Adapting ISO 20462 Softcopy Quality Ruler Method for on-line Image Quality Studies

Peter D. Burns, Jonathan B. Phillips[1] and Don Williams[2]
Burns Digital Imaging, [1]NVIDIA Corporation and [2]Image Science Associates

ABSTRACT

The ISO 20462 method for subjective image quality evaluation relies on a set of reference images, which are calibrated in terms of known absolute quality differences. The original reference images, Standard Reference Stimuli (SRS), were in the form of photographic prints, to be viewed under controlled lighting and fixed viewing conditions. This method was then extended for scenes presented on computer monitors as Digital Reference Stimuli (DRS). This softcopy ruler method was developed as part of the Camera Phone Image Quality (CPIQ) Initiative and has now been adopted as an updated ISO 20462 method (ISO 20462-3:2012). This extended method of using the softcopy rulers was validated by CPIQ using two systems of laboratory lighting and display. That effort, while successful, required significant effort and resources to accomplish. Chief among these was gathering a sufficient number of qualified viewers who could commit to viewing images on a narrowly defined schedule at a limited number of laboratories. In our study, we investigate whether and to what extent the ISO 20462 softcopy ruler method can be adapted to Internet-based subjective evaluations. Our objective is to develop and test a method that uses a commercial online survey service. We describe several limitations to be overcome, including image file size, and a static interface rather than one allowing dynamic updating of the reference image. The method that we developed and tested uses reference anchor images rather than the slider-selected reference. However, our anchors were drawn from the ISO 20462 set, and therefore were taken as calibrated reference images, albeit viewed under uncontrolled conditions. We describe the verification study that was completed using Survey Monkey®, and compare results with the corresponding Softcopy Ruler data. Similar results were obtained for observer ratings and their scene-dependency. We conclude that crowdsourcing is useful, particularly when common non-laboratory image viewing is the intent. When calibrated subjective image quality measures are needed, our adapted method should be considered an efficient alternative to the ISO 20462 standard, provided that common reference images are used.

Keywords: softcopy quality ruler, ISO 20462, subjective image quality, online survey, crowdsourcing, Survey Monkey

1. INTRODUCTION

ISO TC42/ WG18 created a large corpus of work on digital imaging performance metrics over the last fifteen years. One successful model for connecting such objective performance metrics to subjective ones has been the protocols described in ISO 20462 [1]. This ISO method for subjective image quality evaluation relies on a set of reference images, which are calibrated in terms of known absolute quality differences. The original reference images, Standard Reference Stimuli (SRS), were in the form of photographic prints, to be viewed under controlled lighting and fixed viewing conditions. This method was then extended so that the scene content could be presented on computer monitors as Digital Reference Stimuli (DRS). This softcopy ruler method was developed as part of the Camera Phone Image Quality (CPIQ) Initiative and has now been adopted as an updated ISO 20462 method (ISO 20462-3:2012). This extended method of using the softcopy rulers was validated by CPIQ using two systems of laboratory lighting and display, where the image degradations were introduced as adaptive image noise-reduction operations. The attribute of interest was texture loss.

Since the initiation of the CPIQ effort, it has become more common to conduct imaging related surveys via Internet-based participation. In this case, images are not viewed under controlled conditions. However, there may be situations where this is desirable, e.g., to simulate consumer decisions regarding image editing. In addition, such crowdsourcing has the advantage that large numbers of responses can be acquired efficiently. Several studies have addressed the demographics of respondents and validity of crowdsourced responses, and concluded that services such as Mechanical Turk are useful for research, and results are comparable to traditional surveys [2].

It is natural, therefore, to ask whether and to what extent the ISO 20462 softcopy ruler method can be adapted to Internet-based subjective evaluations. In this paper, our objective is to develop and test a method that uses a commercial online service that allows presentation of images. The motivation for this effort was,

- development of efficient methods for gathering image evaluation data that can be related to the above ISO method.
- to do so when realistic (uncontrolled) viewing conditions, hardware, software and lighting are needed.

2. SOFTCOPY QUALITY RULER METHOD

ISO 20462 standard describes a method for subjective image quality assessment. The method is based on comparison of test images to sets of reference images with known levels of subjective image quality degradations. In the original standards, the reference images were photographic prints. Sets of reference prints are referred to in the standard as Standard Reference Stimuli (SRS), and are calibrated in terms of a scale of just-noticeable-differences (JNDs) of image quality. An image quality ruler is formed from a set of prints based on a single scene, which vary from ideal to significantly lower image quality. Test images are compared to SRS in a controlled viewing environment with results reported as primary Standard Quality Scale (SQS_1) JNDs.

More recently this method was extended, in ISO 20462 Parts 3, [1] so that the SRS are presented in softcopy form on a specific calibrated computer monitor. [3] The reference images for a softcopy ruler, and are also calibrated in terms of image quality JNDs, though as secondary Standard Quality Scale (SQS_2) JNDs. The softcopy method has the advantage that new scenes can more easily be generated and used as rulers. In addition distribution of rulers and display software can be easily done, for distribution to various testing sites.

