its ubiquitous nature (e.g., in university psychology experiments) over many years has meant that, in some departments, these images are more familiar than they once were, possibly attenuating their impact. For this reason, in the present study, the less well-known GAPED database [36] is used in the context of a single stimulus numerical categorical IQA experiment designed to test the influence of affective image content on subjective quality assessment.

2. METHOD

A. Participants

Twenty-five unpaid naïve observers with self-reported normal or corrected-to-normal color vision (18 men, 7 women) participated in the study (mean age 27 years, standard deviation 10.26 years). Participants comprised an opportunity sample of university students and staff not experienced in image quality evaluation, and not studying or employed in any area related to image processing, computer graphics, or vision. Participants were treated in accordance with applicable ethics guidelines.

B. Stimuli

One hundred of the 730 affective color images provided in the Geneva Affective PictureE Database (GAPED) served as

| Table 1. Constitution of Common Databases Used in Image Quality Assessment |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Name            | LIVE            | MICT            | CSIQ            | IVC             | TID             | WIQ             |
| Spatial resolution | 768 × 512      | 768 × 512      | 512 × 512       | 512 × 512       | 512 × 384       | 512 × 512       |
| Source images   | 29              | 14              | 30              | 10              | 25              | 7               |
| Distorted images| 779             | 168             | 866             | 235             | 1700            | 80              |
| Distortion types| 5               | 2               | 5               | 4               | 17              | 1               |
| Observers       | 20–29           | 16              | 25              | 15              | 838             | 30              |
stimuli. GAPED images span two affective dimensions: valence (pleasantness, ranging from very unpleasant to very pleasant), and arousal (intensity of emotion, from calm and soothing to exciting/agitating), each with values from 0 to 100, determined experimentally by averaging 60 independent human ratings. [36]. Our subset of 100 images (to be referred to as source pictures) comprised, where possible, an even number from each bin (of width and height 10) along a two-dimensional arousal/valence space, such that only 4:3 aspect ratio images at the highest available spatial resolution (640 × 480), and that did not already possess noticeable compression artifacts or other visual defects, were selected. Of the six GAPED stimulus categories, our subset comprised 4 images of spiders, 23 images of snakes, 28 “positive” images, 19 “neutral” images, 21 images of “human concerns,” and 4 images of distressed animals. The valence/arousal constitution of the GAPED database (organized by category) and the images selected for this study are shown in Fig. 1. A strong relationship between valence and arousal exists, exemplified in the GAPED database by a significant bivariate correlation \( r = 0.85 \) (728), \( p \) (two-tailed) < .01. This correlation is attenuated in the subset of 100 images selected \( r = 0.70 \) (98), \( p \) (two-tailed) < .01, due to both a smaller sample size and the selection of images that are as widely dispersed as possible in two-dimensional arousal/valence space.

For each of the 100 GAPED images (source pictures) selected, four JPEG [37] compressed versions were generated using GIMP 2.6 (Free Software Foundation, Boston, Massachusetts), yielding five versions of each image (500 images in total). Where \( q_5 \) is the reference image (high quality, unprocessed), the GIMP JPEG quality setting (which ranges from 0 to 100) for images at quality levels \( q_4 \) to \( q_1 \) was 30, 20, 15 and 10, respectively. The JPEG setting of each image is to be referred to as its objective quality setting. As compression rate is increased, artifacts such as ringing, contouring, posterizing, staircase noise, and blocking become readily apparent. An example image at each quality setting is shown in Fig. 2. Since the original spatial resolution of the source pictures was 640 × 480, images were divisible by the JPEG macroblock size of 8 × 8, ensuring that truncated blocks were not created.

C. Apparatus
Stimuli were displayed on a 13-inch MacBook Pro (Apple Corp., Cupertino, California) set at a spatial resolution of 640 × 480 pixels and a refresh rate of 60 Hz. This spatial resolution ensured that stimuli spanned the display screen but did not require interpolative rescaling from their native resolution. The monitor was positioned at 20–30 cm from each participant, adjusted for personal comfort. A quiet, well-lit room was used; ambient illumination was held constant between trials and participants.

