Comparing Subjective Image Quality Measurement Methods for the Creation of Public Databases

Judith Redi*[a], Hantao Liu[b], Hani Alers[b], Rodolfo Zunino[a], and Ingrid Heynderickx[b,c]
[a]University of Genoa, DIBE, Via Opera Pia 11a, Genova, Italy 16145
[b]Delft University of Technology, Mekelweg 4, Delft, The Netherlands 2628 CD
[c]Philips Research Laboratories, Prof. Holstlaan 4, Eindhoven, The Netherlands 5656 AA

ABSTRACT

The Single Stimulus (SS) method is often chosen to collect subjective data testing no-reference objective metrics, as it is straightforward to implement and well standardized. At the same time, it exhibits some drawbacks; spread between different assessors is relatively large, and the measured ratings depend on the quality range spanned by the test samples, hence the results from different experiments cannot easily be merged . The Quality Ruler (QR) method has been proposed to overcome these inconveniences. This paper compares the performance of the SS and QR method for pictures impaired by Gaussian blur. The research goal is, on one hand, to analyze the advantages and disadvantages of both methods for quality assessment and, on the other, to make quality data of blur impaired images publicly available. The obtained results show that the confidence intervals of the QR scores are narrower than those of the SS scores. This indicates that the QR method enhances consistency across assessors. Moreover, QR scores exhibit a higher linear correlation with the distortion applied. In summary, for the purpose of building datasets of subjective quality, the QR approach seems promising from the viewpoint of both consistency and repeatability.

Keywords: Image Quality, Subjective Quality Assessment, Psychometrics, Single Stimulus, Quality Ruler

1. INTRODUCTION

The control and enhancement of image and video quality have become crucial in the design and development of displays. With the development of digital imaging and the introduction of Internet as a source of multimedia content, the range of possible distortions affecting quality has widened, calling for new, more effective post-processing video chains. In the process leading to image correction and enhancement, the detection of artifacts and the quantification of their impact on the image-quality level is of paramount importance. Image Quality Assessment (IQA) algorithms aim at consistently reflecting human quality perception in order to provide reliable estimates of the quality level of the images being processed for enhancement.

Typically, the development of IQA algorithms is supported by subjective studies, directly involving humans and aiming at measuring their quality perception. Subjective testing is the most reliable methodology allowing a better understanding of the mechanisms underlying quality perception, providing useful information for the subsequent modeling phase. On the other hand, subjective testing is expensive and time-consuming, and therefore, often performed only for a specific (limited) quality aspect. For testing IQA algorithms, however, large sets of subjective data are needed, especially because, as pointed out in [1], the performance of quality metrics can strongly depend on the database used for testing. Performing subjective experiments is then often mandatory, due to the lack of a sufficient number of publicly available subjective data or because the reliability of the existing data is not easily verifiable. These observations point out the urgency for the community to (1) share more data, in order to allow easier and less expensive testing of IQA algorithms, and (2) define benchmarks, so that new IQA algorithms can be presented together with a standard, reliable validation.

Some publicly available databases of subjective quality data already exist. A first example is the LIVE Database [2]. In its second and more recent release, images affected by several distortions (Jpeg and JPEG 2000 compressions, Gaussian blur, White Noise and Fast fading) are provided together with the respective Difference Mean Opinion Score (DMOS). These scores were obtained through a single stimulus experiment based on a categorical sorting using a continuous quality scale. LIVE release II is nowadays one of the most widely used databases for IQA algorithm validation. Other

Image Quality and System Performance VII, edited by Susan P. Farnand, Frans Gaykema, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7529, 752903 · © 2010 SPIE-IS&T · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.839195

examples of publicly available databases of subjective quality are the IRCCyN/IVC [3] and the Toyama [4] databases. The first one adopted the Double Stimulus Impairment Scale (DSIS) methodology, whereas for the last one the Single Stimulus Absolute Category Rating (SSACR) was used. Finally, recently the Tampere Image Database [5, 6] has been released, containing 17 different kinds of distortion and a total of 1700 evaluated images. For this database, a pair-wise comparison method, followed by a validation through a single stimulus quality rating was used. All the cited examples represent a precious asset for the community. Though, each database was created using different methodologies and environmental settings; hence, subjective scores are neither comparable nor can be merged in a single, larger database allowing more reliable testing of IQA algorithms. In this scenario the need for standardization becomes evident.

In 2002, the International Communication Union reviewed its recommendation for the subjective assessment of the quality of television pictures in [7]. This recommendation mentions five methodologies as being reliable for image and video quality assessment, including both double- and single-stimulus approaches. The ITU organization also advised (in [8]) the most suitable methodologies for the subjective audiovisual quality assessment for multimedia applications. The International Standard Organization (ISO) introduced in 2004 [9] the ISO 20462 standard for psychophysical image quality measurements. The standard presents two methodologies, the Triplet Comparison and the Quality Ruler, to be used for measurements over small and wider ranges of quality respectively. More and more effort is devoted to developing or refining psychometric methodologies (by VQEG, for example [10]) in order to overcome issues such as bias in evaluations or inconsistencies. Despite that, researchers are far from agreeing upon a standard methodology, to be used extensively to create wider and wider subjective quality databases.

