Video Quality Ruler: A New Experimental Methodology for Assessing Video Quality

Pedro Garcia Freitas*, Judith A. Redi†, Mylène C.Q. Farias‡ and Alexandre F. Silva§
*§Department of Computer Science, University of Brasília, Brazil
†Department of Intelligent Systems, Delft University of Technology, Netherlands
‡Department of Electrical Engineering, University of Brasília, Brazil

***Abstract*—In this paper, we propose a subjective video quality assessment method called video quality ruler (VQR) that can be employed to determine the perceived quality of video sequences. The described method is an extension of the ISO 20462, which is a method to assess image quality. The VQR method provides an interface with a set of pictures. The subjects assess the video using these pictures as a scale and compare the subjective perceived video quality with their perceived quality. The pictures are calibrated to form a numerical scale in units of just noticeable differences (JNDs), which allows to analyze and compare both subjective video and image stimuli. To evaluate the effectiveness of the proposed method, we compare the VQR method with a well-used single stimulus (SS) method. The results show that proposed method can be used to quantify the overall video quality with higher efficiency and with a less biased results than the SS method.**

I. Introduction

Subjective visual quality assessments are crucial for designing reliable objective quality metrics. Subjective experiments are necessary to (1) observe perceptual and annoyance mechanisms in users when exposed to an impaired stimulus, to be modeled in objective metrics and (2) collecting data (subjective quality scores) to be used as a benchmark to test the accuracy of these metrics [1]. The reliability of subjective quality data is, therefore, a major precondition for the development of effective quality metrics. To collect subjective quality assessments, psychometric experiments are typically performed, often involving a set of participants (subjects) which are asked to judge the quality of a set of stimuli using a rating scale [2].

When concerned with measuring the quality of video material, several subjective video quality assessment methodologies are available [2, 3]. A main characteristic of subjective methodologies relates to the way in which stimuli are presented to the subjects. In Single Stimulus (SS) methodologies, subjects rate the quality of just one video clip (the test video), without having a reference. In Double Stimulus (DS) methodologies, subjects rate the quality or difference in quality between two or more videos presented simultaneously or closely spaced in time. Methodologies also differ with respect to the type of scale on which the stimulus is rated. Rating scales can be discrete or continuous, labeled or unlabeled, or with numbered rating points or categories [4]. Each methodology type has advantages and disadvantages. As stated by Engeldrum [5], it is practically impossible to cover all factors affecting the results of a scaling task and provide specific recommendations for each of them. There are common pitfalls in standardized quality assessment methodologies, such as the dependency of the scores on the range of quality spanned by the test samples [6] and the difficulty os subjects is to give a numerical (or categorical) value for quality [5], that can lead to imprecision in measurements and subject bias [1, 7]. Imprecision manifests itself as wide confidence intervals that cause problems in the discriminability of pairs of stimuli. Therefore, it is preferable to choose an experimental methodology that minimizes inter-subject variability of scores, hence maximizing confidence.

It has been shown that SS methodologies (e.g. ACR) and DS methodologies (e.g. DSIS) [2] yield similar confidence levels in Mean Opinion Scores (MOS). On the other hand, for image quality assessment, it has been shown that the Quality Ruler (QR) [3] methodology has advantages in this respect. The image Quality Ruler method is based on the use of a set of reference images that are evenly distributed along a pre-calibrated quality scale (the Standard Quality Scale - SQS). The task of the subjects is to find the image in the ruler whose quality matches that of the test image. The position of the matching ruler image on the SQS gives the quality score of the test image. The task of the subject is therefore reduced to a visual comparison (subjects decide whether the qualities of the ruler image and the test image match), which is simpler than giving a quality score [8]. As a result, the image Quality Ruler retains the advantages of methodologies purely based on visual comparison (such as Paired comparison [2]), but is less time-consuming since the set of comparisons to be performed per test stimulus is limited to the number of reference images in the ruler. The Quality Ruler makes it possible to estimate the quality of images within a large quality range [9] with higher confidence than SS methodologies [10]. In addition, the method has been shown to be less prone to context effects [11].