Figure 1 shows the user interface for the softcopy ruler method. The observer is asked to compare the test image on the right with the ruler image on the left. A slider control is used to adjust the reference image that is presented. Moving the slider to the left causes a higher quality version of the scene to be displayed on the left. Movement to the right selects a lower quality image for display. When the observer obtains an image quality match between the test and displayed ruler images, the *Next* button is selected and the slider position is recorded. Mean evaluation times of approximately 15 seconds per trial have been reported. [4].

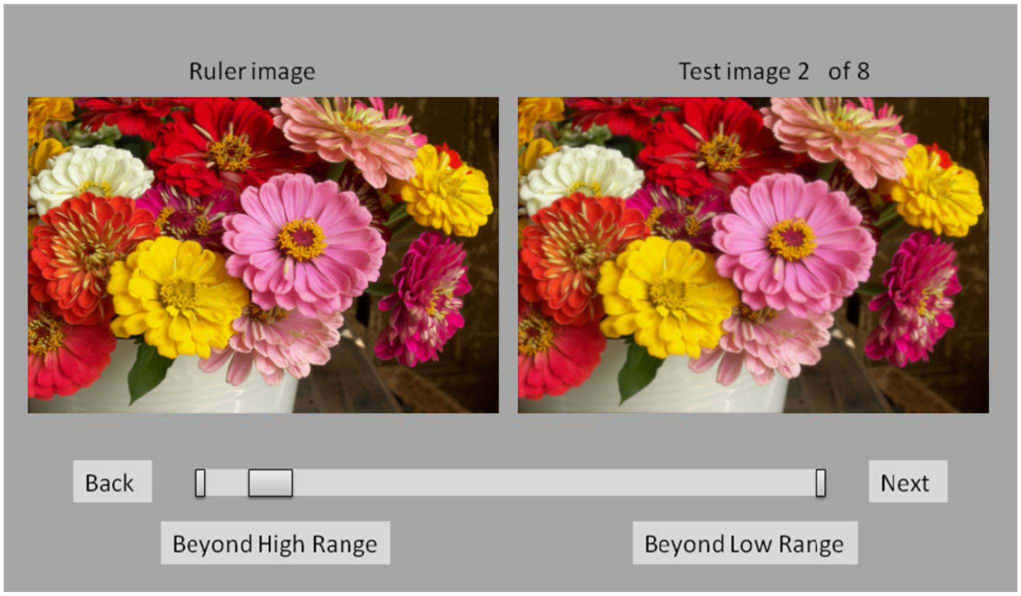


Figure 1: User interface for the softcopy quality ruler. The ruler image is on the left and the test image is on the right.

2.1 Previous Application to Texture Blur

As part of the Camera Phone Image Quality (CPIQ) initiative, the softcopy quality ruler method was validated in a study of image texture loss due to digital image processing [4]. Ten scenes were selected from the ISO 20462 Part 3 available rulers. Rulers are formed by linear (convolution) filtering of high quality baseline digital images, so they vary primarily in a single perceptual attribute, sharpness. The baseline (high quality) ruler images were then processed with eight levels of noise cleaning, as indicated in Table 1. The noise cleaning filters are specified by a kernel size and a C parameter (a noise-table multiplier). The noise table is a linear function of standard deviation versus the pixel value, scaled by C. It is calculated from

$$noise\ value = C\left[0.01p + 2.0\right], \quad (1)$$

where p is the input pixel value. For each N×N kernel block, the central pixel is replaced with the average of all pixels in the block that differ from it by less than the Noise Value.

Table 1: C parameter and N value of the N×N kernel for each noise cleaning filter level used in the Softcopy Quality Ruler validation study. [4]

Noise Cleaning Level:	1	2	3	4	5	6	7	8
C Parameter	2	4	6	10	13	18	22	28
N Value of N×N Kernel	5	5	9	9	9	9	19	19

Sets of digital images were generated by applying a noise-cleaning operation varying in level of severity. Low-level cleaning resulted in low texture loss, and high quality images. Note that the unblurred 'baseline' images were those from the ruler which were degraded by 1 just noticeable difference (JND) of sharpness in order to avoid endpoint concerns. Extreme noise cleaning resulted in near-complete loss of image texture and image quality. The validation study included the use of the softcopy quality ruler method, duplicated in four separate laboratories. Two methods of ruler evaluation were used and a successful correlation with two objective methods was achieved. [4]

Figure 2 shows several of the scenes that were used in the above validation study; each captured using one of two digital cameras. The softcopy quality rulers for each scene varied in image sharpness to achieve the range of image quality spanned by the study. Examples from the softcopy ruler and noise-cleaning image sets are shown in Fig. 3. Note that the ruler varies in image sharpness in a way that is calibrated in SQS_2 JNDs of overall image quality. Results from the texture-loss study indicated the scene-dependent nature of the image quality loss due to the noise cleaning.