To contribute to this effort, in this paper two popular psychometric methodologies are compared. The Single Stimulus method [7] and the Quality Ruler method [9, 11, 12] are selected as representative methodologies, and tested on pictures impaired with Gaussian Blur. Two sets of images impaired with Gaussian Blur are used: a subset of the LIVE dataset for Gaussian Blur (LS), and a set of highly textured images (HTI), spanning roughly the same range of quality (see fig.1). The two sets of images are assessed with the two methodologies by two groups of people, each asked to assess each set of images with a different method. The aim of the paper is twofold: (1) to get more insight in the advantages and disadvantages of both methods for quality assessment, and (2) to make quality data of blur impaired images publicly available.

The Single Stimulus (SS) method with continuous quality scaling is often preferred by researchers over other methodologies, as it is straightforward to implement and well standardized. However, it also has some inconveniences, such as (1) the difficulty for the observers to give a numerical value for quality without having a reference, and (2) the dependency of the obtained values on the quality range spanned by the test samples, which implies that scores obtained in experiments involving different sets of images cannot be accumulated [13]. The Quality Ruler (QR) method claims and in specific cases has already shown to overcome these inconveniences. It is based on the use of a set of reference images of known quality that are evenly distributed along a pre-calibrated quality scale (Standard Quality Scale, SQS). The quality scores assessed for new images then correspond to their position along such a scale, providing a JND-based measure of visual quality.

The remainder of this paper is organized as follows: In section 2, the psychometric methodologies involved in the comparison are presented. Section 3 reports about the practical implementation of the two psychometric tools. The experimental setup and results are reported in section 4. A validation experiment is then described in section 5, and finally, conclusions are drawn in section 6.

Figure 1. Original images for the two sets (LIVE Subset (a) and High Textured Images (b)) involved in the subjective experiments. Each original image was impaired using a circular-symmetric 2-D Gaussian kernel of standard deviation σ_B pixels on the three components R, G and B, for six different values of σ_B.

2. PSYCHOVISUAL EVALUATION METHODS FOR PUBLIC DATASETS CREATION

Producing public databases of subjective image quality data for IQA algorithm evaluation is not a trivial task. Preferably these databases contain stimuli of variable content, impaired by a broad set of artifacts over a wide range of quality levels. As a consequence, exhaustive subjective testing is needed.

The methodology to be adopted for this subjective testing should be selected carefully, evaluating several aspects. According to Engeldrum [14], the confusion level within the set of stimuli and the effort required to the observer for judging are two critical elements when selecting a psychometric method. The confusion level is determined by how closely the stimuli are spaced in quality. The narrower the range of quality is, the higher is the probability of inducing confusion (disagreement) in across-observers judgments. The psychometric method should be chosen according to the required degree of robustness to confusion in the set of stimuli. The effort required to the observers during an experiment is instead determined by the number of judgments needed, and therefore by the number of stimuli involved. Methods requiring a high number of judgments per stimulus are not suitable for experiments involving large datasets. Other aspects to be taken into account are that the methods should allow an easy implementation and a simple (and reliable)

Table 1. Compliance of the methods selected for the study to criteria selected to evaluate the suitability of a methodology for the creation of a wide database of image quality data

Criterion	*Single Stimulus (SS)*	*Quality Ruler (QR)*
Standardization level	High	High
Confusion in the test set	Low	Low
Required effort per sample	Low	Medium-Low
Implementation easiness	Yes	Requires considerable effort for the first implementation
Result analysis easiness	Yes	Yes
Psychophysical significance of the results	Low	High (JNDs)
Confidence in the comparison of different experiment	Low	High

analysis of their results, enabling fast releases of new data. Moreover, the provided results should be consistent and psychophysically significant, i.e. expressed in psychometric relevant quantities [11]. Finally, the level of standardization of the methodology should be high enough to allow repetitions of the experiment under the same conditions. Related to that, a major advantage would be the possibility to merge measurements obtained in different experiments in a single database. Given the above considerations, and the fact that the considered experiments would likely involve a large number of stimuli, reasonably spanning a wide range of quality, the SS with numerical continuous scaling method [7] and the QR [9, 11] method were chosen among the valuable methodologies recently developed and standardized [7, 8, 9, 10]. Table 1 reports the compliance of the chosen methodologies to the proposed criteria.

2.1 Single Stimulus Methodology

The SS method is one of the methodologies most widely used among visual quality researchers evaluating IQA algorithms [2, 4, 5, 15, 16]. Formalized in the ITU-R BT.500-11 and ITU-T P.910 recommendations, the SS method is widely appreciated for its intrinsic simplicity, both in setting it up and in the observer's task. Despite of its simplicity it still allows to obtain reliable results.

