Naturalness and interestingness of test images for visual quality evaluation

Raisa Halonen, Stina Westman, Pirkko Oittinen
Media Technology, Aalto University School of Science and Technology
P.O. Box 15500, FI-00076 Aalto, Finland

ABSTRACT

Balanced and representative test images are needed to study perceived visual quality in various application domains. This study investigates naturalness and interestingness as image quality attributes in the context of test images. Taking a top-down approach we aim to find the dimensions which constitute naturalness and interestingness in test images and the relationship between these high-level quality attributes. We compare existing collections of test images (e.g. Sony sRGB images, ISO 12640 images, Kodak images, Nokia images and test images developed within our group) in an experiment combining quality sorting and structured interviews. Based on the data gathered we analyze the viewer-supplied criteria for naturalness and interestingness across image types, quality levels and judges. This study advances our understanding of subjective image quality criteria and enables the validation of current test images, furthering their development.

Keywords: test image, quality attribute, subjective evaluation, evaluation method

1. INTRODUCTION

The effect of image content in quality evaluation is well-known. This is often taken into consideration by selecting test images of different types in an effort to cover the relevant photo space. Another approach strives to compose a single representative test image. Our research group has been involved in the development of such stand-alone test images for visual print quality evaluation\(^1\)\(^2\).

Preferential high-level quality attributes such as naturalness and interestingness are key issues in visual applications aimed at improved perceived quality and engagement of users. Naturalness refers to the degree of correspondence between reproduced images and internal references\(^3\). It is an aspired characteristic for test images, as they typically portray staged scenes with content chosen to facilitate evaluation. In designing these compositions, naturalness is valued. Interestingness is related to individual and community preferences in image aesthetics\(^4\) and reflects images’ ability of attracting and holding attention\(^5\). Interestingness of test images is important in order to make the evaluation of the image pleasant and effortless. Test images for visual quality evaluation should thus balance naturalness and interestingness.

Studies of high-level image quality attributes typically use a bottom-up approach, addressing e.g. how manipulated changes in attributes such as sharpness or chroma affect for example perceived naturalness. There is little study on how high-level quality attributes break down into lower-level constructs. Also, perceived quality of test images as such has not been evaluated. This study continues our efforts in evaluating test images in relation to relevant quality attributes. We pose the following questions: Which dimensions constitute naturalness and interestingness in test images? Are viewers consistent in their ratings? What is the relationship between naturalness and interestingness?

2. IMAGE NATURALNESS AND INTERESTINGNESS

Image quality attributes may be distinguished on multiple levels. Keelan\(^6\) divides image quality attributes into artifactual (e.g., unsharpness, graininess, redeye, digital artifacts), preferential (e.g., color balance, contrast, colorfulness, memory color reproduction), aesthetic (e.g., lighting quality, composition) and personal (e.g., preserving a memory, conveying a subject’s essence). An artifactual attribute is defined as one that is not always evident, but when apparent leads generally to a degradation of quality. Preferential attributes are nearly always evident in an image and depend both on the taste of the observer and the image content. Aesthetic image quality attributes may be very subjective in nature. However, to some extent, there is agreement on desirable aspects of factors such as composition and lighting. Personal attributes influence first-party assessment, but do not affect third-party evaluation. High-level image quality criteria such as
naturalness and interestingness compliment artifactual, low-level quality criteria. These types of quality criteria have been found to be a requirement for high-quality images.

Naturalness is typically defined as the degree of correspondence between the visual representation of the image and the knowledge of reality as stored in memory. In color reproduction, naturalness can be assessed by the mental recollection of the colors of familiar objects, i.e. memory colors. Naturalness is often considered by the observers when evaluating overall image quality. Several studies that probe image quality judgments report on observers using attributes such as real, life-like, and their opposites. While these attributes have slightly different meanings, they are often used by the observers to describe the naturalness of an image. Due to the interrelatedness of the attributes, they are often combined also by researchers to simplify the setup. As studies of high-level quality attributes typically use a bottom-up approach, naturalness studies are no exception. Naturalness is typically studied by addressing e.g. how manipulated changes in attributes such as sharpness or chroma affect the perceived naturalness of an image. There has been little study on modeling image naturalness; some attempts have, however, been reported.