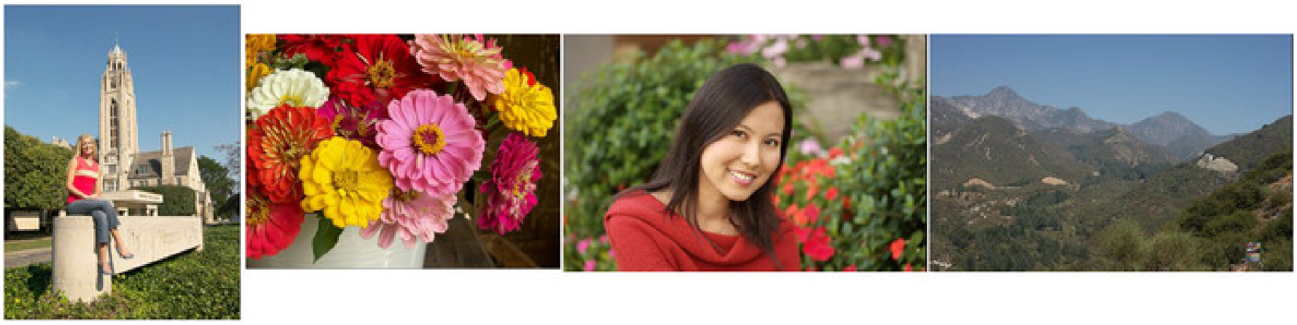


Figure 2: Scenes used in their original form (1088 x 816, or 1253 x 834 pixels); (L-R) Memorial Art Gallery, Flowers, Girl and Mountain.

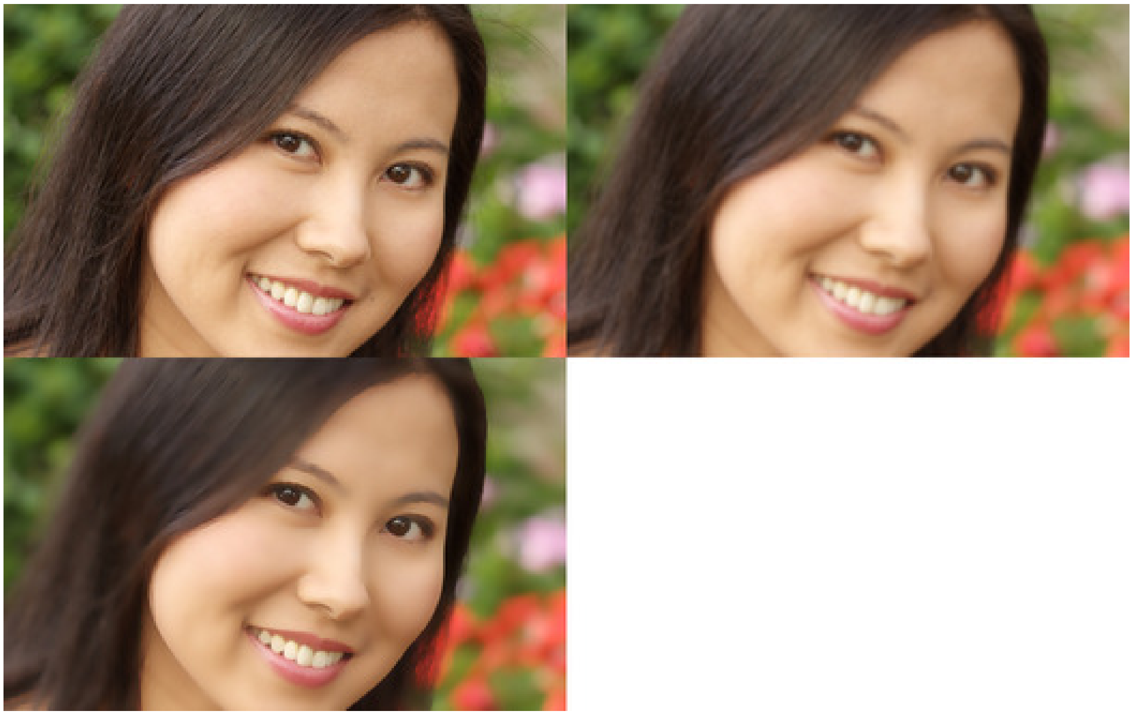


Figure 3: Cropped examples from the softcopy quality validation study. Top-left: high-quality ruler, secondary Standard Quality Scale (SQS_2) 30. Top-right: lower quality ruler, SQS_2 15. Lower-left: noise-cleaned sample treatment level 6 (of eight levels).

Figure 4 shows the secondary SQS JND subjective results from the softcopy quality ruler method. Note that there is a strong scene dependency on the results. For example, the Memorial Art Gallery scene had minimal overall quality loss with the eight levels of noise-cleaning, while the Mountain scene had significant overall quality loss with increased noise-cleaning treatment.

