https://doi.org/10.2352/ISSN.2470-1173.2019.10.IQSP-311

Semantic label bias in subjective video quality evaluation: a standardization perspective

Mihai MITREA, Rania Bensaied*, Patrick Le Callet***

* *Institut Mines-Télécom; Télécom SudParis, ARTEMIS Department ; UMR 5157 – SAMOVAR*

** *Université de Nantes, Polytech'Nantes, LS2N*

Abstract

Recent studies brought to light that the semantic labels (e.g. *Excellent*, *Good*, *Fair*, *Poor*, and *Bad*) commonly associated with discrete scale ITU subjective quality evaluation induce a bias in MOS computation and that such a bias can be quantified by some reference coefficients which are independent with respect to the observers panel. The present paper reconsiders these results from a standard upgrading perspective. First, it theoretically investigates the way in which results obtained on semantically labeled scales can be "cleaned" from such an influence and derives the underlying computation formula for the mean opinion score. Secondly, it suggests a unitary evaluation procedure featuring both semantic free MOS computation and backward compatibility with respect to state-of-the-art solutions. The theoretical and methodological results are supported by subjective experiments corresponding to a total of 440 human observers, alternatively scoring 2D and stereoscopic video content. For each type of content, both high and low quality excerpts are alternatively considered. For each type of content and for each type of quality a 5 level (*Excellent*, *Good*, *Fair*, *Poor*, and *Bad*) grading scales is considered.

1 Introduction

For 100 years already, various research fields (psychology, psychophysics, sociology, marketing, medicine, ...) have considered the use of rating scales in subjective evaluations [1], [2]. Despite this long and fruitful history, no consensus is reached yet on a usage of a specific scale for a specific purpose, and several scale typologies still coexist and contradict each-other: graphic *vs*. numerical *vs*. semantic labeled scales or continuous *vs*. discrete scales. Moreover, the dynamics of numerical scales is varying with the experiment: for instance, continuous scales can range from 0 to 75, to 100, to 120 or even to 200 [3], [4], [5] while the discrete scales can feature between 2 and 11 evaluation classes [1].

For the visual content evaluation, the ITU Recommendations have proven their effectiveness and are already intensively used in several research studies aiming at a large variety of applications (device evaluation/calibration, compression, 3D image reconstruction, watermarking, *etc*.). Some studies report experiments had carried out on 5 quality levels while other on 11 quality levels. Yet, no answer on how to choose either this number or the quality levels themselves is provided.

The relationship between continuous and discrete scales is recurrently addressed in research studies. For instance, [6] shows that assessments on the discrete scale have the highest level of stability, at least for the experiment under consideration (a self-assessment of the previous knowledge in statistics). This study also raises a concern about the very meaning of the *continuous* and *discrete* terms during the subjective evaluations.

The impact of semantic labels is discussed and detailed in various research studies. On the one hand, some studies state that adjacent ITU labels are characterized by non-uniform semantic distances [7], [8]; yet, such a behavior is not quantified. On the other hand, some studies [9] claim the contrary, *i.e.* that the semantic of adjacent ITU labels does not impact the results. While some results correspond to subjective studies carried out for different languages (Japanese, German, English, French and Italian), the ITU-T P.913 explicitly postulates that the *MOS* is invariant with respect to the semantic labels translation, but does not provide any ground for this.

Under this framework, [10], [11] establish a theoretical formula mapping the scores assigned by the observer on continuous scale to an arbitrarily, q levels discrete grading scale and carry out an experimental study evaluating the semantic impact of the ITU semantic labels in the MOS (mean opinion score) computation for experiments carried out on 5 levels (*Excellent*, *Good*, *Fair*, *Poor*, and *Bad*) semantic labeled scales.

The present paper goes one step further and theoretically investigates the possibility of compensating the semantic impact induced in the MOS computation by the ITU labels. The experiments consider 440 human observers, alternatively scoring 2D and stereoscopic video content, and, for each type of content, both high and low quality excerpts (as a priori evaluated by objective quality metrics). It is thus brought to light that the MOS can vary up to 18% from its theoretical value. The paper also includes a discussion about various possible usages of these results.

