DOES INTER-SUBJECT VARIABILITY DEPEND ON TEST MATERIAL?

Stefan Winkler

Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore

ABSTRACT

Quality of experience (QoE) research usually focuses on measuring or predicting mean opinion score (MOS), i.e. the average of ratings across subjects. Here we focus instead on the variability among subjects and explore questions such as: What factors does it depend on? Are these factors consistent across different experiments? Could variability be predicted from the test content?

1. INTRODUCTION

A number of papers have explored the properties of data gathered in subjective experiments. For example, the importance of using non-parametric statistics for modeling user ratings was highlighted in [1, 3]. We previously analyzed the effects of rating scales on MOS standard deviations [7] and studied MOS uniformity and variability across databases in [8].

In this paper, we focus on the variability among subjects. Naturally this depends on the specific subjects in a given experiment. However, there may be other factors of influence, which we attempt to study here. Specifically, we propose a way to detach the standard deviations (SD) of subjective ratings from their dependence on MOS, investigate the effects of source content and test conditions on SD, and test a simple model to predict SD from source content.

2. EXPERIMENTS

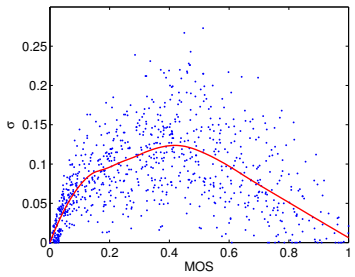
2.1. Databases

We use the following image quality databases: Categorical Subjective Image Quality (CSIQ) Database [2], LIVE Image Quality Assessment Database [5, 6], and Tampere Image Database (TID2013) [4]. These databases were selected because they represent the largest ones available at the moment. Their key parameters are summarized in Table 1. Standard deviations of ratings across subjects (σ) are provided by all three datasets.

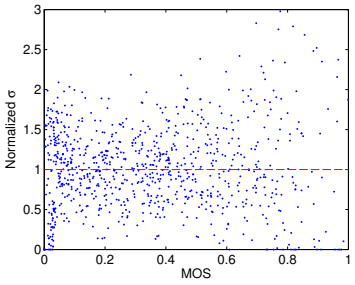
This work is supported by the research grant for ADSC's Human Sixth Sense Programme from Singapore's Agency for Science, Technology and Research (A*STAR).

2.2. SD Normalization

It is well known that the variability of ratings depends on MOS; it is typically highest for MOS in the middle of the rating scale and decreases towards both low and high ends of the scale. This behavior is due to the finite range of the scale and can been observed in virtually all subjective experiments [8].



(a) σ as a function of MOS (red curve: "moving average" $\bar{\sigma}$)



(b) After normalization ($\hat{\sigma}$)

Fig. 1: Normalization of standard deviation (CSIQ database).

The influence of absolute quality on SD needs to be removed in order to perform hypothesis testing and compute useful summary statistics. We therefore normalize the standard deviation (σ) with respect to MOS as follows. We compute a "moving average" $\bar{\sigma}(\text{MOS})$ of the standard deviation as a function of MOS by applying robust local regression, using weighted linear least squares and a 2nd-degree polynomial model.[1] The standard deviation of each test image k is then normalized: $\hat{\sigma}_k = \sigma_k / \bar{\sigma}(\text{MOS}_k)$, which results in a more uniform distribution across the MOS range (see Figure 1).

[1] Matlab function `smooth(MOS,σ,0.5,'rloess')`

Table 1: Summary of datasets.

Database	Year	Images	Scenes	Distortions	Levels	Resolution	Method	Data	Subjects	Ratings
CSIQ [2]	2010	866	30	6	4-5	512×512	Custom	DMOS+σ	25	5-7
LIVE [5, 6]	2006	779	29	5	5-6	∼768×512	ACR	DMOS+σ	161	20-29
TID [4]	2013	3000	25	24	5	512×384	PC	MOS+σ	971	?