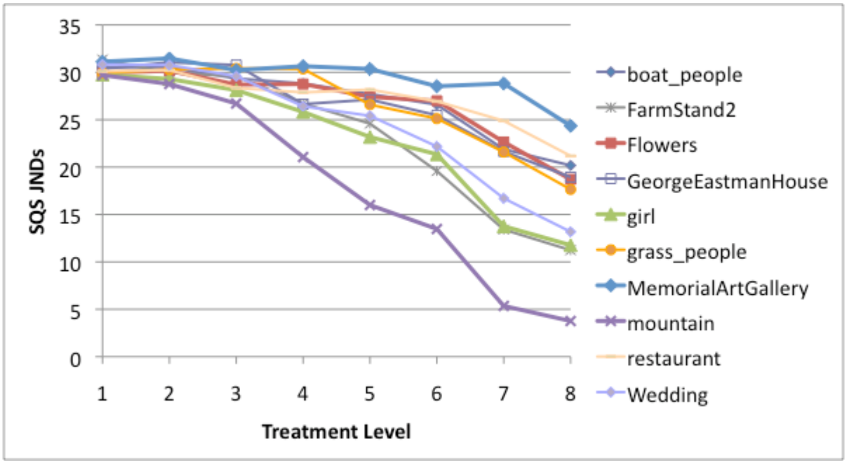


Figure 4: SQS_2 JNDs for each of the 10 scenes in the original experiment following the softcopy quality ruler method in ISO 20462 Part 3. Flowers, Girl, Memorial Art Gallery, and Mountain were selected to represent the span of scene dependency in this paper. (N=17 observers)

For this paper, four scenes were selected to represent the span of the scene-dependency: Flowers, Girl, Memorial Art Gallery, and Mountain. In addition, a subset of the treatment levels were selected to simplify the experiment: levels 1, 4, 6, 7, and 8. Figure 5 shows the original experimental data for these subsets of the scenes and treatment levels, including the standard error of the subjective results. These subsets were chosen in order to provide an online test with an approximate completion time of 15 minutes. Short test times were thought necessary in order to lower the rate of incomplete tests, since the online tests were not monitored by a proctor to ensure completion.

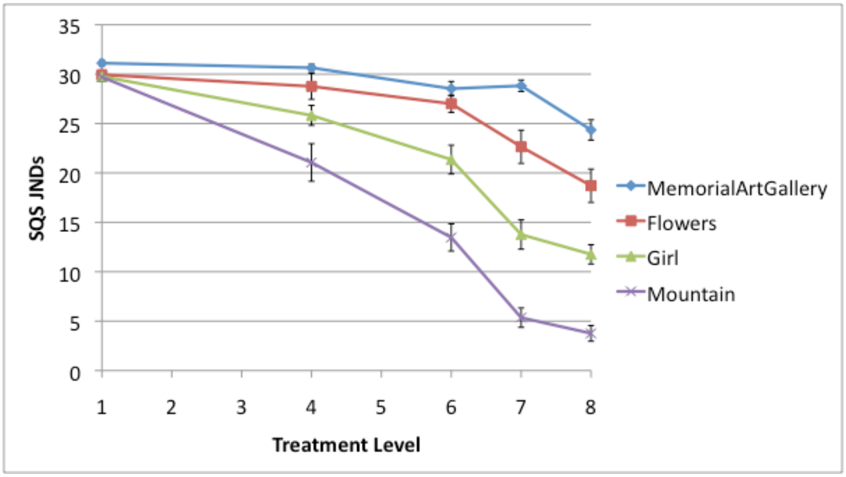


Figure 5: SQS_2 JNDs for the 4 scenes and treatments levels which served as the calibrated reference points for this paper. Error bars are standard errors.

3. ADAPTING THE RULER METHOD FOR DISTRIBUTED ASSESSMENT

Our interest in this study was to investigate whether and how the softcopy quality ruler method could be adapted for a simple Internet-browser based image display and evaluation. We chose to restrict our attention to simple, commercially available on-line survey products. In addition, we chose to evaluate several of the same scenes and treatment images that were used in the softcopy quality ruler study. Adaptation of the method for Internet-based evaluation and uncontrolled display hardware required two types of accommodation; image display and survey method.

3.1 Image Display

Since our image quality study involves the evaluation of image microstructure, control of the formatting of the image display is important. For example, it is necessary that images are not sub-sampled or interpolated when using common browser software programmes. So, while we need to accommodate common computer video display sizes and formats (N x M pixels), some image scrolling was preferable to resizing of image data. Ideally, of course, we wanted the simultaneous display of ruler and treatment images. In striking a balance, therefore, between the display of image content, accommodation of common computer display sizes and minimizing scrolling, it is important to choose a service for which displayed images can cover a large part of the displayed screen. This can be helped by the formatting of image-based survey questions being presented next to one another.

Another potential challenge to image-based visual studies is the limitation of the size of individual files used for the study. Several services that we investigated limit uploaded image files to much less that 1MB. However, linking of larger images by uniform (or universal) resource locator (URL) is common, and we used this method.

Given the above requirements for our study, we chose a service that is widely used, and includes the display of images for both free and paid subscription offerings, Survey Monkey® [5]. The display of modest sized images was done without resizing by the browser, based on our evaluation. Including image files by URL allowed uncompressed BMP-formatted image files to be used. With the limitations of common electronic displays, however, intentional reduction of our images was required. Rather than generate smaller versions of the ruler and noise-cleaned treatment images, we chose to subsample the files used in the original validation study. Figure 6 shows the cropped versions of the four scenes chosen from the validation study. The original images were either 1088 x 816, or 1253 x 834 pixels. We cropped all images to (525 x 330 pixels), based on the design of the user interface (discussed below) and ideal user display of approximately 1600 x 900 pixels.