Interestingness as image quality criterion is employed by Flickr, the popular online photo management and sharing application. Flickr’s interestingness metric is defined through several parameters, including who marks it as a favorite, where the click-throughs are coming from, who comments on it and when, and what are the tags affiliated with it. Katti et al. have defined interestingness as an aesthetic property which arouses curiosity and is a precursor to attention. To determine if an image is interesting different amounts of cognitive processing and prior knowledge might be required. Vaiapury and Kankanhalli specify interestingness as an entity that arises from interpretation and experience, surprise, beauty, aesthetics and desirability. Greisdorf and O’Connor suggest that the gap between looking and seeing is bridged by steps of engagement with what is being looked at. The proposed steps are attention (focus), interest (cognitive awareness of the focus), involvement (meaning attached to the awareness) and attitude (feeling regarding the meaning).

3. TEST IMAGES FOR IMAGE QUALITY EVALUATION

‘Test images’ refer to natural images, i.e., photographs that have been designed for subjective and objective evaluation of image quality. As subjective evaluation of images is strongly affected by image content, test images must contain diverse subject matter. In addition to covering the relevant photo space, standard test images must be well-defined and of high-quality to ensure good quality reproductions, when rendered properly, and provide a reasonable test of a given evaluation task. By requiring the images to look natural, test image development can be challenging, e.g., for color evaluation, as a fairly limited image set should cover the full reference color gamut and still contain subtle color differences required in such images. Several single test images and test image sets have been proposed for imaging research. Here we shortly review commonly used test image sets, and a set of stand-alone images attempting to cover the relevant photo space.

Kodak Lossless True Color Image Suite is a set of 24 lossless, true color photographs released by Kodak for unrestricted usage. Originally, the photographs appeared in a Kodak Photo CD. The popular set, consisting of 22 outdoor and two studio images, has served as a standard test suite for studying various aspects of imaging, including information hiding, gamut mapping, and image quality metrics.

ISO 12640-1:1997 specifies a set of CMYK color images to be used for evaluation of changes in image quality during coding, image processing, film recording or printing. The standard set consists of eight photographed images and ten synthetic, digitally created images. Seven of the photographs are studio compositions, one is an outdoor scene. The photographs include skin tones, fine image detail, highlight and shadow vignettes, a variety of colors, and fine texture. The photographs have been used in several studies, including studies on moiré suppression, image compression, and high-level image quality evaluation.

Sony sRGB standard images are three photographs provided by Sony for CIE TC8-04, a technical committee that investigated visual adaptation under mixed illumination conditions. The set consists of two studio scenes, representing a portrait and a party image, and an outdoor image with a picnic theme. In addition to the research done in the technical committee, the images have been used e.g. in studying color demosaicing, gamut mapping, and image compression.

ISO 12640-2:2004 specifies a set of 15 standard color images encoded as XYZ and sRGB data. The set contains eight photographs (seven studio compositions, one outdoor image), and seven synthetic images. The natural images contain skin tones, images with detail in the extreme highlights or shadows, neutral colors, brown and wood tone colors, memory colors, complicated geometric shapes, fine detail, highlight and shadow vignettes. Several studies have incorporated a
sub-set of the photographs in their test image sets in different research fields, such as gamut mapping\textsuperscript{22,30}, perceived print quality\textsuperscript{31}, and color difference computation\textsuperscript{32}.

I3A identifies six image clusters that cover approximately 70\% of images taken with camera phones\textsuperscript{33}. The clusters contain different combinations of subject illuminance and subject-camera distance. Typical scene descriptions include close ups in dim-dark lighting (indoor/outdoor), close ups in typical indoor lighting (indoor/outdoor), small groups in dim-dark lighting (indoor/outdoor), small groups in typical indoor lighting (indoor/outdoor), small group in cloudy bright to sunny lighting (outdoor), and scenic landscape or large group in cloudy bright to sunny lighting (outdoor).

ISO 12640-3:2007 specifies a set of standard large gamut color images encoded as CIELAB data\textsuperscript{19}. The set consists of eight natural images, eight color charts and two color vignettes. The design of the natural images follows the guidelines of the previous parts of the standard: the photographs include well-defined characteristics for evaluating the rendering of skin tones, neutrals, pastels, high-key tones, metallic objects, textile structures, details and memory colors. So far only few studies\textsuperscript{e.g. 30} report to have employed the set of images in imaging-related research.