D. Procedure
Each participant completed a sequence of 100 no-reference (NR) single stimulus (SS) quality assessment trials. In each trial, first, a single, full-screen stimulus image of random objective quality (from \( q_1 \) to \( q_5 \)) was displayed for a duration of 1950 J. Opt. Soc. Am. A / Vol. 29, No. 9 / September 2012 I. van der Linde and R. Doe

---

Fig. 1. Valence and arousal ratings of 730 GAPED images (red dots, spiders; green dots, snakes; blue dots, positive; cyan dots, neutral; magenta dots, human concerns; gray dots, distressed animals; black circles, 100 selected images). Colored lines represent the convex hull of each image category.

Fig. 2. (Color online) Example image at each objective quality setting \( q_5 \)–\( q_1 \).

3000 ms. Next, a full-screen uniform white noise image was shown (textured mid-gray in appearance, sometimes referred to as a post-exposure field), also for 3000 ms, helping to eliminate afterimages relating to the previous trial and to clearly demarcate each image from its successor. A numerical categorical judgment method [the single stimulus numerical categorical scale (SSNCS)], rather than a rating system that includes explicit labels, such as the ITU-R ACR (absolute category rating) scale of bad, poor, fair, good, and excellent, is the preferred method of assessment where no reference image is presented [7,38], and where the granularity of perceptual differences are anticipated to be small [39]. Participants stated a numerical subjective quality assessment (sometimes referred to as a vote) for each image on an integer scale of 1 (bad) to 10 (excellent) verbally, which was recorded by the experimenter; in essence, a 10AFC psychophysics procedure [40], with greater discriminative resolution than the more common ITU-5 point scale. Image quality assessments were accepted during the post-exposure field that followed each stimulus. The basic experimental procedure is illustrated in Fig. 2. Prior to data collection, participants were instructed (both verbally and in writing) that their task was to rate image quality, and it was emphasized that the aesthetic quality and affective content of the images was not relevant. Participants were forewarned that they may find some of the images presented to be disagreeable. Participants were walked through a training sequence, observing example images (not used in the main experiment) at very high and very low objective quality settings, for which visual artifacts were highlighted; the number of objective quality settings for each source pictures (here 100), and – denotes by the variables under analysis (valence and arousal),


E. Design and Data Screening

To confirm participants’ ability to discriminate between images compressed to different degrees (objective quality setting), the Pearson product moment correlation coefficient between subjective IQA score (1–10) and objective quality setting (q1–q5) is calculated. To test whether participant image quality ratings are influenced by the affective content of each image, the Pearson product moment correlation coefficient between mean subjective IQA score for each image (1–10) and the valence and arousal scores (1–100) for the corresponding image is calculated.

Coherence analysis of raw subjective scores to eliminate (ostensibly) spurious data from participants that did not rate images consistently (e.g., due to inattention) is not required, in agreement with current procedure (recommendation ITU-R BT.500-13 [41]), since the number of participants is ≥20. Furthermore, data cleaning may eliminate eccentric ratings caused by the variables under analysis (valence and arousal), since, just as aesthetic judgments evoke strong individual differences, aversions to specific images may be inherently idiosyncratic [42]. However, for comparison, the full screening procedure (outlined below) is run to establish the number of participants that would have been rejected had full screening been applied.

Let K denote the number of source pictures (here 100), and J denote the number of objective quality settings for each source picture (here 5, yielding 500 stimulus images in total). Each of N participants (here 25), denoted i, rates L stimulus images (here 100, the length of each of the five sequences, each containing all K source pictures). Where uijk denotes a single rating issued by a specific participant for a specific source picture and objective quality setting (collectively forming a 100 × 5 × 25 sparse matrix containing 2500 ratings, since participants rated each source picture at one objective quality setting only), the mean subjective quality rating for each source picture presented at a particular objective quality setting collapsed across participants (i.e., the MOS) is given by

\[
\bar{a}_{jk} = \frac{1}{N} \sum_{i=1}^{N} u_{ijk},
\]

yielding 500 values. The mean subjective quality rating for each source picture collapsed across participants and objective quality settings (collectively forming a 100 × 5 × 25 sparse matrix containing 2500 ratings, since participants rated each source picture at one objective quality setting only), the mean subjective quality rating for each source picture presented at a particular objective quality setting collapsed across participants (i.e., the MOS) is given by

\[
\bar{a}_k = \frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{N} u_{ijk},
\]

yielding 100 values, and the mean subjective quality rating for each presentation condition collapsed across source pictures and participants is

\[
\bar{a} = \frac{1}{NK} \sum_{j=1}^{J} \sum_{k=1}^{K} \bar{a}_{jk},
\]

yielding 5 values. In keeping with ITU-R BT.500-13, all means are to be reported with 95% confidence intervals (to be denoted CI(95%), e.g., for each of the 500 mean subjective quality ratings for each source image at a particular objective quality setting collapsed across participants,