2 State of the art

To the best of our knowledge, for visual quality assessment applications, the problem of the semantic impact of the labels in the overall quality evaluation was first raised

by [7]. Two panels are considered for experiments: 49 persons compose a panel of English speakers (native USA speakers) while 24 persons compose a panel of Italian speakers. The experiments are conducted in parallel for English and Italian, but they will be illustrated here only for English. During the experiments, a continuous scale featuring no intermediate labels but having its two extremities marked with *Best imaginable* and *Worst imaginable* is presented to the panel. The panel members are asked to place, on that continuous scale, according to their own understanding, 15 adjectives (labels): *Superior*, *Ideal, Excellent, Good, Fine, OK, Fair, Passable, Marginal, Not Quite Passable, Poor, Inferior, Bad, Not Usable, Awful*. For both languages, the results showed that the ITU labels were not evenly distributed along the graphic scale suggesting a non-uniform semantic distance between adjacent ITU labels. Specifically, a kind of compression at the end points of the scale was identified and explained as a reluctance of the observers in using the continuous scale extremities. The results also show a clustering tendency, with 9 classes (e.g. the adjectives *Ideal*, *Excellent* and *Superior* are very close each-other). These results were later corroborated for other languages, such as Swedish, Dutch and British English, but not for Japanese and German, which exhibited uniformly distributed labels along the scale [8], [12], [9].

The semantic impact of the ITU Japanese descriptive terms for quality and impairment is investigated in [8]. The two experiments follow the principles in [7]. A panel of 40 Japanese speakers is inquired. They are asked to alternatively position on a continuous scale with no intermediate but extremities labels either 13 quality terms or 12 impairment terms. It is thus brought to light that the perceived quality intervals are non-uniform. Yet, in Japanese, they are distributed more evenly than in English, French or Italian, following a similar trend as in the German case. The impairment experiment shows the Japanese terms have lower semantic impact than the corresponding terms in other languages. Note that the underlying experiments for German are presented in [12].

The two ideas of semantic impact and language dependency related to the ITU labels are considered as a starting point for the research study presented in [13]. In order to refine the precision and the stability of the results, the subjective quality evaluation is considered to be a multidimensional process and some means for identifying the different dimensions and the appropriate vocabulary are advanced. In this respect, the use of unlabeled continuous rating scale is considered as a ground for investigation. A panel of 24 subjects is asked to score audio-visual content on a continuous scale whose extremities are labeled by + and – signs. The results show that such a quality rating procedure is "*remarkably consistent*", thanks to the fact that the subjects set their own criteria. The results also show that using an unlabeled scale reduces the tendency of subjects to avoid the end points of the scale.

The SSCQE and DSIS methods are compared on the same test material in [14]. In order to avoid the semantic impact of the *Excellent*, *Good*, *Fair*, *Poor* and *Bad* labels, the SSCQE method is presented to the observer as a vertical slider with only the two labels *Good* and *Bad* at the top and bottom ends of the slider. On the contrarily, the DSIS scale follows the general ITU recommendations and is presented to the observer on a discrete, 5 levels scale ranging from *Imperceptible* to *Very annoying*. Results obtained from 20 observers show that the two methods are highly correlated and produce comparable quality results.

Three subjective audio quality evaluation tests are presented in [9]. Each of these three tests considers a different scale: the standard continuous scale with the 5 ITU labels, a 5-point continuous impairment scale, and a label-free continuous scale. The results showed a high similarity between the scores obtained with the ITU labeled scale and the label-free quality scale, with an almost perfect linear regression between them. Hence, this study supports the idea that the ITU quality scale is indeed an equal-interval scale.

The study in [15] investigates the suitability of SAMVIQ assessment methodology; in this respect, two psychovisual experiments are carried out in two different laboratories. The subjective video quality evaluation follows the ITU-T P.910. The observers (whose number is not presented) assign their scores on a continuous 0-100 scale labeled by *Excellent*, *Good*, *Fair*, *Poor* and *Bad*. The experimental results indicate that the SAMVIQ methodology provides results comparable to other existing methods, such as the single stimulus ACR methodology.

The ACR and SAMVIQ subjective quality assessment methodologies are compared in [16]. The ACR is presented to the observers with a 5 levels discrete scale associated to the labels *Excellent*, *Good*, *Fair*, *Poor* and *Bad* while the SAMVIQ with a continuous scale ranging from 0 to 100, yet featuring the same labels. The viewing conditions are not précised; the number of observers participating in the test is 43. The results of this study show that the ACR uses 96.3% of the available range while SAMVIQ uses only 82%. It is thus demonstrated that the two assessment methodologies have different behaviors; it is also shown that the relation between their results depends on the evaluated content quality and it is subsequently stated that, for a given number of observers, SAMVIQ is more precise than ACR.