Jin et al.\textsuperscript{34} implemented a softcopy quality ruler method to extend ISO 20462-3, a standard on psychophysical experimental methods for image quality estimation, by creating digital reference images of known subjective image quality, complimenting the hardcopy stimuli. Ruler images were created for 16 photographs (five indoor, eleven outdoor images) depicting a variety of subject matter including people and landscape scenes. A subset of the ruler scenes has been incorporated in related research developing texture appearance metrics for camera phone imaging\textsuperscript{35,36}.

DigiQ test images\textsuperscript{1,2} are a set of stand-alone test images for print quality evaluation designed to cover the relevant photo space. The set consists of three photographs, corresponding to different development versions of the same image, portraying a breakfast scene. In designing the images, the aspects taken into consideration included a recognizable theme, memory colors and shapes, versatile surface materials, amount of detail and salience of objects. Moreover, perceptual and aesthetical aspects were considered. In the development process, emphasis was placed on the naturalness of the situation and setting. From the technical point of view, important perceptual characteristics for the evaluation of specific quality attributes were also included in the images.

4. METHODOLOGICAL APPROACHES TO SUBJECTIVE QUALITY EVALUATION

Psychophysics in an essential tool in understanding image quality as humans perceive it. The questions dealt with psychometric scaling studies are: is the quality impairment visible and how large is it and is it acceptable (to a certain end user or for a certain end application)? Image quality may be assessed separately for a single type of impairment (e.g. unsharpness) or in a summative manner, assessing overall image quality. There exist a variety of psychometric techniques for assessing a series of samples. Here we review common methods based on Keelan\textsuperscript{6} and Engeldrum\textsuperscript{37}.

**Paired comparison** is the most commonly employed technique for determining small quality differences. In paired comparison each test sample is paired with every other test sample and the observer must select one. The method is often called forced-choice paired comparison because “no difference” or “no preference” responses are not allowed. The downside is that many comparisons and many observers are required for valid judgments and the analysis breaks down if sample differences are large enough to give unanimous judgments.

**Rank order** is related to paired comparison. In ranked order experiments participants view all samples at once and rank them according to the attribute at hand. The procedure is more complex for the observer but faster and easier to implement. The data, while in principle equivalent to a full round of paired comparison, may suffer from the complexity of the sorting task.

**Ratio scaling** refers to a family of direct scaling methods in which observers directly estimate or adjust the magnitude of stimuli. In magnitude estimation the observer is to provide a numerical response in proportion to how they perceive the quality attribute. This number is expressed as a ratio compared to a reference stimuli: e.g. a rating twice as high as the reference value is expected if the sample quality is twice as good as the reference. In magnitude production the observer is presented with the numerical value and adjusts the magnitude of the stimulus to match the numerical value.

**Category scaling** consists of sorting of the individual samples into a finite number of categories (typically five to nine). The categories are labeled (e.g. from excellent, very good, good, fair, poor, not worth keeping). Most category scales are nominal or ordinal. Category scaling requires a large number of observers for reliability and may not result in an interval scale despite best efforts in category labeling.
Quality rulers are a type of categorical scale with images used as category markers (anchored scaling). Participants are to assess the quality of one test sample at a time by placing them on the ruler. The ruler provides a numerical score based on the anchor images being calibrated to certain perceptual attribute differences. The data may suffer if test samples have content different from the anchor images. Also, test samples are not compared directly.

In addition to these well established techniques novel methods have started to emerge, taking a qualitative approach to studying image quality. The overarching question dealt with these qualitative methods is what matters to the viewer in terms of quality? The interpretation-based quality approach (IBQ)\textsuperscript{10,34} combines qualitative and quantitative measurements with the aim of defining the perceptual image quality dimensions. In the IBQ procedure the instruction is simply to find the most crucial features that appear to be changing in the images across the sample. The procedure includes the quality assessment task which is followed by a form interview designed to elicit subjective quality criteria. The IBQ approach can complement psychometric approaches and objective measurements by defining the subjective meaning of image quality attributes and characteristics; in other words, it reveals how important they are for the overall perceived quality. The method has been used in defining subjective quality dimensions related to sharpness changes that mimic the camera lenses\textsuperscript{10}, the comparison of digital still\textsuperscript{8} and video cameras\textsuperscript{39}, print quality with natural image contents\textsuperscript{38}, and stereoscopic 3D image quality\textsuperscript{40}.