The method is based on the presentation of a set of stimuli one at a time, with the possibility of including a reference image in the set, without explicitly informing the observer of its presence (hidden reference). Observers are asked to evaluate either the quality or the impairment level of each stimulus. Scores can be expressed as a predefined category (Absolute Category Rating, Numerical Category Rating), or as a position along a continuous scale. Hence, the setup of the method is straightforward requiring a single device to show the stimuli and some tool to allow scoring. The observer task is very intuitive and easy to understand, although it might be difficult for the observers to calibrate their judgment criterion without a reference, if not well trained in advance. Since a single judgment is required per assessment, the effort required to the observer is relatively low, allowing the inclusion of a higher number of stimuli.

The analysis of the results, performed as in [7], brings an average score per stimulus, expressed in the scale used for the experiment. These scores reflect human preference, though do not have a precise psychophysical meaning. Indeed, the obtained scores may vary with the definition of the scale [14], as well as with the composition of the set of stimuli (i.e. the quality range spanned by the stimuli) [17]. This suggests that comparing result of different experiments might be problematic, possibly inducing inconsistencies when merging these data in a single, larger database.

2.2 Quality Ruler Method

The QR method was first described by Keelan in 2002, and subsequently adopted as an international ISO standard for psychometric experiments for image quality estimation [9]. The method is an evolution of the classic linear scaling approach, and aims at overcoming what Parducci [17] called the range-effect, which is the tendency of the observers to judge each stimulus relative to the range spanned by the whole tested set. The core idea of the QR method is to provide the observer with a set of reference images, anchored along the scoring scale so that they are closely spaced in quality but together span a wide range of quality. This aims at guaranteeing reliable results when assessing large sets of stimuli spanning a wide range of quality. The close spacing of the reference images should allow the observer to score with higher confidence, decreasing the risk of inversions and range effects. A quality ruler has three main characteristics:

1. It is composed of a series of reference images, whose scale value is known, and that are closely spaced in quality, but span a wide range of quality altogether;
2. The references are presented in a way that easily allows detection of the quality difference between them, allowing the observer to find the reference image closest in quality to the test stimulus by visual matching;
3. The reference images depict a single scene and vary in only one perceptual attribute.

Underlying the reference images is a scale, called the Standard Quality Scale (SQS), which unit is one JND in quality and its zero point corresponds to an image with little informative content. Observers are asked to position the test stimulus on the SQS by visually matching its quality level to one of the reference images. In that way, the observer actually performs several comparisons (i.e. depending on the number of reference images included in the SQS) to complete a single assessment. An average SQS score per stimulus can be easily computed, and yields a quality value directly expressed in JNDs. Moreover, scores derived from different experiments but based on the same SQS, can be easily compared and merged into a larger database. On the other hand, the process that leads to the SQS definition is complex and delicate, and requires a considerable effort in the implementation stage.

3. METHODS IMPLEMENTATION

Both methods were implemented in a softcopy version. Details on the implementation are given below. The quality ruler was implemented from scratch, repeating the SQS calibration process.

3.1 Single Stimulus Method implementation

In our study, the quantity to be measured was overall quality. A continuous numerical scale of quality ranging from 0 (worst quality) to 100 (best quality) was used. A graphical user interface was developed in Java Swing to facilitate the scoring task. The quality scale, including numerical labels and additional semantic labels (i.e. "very low", "medium", and "very high") at intermediate points for reference, was shown by the graphical user interface below each stimulus, as illustrated in Figure 2.a. Ticks representing units were depicted along the scales, decades were instead marked with numbers.

3.2 Quality Ruler Implementation

Keelan (in [11], and subsequently in ISO 20462) provides technical details for the definition of the SQS, the generation of the reference images, and the implementation of a hardcopy and a softcopy quality ruler. Although the standard advises to generate reference images by varying the Modulation Transfer Function according to a predefined function, for this study the whole empirical procedure described in [11] to create the SQS was repeated from scratch. This choice was based on a twofold motivation: (1) the study partly focuses on testing the validity of the method itself, independent of the quality scale adopted, and (2) at the moment of our quality ruler implementation, no softcopy (i.e. digital) reference images were available[1]. Also, repeating the SQS calibration was useful to test the reliability of the SQS definition procedure itself.