Considering these important advantages, we investigated the opportunity to extend the image Quality Ruler for *video* quality assessment. The main challenge to tackle here is how to allow the comparison of a video, which is dynamic, with a set of still images. Comparing pairs of images is straightforward since they are static and no details are missed when moving the focus of attention from one to the other, which cannot be recuperated by focusing back on the first image. This is not necessarily the case for video. Specifically, questions arise whether (1) subjects can match the quality of an image with that of a video and (2) the use of a set of images for comparison distracts the subject's attention from the video. In this paper, we report how we addressed this challenge and implemented a 'Video Quality Ruler' (VQR). To validate the method, we

978-1-4799-8958-4/15/$31.00 ©2015 IEEE

conducted an experiment in which we evaluated a set of videos with both the VQR and SS. Our results show how the VQR is a promising methodology to evaluate the subjective quality with high confidence.

The remainder of this paper is organized as follows. In Section II we describe the original image quality ruler methodology. In Sec. III, we describe the proposed VQR. The experimental setup and results of both experiments are reported in Sections IV and V, respectively. Finally, in Section VI, we present our conclusions and future work.

II. Standardized Quality Ruler Method

The QR method was first described by Keelan [8] and, subsequently, adopted as an international ISO standard for image quality assessment [3]. In this method, the subject compares a test image with a set of reference (ruler) images, anchored along a calibrated quality scale. Ruler images depict a single scene, varying in one perceptual attribute (e.g., blur). The images are closely spaced in quality, but the complete set spans a wide range of quality. Their presentation allows an easy detection of the quality differences between pairs while their close spacing allows subjects to score with higher confidence, what decreases the risk of inversions and range effects [8]. Ruler images differ in one Just Noticeable Difference (JND) of overall quality and are anchored along the so-called Standard Quality Scale (SQS). The zero point in the scale corresponds to an image with little informational content (i.e., extremely distorted), and its top end corresponds to a high-quality, unimpaired image. Obtaining ruler images is a complex process that requires delicate calibration [3, 8].

In QR, the subject task consists of positioning the test image on the SQS by visually comparing it to ruler images and deciding which ruler image matches the quality of the test image. As a result, the subject performs several comparisons to complete a single assessment, taking longer than traditional single or double stimulus methodologies, where the stimulus is compared with at most one reference and a quality judgment is performed. When the subject has found a match, a JND score corresponding to the numerical position of the match in the SQS is attributed to the test image. MOS can be computed by averaging the SQS values across the pool of subjects.

It has been shown that reducing the scoring task to a visual comparison decreases variability in the subjects' judgments, generating higher confidence levels for the MOS [10]. In addition, as long as the ruler images are kept the same, subjective scores obtained from a quality ruler experiment always refer to the range of quality expressed by the SQS, and not to the specific range spanned by the test stimuli. This limits the context effects [11], making the quality ruler also very appealing for video quality assessment.

Fig. 1 shows the interfaces for SS and QR-based *image* quality assessment. The SS interface (Fig. 1 (a)) presents a single image and a numerical scale in which the subject gives a numerical score. Fig. 1 (b) shows the QR interface. The left picture is the ruler image and the right picture is the test image. Subjects must compare a set of images on the left side with a single image on the right side and decide whether they are equivalents in terms of quality. Once a match has been found, the next test image is shown.

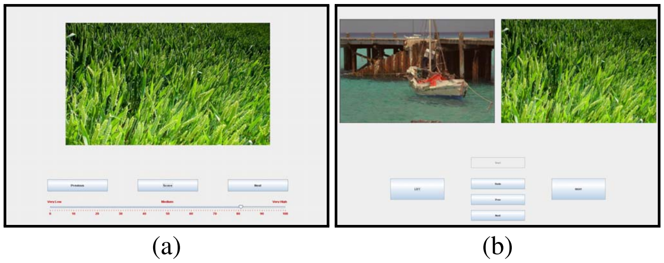

(a) (b)