The variability of subjective ratings obtained with different scales (0-100 continuous scale and 5, 9, and 11 discrete scales) is investigated in [17]. The study relays on simulated data instead of real experimental data, since it is considered that the differences among experiments available in the literature are too large for reliable direct comparison. It is concluded that although an increased discretization level of the scale leads in theory to an increase of the standard deviation of the scores (and therefore to a decrease of precision), practical proof of this effect remained

inconclusive. He also found that the number of subjects may not need to be as high as generally assumed; in fact, the minimum of 15 recommended by ITU appears to be a very reasonable suggestion.

The study [18] compares 4 different ITU grading scales with labels: two of them are discrete (with 5 and 9 levels) while the other two are continuous, yet with 11-point and 5-point grades. A subjective assessment test following the ITU-R BT.500-11 is conducted. 92 observers assign their scores according to the ACR method using the 4 different scales. The total evaluated content is composed of 128 sequences of 12s each. The results show that no significant statistical difference is found among subjective results obtained with the different four scales.

The studies in [10] and [11] establish a theoretical framework for investigating the semantic impact of the labels in the overall MOS computation. The experiments considers a SSCQE (Single Stimulus Continuous Quality Evaluation) method and both continuous and discrete, semantically labeled scales. It is thus brought to light that the semantic impact of the 5 ITU levels (*Excellent*, *Good*, *Fair*, *Poor*, and *Bad*) can be evaluated by a set of reference coefficients, which are independent with respect to the observers and solely depends on the type of content (3D or 2D video, high or low quality).

The present paper extends these previous results [10], [11] by providing a methodological formula for a posterior canceling the semantic label impact.

3 Method presentation

The methodological framework for handling the joint impact of discretization and semantic labeling the evaluation scales is defined based on our previous studies and is structured in three main steps which will be subsequently detailed:

1. Continuous to discrete unlabeled scales mapping [10],
2. Semantic label assessment [11],
3. Semantic label compensation.

3.1 Continuous to discrete unlabeled scales mapping

Be there a subjective quality evaluation experiment carried out on a continuous, unlabeled grading scale and be X the r.v. (random variable) theoretically modeling the observer's inner appreciation about the content under evaluation.

Let assume that X is continuously distributed in the interval $[0; M]$, according to a probability density function (*pdf*) $p_X(x)$ and be *MOS* and σ the mean value and standard deviation, respectively.

Assume now the case in which an evaluation on a discrete scale with q quality levels, evenly distributed would be required. The scores would be distributed according to a new *r.v.* Y, whose values y are obtained from the x values according to a non-linear transformation $f(x)$:

$$y = f(x) = \\ = \begin{cases} 0, & x \le 0 \\ i, & (i-1)M/q < x < iM/q, i \in \{1,2,\dots,q\} \\ 0, & x > M \end{cases} \quad (1)$$

Hence, the $p_Y(y)$ *pdf* can be computed as follows:

$$p_Y(y) = \sum_{i=1}^{q} \delta(y-i) \int_{(i-1)M/q}^{iM/q} p_x(x) \quad (2)$$

where $\delta(.)$ denotes the Dirac's Delta distribution.

The mean value of Y, denoted by MOS_q, represents the mean opinion score corresponding to the evaluation on a q quality level grade scale (with uneven, unlabeled gradations):

$$MOS_q = \sum_{i=1}^{q} i \int_{y_{i-1}}^{y_i} p_X(x)dx \quad (3)$$

The standard deviation of Y, denoted by σ_q, is:

$$\sigma_q = \sqrt{\sum_{i=1}^{q} i^2 \int_{y_{i-1}}^{y_i} p_X(x)dx - MOS_q^2} \quad (4)$$

3.2 Semantic label assessment

Be there a subjective quality evaluation experiment carried out on a q level, semantically labeled grading scale; for instance, for $q = 5$, the labels can be *Excellent*, *Good*, *Fair*, *Poor*, and *Bad*.

Assume $[n_1, n_2, \dots, n_q]$ the number of times each of the classes was scored. The r.v. modeling the scores assigned by the observers is denoted by Z and its *pdf* by $p_Z(z)$.