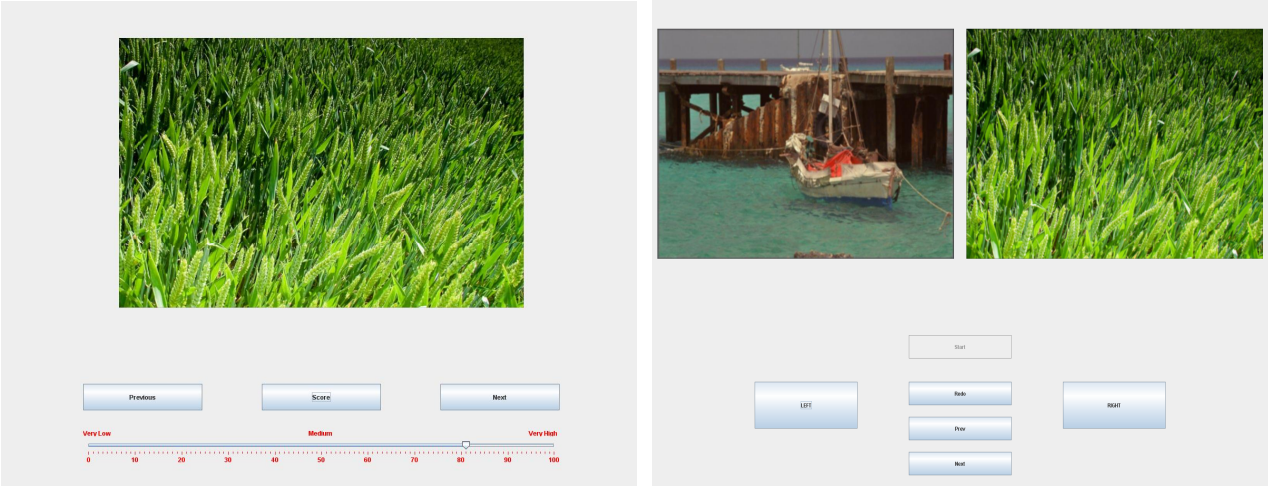
Two preliminary experiments were performed to address the effect of image content and artifact variation on the reliability and consistency of the quality ruler. Several un-calibrated soft-copy rulers were created by degrading different source images (of different content) with different artifacts at ten quality levels. Position of the reference images along the quality scale were defined only on the basis of the objective distortion value (e.g. the width of the Gaussian kernel in the case of blur, or the standard deviation of the White Gaussian Noise added to the stimuli). A set of test stimuli including multivariate images was designed to be scored by the observers (naïve for the first experiment, experts for the second) with the different un-calibrated rulers. Although a learning and stress effect was detected, its influence on the data was considered sufficiently small, and the data were reliable enough to draw conclusions for the further design and implementation of the quality ruler. The ruler based on the scene *Sailing_4* from LIVE and varying in sharpness was most effective in terms of confidence in the obtained ratings. Thus, the final ruler was constructed with reference images based on these two characteristics, and using the blurring kernel width (σ_B) as objective metric for degrading the quality.

Keelan advises a seven-step procedure to create the SQS, needed to select the reference images for the quality ruler. For our implementation we decided to reduce the procedure to five steps, since we already had some reliable data from the experiments mentioned above. The quality ruler was then generated by performing the following steps:

1. Preparation of a set of stimuli varying in the degree of Gaussian blur applied; to ensure a multivariate JND calibration, multivariate stimuli from the preliminary experiments were also included in the test set.
2. Design of a paired comparison experiment, to define the multivariate JND interval. To minimize the number of required comparisons, pairs that certainly would be unanimous decisions were excluded from the experiment (data were derived from previous experiments).
3. Paired comparison experiment performance and extraction of the angular deviates z_a.
4. Interval scale extraction from the paired comparison outcomes. JND increments ($\Delta\Omega_J$) in objective metric units were found by applying the following equation

$$\Delta\Omega_J = \frac{z_a(p_c)}{\left(\frac{\partial z_a}{\partial \Delta\Omega}\right)} = \frac{0.656}{\left(\frac{\partial z_a}{\partial \Delta\Omega}\right)} \quad (1)$$

[1] A softcopy version of the ISO 20462 SQS is now available for purchase



(a) *(b)*

Figure 2. Screenshots of the Graphical User Interfaces used for the Single Stimuli experiments *(a)* and for the Quality Ruler Experiments *(b)*

where $\frac{\partial z_a}{\partial \Delta\Omega}$ is the slope of the linear regression of the angular deviates z_a against the objective metric difference $\Delta\Omega$.

The JND increment was deduced for small groups of stimuli closely spaced in quality, and then the values were interpolated by a first order polynomial obtaining the JND increment function Δs_J. Finally, the SQS parameterized on the objective blur distortion value was computed, by applying the equation

$$\iota(s) = \iota_r + \Delta\iota_J \cdot \int_{Sr}^{S} \frac{ds'}{\Delta s_J(s')} \tag{2}$$

choosing as reference value $\iota_r = 0$, a constant JND increment of $\Delta\iota_J = 1$, and integrating the JND increment function Δs_J between the reference scale value 0 and the metric value of interest. The zero point of the scale was set in correspondence of the lowest value computed (-16 JNDs), in order to get positive JND values along the final SQS.

5. Generation of the reference images for the final quality ruler according to the SQS values found.

The obtained quality ruler spanned in total 16 JNDs of overall quality. A JAVA Graphical User Interface was developed compliant to the standard specifications for the softcopy ruler, with the exception that reference and test images were shown on the same display. This choice proved to be appropriate, since the standard update for softcopy rulers [12] adopted this strategy as well.

