A Statistical Paradigm for Assessment of Subjective Image Quality Results

Matthew D. Cutone[1], Marc Dalecki[2], James Goel[3], Laurie M. Wilcox[1], Robert S. Allison[4]

[1] Department of Psychology, Centre for Vision Research, York University, 4700 Keele St., Toronto, ON, Canada, M3J 1P3; +1 416-736-5659; cutonem@yorku.ca; lwilcox@yorku.ca
[2] School of Kinesiology, Louisiana State University, Baton Rouge, USA; +1 225-578-6087; mdalecki@lsu.edu
[3] Qualcomm Canada Inc., 100-105 Commerce Valley Drive West, Markham, Ontario, Canada L3T 7W3; +1 (905) 482-5765; jgoel@qti.qualcomm.com
[4] Department of Electrical Engineering & Computer Science, Centre for Vision Research, York University, 4700 Keele St., Toronto, ON, Canada, M3J 1P3; +1 416-736-5659; allison@cse.yorku.ca

Abstract

ISO/IEC 29170-2 outlines a subjective procedure for assessing codec quality for near-threshold artifacts. Here we outline a statistical method for analyzing these data using Generalized Linear Mixed-Models (GLMMs). This procedure provides insightful metrics concerning the relative performance of two or more codecs that may aid in the perceptually-guided development and selection of novel codec technologies.

Author Keywords

subjective quality assessment; image compression; statistical modeling

1. Objective and Background

Objective image quality assessment metrics such as Peak Signal-to-Noise ratio (PSNR) and S-CIELAB [1] are often used to quantify the reconstruction error of lossy codecs. While the results of such measures can be useful for codec evaluation, they do not necessarily or reliably predict human sensitivity to artifacts. In fact, subjective preference is sometimes non-intuitive, because distortion introduced during signal reconstruction may be aliased or attenuated by the sensory system. As humans are the likely end-users of visual media, subjective image quality assessment methods which are sensitive to the nuances in human visual perception are essential. Hoffman and Stolitzka [2] presented a psychophysical method to assess the detectability of barely visible image artifacts; forming the basis of the ISO/IEC 29170-2 [3] standards document. The procedure involves temporally interleaving a source picture with its decompressed version, alternating at some fixed frequency, where any perceptually relevant distortions will appear to scintillate. Hoffman and Stolitzka claim this better reflects real-world media viewing, where frames compressed at different bitrates are temporally interleaved in-stream. The ISO/IEC 29170-2 document provides a reporting guideline with recommended descriptive statistics and graphs for analyzing and presenting results. This method was employed recently to conduct a large-scale evaluation of the VESA Display Stream Compression 1.2 [4]. While such an approach is useful for assessing the visibility of artifacts over a population of users. In other contexts, one of the major challenges using subjective measures like this is inter-observer variability. The ISO/IEC standard describes a means of determining whether a codec is 'lossy' for a given use case but provides no guidelines for comparing relative artifact detectability between codecs. In this paper we introduce a within subjects' analysis approach to improve the precision and power of subjective codec comparisons. Applying statistical models to the data addresses this question, allowing one to test hypotheses and make statistical inferences. We report a logistic-regression procedure that applied the General Linear Mixed-Effects Model (GLMM) to subjective response data (obtained using the ISO/IEC standard), with intent to determine artifact detectability differences between two anonymous codecs.

2. Methods

We considered the task of comparing the relative performance of two codecs, designated A and B. Without loss of generality we assume codec B is a reference codec to which codec A compared. For each image, the following hypotheses regarding relative log-likelihood differences were tested:

- **H_0**: Codec A has similar or lower detection rate relative to B ($A - B \leq 0$)
- **H_1**: Codec A has a higher detection rate relative to B ($A - B > 0$)

Note that here we are not assessing whether a codec is lossy or not, instead we ask if it performs poorly relative to the reference codec tested within the same experiment. Such a scenario may arise during codec development where one wishes to know if changes to an underlying algorithm made artifacts more conspicuous, or if one codec is interchangeable for another in a similar application. While the focus here is on codec comparison, the techniques described here can be used to model the effects of parameter settings such as bitrate on the expected visibility of image artifacts or make other quantitative predictions.