$p_Z(z)$ can be estimated from the scores by any discrete *pdf* estimation method; for instance, in the present study, a frequency based estimation is considered:

$$p_Z(z) = \sum_{i=1}^{q} p_Z(i)\, \delta(z-i) \quad (5)$$

where $p_z(i)$ is the relative frequency of the scores assigned to i^{th} quality class:

$$p_Z(i) = \frac{n_i}{\sum_{j=1}^{q} n_j} \quad (6)$$

The mean opinion score corresponding to this experiment is:

$$MOS_Z = \frac{\sum_{i=1}^{q} i \cdot n_i}{\sum_{i=1}^{q} n_i} \quad (7)$$

Assuming the semantic labels have no impact, the Y and Z r.v. would be identical: that is, the evaluations on discrete, unlabeled and labeled scales would yield the same results. Conversely, differences in the *pdf* describing the Y and Z r.v. bring to light a semantic influence of the labels.

Consequently, in order to evaluate the semantic impact of the labels, the $[0 = y_1, y_2, \dots, y_q = M]$ partition ensuring identity between the Y and Z random variables is searched

for. In this respect, the $p_Y(y)$ and $p_Z(z)$ are compared through a binomial test.

The semantic impact is assessed by relative variation of the partition intervals with respect to the uniform partition. The set of coefficients $\rho_{q-i}, i = 0,1, \dots q-1$ are computed in this respect:

$$\rho_{q-i} = \frac{y_{q-i} - y_{q-i-1}}{M/q} \tag{8}$$

A unitary value for such a coefficient demonstrates that the related semantic label does not modify the evaluation - that is, an even partition $[0 = y_1, y_2, \dots, y_q = M]$ ensures the identity between *Y* and *Z*. A value larger than 1 indicates that the related semantic label makes the observer more likely to score that way while, conversely, a value lower than 1 shows that the related label makes the observers more reluctant in assigning that label when scoring.

3.3 Semantic label compensation

The presence of a semantic impact associated to the label can be translated into some "errors" occurred in the scores assigned by the observers: the semantic impact of the labels makes the observer score by influences by factors externals to the evaluated content itself.

Assume now the case in which an experiment of the type described in Section 3.2 is performed and an experimenter, knowing the set of ρ coefficients would like to post-process the $[n_1, n_2, \dots, n_q]$ scores for canceling the semantic impact. The principle is to adjust the scores according to a set of γ coefficients related to the probability that a score would be assigned in a class because of the semantic impact and not because of the content quality, as detailed below and illustrated in Fig. 1 for the case of $q = 5$.

Be U the r.v. obtained from X by a scaling of the variable x to $u = qx/M$. Hence, the subjective quality evaluation process can be now evaluated based on three related r.v.:

- the U r.v., continuously taking values between $[0, q]$;
- the Y r.v. taking the values $[1,2, \dots q]$ in the lack of semantic impact
- the Y^{sem} r.v. taking the values $[y_{sem,1}, y_{sem,2}, \dots y_{semq,}]$ where:

$$y_{sem,i} = q - \sum_{j=i+1}^{q} \rho_j, \tag{9}$$

where $i \in \{1,2, \dots, q-1\}$ and $y_{sem,q} = q$.

The weighting coefficients γ_i, $i \in \{2, \dots, q\}$ are defined as follows:

$$\gamma_i = \frac{\int_{i-1}^{i} p_U(u)}{\int_{y_{sem,i-1}}^{y_{sem,i}} p_U(u)} \tag{10}$$

The scores $[n_1, n_2, \dots, n_q]$ can now be post-processed so as to cancel the semantic impact and obtain the new set $[s_1, s_2, \dots, s_q]$, where:

$$s_i = n_i + \gamma_i n_{i-1}, \tag{11}$$

where $i \in \{2, \dots, q\}$.

Finally, the semantic compensated mean opinion score can be obtained as follows:

$$MOS_{comp} = \frac{\sum_{i=1}^{q} i \cdot s_i}{\sum_{i=1}^{q} s_i} \tag{12}$$

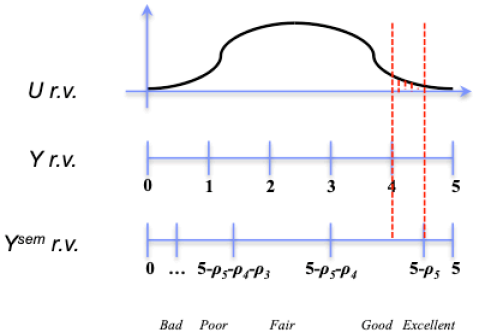


Figure 1 – Principle of the semantic impact cancelation

4 Experimental setup

The evaluation has been conducted at the Advanced Research &Techniques for Multimedia Imaging Systems (ARTEMIS) Department at Telecom SudParis engineering school in France.