4. EXPERIMENTAL COMPARISON OF METHODS

4.1 Image Material

Two sets of images were used: a set of highly textured images (HTI), and a subset of the LIVE dataset (LS). In both cases, the artifact under investigation was Gaussian blur. The width of the Gaussian blurring Kernel σ_B was defined as the objective distortion value. Both dataset spanned roughly the same range of quality.

A set of highly textured images was included with the purpose of creating a new benchmark for blur based IQA algorithms. Blur based IQA algorithms are often based on edge detection, followed by a measurement of the spread of such edges (assumed to be representative for the perceived level of sharpness). For critical image material, such as highly textured images, these metrics could partially fail, since edges are not well-defined and difficult to isolate. Available databases with blur impaired image content seldom include highly textured images. Hence, to alleviate this lack, we included in our experiment a set of images with highly textured content.

The HTI dataset contains 12 images, 8 of which have texture all over the image, while 4 combine texture with smooth areas (fig 1.b). All 12 original images are 768x512 pixels in size. To blur an image, the R, G, and B components were filtered using a circular-symmetric 2-D Gaussian kernel of standard deviation σ_B pixels. The values of σ_B ranged from 0.42 to 15 pixels. Each original image was blurred at five different levels, yielding to 72 stimuli, including the originals.

For the second dataset (LS) nine out of the 29 original images of the LIVE database were selected (see fig 1.a). The selection was based on the Pearson's correlation coefficient between the magnitude of the distortion σ_B and the corresponding subjective scores provided by LIVE. The six image contents with the lowest correlation (i.e. Caps, Church_and_capitol, Lighthouse, Painted_house, Rapids and Sailing_2) were selected. Three images were added to enlarge the quality range spanned (i.e. Bikes, Dancers and Woman_Hat). The 5 blurred versions of each original image as provided by LIVE were included in the LS dataset, which counted to 54 stimuli, including the original versions.

4.2 Experimental setup and general methodology

The participants of the study were recruited from the Delft University of Technology. A total of 36 subjects aged 23 – 40 years participated. They were divided into 2 groups (A and B) of 18 people each. Group A was asked to evaluate the LS dataset with the SS method (session A1) and the HTI dataset with the QR method (session A2); group B instead assessed the HTI dataset with the SS method (session B1) and the LS dataset with the QR method (session B2) (as reported in table 2). Each participant therefore did not assess the same stimulus twice. To avoid stress the assessment sessions 1 and 2 were performed at different times, separated by at least one hour. The order of execution of session 1 and 2 was balanced over participants. For all sessions, the images were presented in a different (randomized) order for every subject.

Table 2. Arrangement of the experimental tasks for groups A and B: each group assessed one dataset with the Quality Ruler method and one dataset with the Single Stimulus method

		Psychometric Method	
		SS	**QR**
Dataset	**LS**	A1	B2
	HTI	B1	A2

All subjects were orally instructed on their tasks and on the experimental procedure before they started the first session of the experiment. A written summary of the experimental outline and the software usage was also provided.
For all the experiments a Dell 24" LCD screen (native resolution of 1920 x 1200 pixels) was used. Subjects observed the images from a distance equal to twice the height of the screen, hence 70 cm. A chinrest was used to keep the distance fixed. The lighting settings were compliant to ITU BT.500 specifications. Environmental settings were kept consistent for all sessions.

4.3 Single Stimulus experimental procedure

Observers were asked to express their judgment by moving a slider along a scale ranging from 0 to 100. In order to minimize the range-frequency effects, subjects were accurately trained before performing their task. The training session was divided into two phases. First, a set of ten images, covering the same range of blur annoyance as used in the actual study but not included in the test set, were presented to the subject in order to familiarize him or her with how to use the range of the scoring scale. Then, a second set of 7 images was shown to the participant with the request to actually score them on the scoring scale. Also these images were different from those used in the actual experiment. After the training, the test images were shown in a random order to each subject in a separate session.

4.4 Quality Ruler experimental procedure

The observer task in this case consisted of a series of paired comparisons, between the reference images of the Quality ruler and each of the test images. The user was asked to select the sample with the highest quality between the reference and test image (shown simultaneously), by using the right and left arrow keys. An automated procedure based on a binary sorting algorithm selected then the next reference image to be shown to allow the refinement of the assessment. The procedure was repeated until the test stimulus was judged better than one reference image, but worse than its adjacent reference image in the SQS. The test sample was finally scored as the average quality value of the two adjacent reference images. A training session, consisting of several trial samples, was first performed by the observer to familiarize him/her with the task.

4.5 Experimental results

The data analysis was performed following the ITU-R BT.500 specifications, separately for each session (A1 and B1 for the SS method, and A2 and B2 for the QR method). No outlier observer was detected.