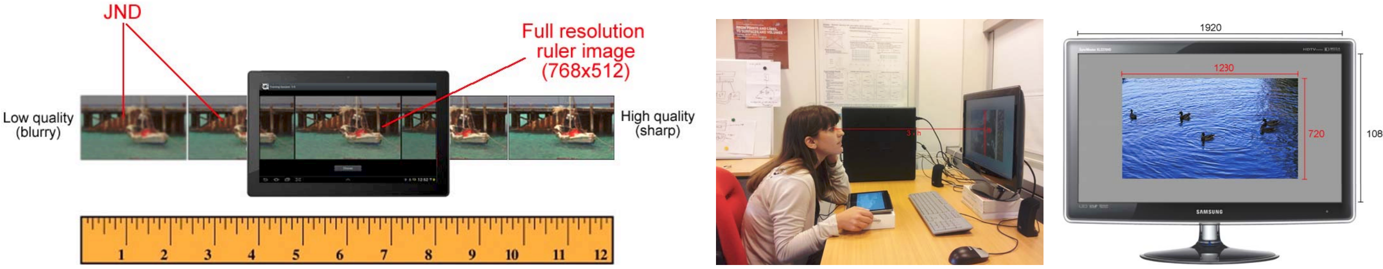
Fig. 1: Example of Graphical User Interfaces used for assessing image quality via (a) Single Stimulus and (b) Quality Ruler methods.

By looking at these interfaces, it seems clear that, whereas the transposition of SS methodologies to video quality assessment is straightforward, this is not the case for the QR. Showing the QR image along with a test video on the same screen would imply serious restrictions in terms of the maximum resolution of the videos to be evaluated. Showing them on two separate, adjacent screens, on the other hand, may force the subject to switch the focus of attention too often, between ruler images and test stimulus, possibly resulting in a lower noticeability of artifacts in the test video. Finally, a major question in whether subjects could match the image and the video quality in a meaningful way. Therefore, the construction of an interface which enables the strategy of QR method for video stimuli is challenging, and, for this reason, an adaptation to video of the QR has never been specified.

III. The Video Quality Ruler

The first challenge in the implementation of a Video Quality Ruler is related to the ruler stimuli and the SQS. Given that ruler images are the core of the QR methodology and that the calibrating process of the SQS is delicate and time-consuming, we use a set of calibrated ruler images used in a previous work of Redi et al. [10]. This ruler [10] includes 16 images, spanning a range of 15 JNDs (notice that JNDs in image quality may not directly map into JNDs of video quality; this remains an open question that demands further work). The ruler images depict the "sailing boat" shown in Fig. 1 (b), and their quality varies depending only on the amount of Gaussian blur applied to them. The choice of using only one artefact to vary the quality of the ruler images is based on the recommendation of Keelan [3], who showed how people are able to assess image quality using distinct artifacts on ruler and test images, and that Gaussian blur was most suitable for varying the ruler images. It is of course an open question whether this holds for video quality assessment, which we are going to investigate in our study.

Concerning the presentation of the ruler images, we chose to visualize them on a separate screen from the test videos to allow both to be displayed at their full resolution. To avoid creating issues in the stimuli visualization (and thereby in the artifact visibility) related to peripheral vision and viewing angle, we displayed the ruler on a tablet. This choice presents a set of advantages. First, a tablet can be placed outside the field of view of the subject while he/she is evaluating the video stimulus. Second, the interface is far more engaging than the traditional keyboard and mouse, allowing subjects to look through the ruler images by swiping the touchscreen. On the tablet, the ruler is implemented as depicted in Fig. 2 (a).



(a) (b) (c)

Fig. 2: Parts of the VQR apparatus. (a) Quality ruler implemented on a tablet screen. The images going out to the tablet represent the next or previous reference images. (b) Experimental setup of Video Quality Ruler. (c) Video playing on screen monitor.

The ruler images are sorted from the worst (left) to the best (right) quality. To perform the video evaluation, the subjects choose only one reference image per video. To find this image, subjects scroll the ruler to the right to increase the quality of the ruler images, and to the left to decrease it. When the subject judges that the quality of the video and the image is the same, she presses the button presented on the interface. This is repeated for each video evaluation.