The viewing conditions are set in concordance with ITU-R BT.1788, ITU-R BT.500-11, ITU-R BT.500-13, ITU-T P.913.

A 47" LG LCD, full HD 3D monitor (1920 x 1080 pixels) and a 400cd/m² maximum brightness are used in the experiments.

The experiments involve maximum 2 subjects per session who are presented a SSCQE (Single Stimulus Continuous Quality Evaluation) scale. The continuous scale features numerical labels, between the minimal and maximal values (0 and M = 100, respectively), with a precision of 10. The discrete, semantically labeled scale alternatively considers 5 levels (labeled *Excellent*, *Good*, *Fair*, *Poor* and *Bad*) and 3 levels (labeled *Good*, *Fair* and *Bad*). Two training sessions are considered for each viewing session.

The stereoscopic video content processed in the present study is produced under the framework of the 3DLive French national project. The 3DLive corpus sums-up 2 hours, 11 minutes and 24 seconds of stereoscopic video sequences (197000 stereoscopic pairs encoded at 25 frames per second), representing 10 minutes of a rugby match, 10 minutes of a dancing performance, 1 minute of a private gig of rock band "Skip the Use", one hour and 45 minute and 24 seconds of a volley-ball match and 5 minutes of a theater

play "Les Fourberies de Scapin" by Molière. These sequences are full HD encoded (1920×1080 pixels). This corpus is subsequently watermarked by 4 different methods, with 4 different configurations. Despite their peculiarities (which are irrelevant for the present paper) all these watermarked sequences features a high *a priori* quality expressed by values 35dB < PSNR < 40dB, by SSIM (Structural SIMilarity) values larger than 0.98 and by NCC (Normalized Cross Correlation) values larger than 0.98. From the 3DLive corpus, 16 sequences with individual durations between 40 sec and 80 sec, summing up about 20 minutes are randomly sampled, thus obtaining the high quality stereoscopic video corpus.

In order to obtain the low-quality stereoscopic video content, the high quality stereoscopic video corpus is compressed (while keeping the frame resolution and rate constant) so as to obtain 25dB<PSNR<30dB. It should be noticed that the low-quality corpus is downgraded 25db – 30dB with respect to the high quality corpus which is, in its turn, downgraded 35db - 40db with respect to the original content. It is also *a posteriori* verified that the values corresponding to the SSIM and NCC (with respect to the high quality corpus) range between 0.97 – 0.98 and 0.95-0.97, respectively.

The high quality 2D video corpus corresponds to the left view from the high quality stereoscopic video corpus.

The low-quality 2D video corpus is organized under the framework of the MEDIVALS (waterMarking et Embrouillage pour la DIffusion et les Echanges Vidéos et Audios Legalisés) French national project. The video content is encoded at 640x480 pixels, 25 fps. An MPEG-4 AVC encoder is considered, with the baseline profile and a 512 kbps rate. The corpus has a total duration of 1h30 minutes and is composed of 4 types of professional TV content: news, documentary, movies and talk-shows.

In order to obtain the excerpts to be presented to the observers, from each type of content, a sequence with a duration between 50 and 60 sec is randomly extracted. Then, each of this sequence is downgraded with 7 distortion configurations. These 28 sequences, summing up to 20 minutes, are then shuffled prior to their presentation to the observer.

A total of 440 non-expert viewers (160 viewers for each of the 4 types of content) are involved in the experiments. The age distribution ranged from 20 to 37 with an average of 23. All the subjects were screened for visual acuity using Snellen chart and color vision using the Ishihara test.

For each type of content, the 160 viewers are grouped in three types of panels. First, the *reference* panel is composed of 60 observers scoring on a continuous scale and allows the reference (theoretical) model for X and Y r.v. to be computed. Secondly, the *cross-checking* panel was composed of 25 observers scoring on $q = 5$ quality level semantic labels, namely *Excellent*, *Good*, *Fair*, *Poor*, and *Bad*. Finally, the so-called *a posteriori* validation panel is also composed of 25 observers scoring on $q = 5$ quality level scales, and is considered in order to validate the generality of the results. The outliers are detected and eliminated according to a kurtosis coefficient criterion, according to the ITU BT.500-11/13.

