Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 69 (2015) 55 – 65

Procedia Computer Science

7th International Conference on Advances in Information Technology

Enhancing Reliability through Screening and Segmentation: An Online Video Subjective Quality of Experience Case Study

Mark Chignell[a], Weiwei Li[a], Alberto Leon-Garcia[a], Leon Zucherman[b], and Jie Jiang[b]

[a]*University of Toronto, 27 King's College Circle, Toronto, M5S 1A1, Canada*
[b]*TELUS Communications Company, 2455 Cawthra Road #55, Mississauga, L5A 3P1, Canada*

Abstract

In this paper we examine the reliability of subjective rating judgments along a single dimension, focusing on estimates of technical quality produced by integrity impairments and failures (non-accessibility, and non-retainability) associated with viewing video. There is often considerable variability, both within and between individuals, in subjective rating tasks. In the research reported here we consider different approaches to screening out unreliable participants. We review available alternatives, including a method developed by the ITU, a method based on screening outliers, a method based on strength of correlations with an assumed "natural" ordering of impairments, and a clustering technique that makes no assumptions about the data. We report on an experiment that assesses subjective quality of experience associated with impairments and failures of online video. We then assess the reliability of the results using a correlation method and a clustering method, both of which give similar results. Since the clustering method utilized here makes fewer assumptions about the data, it may be a useful supplement to existing techniques for assessing reliability of participants when making subjective evaluations of the technical quality of videos.

Peer-review under responsibility of the Organizing Committee of IAIT2015

Keywords: Reliability; Quality of Experience; Subjective Assessment; Cluster Analysis

1. Introduction

Monitoring and control of quality is an important aspect of many services. In some cases, quality of experience may be predicted algorithmically (e.g., [1]). However, in the case of video impairments it is not clear that an algorithmic prediction of quality of experience is feasible. For instance, people might feel that one or two instances of freezing of the video are ok, but perceive a large decrement in experienced quality if further instances of freezing

Peer-review under responsibility of the Organizing Committee of IAIT2015
doi:10.1016/j.procs.2015.10.006

occur. Since there are no obvious algorithms for predicting quality of experience in the face of video impairments it seems natural to use subjective ratings such as the mean opinion score (MOS) developed by the ITU[2] and associated researchers.

Since individual subjective ratings are subject to error, the judgments of a number of participants are typically averaged to obtain estimates of the "true" values of the construct being judged. However, there may be participants who are unmotivated, incapable of judging the construct accurately, or whose judgments may be unreliable (e.g., they are making judgments on two dimensions rather than an assumed single dimension).

In this paper we address the issue of how participants should be screened when subjectively rating aspects of services such as online video. We focus in particular on the task of subjectively rating quality of experience for videos that have impairments and failures. After reviewing a number of approaches we report on an experiment where participants rated the technical quality of online videos that had associated impairments and failures. We present a case study on using correlational, and cluster, analysis to identify unreliable participants within a sample.

2. Background

The literature on scaling in psychometrics and psychophysics has been dominated by three main tasks, namely judgments of intensity or magnitude, judgments of proximity or similarity, and hedonic (preference, or liking) judgments. Depending on one's perspective subject quality of experience (SQE) judgments can be seen as involving intensity (e.g., what was the overall quality of the experience, what was the technical quality of the experience) and/or hedonic (e.g., how much did I enjoy the experience, what is my preference for the experience vs. other experiences) components. [3] provides a relatively early review of intensity scaling methods, while [4]'s review reflects more recent interest in hedonic scaling, and [5] provide a review of multidimensional scaling of proximities, similarities and preferences.

Research on SQE (see [6] for a recent review) has generally assumed that it is possible for people to give relatively accurate ratings of their experience. As noted by [7], video quality is usually measured using a five-point scale, where a score of 1 means lowest video quality and a score of 5 means highest video quality. The justification [4]for treating humans like measuring rulers is that it often seems to work. For instance, [8] found that absolute ratings of videos presented one at a time produced "repeatable subjective results, even across different scales and different groups of participants."

However, humans are not always perfect measurement rulers and the question then arises of how to deal with inconsistencies in rating. There have been a number of attempts to deal with such inconsistencies in subjective quality of experience experiments. [9] discussed methods for dealing with inconsistency. One frequently used approach is to remove statistical outliers. This can be done on a per-trial basis or at the level of the study participant. The idea is to characterize a collection of results as a distribution (typically a normal distribution) and then to remove results as outliers if they are in the tail of the distribution (e.g., at a percentile of 97.5%, or 99%, or greater). The removal of statistical outliers can be problematic because it doesn't take into account the accuracy or consistency of judgment, but simply removes trials or participants based on statistical departures from average performance.

Another approach is to explicitly model participants as being "reliable" or "unreliable". [9] examined the issue of reliability in a challenging observational setting where participants made judgments of SQE in their own ("crowdsourced") environment (i.e., in the absence of a supervising experimenter, with relatively anonymous participants, and with no control over where the participant chooses to carry out the experiment). They asked screening questions to check if the participant was paying attention. An example of a content related screening question was "which of the following animals did you see in the video", and a quality related question example was "did you notice any stops in the video that you just watched?".

Another method for assessing reliability cited by [9] included identifying participants as reliable if their judgments had relatively high correlations with mean scores for the entire sample participants. In this paper we develop a clustering approach to differentiating participants that is distribution free and that can be carried out automatically so as to remove the possibility of bias. Our ultimate goal is to identify likely causes of unreliability and to develop a method for screening participants so as to increase experimental efficiency. We compare our clustering approach