The tablet was placed on a platform on the same plan where the monitor was standing. It was located perpendicularly to the monitor at a fixed distance of two and a half times the video's height. Fig. 2 (b) shows the schematic experimental setup of Video Quality Ruler experimental methodology. From this figure, we can notice that, when observing the video, the ruler is outside the subject's field of view and vice-versa. It should be noted that the tablet in this figure shows the instructions on how to proceed in the experiment ("guided practicing" stage). This figure illustrates the arrangement of the apparatus, but it does not display all original setup. The lights were brighten to take this photo.

When performing the scoring, the subject divides his/her attention between the ruler (tablet) and the test video (monitor screen), which may cause parts of the video stimulus to be missed. To avoid this, we disable the ruler (turning off the tablet screen) every time the subject sees a new video stimulus. Then, the ruler is enabled when the subject has completed an entire visualization of the video. At this point, the subject is allowed to distribute attention between the tablet and the monitor where the test video is repeated in a loop. Once a match is found, the tablet emits a sound before a new video is played in order to redirect the subjects' attention to the main monitor.

It is important to point out that the usage of the ruler in its tablet implementation may be too complex for subjects to understand. For this reason, along with an instruction phase, we introduced a "guided practicing" stage before the beginning of the experiment. This was split into three parts. First, a video tutorial describing how to use the interface was shown. Second, subjects tried out the tablet interface and inspected all reference images in the ruler. Third, subjects used the interface on the tablet to evaluate a few practice videos displayed on the monitor screen. Assessment provided in this practicing stage were not used in analysis and the videos used for training were not used in the actual experiment.

IV. Experimental Setup

We performed a subjective experiment to evaluate the proposed VQR methodology. The experimental test-videos were impaired with compression artifacts (i.e. blockiness and blurriness) combined with network impairments (packet-loss) at different strengths. Assessing the quality of videos impaired by multiple artifacts is a difficult task for subjects, as the annoyance of one artifact may be very difficult to compare to that of another one [8, 10]. We also performed the same experiment using a classic SS setup and a different pool of users. Our goal was to compare (between subjects) for the two methods their ability to provide MOS with a similar value and a sufficiently high confidence level. This section describes the generation of test sequences, the experimental setup, and the experimental protocols used for both experiments.

A. *Stimuli (test videos)*

We used seven high-definition videos from the VARIUM database [12] for generating the test sequences. These videos have a spatial resolution of 1280×720px (720p) and a temporal resolution of 50 frames per second (fps). The first frames of the originals are depicted in Fig. 3. Videos are all ten seconds long and chosen with the goal of generating a diverse content with diverse spatial and temporal characteristics (see [12] for details).

Test sequences were impaired with combinations of blurriness, blockiness, and packet-loss. These artifacts are commonly present in real applications (e.g. video compression and digital transmission). We used a subset of the stimuli created by Silva *et al.* [12] (see details also in Farias *et al.* [13, 14]), whom generated the test sequences in conformity with ITU Recommendation P.930 [15]. From the 140 test sequences generated by Silva *et al.* [12], we chose 49 videos with 7 distinct combination of artifacts (see Table I). This number of sequences made it possible to run the QR experiment with session of up to 50 minutes.

B. *Equipment and Methodology*

For both VQR and SS, experiments were run with one subject at a time using a PC computer and a Samsung LCD monitor of 23 inches (Sync Master XL2370HD), with resolution 1920×1080@60hz (FullHD, 1080p). The dynamic contrast of the monitor is turned off and contrast is set to 100 and brightness to 50. The measured gamma of the monitor is approximately 1.94, 1.57, 1.91, and 1.17 for luminance,

(a) Park Joy (b) Into Trees (c) Park Run (d) Romeo and Juliet (e) Cactus (f) Basketball (g) Barbecue

Fig. 3: Screenshots of the first frame of the sequences used in the experiments.