We employ a form of category scaling with aspects of IBQ built into it. We chose category scaling due our goal of exploring the types of naturalness/interestingness criteria utilized by observers, the variation in the image content in our test samples, and the further utilization of the categories to simplify the interview procedure from IBQ. Various image sorting procedures and subsequent category naming have been utilized in studies of image similarity\textsuperscript{20}, e.g.\textsuperscript{41} and may be applied to image quality evaluations as well. Engeldrum\textsuperscript{37} discusses different approaches to defining the quality attributes used in scaling: defining the attribute in the textual/verbal instructions, use of visual references as anchors and letting the viewer define the attribute using some internal criteria. In this study we focus on the last option and, most importantly, gather information on the criteria employed by the participants in relation to the attributes used and categories created.

5. METHODOLOGY

5.1 Aims

The aims of the study were to:

- find out the dimensions which constitute naturalness and interestingness in the context of test images,
- clarify the relationship between naturalness and interestingness and 
- evaluate the consistence of subjective ratings on naturalness and interestingness.

Our data encompasses the subjective perception and interpretation of image quality, furthering our understanding of subjective image quality criteria across content types, quality continuum and viewers. The context of test images enables validation and further development of current test images.

5.2 Participants

We recruited 24 participants (12 female) through university newsgroups. Participants were undergraduate and graduate students in engineering, humanities and arts. Their average age was 23 years (sd = 3, min = 20, max = 32). Their vision was normal or corrected to normal. All had normal color vision.

5.3 Material

We employed existing collections of test images intended to evaluate a wide range of color imaging systems, color image output devices, image processing algorithms and coding technologies. We used a total of 76 images: Kodak Lossless True Color Image Suite\textsuperscript{20} (24 images), ISO12640 (CMYK\textsuperscript{24}, sRGB\textsuperscript{18}, CIELAB\textsuperscript{19}) natural images (24), Sony sRGB standard images\textsuperscript{27} (3), images for evaluation of Nokia camera phones, captured based on I3A image clusters\textsuperscript{33} (6), softcopy quality ruler images by Jin et al.\textsuperscript{34} (16), and DigiQ test images\textsuperscript{2} (3).

The test images were printed on a semigloss photo paper (251g/m\textsuperscript{2}) with a desktop ink-jet printer, Epson Stylus Pro 3800. In the color-managed process, optimal print settings were chosen and a paper-specific ICC profile was created for the settings. Each image was prepared for printing using Adobe Photoshop CS5. From the ISO 12640 images, text was removed by content-aware fill, and cloning if needed. Five of the ISO 12640-3 images were cropped in order to exclude the text, due to surrounding complicated texture. Each image was resized to 360 dpi and short side set to 10 cm. Finally,
the images were converted to the printer color space using perceptual as the rendering intent. The images were printed using Adobe Photoshop CS and visually checked to have no visible artifacts. The printed test images were glued to gray cardboard, and a randomly assigned ID number was marked in the lower right corner of the cardboard.

5.4 Procedure

Participants performed two sorting tasks, sorting the test images according to their naturalness and interestingness. The order of these tasks was counterbalanced for participant blocks. After each sorting task participants were interviewed on their sorting behavior using an interview procedure modified from the Interpretation Based Quality method. The procedure was piloted twice to ensure smooth running of the experiments.

At the beginning of the naturalness sorting task, participant was handed the pile of 76 test images in random order. They were asked to browse the images in order to get an idea about the naturalness of the images. After that they were to categorize the images according to their naturalness on a scale from 1 (very unnatural) to 7 (very natural). This was done by sorting the images into piles on top of a paper scale with the explanations of the end points visible. They were not required to use all seven categories or the scale endpoints. There was no time limit for the sorting. The procedure for the interestingness sorting was analogous. The scale ranged from 1 (very uninteresting) to (very interesting).

Each sorting task was followed by a form-based interview. During the interview participants went through one image category at a time (starting from category 1), read out loud the numbers of the images therein (essentially going through the images once more), and were then asked the following questions about the category: How would you describe the naturalness of the images in this category? What about the images is natural? What about the images is unnatural? How does this category differ from the previous category in terms of naturalness? Participants were encouraged to give examples through particular images. If a particular image or content detail was mentioned, the participant was further asked what about that particular image or detail was natural or unnatural. After discussion on a category ended, the interview continued to the next category until all categories utilized by the participant had been covered. The interview ended with a closing question: How would you define image naturalness? The interview procedure after the interestingness sorting task was analogous.