2.1 Data: To illustrate the approach we used extant data collected from undergraduate students (N = 21) who met the observer selection criteria outlined in ISO/IEC 29170-2. Static reference images were presented alongside the temporally interleaved image; in a forced-choice paradigm, participants were tasked to indicate which of the pair appeared to be scintillating. Each image and codec combination were presented to each observer 30 times, resulting in a total of 360 trials per codec per subject. Each trial had an associated dichotomous response corresponding to whether the compressed image was detected or not. From pure chance one expects a correct answer on 50% of trials and if an artifact were highly visible we should see nearly 100% correct responses.

2.2 Analysis: The use of repeated-measures in the flicker paradigm resulted in non-independent clusters of observations grouped by participant. Due to individual differences in sensitivity to artifacts, subject-wise clusters have their own statistical moments that may vary between clusters, potentially leading to the logistic-regression model being dispersed; where the observed residual variance is not well predicted by the model. Here, we performed a logistic-regression using the Generalized Linear Mixed-Effects Model (GLMM). GLMMs are like a conventional Generalized Linear Models (GLMs); modeling the log-odds of some non-normal response given experimentally manipulated predictor variables (fixed-effects). However,

ISSN 0097-996X/18/4703-1312-$1.00 © 2018 SID

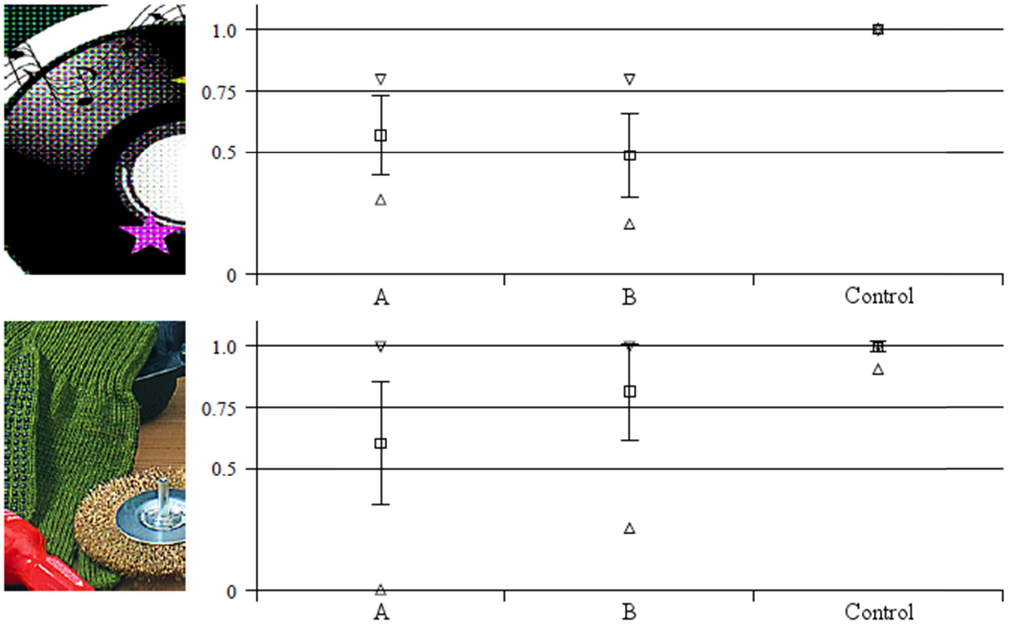

Figure 1. Plots of descriptive statistics for 'HintergroundMusik' (top) and 'Tools' (bottom) test images in the reporting format specified in ISO/IEC 29170-2. The square represents the mean of averaged scores which is interpreted as the estimated artifact detection rate for that codec. The triangles present the range of the data; as per the ISO/IEC standard, the codec is lossy for that image if the upper range falls somewhere above the 0.75 line. The error bars show the standard deviation of the data.

GLMMs can also incorporate random effects, which are uncontrolled correlation/variance inherent to a cluster of samples [5]. In the present case, random effects model within-observer parameters that are used to estimate common fixed effects parameters, assuming they arise from the same population distribution [6]. In practice, these fixed-effects facilitate generalizations as the parameters better capture the nature of the population rather than just the sample. The analysis was conducted with R (v3.4.2), a statistical computing environment; GLMMs were fitted to the data using the 'glmer' routine supplied by the 'lme4' package [7]. Pairwise comparisons were done using routines in the 'lsmeans' package [8]. A binary response (correct) was modeled with image and codec specified as categorical predictor variables (fixed effects) whose interaction was also modeled. Within-subject clusters were treated as non-independent groups in the random effects term.