5 Experimental results

5.1 Evaluating the semantic impact of the labels

The first experiment brings to light the semantic impact of the ITU labels by computing the set of ρ coefficients defined by (7) under the experimental framework presented in Section 4. The quantitative results are presented in Tables 1-4 below.

		Semantic-impacted label limit	ρ
Bad	*0 - 20*	0 - 20	1
Poor	*21 - 40*	21 – 40	1
Fair	*41 - 60*	41 – 60	1
Good	*61 - 80*	61 – 87	1.35
		61 – 88	1.4
Excellent	*81 - 100*	88 – 100	0.65
		89 – 100	0.6

TABLE 1 SEMANTIC IMAPCT WHEN EVALUATING HIGH-QUALITY 3D VIDEO CONTENT

		Semantic-impacted label limit	ρ
Bad	*0 - 20*	0 - 20	1
Poor	*21 - 40*	21 – 29	0.45
		21 – 30	0.5
Fair	*41 - 60*	30 – 60	1.55
		31- 60	1.5
Good	*61 - 80*	61 – 80	1
Excellent	*81 - 100*	81 – 100	1

TABLE 2 SEMANTIC IMAPCT WHEN EVALUATING LOW-QUALITY 3D VIDEO CONTENT

		Semantic-impacted label limit	ρ
Bad	*0 - 20*	0 – 20	1
Poor	*21 - 40*	21 – 38	0.9
Fair	*41 - 60*	39 – 60	1.1
Good	*61 - 80*	61 – 87	1.35
		61 – 88	1.4
Excellent	*81 - 100*	88 – 100	0.65
		89 – 100	0.6

TABLE 3 SEMANTIC IMAPCT WHEN EVALUATING HIGH-QUALITY 2D VIDEO CONTENT

		Semantic-impacted label limit	ρ
Bad	*0 - 20*	0 - 20	1
Poor	*21 - 40*	21 – 31	0.55
Fair	*41 - 60*	32 – 60	1.45
Good	*61 - 80*	61 – 83	1.15
Excellent	*81 - 100*	84 – 100	0.85

TABLE 4 SEMANTIC IMAPCT WHEN EVALUATING LOW-QUALITY 2D VIDEO CONTENT

As a general trend, it can be noticed that for high quality content (both 3D and 2D video), the *Excellent* label has a reluctance impact. Conversely, for low quality content (both 3D and 2D), the *Poor* label has a reluctance impact, the observers being prone to rather assign *Fair*.

It can also be noticed that the semantic impact is quite large: the grading scale is characterized by non-uniform grading distances which vary, in some cases, up to 55% from their reference value.

5.2 Estimating the cancelation effect

As the semantic impact assessed in the first experiment is large, the second experiment evaluates the possibility of canceling it by using the method advanced in Section 3.3.

In this respect, the relative error between the semantic impacted mean opinion score and the semantic compensated mean opinion score are computed. That is, we compute the $\varepsilon_{r,comp}$ between the values of MOS_Z computed by (7) and of MOS_{comp} computed by (12), as follows:

$$\hat{\varepsilon}_{r,comp} = \frac{MOS_{comp} - MOS_Z}{MOS_Z} \tag{13}$$

Note that a positive value for $\hat{\varepsilon}_{r,comp}$ indicates an overall reluctance effect: the mean opinion score is reduced by the semantic impact. Conversely, a negative value for $\hat{\varepsilon}_{r,comp}$ brings to light that the mean opinion score was increased by the semantic impact of the labels.

By its very nature, $\hat{\varepsilon}_{r,comp}$ can only be approximated by simulation but cannon by estimated through direct estimation. In this respect, for each of the 4 types of content, we simulated 25 virtual panels of 25 observers and the quantitative results are presented in Table 5.

	Relative error in MOS		
	min	average	max
3D high quality	0.05	0.11	0.19
3D low quality	-0.06	- 0.14	- 0.16
2D high quality	0.06	0.13	0.19
2D low quality	- 0.01	- 0.05	- 0.15

TABLE 5 RELATIVE ERROR IN MEAN OPINION SCORE COMPUTATION CANCELED BY THE ADVANCED PROCEDURE

The values reported in Table 5 demonstrate both the need for and the efficiency of the advanced method: between 5% and 15% absolute relative variations in the mean opinion score can be related to the semantic impact.