<table>
<thead>
<tr>
<th>Trial Number</th>
<th>Computer Display</th>
<th>Time (s)</th>
<th>Participant Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td><img src="image1" alt="Study Image" /></td>
<td>0</td>
<td>Study Image</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Rate Image</td>
</tr>
<tr>
<td>002</td>
<td><img src="image2" alt="Study Noise" /></td>
<td>6</td>
<td>Study Noise</td>
</tr>
<tr>
<td></td>
<td><img src="image3" alt="Study Image" /></td>
<td>9</td>
<td>Rate Image</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Study Noise</td>
</tr>
<tr>
<td>588</td>
<td><img src="image4" alt="Study Image" /></td>
<td></td>
<td>Study Image</td>
</tr>
<tr>
<td></td>
<td></td>
<td>594</td>
<td>Rate Image</td>
</tr>
<tr>
<td>100</td>
<td><img src="image5" alt="Study Noise" /></td>
<td></td>
<td>Study Noise</td>
</tr>
</tbody>
</table>

Fig. 3. (Color online) Schematic of experimental procedure.
\[ \tilde{u}_{jk} \pm 1.96 \frac{S_e}{\sqrt{N}}, \text{ where } N = 25 \text{ and } S \text{ denotes the standard deviation of the distribution of ratings averaged to generate each cell (here, 5).} \]

The distribution of the five ratings in each of the 500 image-quality cells is subject to a \( \beta_2 \) kurtosis test, wherein normality is assumed if \( 2 \leq \beta_2 \leq 4 \), such that \( \beta_2 = \frac{\sum_{i=1}^{N} (u_{i}-\tilde{u})^2}{\sum_{i=1}^{N} (u_{i}-\bar{u})^2} \); i.e., the ratio of the fourth order moment to the second order moment squared. Next, we calculate, for each participant, \( i \) two integer values, \( P_i \) and \( Q_i \), using the following algorithm, which bifurcates trial-to-trial according to the outcome of the preceding \( \beta_2 \) test for each stimulus image (source picture at a given objective quality setting), incrementing each time the present participant’s rating is eccentric to the distribution of ratings for all participants.

IF \( 2 \leq \beta_2 \leq 4 \) THEN
   IF \( u_{ijk} \geq \tilde{u}_{jk} + 2S_{jk} \) THEN
      \( P_i = P_i + 1 \)
   ELSE IF \( u_{ijk} \leq \tilde{u}_{jk} - 2S_{jk} \) THEN
      \( Q_i = Q_i + 1 \)
   END
ELSE
   IF \( u_{ijk} \geq \tilde{u}_{jk} + \sqrt{20S_{jk}} \) THEN
      \( P_i = P_i + 1 \)
   ELSE IF \( u_{ijk} \leq \tilde{u}_{jk} - \sqrt{20S_{jk}} \) THEN
      \( Q_i = Q_i + 1 \)
   END
END

Finally, participant \( i \) is flagged as having responded inconsistently if
\[
\left[ \frac{P_i + Q_i}{L} > .05 \right] \land \left[ \frac{(P_i - Q_i)}{(P_i + Q_i)} < .3 \right].
\]

3. RESULTS

Five of 25 participants were found to violate the ITU-R BT.500-13 screening procedure (i.e., to have supplied responses that, to some degree, deviated from the majority). However, since raw ratings were found to be broadly contingent upon objective quality setting (Fig. 4), and since \( N \geq 20 \), all participants were retained. The grand mean subjective quality rating for all participants and images at all objective quality settings was 5.50 (CI[95] = 0.10).

The mean subjective quality rating for each of the five image-order sequences was 5.30 (CI[95] = 0.35), 6.18 (CI[95] = 0.37), 5.65 (CI[95] = 0.38), 4.79 (CI[95] = 0.29), and 5.58 (CI[95] = 0.48), shown in Fig. 5A.