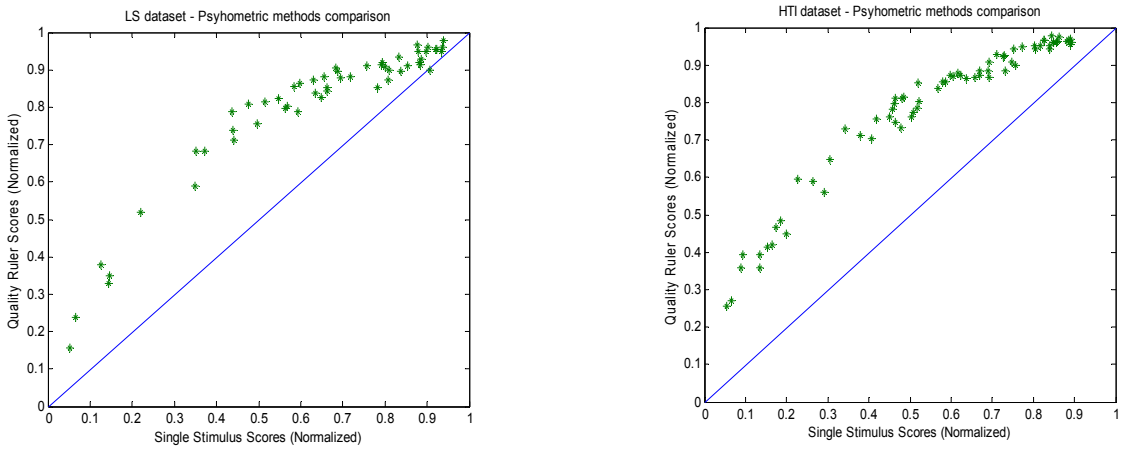
For comparison purposes, each QR score was remapped into the range [0, 100]. To evaluate the reliability of the measurements, the 95% confidence interval was calculated as an indicator of the agreement across observers on the judgment given to a single image. Two additional quantities were computed, namely the correlation of the subjective scores with the amount of objective distortion, and the RMSE for linear regression. Table 4 shows these quantities. A first, relevant outcome is that, for both datasets, the width of the confidence interval (averaged over all images in a dataset) is smaller for the QR scores than for the SS scores. The QR scores are also more correlated to the objective measure of distortion (i.e. the width of the blurring kernel) than the SS scores, and the RMSE for linear regression is smaller for the QR method than for the SS method.

Table 3.Relevant outcomes of the main comparison experiment. Average width of the confidence interval per score, RMSE for linear regression and Correlation of the scores with the objective distortion amount are reported for both datasets and both methods. QR scores were remapped into the range [0, 100] for comparison purposes.

		SINGLE STIMULUS	QUALITY RULER
LS Dataset	95% Confidence Interval avg. width	4.63	3.59
	RMSE for linear regression	13.33	4.21
	Pearson Correlation with objective distortion amount	-0.85	- 0.97
HTI Dataset	95% Confidence Interval avg. width	5.80	4.21
	RMSE for linear regression	13.82	7.97
	Pearson Correlation with objective distortion amount	- 0.84	- 0.92

All three indicators seem to favor the QR method in terms of higher measurement accuracy. Additionally, it should be noted that observers scored faster with the QR method than with the SS method. The majority of the observers also mentioned difficulties in scoring on a numerical scale without a reference, while they explained to be more confident in assessing quality with the QR method.

Fig. 3.a and 3.b show the scatter plot of the QR scores vs. the SS scores, for the LS and HTI datasets, respectively. These plots demonstrate that the QR method seems to overestimate the quality level of the stimuli with respect to the SS method. There are two possible explanations for this behavior. The overestimation could be due to inaccuracies in the SQS calibration. On the other hand, the observed behavior could also be a consequence of the range effect. When scoring without a reference as for the SS method, observers tend to use the whole scale for the whole test set of images. In the case of the QR method they use the reference images to assess the quality of a test image. Since these references



(a) LS dataset *(b) HTI dataset*

Figure 3. Scatter plots of the scores obtained with the two methods for the LS dataset *(a)* and for HTI dataset *(b)*

images span a broader range of quality than present in the test set, higher quality values for the QR method may be the result.

In conclusion, the results of the comparison experiment outline that:

a) The QR method seems to provide more consistent judgments across observers
b) Scores obtained with the QR method are highly correlated with the objective amount of blur distorting the image
c) The QR method seems to provide overestimates of the scores, or, from the opposite point of view, the SS method seems to underestimate the quality of the stimuli

These outcomes are of particular interest for one of the aims of this paper, which is establishing the reliability of the methods. Since it was not possible to draw definite conclusions from the available data, further empirical evidence was needed at this stage of the study. Moreover, from the outcome of this comparison experiment no confirmation was obtained on the robustness of the methods in providing consistent scores in different experiments. Hence, a second validation experiment was designed to better address the reliability and consistency issues.