The results reported in Table 5 are in agreement with the ones reported in Tables 1-4: for high quality content, the semantic impact is associated to a reluctance effect (i.e. the mean opinion score is reduced because of the *Excellent* label) while for low quality content the overall effect is of increasing the mean opinion score (the avoidance of the *Poor* label).

6 Conclusion

The present investigates the possibility of canceling the semantic impact induced in the mean opinion score computation by the semantic labels generally associated to the discrete grading scales: *Excellent*, *Good*, *Fair*, *Poor*, and *Bad*. In this respect, two different approaches can be considered.

First, the subjective evaluation can be carried out on a continuous, unlabeled scale and the scores thus obtained can then be post-processed by using (1)-(4) so as to compute the mean opinion score (3) and its confidence limits – by means of (4) – on any q level discrete grading scale. This way, a more flexible and versatile evaluation framework is obtained and the controversial issue of the numbers of evaluation levels on a discrete scale is intrinsically bypassed.

Alternatively, the experiments can be carried out on discrete, semantically labeled grading scale (*e.g.* as the ITU SSCQE) and the scores thus obtained can be cleaned from the semantic impact by using (10)-(12). This way, although the experiments are carried out according to conventional (current day) standards, more reliable mean opinion scores are obtained: although they are intrinsically subjective, the scores will no longer be impacted by the cultural influence of the labels semantics. Note that, as shown by the experiments reported in Table 5, such a semantic influence is as large as 5% to 14%.

These results may pave the way towards either updating the standard evaluation procedure (in the sense of modifying the grading scale) or of post-processing the scores obtained in traditional way. Of course, future work is required in this respect. First, we shall investigate more accurate (in the quality evaluation sense) possibilities of detecting outliers based on continuous scale evaluation as well as on different local scoring statistical behavior [19]. Secondly, we shall extend our methodological framework presented in Section 3.3 for computing confidence limits for the mean opinion score formula (12).

Besides the direct, standard updating perspective, our results can find their usefulness in other related applicative fields. For instance, they can be also considered for completing some objective visual quality metrics (which are generally continuous) with a post-processing step allowing them to be better matched to the subjective evaluations.

Note that the problem of finding the objective continuous quality metrics featuring the highest correlation with the subjective scores (generally assigned on discrete, semantically labeled scales) is still an open research topic [19]: hence, future work will be devoted to reconsider state of ten art study and to investigate whether and at what extent such results varies with the semantic impact (*i.e.* with the ρ coefficients) and/or with the cleaning of their influence (*i.e.* with the γ coefficients.

References

[1] Freyd M., "The graphic rating scale". Journal of Educational Psychology 14, 83-102, (1923).

[2] Froberg D.G, and Kane R., "Methodology for measuring health-state preferences-ii: scaling methods", Division of Human Development and Nutrition, School of Public Health, University of Minnesota, Minneapolis, MN 55455, U.S.A., (1988)

[3] Aitken R.C., "Measurement of feelings using visual analogue scales". Proc R Soc Med, 62989- 993, (1969).

[4] Bond A. and Lader M., " The use of analogue scales in rating subjective feelings". British Journal of Medical Psychology 47, 211–217, (1974).

[5] McGuire D.B., "The measurement of clinical pain". Nurs Res, pp. 152-156, (1984).

[6] Svensson E., "Comparison of the quality of assessments using continuous and discrete ordinal rating scales". Biometrical J., vol. 42, no. 4, pp. 417–434, (2000).

[7] Jones B.L., and McManus P.R., "Graphic scaling of qualitative terms". SMPTE Journal, 1166–1171, (1986).

[8] Narita N., "Graphic scaling and validity of Japanese descriptive terms used in subjective evaluation tests". SMPTE J., vol. 102, no. 7, pp. 616–622, (1993).

[9] Zieliński S., Brooks P., and Rumsey F., "On the use of graphic scales in modern listening tests". Proc. 123rd AES Convention, New York, (2007).

[10] Bensaied R., Mitrea M., Chammem A., Ebrahimi T.,"Continuous vs. discrete scale stereoscopic video subjective evaluation: case study on robust watermarking", in Proc. of Quality of Multimedia Experience (QoMEX),Sixth International Workshop on, pp. 238 - 244, DOI: 10.1109/QoMEX.2014.