3. Results

Figure 1 shows plots in the reporting format prescribed by the ISO/IEC 29170-2 standard. The plot cannot be used for comparing codecs directly as the variability shown is the inter-subject variability; however, it does indicate the extensive range of variation in response across subjects. This variability reduces statistical power in comparing the codecs unless it is accounted for.

3.1 Overall Effects: A Wald chi-squared test applied to the fitted GLMM model shows a significant main effect of image (χ2(11, N = 21) = 120.681, p < 0.001), where detection rates differed on average. This was expected as some images are more challenging to encode than others, biasing detection rates regardless of codec. However, there was no significant main effect of codec (p = 0.681), expressing artifact detection probabilities between codec A and B were on average similar when collapsing across all images. There was a significant interaction effect between image and codec (χ2(11, N = 21) = 43.067, p < 0.001), showing an association between codec and image which varies relative detection rates. This may be attributed to certain images being more challenging for one of the codecs used.

Table 1. Table showing log-odds differences between codecs A and B. Positive *β*- and *z*-values indicate cases where codec A is more likely than B to produce detectable artifacts. Whether the difference is statistically significant (satisfies H_1: A – B > 0) at the 0.05 level is indicated with an asterisk (*).

Image	*β*	*z-score*	*p*
Barbara	-.085	-.411	.659
CircuarPattern26	-.021	-.103	.541
Clipboard	.043	.207	.418
FemaleHorseFly	-1.077	-4.424	1.000
HintergroundMusik	.362	1.750	.040*
Landscape102	-.381	-1.302	.904
Mandrill	-.585	-2.684	.996
MosaicBroadcom	-.093	-.432	.667
MysticMountain	-.311	-1.420	.922
Noise	.228	.673	.250
Peacock	-.465	-2.201	.986
Tools	-1.123	-4.619	1.000

3.2 Pairwise Comparisons: Multiple pairwise comparisons determined which cases had a significant difference in detection rates. Tests were parametrized to test the specific hypotheses mentioned previously; they were one-tailed (right) with a confidence level of 0.95. The resulting *p*-values were false detection rate corrected [9] to control for error inflation from multiple comparisons. A single significant difference between in codec A and B means was found for 'HintergroundMusik' ($p = 0.0392$), where the fitted parameter ($\beta = 0.362$, $SE = 0.207$) indicated that codec A was about 44% more likely to have detectable artifacts than B. Codec A was not significantly worse than codec B on any other image condition (see Table 1 for full results).

4. Impact

We have demonstrated a procedure to assess subjective differences in artifact detectability between codecs using GLMMs. It was shown that codec A had a significantly higher artifact detection rate than B for a single image in our set. In all other cases, codec A and B may be interchangeable for a given application. This statistical procedure can increase the utility of extant subjective quality assessment data, permitting one to make statistical inferences that can inform a codec's development.

5. References

[1] X. Zhang and B. Wandell, "A spatial extension of CIELAB for digital color reproduction" Journal of the Society for Information Display **5**, 731-734 (1996).

[2] D. M. Hoffman and D. Stolitzka, "A new standard method of subjective assessment of barely visible image artifacts and a new public database" Journal of the Society for Information Display **22(12)** 631–643 (2014).

[3] "Information technology — Advanced image coding and evaluation — Part 2: Evaluation procedure for nearly lossless coding," International Organization of Standards, Geneva, Switzerland, ISO/IEC 29170-2:2015, (2015).

[4] R. S. Allison, L. M. Wilcox, W. Wang, D. M. Hoffman, Y. Hou, J. Goel, L. Deas, D. Stolitkza, "Large Scale Subjective Evaluation of Display Stream Compression" Society for Information Display – Digest **75(2)** 1101-1104 (2017).

[5] A. Agresti, "An introduction to categorical data analysis." John Wiley & Sons Inc. (2007).

[6] F. Tuerlinckx, F. Rijmen, G. Verbeke and P. Boeck, "Statistical inference in generalized linear mixed models: A review." British Journal of Mathematical and Statistical Psychology, **59(2)** 225–255 (2006, nov).

[7] D. Bates, M. Mächler, B. Bolker and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4" Journal of Statistical Software **67(1)** 1-48 (2015).

[8] R. V. Lenth, "Least-Squares Means: The R Package lsmeans" Journal of Statistical Software **69(1)** 1-33 (2016).

[9] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. 289–300 (1995).