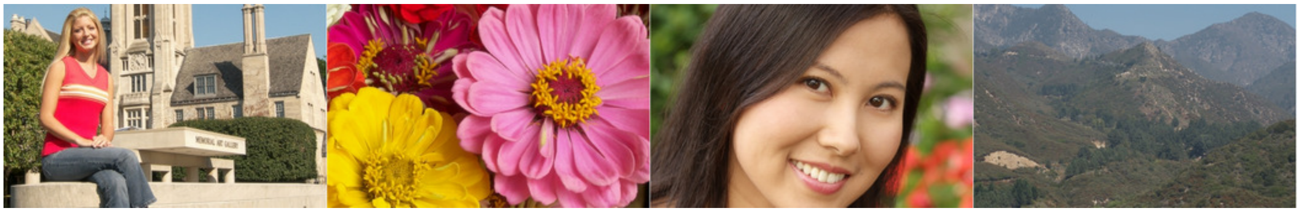
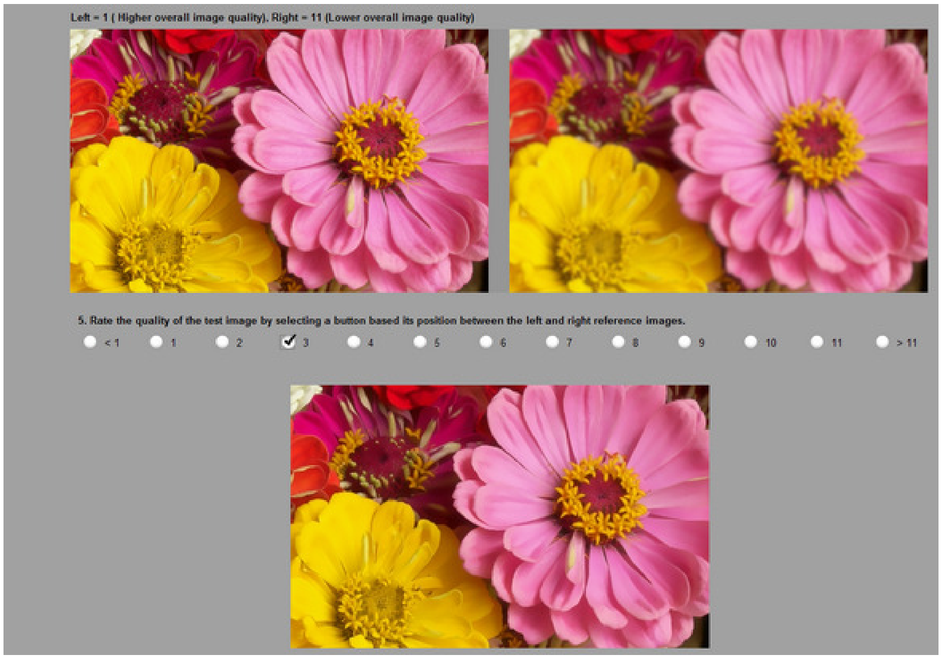


Figure 6: Cropped scenes used in the current study (525 x 330 pixels per scene).

3.2 Method Using Survey Monkey

As shown in Fig.1, the user interface for the softcopy quality ruler includes a slider for the control of the ruler image that is displayed. This was implemented using Matlab GUI programming with software callback functions. Being restricted to static image display, we chose to present two anchor ruler images per trial. From the question types that are available in Survey Monkey, we chose a forced-choice rating based on selection of a radio-button. Several buttons defined a scale that relates to two reference image quality ruler images. The user interface is shown in Fig. 7, with eleven button-levels defining the rating. The selection of a button level is intended to approximate the selection of a continuous value on the range defined by the slider position in Fig. 1. Note that the meaning of the number scale is for 1 being high quality and 11 being lower quality. In other words, they are in increasing amounts of image quality degradation.



Left = 1 (Higher overall image quality), Right = 11 (Lower overall image quality)

5. Rate the quality of the test image by selecting a button based its position between the left and right reference images.

○ < 1 ○ 1 ○ 2 ☑ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 ○ 11 ○ > 11

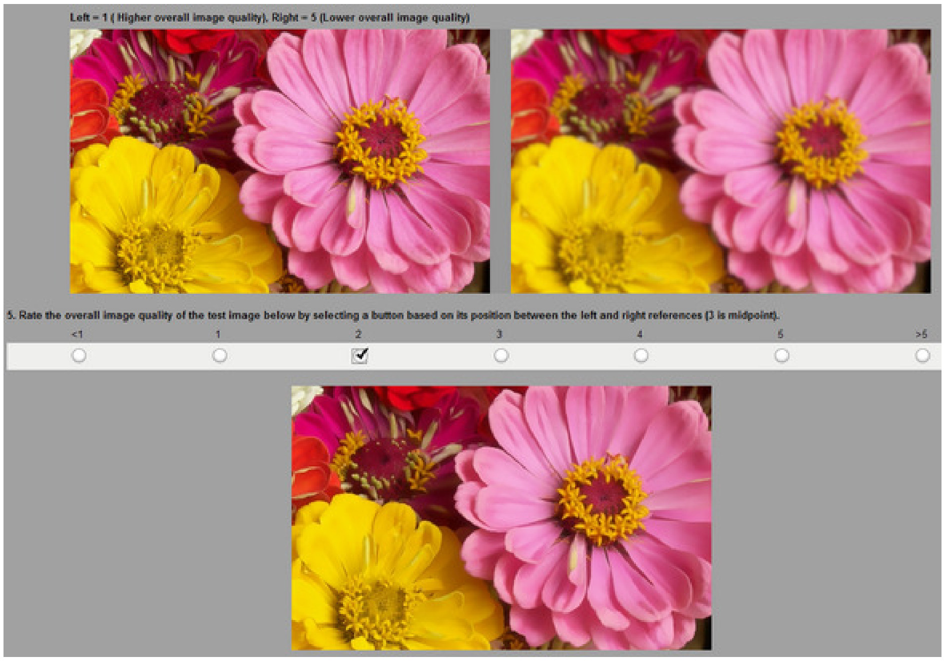
Figure 7: User interface for the browser-based study based on Survey Monkey. Here button 3 was selected because the (lower) test image was judged closer to the left reference than the one on the right in overall image quality. This is the 21-button method (see text).