Combination	Packet loss	Blockiness (α)	Blurriness (β)
1	0.0	0.0	0.0
2	0.0	0.6	0.0
3	8.1	0.0	0.0
4	0.7	0.0	0.4
5	8.1	0.0	0.6
6	8.1	0.4	0.6
7	8.1	0.6	0.6

TABLE I: Combinations of the parameters (packet loss, blurriness and blockiness) used to generate the sequence of testing videos. Packet loss values indicate the percentage of lost packets for a whole video. Blockiness and Blurriness values indicate the strenght of the artifacts signal added [16].

red, green, and blue, respectively. Since the resolution of the monitor screen (1080p) was higher than the video resolution (720p), we filled the space between video and monitor edges with a gray border to preserve the original video proportions (see Fig. 2 (c)). For the VQR test, we used a Galaxy Tab 2 10.1in WXGA PLS TFT with 1Ghz dual-core processor and an Android 4.0 (API level 14) operating system. This tablet has a screen resolution of 1280×800px, what allows showing the ruler reference images (768×512px) in full resolution.

For both VQR and SS, the room had the lights dimmed to avoid reflections on the monitor (i.e. constant illumination of approximately 70 lx). This illumination conditions are compliant to ITU-T Recommendation BT.500-11 [2]. The subjects were seated straight ahead of the monitor, centered on slightly below eye height for most subjects. The distance between the subject's eyes and the video monitor was 3 times the height of videos. We used a chin rest in SS experiments to guarantee that the distance between the subject's eyes and the monitor remained constant. The chin rest was not used in VQR experiments. Before starting the experiment, the experimenter checked if the subjects were properly seated at the adequate distance. The experimenter gave oral instructions describing the experimental task. For both methodologies, before proceeding to the experimental task, subjects had to (1) visualize freely a set of 7 videos (not included in the actual test set), showing different levels of quality similar to the ones in the experiment, and (2) try-out the scoring interface to get acquainted with the task. For the VQR, they followed the tutorial specified in Section III. Next, test videos were presented in random order. To avoid sampling bias, in the VQR test the ruler image displayed when enabling it after the full visualization of the test video, was also randomly selected for every stimulus.

The VQR experiment was performed with 17 subjects following the protocol described in Section III. The SS experiment was performed with 24 subjects following the protocol and conditions described by Silva *et al.* [12]. The SS scoring was performed on a continuous scale ranging from 0 to 10. In both experiments, subjects were volunteers from the institutions where the tests were performed. They were considered naive of most kinds of digital video defects and the associated terminology. No vision test was performed on the subjects, but they are asked to wear glasses or contact lenses if they need them to watch TV. In order to avoid fatigue of the subjects, the experiments were split into three sessions, between which subjects could rest.

V. Results

To check the reliability of the quality judgments provided with VQR, we compared the MOSs [1] obtained from VQR with those obtained with SS. The VQR MOS for a video v_i is computed by averaging (across all subjects) the SQS values of the ruler images chosen to match the quality of v_i. The SS MOS are computed by simply averaging the numerical values given by subjects to v_i. To allow a comparison between our results and other results in the literature [17]–[19], MOSs obtained for each method result were linearly scaled into a continuous scale ranging from 1 to 5 in ascending order of quality using.

We evaluate the reliability of VQR in two stages. First, we compare SS MOS and VQR MOS. Since SS experiments are widely used, we assume SS provides acceptable subjective quality measures. We test if VQR can provide quality assessments close to the ones given by the SS. Second, we compare the inter-subject variability observed in the judgments expressed with both methods. The idea is to verify which experimental procedure produces the most reliable results in terms of MOS confidence.

A. Parallel-Forms Reliability

In test theory, parallel form reliability is used to assess the consistency of the results of two psychophysical experimental methodologies [20]. Fig. 4 (a) shows the plot of SS MOSs versus VQR MOSs. Notice that QR and SS MOSs are very similar to each other. This suggests that MOSs are consistent with the two different experimental methodologies.