5.5 Viewing conditions

Illumination of the experiment room (150 lux) was controlled with halogen lamps and dichroic glass filters producing 5000K light. The lamps were positioned so that the light was uniform and indirect (reflected from the ceiling). The sorting was done on a light gray tabletop and the back wall was covered with a gray curtain.

5.6 Analysis

The experimental sessions were recorded and transcribed. We analyzed the sorting criteria for naturalness and interestingness as well as the definitions provided by the participants qualitatively. Qualitative coding was done in ATLAS.ti version 6.2.13. Two of the authors performed the coding, reviewing each others’ coding and discussing the coding scheme and its application to the data until consensus was reached. For the quantitative analysis we first cross tabulated the frequencies of each attribute for each naturalness/interestingness category per participant. We then normalized this data so that the data was binary: either an attribute was mentioned by the participant for the particular category or not. Further analysis was conducted on a contingency table (attributes x categories) calculated based on the individual participants’ binary data tables. In our analysis we treat the scale (1-7) as an interval scale, enabling us to present averages of the attribute values for each image. The qualitative analysis is naturally not contingent on scales and correspondence analysis only requires nominal data. Correspondence analysis (CA) is used to show the relationship between categorical variables in a spatial map, illustrating the similarities and differences between categories as well as the associations between them in a way that is similar to factor analysis. CA reduces the dimensionality of the data matrix allowing visualization in a lower-dimensional subspace. We use CA to visualize the connections between different quality attributes based on the contingency table of attributes and categories.

6. RESULTS AND DISCUSSION

6.1 Naturalness and interestingness scores

The average naturalness and interestingness scores for the test image are plotted in Figure 1. On the scales from 1 to 7, the average naturalness of the images varied between 2 and 6.2 and average interestingness was between 2.5 and 4.8. The standard deviations for naturalness were somewhat lower (min = 0.9, max = 1.9) than for interestingness (min = 1.1,
We calculated the correlation between the naturalness and interestingness scores. Correlation for different participants ranged from -.24 to .81 and for the complete participant set the correlation was .28 (df = 74, p < 0.01).

In order to evaluate if participants agreed on the naturalness and interestingness of the test images, we calculated an intraclass correlation coefficient (equal to average kappa with quadratic weights) for both attributes across the 24 participants. For naturalness the value of the coefficient was .40 and for interestingness .07. These figures reflect fair to almost no agreement. This was to be expected for the evaluation of interestingness which relied more on subjective interests and preferences in image content. The coefficient value for naturalness indicates inter-rater agreement on a level which is statistically significant (F = 19.559, df = 75, p < .001).

Figure 1. Image interestingness as function of naturalness

6.2 Naturalness attributes and dimensions

The qualitative analysis resulted in 52 distinctive naturalness codes, corresponding to attributes which either decreased or increased the perceived naturalness of the test images according to our participants. The attributes are listed in Table 1. Naturalness attributes included artifactual attributes (related to e.g. low-level quality), preferential attributes (e.g. color balance, preferred colors), aesthetic attributes (e.g. lighting, perspective, cropping), and personal attributes (e.g. personal connection, familiarity).

We calculated the average number of distinct naturalness attributes utilized by a participant for each naturalness category (Figure 2). The attributes contributing to the naturalness or unnaturalness of the images differed for the different categories (1-7). At the unnatural end of the scale even the presence of natural objects demonstrated naturalness while at the high end personal connotations and familiar imagery was deemed to be criteria for naturalness. The realism of the situation portrayed was the most agreed upon naturalness criteria. Unnaturalness was brought upon by staged setups throughout the scale, and by posed portraits especially in the middle of the scale.
Table 1. Attributes contributing to image naturalness