3.3 Study Design

For our study, we chose four of the scenes used in the texture blur validation study, as shown and as cropped in Fig. 6. As described above, we selected five of the eight noise-cleaning treatment levels used in the previous study. Each test sample was rated using two pairs of reference anchors such that one pair was the higher overall quality half of the SQS_2 and another pair was the lower overall quality half of the SQS_2. The anchor images were selected from the previously developed softcopy quality ruler set for SQS_2 levels, 1 (low quality), 15, and 30 (high quality). So each test image was compared with paired ruler levels 1 and 15, and with the pair 15 and 30. For a first set of observers, we used scales with eleven levels. When a sample is compared with the higher-quality references, the buttons are labeled <1, 1,…,11,>11. When a sample is compared with the lower quality anchors, they are labeled <11, 11,…, 21, >21. Thus, we refer to this as the 21-button method. In addition to these test images, the reference anchor images were included in the study trials as null images. This was used to evaluate whether observers understood the visual task well enough to closely match the image quality of identical images. Almost all observers who completed the study were able to match these anchor test images closely. A second experiment used fewer buttons, as explained next.

3.4 Modified On-line Survey

Following the initial 21-button rating test, several observers felt having fewer buttons would make the task easier. A modified user interface with a five-button rating scale per trial was then run for a second group of both expert and non-expert observers. The interface is shown in Fig. 8. Note that the anchor images remained the same, so the results from both experiments could easily be compared. Since each test image is compared to two pairs of anchors (the high, mid quality pair, and the mid, low quality pair), the range of button labels is 1–5, and 5-9, respectively. Therefore we refer to this as the 9-button method.



Left = 1 (Higher overall image quality), Right = 5 (Lower overall image quality)

5. Rate the overall image quality of the test image below by selecting a button based on its position between the left and right references (3 is midpoint).

Figure 8: User interface presented with fewer buttons. This is the 9-button method (see text).

4. RESULTS

Results are described as follows:

a. Initial subjective results with 21 overall quality buttons for the observers to select. These observers represented a range of population with and without involvement in the imaging industry.

b. Subjects results with the reduction to 9 overall quality buttons for the observers to select. These observers also represented a range of population with and without involvement in the imaging industry, though not the same observers as in (a).

c. Subjective results with the 9 overall quality buttons for observers with direct ties to the imaging industry and research arena.

For each of the online methods, the average button values were converted to inferred SQS_2 JND values by utilizing the reference SQS_2 JNDs associated with a 25-inch viewing distance.

Figure 9 shows the subjective results for the first online method with 21 overall quality buttons. There were 21 observers for this data set, representing a range of population with and without involvement in the imaging industry. Note that these results clearly exhibit the scene dependency observed in the reference experimental data; four distinct responses exist for the four scenes. As compared to the reference values in Figure 5, the online experimental data show similar order in that the Memorial Art Gallery has the least degraded overall quality as the noise cleaning treatments increase and the Mountain scene has the most degraded overall quality.

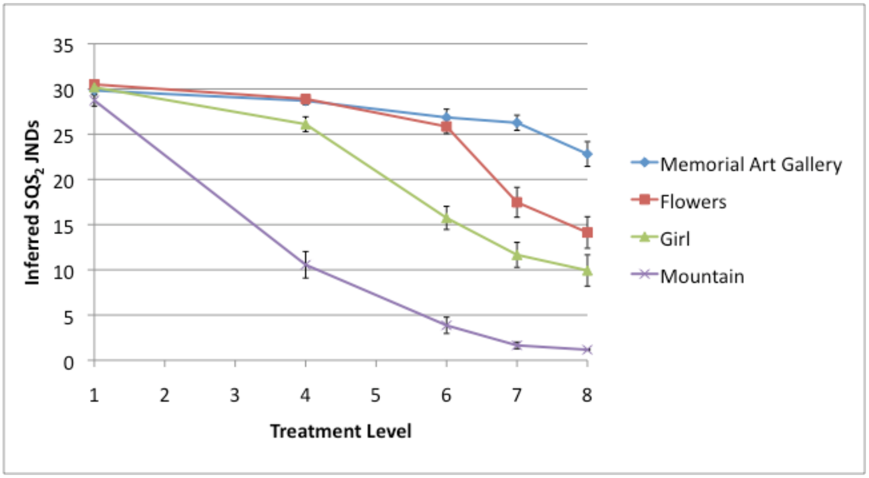


Figure 9: Inferred SQS_2 JNDs for the 4 scenes and treatments levels for 21 buttons of overall quality. 21 observers representing a range of population with and without involvement in the imaging industry participated in this portion of the experiment. Error bars are standard errors.