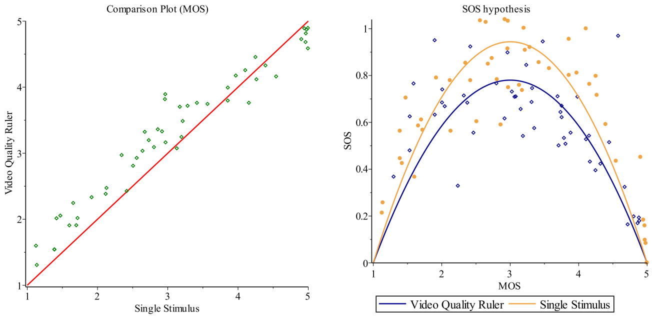
	PCC	SROCC	OR	RMSE	KCF
Metric	0.9663	0.9643	0	0.3871	0.8511

TABLE II: Statistical metrics measuring the correspondence between SS MOS and VQR MOS.

According to the "Final report from the video quality experts group on the validation of objective models of video quality assessment" [1], statistical metrics can be used for evaluating the performance of video quality assessment metrics in predicting subjective scores. These include the Pearson Correlation Coefficient (PCC), the Spearman Rank Order Correlation Coefficient (SROCC), the Outlier Ratio (OR), Kendall Correlation Factor (KCF) [21], and the Root Mean Square Error (RMSE). We use these statistical metrics as performance metrics [4] to evaluate whether VQR MOS can predict SS MOS. Table II shows the PCC, SROCC, OR, RMSE, and KCF values computed for the SS and VQR MOSs. As expected, PCC and SROCC are high (comparable to what was found by Tominaga for standard methods [22]). OR is zero, RMSE is relatively low, and the KCF is high. In other words, these metrics suggest that VQR MOSs are highly similar to SS MOSs. This also suggests that at a global level, subjects are able to match the quality of an image with that of a video, as evaluations provided seem to be consistent with those that obtained with a standard methodology (SS).

B. Inter-Observer Reliability

From the previous statistical metrics, we can observe that both experimental methodologies provide mean opinion scores that are clearly correlated. Nevertheless, it is also important to verify the level of confidence with which those MOS are expressed. High variability in the scores given by different subjects to the same video would indicate low confidence in the MOS, thereby hampering its reliability. Thus, to evaluate the reliability of VQE, it is necessary to evaluate whether it provides sufficient agreement among subjects.



(a) (b)

Fig. 4: Comparisons of VQR and SS. (a) Correlation between Mean of Opinion Scores (MOS). (b) SOS hypothesis for SS and VQR experimental data.

The most straightforward way to quantify the confidence around the MOS obtained with the two methods would be to compute the width of the 95% confidence interval around MOS [22]. Nevertheless, the width of confidence intervals depends on the number of subjects involved in the video evaluation: having used two subject pools of different sizes, the comparison through this measure would be unpractical. We resort therefore to the calculation of Hossfeld's α [19], which measures the width of the standard deviation of opinion scores (SOS) in relationship with the magnitude of the Mean Opinion Scores. Specifically, alpha is a parameter describing the squared relationship between MOS and SOS. The higher Hossfeld's α value, the worse reliability of the methodology [18].

Fig. 4 (b) shows the SOS-MOS relationship corresponding to each experimental methodology. The values of Hossfeld's α are 0.1950 and 0.2360 for VQR and SS, respectively. These values are consistent with alpha values reported by Hossfeld *et al.* [19] for subjective video quality assessment experiments. Notice that the SOS is higher for the SS method than for VQR. This indicates that the opinion scores obtained using VQR present a higher agreement among subjects.