<table>
<thead>
<tr>
<th>Attribute class</th>
<th>Attributes which decrease perceived naturalness</th>
<th>Attributes which increase perceived naturalness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colors</td>
<td>studio lighting, poor color balance, nonpreferred colors, unnatural colors, strong colors</td>
<td>natural lighting, good color balance, natural colors, preferred colors</td>
</tr>
<tr>
<td>Composition</td>
<td>unrealistic perspective, poor cropping</td>
<td>realistic perspective, good cropping</td>
</tr>
<tr>
<td>Genre</td>
<td>controlled</td>
<td>typical, effortless</td>
</tr>
<tr>
<td>Manipulation</td>
<td>staged, artificial, too neat, manipulated, synthetic</td>
<td>unmanipulated</td>
</tr>
<tr>
<td>Objects</td>
<td>nonliving, unnatural combination</td>
<td>living, natural objects, natural combination</td>
</tr>
<tr>
<td>People</td>
<td>posed, fake expression</td>
<td>relaxed people, genuine expression</td>
</tr>
<tr>
<td>Personal</td>
<td>unfamiliar object</td>
<td>personal connection, affective, familiar</td>
</tr>
<tr>
<td>Quality</td>
<td>too high quality, too low quality</td>
<td>low quality, high quality</td>
</tr>
<tr>
<td>Reality</td>
<td>lack of entropy, temporary setup, unconvincing image</td>
<td>real, reality unaltered, entropy, credible</td>
</tr>
<tr>
<td>Setting</td>
<td>unrealistic situation, fake setting</td>
<td>natural setting, realistic situation, natural action, outdoors</td>
</tr>
</tbody>
</table>

Figure 2. Average number of positive and negative naturalness criteria used by category, number of participants who employed the category, and most common criteria (at least one in each direction) with number of participants who used it

We conducted a correspondence analysis on the cross tabulation of naturalness categories 1-7 and the attributes. Data was included for the 36 attributes which were used by 6 participants or more. The analysis was aimed to show interrelations between attributes and explain the naturalness dimensions in the data. We present results for the first three dimensions (Figures 3 and 4) which together account for 80% of variance in the data. Dimension one ("control") was significantly correlated with the image having been manipulated (r = .94), staged (r = .88) and it not being familiar (r = -.95) or effortless (r = -.89). Dimension two ("reality") was correlated with the image not being posed (r = -.73), including natural objects (r = .51) and depicting reality unaltered (r = .39). Dimension three ("no life") was correlated with the image not portraying natural action (r = -.81) or genuine expressions (r = -.55) and it showing nonliving objects (r = .68). These correlations are all significant at the p < .01 level (df = 34).

When asked to define naturalness at the end of the experiment, participants reflected upon several aspects of the images. These are listed in Table 4 together with examples from definitions that further shed light on the criteria employed.

Table 2. Aspects found in participants’ definitions of image naturalness and the number of participants who referred to them

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Examples</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>No staging</td>
<td>The scene is not assembled solely for the purpose of photographing it, no one has staged the image</td>
<td>15</td>
</tr>
<tr>
<td>Natural colors</td>
<td>No single color is overemphasized, colors are natural and balanced</td>
<td>10</td>
</tr>
<tr>
<td>Posing, expressions</td>
<td>People have natural expressions and are not consciously posing or posed for the image</td>
<td>9</td>
</tr>
<tr>
<td>Natural situation</td>
<td>Image captures a moment which really exists and where people are engaged in</td>
<td>8</td>
</tr>
<tr>
<td>Reality-like</td>
<td>Could be a real view instead of a mediated picture, one could turn their gaze to the side and see the same thing, one can picture themselves in the picture, the subject is represented just as it is</td>
<td>7</td>
</tr>
<tr>
<td>Quality</td>
<td>Image not modified, not more detailed than the eye can see, naturalness means smudginess and unsharpness</td>
<td>5</td>
</tr>
<tr>
<td>Genre</td>
<td>Naturalness is relative to the imagery one is accustomed to, compared within an image genre</td>
<td>5</td>
</tr>
<tr>
<td>Reality unaltered</td>
<td>Image has been taken without the environment noticing or reacting to it</td>
<td>5</td>
</tr>
<tr>
<td>Random moment</td>
<td>Image has been taken without planning, a moment where the environment does not react has been captured</td>
<td>4</td>
</tr>
<tr>
<td>Face value</td>
<td>The message conveyed through the image has not been enhanced, modified or falsified</td>
<td>4</td>
</tr>
</tbody>
</table>
6.3 Interestingness attributes and dimensions

The qualitative analysis resulted in 50 distinctive interestingness codes, corresponding to attributes which either decreased or increased the perceived interestingness of the test images according to our participants. The attributes are listed in Table 3. Interestingness attributes included artifactual attributes (related to e.g. low-level quality), preferential...
attributes (e.g. colorfulness), aesthetic attributes (e.g. aesthetic, composition), and personal attributes (e.g. personal connection). They also referenced affective, cognitive and genre factors.