Mean subjective quality ratings averaged over stimulus images and participants for each of the five objective quality settings (JPEG compression level, q1-q5), shown in Fig. 5B, were 3.21 (CI[95] = 0.25), 4.87 (CI[95] = 0.30), 5.73 (CI[95] = 0.28), 6.43 (CI[95] = 0.28), and 7.24 (CI[95] = 0.30), respectively, highlighting that participants did not employ the entire subjective quality range (1–10), but that the mean subjective quality rating does increase monotonically as the objective quality setting rises, confirming participants’ generally successful discrimination between objective quality settings. This is exemplified by a significant “large” \( \hat{r} \) correlation between subjective quality setting and mean subjective quality assessment \( [r = 0.54 (2548), p(\text{two-tailed}) < 0.001] \).

Moreover, a significant “medium” correlation between the mean subjective quality rating for each image and its corresponding image valence rating was found \( [r = 0.30 (98), p(\text{two-tailed}) < 0.01] \); conversely, no significant correlation between median IQA rating and arousal was found \( [r = -0.07 (98), p(\text{two-tailed}) = 0.44] \), indicating that relative to valence, it did not play a critical role in influencing participants’ subjective quality assessment.

To evaluate the impact of practice or fatigue, mean subjective rating over time (trial number) was calculated (Fig. 6), for which reverse arrangement test trend analysis does show a significant overall trend (3432 reversals of 4900, \( z = 5.70, p < .01 \)), indicating increasingly liberal responses towards the end of the session, possibly as a consequence of diminished surprise or habituation to less pleasant images seen later in the presentation sequence, or calibration refinement following exposure to a range of images at different objective quality settings.

4. DISCUSSION

As a consequence of “top-down” visual processing \([44]\), contextual/semantic information bears a direct influence on scene perception. In our natural visual environment, this may expedite the recognition of objects and the nature and composition of the environment in which we find ourselves (e.g., the degree of personal threat/danger, if there are objects available to be used to satisfy ongoing goals, and so on). However, in circumstances where we are asked to make objective judgments, strong top-down influences may bias our evaluation of ostensibly unconnected scene parameters. Thus, in addition to the inherent “observable” components of a visual stimulus, interpretation of visual imagery is affected by expectations, prior knowledge, mood, and motivation. For example,
classical work examining biases in the interpretation of visual information found that poor children estimated the size of coins to be larger than wealthier children [45]. Similarly, a group of thirsty experimental participants were found to rate the size of a glass of water as being larger than a second group of satiated participants [46]. In a related experiment, it was found that participants who were old, weary, or asked to wear a heavy backpack tended to overestimate the incline of a hill [47]. In work that focuses upon nonvisual judgments, Schwarz and Clore found that human respondents rated themselves as being happier and more satisfied with their lives when questioned on a sunny day, relative to a rainy day; however, when first being reminded of the weather, this effect disappeared [48]. It has also been shown that mood (induced by listening to cheerful or gloomy music) can modify the interpretation of advertisement imagery with ambiguous affect, leading to significant differences in judgments [49], and can affect our interpretation of facial emotion [50]. The interaction between perceptual and personal judgments, emotional state, prior knowledge, expectations, and motivation is undoubtedly tremendously complex but is also likely to pervade all decision making, and therefore demands consideration when drawing inferences from perceptually driven human decisions.

In the present study, a standard single stimulus IQA procedure with human participants was run, in which both the objective quality setting (true severity of compression artifacts) and affective impact (in valence and arousal dimensions) of test stimuli were known a priori. Despite submitting subjective image quality assessment responses that confirmed broad adherence to task instructions (on average, correctly categorizing images according to their objective quality setting and subjective rating), our human participants were found to be unable to disentangle affective image content (in particularly, image valence; see below) from their judgment, despite precise instructions to rate JPEG image quality only (without considering image content) and an explicit forewarning that some of the images to be rated may be disagreeable.

The “medium” sized correlation between mean subjective image rating and image valence, but nonsignificant correlation with arousal, suggests that the “pleasantness” of the image viewed is a significant factor that influences subjective rating, whereas the intensity of emotion evoked (from calm to exciting/agitating) is a less critical factor (though we concede that our stimulus set is relatively small, though still larger than all commonly used IQA databases). Arousal ratings do not possess a directional component that links to valence, so it is possible that images that are particularly unpleasant or particularly pleasant may possess the same arousal value (e.g., exciting and pleasant, or exciting and unpleasant), potentially nullifying any effect. In contrast to the finding in

Fig. 5. (Color online) A, box plot showing the relationship between presentation sequence (1–5) and mean subjective quality rating (1–10); B, box plot showing the relationship between objective quality setting (q1–q5) and mean subjective quality rating (1–10). Plus symbols denote outliers.