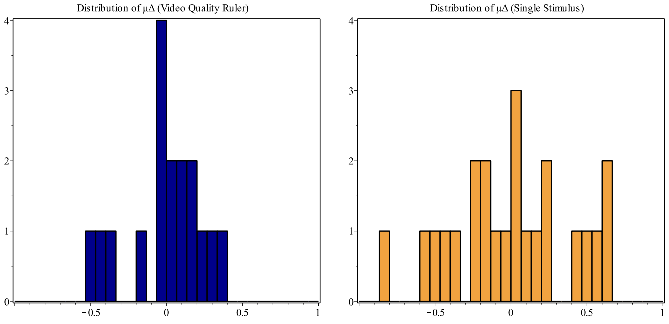
C. Subject Bias

As an additional method to investigate the reliability of MOS provided by the VQR, we investigate subject scoring behavior. It is known that each subject has a tendency to resort to an individual strategy to score video sequences, using the scoring scale in different ways [5, 7]. In the SS method, subjects give an overall numerical value for the quality of the video. Some of them might be more "forgiving", giving higher scores while other subjects may tend to use the lower end of the scale more. Since the VQR method requires from subjects to perform a visual matching between stimuli, this variability in scale usage may be reduced.

Janowski and Pinson [7] proposed a model to estimate subject bias. Subject bias ($\mu\Delta_i$) is estimated for each subject i and each test video j, based on the observed individual ratings (o_{ij}). When $\mu\Delta_i$ is subtracted from o_{ij}, we obtain the 'unbiased' opinion scores (r_{ij}). In other words, we can analyze the experimental data with a minimized influence of $\mu\Delta_i$. The MOS computed from o_{ij} and r_{ij} scores are the same, but the standard deviation around scores decreases. Therefore, we compare the differences of standard deviation before (SOS_j) and after ($\hat{SOS}_j$) we remove $\mu\Delta_i$. Table III shows the impact of $\mu\Delta_i$ removal on standard deviation scores when a Student's t-test analyzes the hypothesis that a more reliable method produces a difference in the data after removing the theoretical subject bias. This table depicts the difference of means (DM), difference of standard deviations (DSTD), the computed statistic (t-statistic), and the computed p-value.

s_1	s_2	DM	DSTD	t-statistic	p-value
SOS_{SS}	$\hat{SOS}_{SS}$	0.07645	0.19162	2.79276	0.00748
SOS_{VQR}	$\hat{SOS}_{VQR}$	0.04238	0.08019	3.69967	0.00055

TABLE III: Impact on Student's t-test sensitivity of removing $\mu\Delta_i$ from each experiment dataset.



(a) (b)

Fig. 5: Comparisons of VQR and SS. (a) Distribution of subject bias for VQR. (b) Distribution of subject bias for SS.

These results indicate that removing subject bias significantly reduces the standard deviation of MOS for both methodologies. However, DM is smaller for VQR than for SS, suggesting that MOSs collected using VQR method are less biased than those collected using SS. This is also clear when we look at the distribution of subject bias for all subjects in the pools of the two experiments. As can be noticed from Fig. 5 (a) and (b), the $\mu\Delta_i$ is more likely to be found close to zero for VQR (unbiased). In other words, the subject bias and the user disagreement are more noticeable in the SS data.

VI. Conclusions

In this paper, we presented a new experimental methodology to assess video quality: the Video Quality Ruler (VQR). VQR is based on the image Quality ruler – a popular psychometric methodology. The statistical analysis of the experimental results shows two important results. First, the VQR successfully adapts the Quality Ruler methodology (ISO 20462), designed for image assessment, to the assessment of video quality, providing Mean Opinion Scores which are very close to those that would be obtained with a standard methodology such as Single Stimulus assessment. Second, when compared to the SS methodology, VQR seems to convey Mean Opinion Scores with a higher confidence and less prone to subject bias. The scoring strategy based on visual matching of quality seems, therefore, to be beneficial in allowing higher agreement among individual subject evaluations.

Clearly, further work is needed to fully evaluate the potential of VQR to become a fully reliable methodology, appealing to research in video quality assessment. The visual matching scoring strategy is time-consuming, and a question is still open whether VQR provides a good enough trade-off between reliability of experimental results and expensiveness of the experiment in terms of time. In this sense, an experimental comparison of the VQR with double stimulus methodologies such as DSIS or Paired Comparison is to be envisioned. Moreover, given that the subjects have to switch multiple times between the tablet and the monitor screen, the proposed methodology may be tiresome for sequences longer than 10 seconds. Further studies should investigate the suitability of the VQR for the quality assessment of long video sequences. In addition, in this work we adopted a SQS and the related ruler images which were calibrated for image quality assessment. Although this choice paid off in terms of reliability of the quality assessments made, it is an open question whether a SQS calibrated on purpose for video quality assessment would be even more appropriate for a Video Quality Ruler. Another open question concerns the case in which temporal variation of quality occurs, as in the case of adaptive streaming. Since video artifacts used in this study were blurring, blockness, and packet loss, further studies considering temporal variation of quality must be performed.