2.3. SD Factors

Is there more or less agreement between subjects for certain types of source content (SRC) or test conditions (a.k.a. hypothetical references circuits or HRC)? In order to examine this, we perform the Kruskal-Wallis test – a one-way analysis of variance by ranks – on $\hat{\sigma}$. As this test is nonparametric, it does not assume normality in the data and is much less sensitive to outliers.

The results are reported in Table 2. For all three databases, both HRC and SRC have significant impact on $\hat{\sigma}$, as implied by $p \approx 0$. Using MOS for comparison, only HRC has a consistent impact on MOS (as one would expect), whereas SRC does not.

Table 2: Results of Kruskal-Wallis test on $\hat{\sigma}$ and MOS. Cells shaded in gray highlight significant differences in medians.

			$\hat{\sigma}$		MOS	
Database	Variable	*df*	χ^2	p	χ^2	p
CSIQ	HRC	28	80.8	0	779	0
	SRC	29	136	0	35.4	0.19
LIVE	HRC	28	282	0	110	0
	SRC	28	48.5	0.0095	17.4	0.941
TID2013	HRC	119	677	0	2584	0
	SRC	24	587	0	0.49	1

2.4. SD Predictability

Based on the results in the previous section, an intriguing question arises: can the variability of ratings be predicted from the image content alone? HRCs are usually designed to create different quality levels, but could SRCs be selected to target different rating variabilities?

Table 3: Correlation coefficients between SI and median $\hat{\sigma}$.

Database	Pearson	Spearman
CSIQ	−0.249	−0.172
LIVE	0.086	0.022
TID2013	0.730	0.816

In an attempt to address this question in a very basic way, we use spatial information (SI) for each reference image (SRC) in the databases as a simple feature for predicting variability. We then compute the median $\hat{\sigma}$ across HRCs for each SRC and correlate those with SI. The results (Table 3) are somewhat inconclusive: While a strong positive correlation is observed for the largest database (TID2013), there is practically none for the other two, despite the fact that they share some source images. One possible reason for this may be the much larger number of HRCs in TID2013. Furthermore, different subjective rating methods were used to gather the data in each database (only LIVE DMOS comes from a straightforward ACR rating method), which could have an impact on the results.

3. CONCLUSIONS

We investigated factors determining inter-subject variability for quality assessment experiments using three large image quality databases. We found that both SRC and HRC have a statistically significant effect on rating variability. Our first and very basic attempt to model this effect in a predictive manner only works for one of the three databases; other features or more sophisticated models may yield better results.

4. REFERENCES

[1] L. Janowski, Z. Papir: "Modeling subjective tests of quality of experience with a Generalized Linear Model." in *Proc. QoMEX*, 2009.

[2] E. C. Larson, D. M. Chandler: "Most apparent distortion: Full-reference image quality assessment and the role of strategy." *J. Electronic Imaging* **19**(1), 2010, http://vision.okstate.edu/index.php?loc=csiq.

[3] M. Mu, A. Mauthe, G. Tyson, E. Cerqueira: "Statistical analysis of ordinal user opinion scores." in *Proc. IEEE Consumer Communications and Networking Conference (CCNC)*, 2012.

[4] N. N. Ponomarenko et al.: "A new color image database TID2013: Innovations and results." in *Proc. ACIVS*, vol. 8192, 2013, http://www.ponomarenko.info/tid2013.htm.

[5] H. R. Sheikh, M. F. Sabir, A. C. Bovik: "A statistical evaluation of recent full reference image quality assessment algorithms." *IEEE Trans. Image Processing* **15**(11):3440–3451, 2006.

[6] H. R. Sheikh, Z. Wang, L. Cormack, A. C. Bovik: "LIVE image quality assessment database release 2." 2006, http://live.ece.utexas.edu/research/quality/subjective.htm.

[7] S. Winkler: "On the properties of subjective ratings in video quality experiments." in *Proc. QoMEX*, 2009.

[8] S. Winkler: "Analysis of public image and video databases for quality assessment." *IEEE J. Sel. Topics Sig. Proc.* **6**(6):616–625, 2012.