[11] R. Bensaied, M. Mitrea, "Assessing the impact of the semantic labels in subjective video quality evaluation", in 11th IMA International Conference on Mathematics in Signal Processing, December 2016, Birmingham, UK.

[12] Teunissen K., "The validity of CCIR quality indicators along a graphical scale", SMPTE J., vol. 105, no. 3, pp. 144–149, (1996).

[13] Watson A., and Sasse A., "Measuring perceived quality of speech and video in multimedia conferencing applications". Proc. ACM Multimedia Conf., pp. 55–60, (1998).

[14] Winkler S. and Campos R., "Video quality evaluation for Internet streaming applications". Proc. SPIE Human Vision and Electronic Imaging, Santa Clara, CA, vol. 5007, pp. 104–115, (2003).

[15] Q. Huynh-Thu, M. Brotherton, D. Hands, K. Brunnström, and M. Ghanbari, "Examination of the SAMVIQ methodology for the subjective assessment of multimedia quality," in Proc. 3rd Int. Workshop Video Process. Consum. Electron., Scottsdale, AZ, USA, Jan. (2007).

[16] Péchard S., Pépion R., Le Callet P., "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm". Proceedings of the Third International Workshop on Image Media Quality and its Applications, IMQA2008, (2008).

[17] Winkler S., "On the properties of subjective ratings in video quality experiments". In Proc. Int. Workshop Quality Multimedia Exper.(QoMEX), San Diego, CA, (2009).

[18] Huynh-Thu, Q., Garcia, M., Speranza, F., Corriveau, P., Raake, A.: Study of rating scales for subjective quality assessment of High-Definition video, IEEE Transactions on Broadcasting 57(1), 1–14 (2011).

[19] Li J., Mantiuk R., Wang J., Ling S., Le Callet P., Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation, NIPS 2018

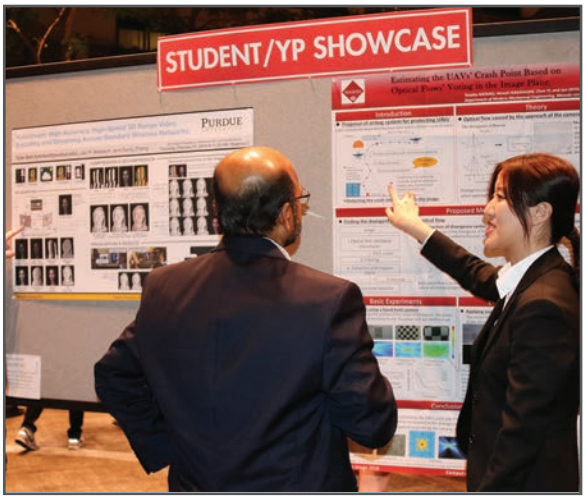
JOIN US AT THE NEXT EI!

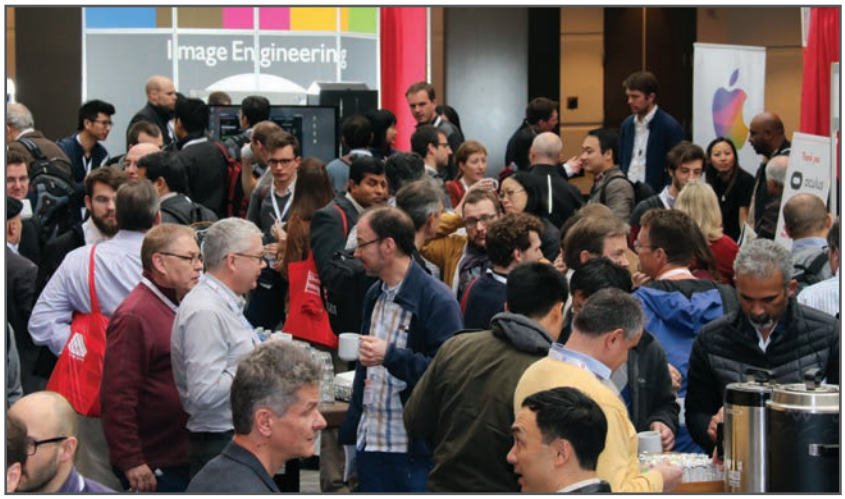
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org