5. VALIDATION EXPERIMENT

An additional experiment was planned to further evaluate the consistency of both methods and to test the suitability of both methods for accumulating results of different experiments (i.e. performed at different moments in time and involving different image material). The images selected for this second step spanned a reduced range of quality, corresponding to the higher part of the quality scale.

5.1 Image Material

The image material was selected from the Live Database Subset (LS) and the Highly Textured Images set (HTI), keeping the distinction between the two groups (from now on referred to as LS_V and HTI_V, respectively). To reduce the quality range, patterns scored as medium to high quality were selected, for two main reasons: (1) the overestimation phenomenon was more evident in that part of the scale (see fig 3), (2) the Spearman correlation between the scores obtained with the two methods was consistently lower for mid-to-high quality images.

The following procedure was adopted to perform the selection, for each dataset separately:

(I) The 40^{th} percentile of the distribution of the SS scores was computed

(II) All the images rated lower than the 40^{th} percentile were excluded from the list of candidate images, leaving the top 60% of the images to be considered for inclusion in the 2^{nd} experiment.

(III) Roughly half of the images in the subsets obtained in step (II) were selected to be included in LS_V and HTI_V, trying to keep a representative sampling of the quality range.

Additionally, the five distorted versions of the image "stream" were included in both reduced datasets, in order to evaluate a possible context effect. As a result, the HTI_V dataset counted 27 images, and the LS_V dataset 23 images.

5.2 Methodology

The experimental methodology, conditions and procedure (including the training) were identical to those adopted for the main experiment, for comparison purposes. The number of subjects involved was reduced to two groups of 10 subjects, all different from the previous experiment. This number was considered sufficient for the purposes of this second step. The two groups again performed different tasks, following the scheme given in Table 2.

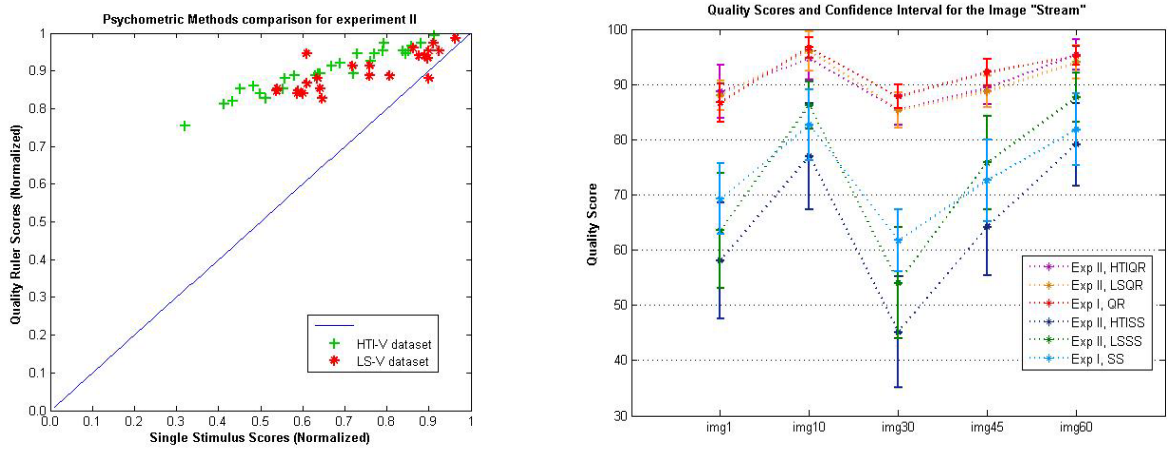
5.3 Experimental Results

The results of this validation experiment confirm several of the tendencies observed in experiment I. As shown in table 4, the consistency among observers (as indicated by the 95% Confidence Interval) is higher for the QR method than for the SS method. Conversely, the decrease in the correlation between the QR scores and the objective distortion value could indicate a lack of accuracy in the calibration of the higher part of the SQS. The QR quality overestimation is again observed (fig. 5.a). This behavior could be either a consequence of inaccuracies in the SQS or an outcome of a possible range-frequency effect in the SS scores. To clarify this point, the RSME of the subjective scores obtained in experiments I and II is evaluated, and found to be consistently lower for the QR method. Hence, while the same images are scored

Table 4.Relevant outcomes of the validation experiment. Average width of the confidence interval per score, RMSE for linear regression and Correlation of the scores with the objective distortion amount are reported for both datasets and both methods. QR scores were remapped into the range [0, 100] for comparison purposes.