As described above, the online interface was modified to 9 buttons representing overall image quality. Figure 10 shows the subjective results for this portion of the experiment. There were 11 observers for this data set, representing a similar range of population with and without involvement in the imaging industry, though none of the same observers as for the 21-button interface portion of the experiment. Note the similarities between Figures 9 and 10. However, the 9-button approach did result in a quantization of the subjective responses. Three elements are highlighted: the larger spread in the subjective response for level 1 treatment of the test scenes, the lack of consistent monotonic loss in overall quality with treatment level increase for the Girl and Mountain scenes, and the globally increased standard error.

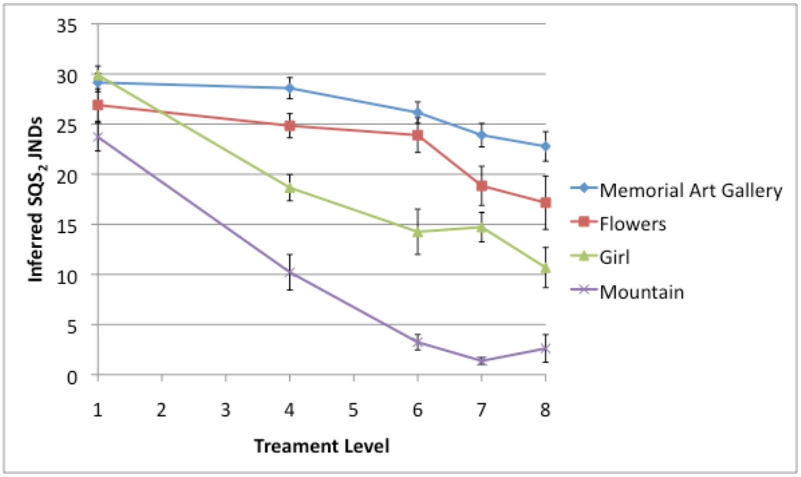


Figure 10: Inferred SQS_2 JNDs for the 4 scenes and treatments levels for 9 buttons of overall quality. 11 observers representing a range of population with and without involvement in the imaging industry participated in this portion of the experiment. Error bars are standard errors.

Finally, the results of the 31 observers with direct ties to the imaging industry and research arena appear in Figure 11. Notice the similar trends to the results from the population of observers with and without involvement in the imaging industry in Figures 9 and 10—similar separation and order of the four scenes. But, also note that the results in Figure 11 show stronger degradation in overall image quality as the strength of the noise cleaning treatment increases. This is discussed further with respect to Figure 12 and following.

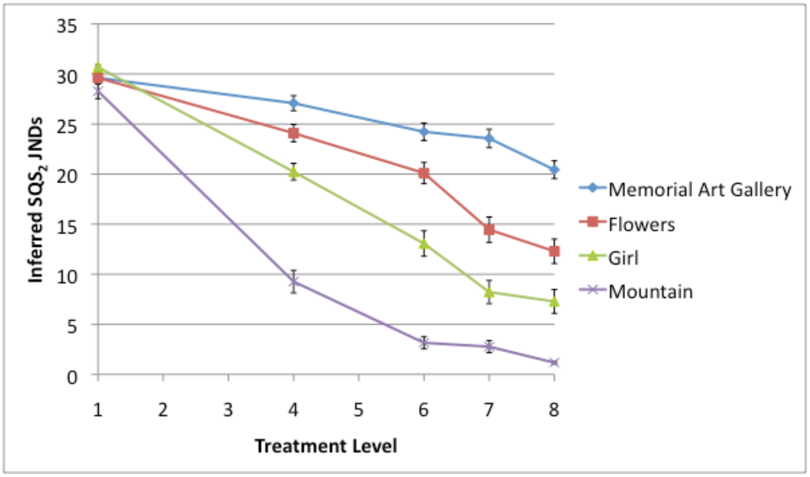


Figure 11: Inferred SQS_2 JNDs for the 4 scenes and treatments levels for 9 buttons of overall quality. 31 observers with direct ties to the imaging industry and research arena participated in this portion of the experiment. Error bars are standard errors.

In order to compare our study's results with those of the original ISO 20462 softcopy quality ruler method, the data for the four scenes were combined for each treatment level (This averaging approach was utilized in the original reference paper [4] and was found to be useful and clearer for visual comparison). Figure 12 is a comparison of the averaged values for the reference SQS_2 JND values those from our Internet-based method. One might have expected the uncontrolled viewing distance, unspecified monitor pixel pitch, and variable viewing environment to result in less perceived overall quality degradation compared to the control experiment. However, all of our Survey Monkey results point to more degraded overall quality compared to the original ISO 20462 results with the controlled quality ruler method.