Acknowledgment

This work is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), by the University of Brasília (UnB), and by the Delft University of Technology (TU Delft).

References

[1] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment - phase ii," http://www.vqeg.org/, august 2003.

[2] ITU-R, "500-11, methodology for the subjective assessment of the quality of television pictures," 2002.

[3] B. W. Keelan and H. Urabe, "Iso 20462: A psychophysical image quality measurement standard," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2003, pp. 181–189.

[4] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 165–182, 2011.

[5] P. G. Engeldrum, *Psychometric scaling: a toolkit for imaging systems development*. Imcotek Press, 2000.

[6] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Visual Communications and Image Processing 2003*. International Society for Optics and Photonics, 2003, pp. 573–582.

[7] L. Janowski and M. H. Pinson, "Subject bias: Introducing a theoretical user model," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, 2014, pp. 251–256.

[8] B. Keelan, *Handbook of image quality: characterization and prediction*. CRC Press, 2002.

[9] D. R. Rasmussen, K. D. Donohue, Y. S. Ng, W. C. Kress, F. Gaykema, and S. Zoltner, "Iso 19751 macro-uniformity," in *Electronic Imaging 2006*. International Society for Optics and Photonics, 2006, pp. 60 590K–60 590K.

[10] J. Redi, H. Liu, H. Alers, R. Zunino, and I. Heynderickx, "Comparing subjective image quality measurement methods for the creation of public databases," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2010, pp. 752 903–752 903.

[11] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," *Signal processing*, vol. 77, no. 1, pp. 1–9, 1999.

[12] A. F. Silva, M. Farias, and J. A. Redi, "Assessing the influence of combinations of blockiness, blurriness, and packet loss impairments on visual attention deployment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2015, pp. 901 402–901 402.

[13] M. C. Farias, I. Heynderickx, B. Macchiavello Espinoza, and J. Redi, "Visual artifacts interference understanding and modeling (varium)," in *Seventh international workshop on video processing and quality metrics for consumer electronics*, vol. 1, 2013.

[14] M. C. Farias and S. K. Mitra, "Perceptual contributions of blocky, blurry, noisy, and ringing synthetic artifacts to overall annoyance," *Journal of Electronic Imaging*, vol. 21, no. 4, pp. 043 013–043 013, 2012.

[15] P. ITU, "930-principles of a reference impairment system for video," *International Telecommunication Union*, 1996.

[16] M. Leszczuk, M. Hanusiak, M. C. Farias, E. Wyckens, and G. Heston, "Recent developments in visual quality monitoring by key performance indicators," *Multimedia Tools and Applications*, pp. 1–23, 2014.

[17] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *Broadcasting, IEEE Transactions on*, vol. 57, no. 1, pp. 1–14, 2011.

[18] E. Siahaan, J. Redi, and A. Hanjalic, "Beauty is in the scale of the beholder: Comparison of methodologies for the subjective assessment of image aesthetic appeal," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, Sept 2014, pp. 245–250.

[19] T. Hobfeld, R. Schatz, and S. Egger, "Sos: The mos is not enough!" in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*. IEEE, 2011, pp. 131–136.

[20] W. M. Rogers, N. Schmitt, and M. E. Mullins, "Correction for unreliability of multifactor measures: comparison of alpha and parallel forms approaches," *Organizational Research Methods*, vol. 5, no. 2, pp. 184–199, 2002.

[21] A. Stuart, K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics, Vol. 2A: Classical Inference and the Linear Model, 6th ed.* Willey, 2009.

[22] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, June 2010, pp. 82–87.