		SINGLE STIMULUS	QUALITY RULER
LS Dataset	95% Confidence Interval avg. width	7.62	3.09
	RMSE for linear regression	7.47	2.92
	Pearson Correlation with objective distortion amount	-0.86	- 0.81
	RMSE Experiment I and II scores	6.60	5.03
HTI Dataset	95% Confidence Interval avg. width	9.84	4.29
	RMSE for linear regression	8.80	3.53
	Pearson Correlation with objective distortion amount	- 0.85	- 0.81
	RMSE Experiment I and II scores	8.78	2.31

quite consistently in different experiments, the SS method seems to be more susceptible to the composition of the dataset, since its RMSE of the linear regression of the scores of experiment I to the scores of experiment II is larger. This also confirms a higher suitability of the QR method for experiments conceived as sub-sessions of larger experiments, being less sensitive to the context (i.e. range effects and confusion in the samples), and hence, allowing a more reliable combination of data. This conclusion can be further validated analyzing the scores obtained for the five versions of the image "stream" included in both LS_V and HTI_V, as well as in the original HTI dataset of experiment I (see fig. 5.b). Across three different experiments (based on datasets covering different ranges of quality), the QR method provides reasonably similar scores, with acceptable confidence. The results obtained with the SS method instead show larger variations, spanning up to 1/6 of the quality scale.



(a) *(b)*

Figure 5. (a) Comparison of the QR and SS scores for the validation experiment. (b) Comparison of the subjective scores obtained for the five version of the content "Stream" when assessed with the LS_V and HTI_V datasets in experiment II and with the HTI dataset in experiment I, for both methods. Error bars refer to the 95% confidence interval.

6. CONCLUSIONS

Two popular psychometric methodologies, i.e. the Single Stimulus method and the Quality Ruler method, were compared, with the aim of establishing their suitability for experiments aimed at creating public databases of images with their quality score. Both methods were compared in subjective experiments with two sets of images varying in Gaussian Blur. The study revealed that, for the purpose of building publicly available data of subjective quality, the Single Stimulus method presents several drawbacks such as low confidence in the scores and susceptibility to range effects. Conversely, the Quality Ruler method is worth the implementation effort from a point of view of consistency and repeatability of the scores. On the other hand, its dependency on the calibration of the Standard Quality Scale may

represent a limitation and introduce inaccuracies (e.g. overestimation) in the measurements. All the subjective data collected during this study will be soon available for the community usage.

REFERENCES

1. Tourancheau, S., Autrusseau, F., Parvez Sazzad, Z.M., and Horita, Y., "Impact of subjective dataset on the performance of image quality metrics", Proc. of IEEE Int. Conf. on Image Processing 2008, 365-368 (2008)
2. Sheikh, H. R., Wang, Z., Cormack, L., and Bovik, A. C., "LIVE Image Quality Assessment Database Release 2," http://live.ece.utexas.edu/research/quality/subjective.htm (2005)
3. Le Callet, P., and Autrusseau, F., "Subjective quality assessment IRCCyN/IVC database," http://www.irccyn.ec-nantes.fr/ivcdb/ (2005).
4. Horita, Y., Kawayoke, Y., and Parvez Sazzad, Z. M., "Image quality evaluation database," ftp://guest@mict.eng.utoyama.ac.jp/.
5. Ponomarenko, N., Carli, C., Lukin, V., Egiazarian, K., Astola, J., Battisti, F. "Color Image Database for Evaluation of Image Quality Metrics", Proc. International Workshop on Multimedia Signal Processing 2008, 403-408 (2008)
6. Ponomarenko, N., Carli, C., Lukin, V., Egiazarian, K., Astola, J., Battisti, F., "Tampere Image Database - TID2008" http://www.ponomarenko.info/tid2008.
7. ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," Geneva, 2002.
8. ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications", Geneva (1998)
9. Keelan, B., and Urabe, H., "ISO 20462, A psychophysical image quality measurement standard," Proc. SPIE 5294, 181-189 (2004).
10. Baroncini, V., "New Tendencies in Subjective Video Quality Evaluation", IECIE Transactions on Fundamentals, vol. 11(89), 2933-2937 (2006).
11. Keelan, B., "Handbook of image quality: characterization and prediction," Marcel Dekker, Inc., New York, 2002.
12. Jin, E.W., Keelan, B.W., Chen, J., Phillips, J.B., and Chen, Y., "Softcopy quality ruler method: implementation and validation", *Proc. SPIE 7424*, 724206 (2009).
13. De Ridder, H., "Cognitive Issues in image quality measurement", *J Electronic Imaging*, 10(1), 47-55 (2001).
14. Engeldrum, P. G., "Psychometric Scaling: A Toolkit for Imaging Systems Development", Imcotek Press, Winchester, MA (2000)
15. D'Angelo, A., Pacitto, M., and Barni, M., "A psychovisual experiment on the use of Gibbs potential for the quality assessment of geometrically distorted images", Proc. SPIE 6806, pp. 16-680616-10 (2008)
16. Liu, H., Klomp, N, and Heynderickx, I., "A No-Reference Metric for Perceived Ringing", Proc, Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-09, Scottsdale (2009)
17. Parducci, A., and Perrett, L.F., "Category rating scales: Effects of relative spacing and frequency", Journal of Experimental Psychology Monograph, 89, 427-452 (1971)