We calculated the average number of distinct interestingness attributes utilized by a participant for each interestingness category from 1 to 7 (Figure 5). The attributes contributing to the interestingness or uninterestingness of the images differed for the different categories. At the uninteresting end of the scale, colors, presence of people and standalone objects were interesting while at the high end were criteria of interestingness dealt with personal connections with the subject matter and its ability to provoke thoughts and inspire action. Uninterestingness was brought upon by a variety of attributes along the scale, ranging from uninteresting subjects to unknown people and to lack of action and novelty in the images. There was less agreement on the criteria of very uninteresting or very interesting images, as demonstrated by the lower number of participants using the most common criteria in categories 1 and 7, compared to the numbers in the naturalness results.

We conducted a correspondence analysis on the cross tabulation of interestingness categories 1-7 and the attributes. Data was included for the 35 attributes which were used by 6 participants or more. The analysis was aimed to show interrelations between attributes and explain the interestingness dimensions in the data. We present results for the first two dimensions which together account for 76% of variance in the data (Figure 6). The third dimension explained less than 10% of variance. Dimension one (“no information”) was significantly correlated with uninteresting content (r = .92) and no cause for action (r = .94) as well as not requiring longer viewing (r = -.93) or provoking thought (r = -.89). Dimension two (“affective”) was correlated with the image not being eye-catching (r = -.94) and not having people in it (r = -.66) but being fun (r = .32). These correlations are all significant at the p < .05 level (df = 33).
When asked to define interestingness, participants reflected upon several aspects of the test images. These are listed in Table 4 with examples from definitions that further shed light on the criteria employed by the participants.

Table 4. Aspects found in participants’ definitions of interestingness and the number of participants who referred to them

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Examples</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Captures attention and creates ideas</td>
<td>Image catches your attention, makes you react or think, want to know more about the subject</td>
<td>11</td>
</tr>
<tr>
<td>Positive affect</td>
<td>Image awakens emotions, memories, pleases the eye, creates a good mood</td>
<td>8</td>
</tr>
<tr>
<td>Content</td>
<td>Content of the image is interesting, its conveys a topic</td>
<td>8</td>
</tr>
<tr>
<td>Can look for a long time</td>
<td>One can and has to focus on the image for a while, it does not reveal itself at once</td>
<td>6</td>
</tr>
<tr>
<td>Real story conveyed</td>
<td>Something is happening in the image, it tells a true story</td>
<td>5</td>
</tr>
<tr>
<td>Would like to see for real</td>
<td>Viewer would one like to be in the situation or place depicted, see it with own eyes</td>
<td>4</td>
</tr>
<tr>
<td>Personal memories and interests</td>
<td>Image is meaningful or important to the viewer based on her experiences or preferences</td>
<td>4</td>
</tr>
</tbody>
</table>

7. CONCLUSIONS

The categorical sorting and interviewing procedure described was able to produce rich data on quality attributes in a fairly short experiment. Participants described several simultaneous criteria for both naturalness and interestingness of the test images. The criteria spanned all four levels of image quality attributes (artifactual, preferential, aesthetic, personal) and further included affective, cognitive and genre factors. The relative importance of the different criteria seemed to depend on the current level of the attribute to be evaluated. Interestingness attributes were more prone to subjective variation (centralized averages on the scale, less agreement over important attributes).

The primary dimensions for naturalness in this study were the control taken during photography, the reality of the scene depicted and the living subjects portrayed. The primary dimensions for the interestingness were the amount and type of content present and the affective aspects of the content. These dimensions do not fully explain the variance in the data, indicating that both quality attributes are multidimensional in nature.

Naturalness and interestingness were significantly correlated and we found fair inter-rater agreement on the naturalness scores assigned to individual images by the participant set. More work is needed to find the possible additional interdependencies between the attributes, and to narrow down the attribute space. Furthermore, these results were obtained using test images and further studies are needed to compare to natural images.
ACKNOWLEDGEMENTS

We would like to thank Nokia Corporation for sharing their images for the evaluation of camera phones.

REFERENCES