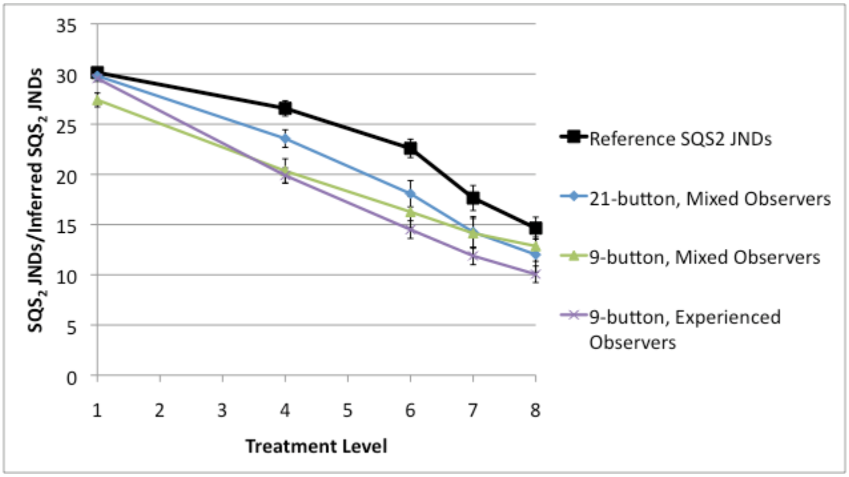


Figure 12: Comparison of the inferred SQS_2 JND results in this paper to the reference experiment with calibrated SQS_2 JNDs. Error bars are standard error.

Several factors could be contributors to this. In order to fit ruler pairs and a test image on the same screen view, image space was limited. To minimize the difference of the spatial component of the stimuli between reference stimuli and the Survey Monkey presentation, we choose to avoid resampling the quality ruler images by cropping out the salient regions of each scene. However, this subsequently forced the observers to make their assessments solely on the salient portions of the four selected scenes with respect to texture blur. Thus, the visual impact of the less salient regions was not incorporated into the judgments in Survey Monkey, were in the control experiment. Another possible contributing factor is the uncontrolled viewing distance. The observers could move closer to their monitors when needing to make judgments. While we did use JND values associated with the closest distance provided in the ISO method, *i.e.*, 25-inch viewing distance, our observers may have been viewing the images at considerably closer distances. Thus image degradation could have been more apparent in the test stimuli.

5. CONCLUSIONS

The extended ISO 20462 method for image quality evaluation presents reference and test scenes on computer monitors as Digital Reference Stimuli (DRS). The use of this validated method requires calibrated monitors and controlled viewing conditions. In this study, we have adapted the ISO 20462 softcopy quality ruler (SCQ) method for presentation in Internet-based subjective evaluations. This was done using a commercial online survey service, Survey Monkey. In developing our method we need to address several limitations, including image file size and static image display. Our method, which uses pairs of anchor images, and a set of rating levels rather than a variable ruler, was applied to a subset of test images from a previous texture-blur experiment.

Despite the use of cropped scenes and uncontrolled display hardware and viewing conditions, we obtained results that were comparable to those from the ISO SCQ study. For the four scenes used, very similar scene-dependent image quality differences were observed. Our results, however, indicated lower absolute quality judgments when compared to the ISO method. Several factors could be contributors to this including the use of cropped scenes, and the uncontrolled viewing distance. When the number of rating (button) levels was reduced, we observed an apparent quantization, or rounding, error being in introduced into the results. Nevertheless, our results appeared consistent (with low standard errors) with image-knowledgeable observers indicating a somewhat lower image quality rating than non-experts, for the same treatments. We conclude that Internet-based evaluation of this type is useful and can be related to the ISO method, particularly when common ‘field’ image viewing is the intent. When calibrated subjective image quality measures are needed, our adapted method should be considered an efficient alternative to the ISO 20462 standard, provided that common reference images are used.

Further exploration of this Internet-based evaluation for non-spatially dependent attributes such as color quality would provide an interesting comparison. For such attributes, the impact of the viewing distance variability amongst users would be minimized. Presumably, this would result in observer results even more comparable to the SQS_2 JNDs than those we observed with our study of the spatially dependent texture blur attribute.

REFERENCES

[1] ISO 20462-3, “Photography – Psychophysical experimental methods for estimating image quality – Part 3: Quality ruler method”, ISO, (2012).

[2] Ribeiro, F., Florencio, D. and V. Nascimento, V., “Crowdsourcing subjective image quality evaluation,” Proc. 18th IEEE Internat. Conf. on Image Processing, 3158-3161, (2011).

[3] Jin, E., Keelan, B., Chen, J., Phillips, J., Chen, Y., “Softcopy quality ruler method: implementation and validation,” Proc. SPIE 7242, 724206 (2009).

[4] Phillips, J., Coppola, S., Jin, E., Chen, Y., Clark, J. and Mauer, T., “Correlating objective and subjective evaluation of texture appearance with applications to camera phone imaging,” Proc. SPIE 7242, 724207-1 (2009).

[5] http://www.surveymonkey.com