Fig. 6. Mean subjective quality rating over time (trial number). Thin black curves represent the ratings for each presentation sequence (1–5); the thick black curve represents the mean rating.
the present study of a valence effect with no corresponding effect for arousal, a seminal study of the impact of these image dimensions on image recall found that only (high) arousal served to increase memory performance (higher accuracy, lower reaction time), both immediately and following a 1-year intermission. However, in the same study, valence did affect reaction times for novel images, leading the authors to propose that both arousal and valence are encoded upon initial display of an image, but that only arousal has a long-term effect on memory fidelity. In the present study, the nature of the task and immediacy of response following each stimulus may have masked any longer-term effect of arousal in detailed scene assimilation.

Due to the time, expense, and difficulty of gathering subjective image quality assessment data, a concerted effort has been made to develop automated systems that accurately simulate the quality ratings that human observers would elicit. Certainly, the task of developing automated image quality assessment techniques that can interpret image affect in formulating a statistical quality judgement is presently insurmountable, although the use of prerated images is quite feasible, as is the collection of affective measures from participants in addition to quality ratings. Thus, in addition to employing a large and varied database in IQA experimentation that involves collecting subjective quality ratings from human participants, images that span a range of valence settings, rather than all possessing “positive” valence should be used, since such images are likely to attract liberal quality ratings (for instance, in IQA with the standardized images found in the IVC database, one may feel inclined to be more sympathetic to “Lena,” relative to “Baboon”). Failing the creation of new databases containing images selected to span a range of arousal and valence settings, for practical purposes, the affective image descriptors of images in extant databases could be established and employed as regressors, since these are likely to contribute a proportion of the variance in subjective quality ratings.

The visual media we consume, for example, in television and film, spans the full gamut of human emotion, containing both pleasant and unpleasant scenes and everything in between; quality assessment databases, which tend to use small numbers (typically under 30) of generally affectively positive or neutral images, should reflect this. Although it is tempting to believe that subjective image quality assessment is a mechanical process, unencumbered by participant memory, emotion, and belief, the present study shows that this is unlikely to hold in all cases. This issue may be especially problematic where laboratories compare results using different databases, since one may contain many more affectively positive images than the other, leading to MOS ratings that vary independently of the image distortion under analysis.

Single stimulus approaches are popular in IQA, due to requiring less time to complete that double stimulus techniques, and because they mimic familiar real-world applications, such as viewing compressed images/video and using mobile communications devices, in which degraded and undegraded stimuli are typically not available to be compared. Relative to single stimulus quality assessment, the procedure used in this study, affective image content may be less problematic in 2AFC or 2IFC procedures, such as in degradation category rating (DCR), also known as the double stimulus impairment scale (DSIS). In these techniques, participants select the highest quality version of two images shown either simultaneously or sequentially, or rate the second image relative to the first image (full quality reference). However, even in this case, participant motivation to discriminate fine quality differences may be also be affected by valence setting.

A recent review included a discussion of the potential future importance of aesthetic image quality assessment, highlighting that in existing attempts to assemble rated image databases, the female nude was found to attract high aesthetic quality ratings. Following from the work of the present study, it is also likely that such images will confound pure image quality assessment ratings, leading to excessively liberal responses. Indeed, an unintentional bias to select pleasant images and video sequences by experimenters for inclusion in test databases may lead to more liberal ratings than a set of images that is actually representative of typical broadcasting, which, even if not graphically unpleasant, may be broadcast with narrative-induced low mood or reporting overtones (e.g., in news broadcasting) that make even neutral images feel unpleasant and liable to attract more conservative ratings if viewed in context. Though the present study focuses upon static images, it is intuitive to suppose that media affect may likewise influence the subjective quality assessment of video; however, with its opportunity to evolve a complex narrative, leading to the modulation of the mood and emotion of the viewer over time, this may prove even more challenging to quantify.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to U. Rajashekar and two anonymous reviewers for their helpful suggestions.

REFERENCES