'MEAN TIME BETWEEN FAILURES': A SUBJECTIVELY MEANINGFUL VIDEO QUALITY METRIC

Nitin Suresh and Nikil Jayant
Department of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

As digital communication of television content becomes more pervasive, the long-standing problem of assessing video quality becomes particularly important. This work describes some innovative guidelines for easy and reliable determination of a quality metric that is subjectively meaningful: Mean Time Between Failures (*MTBF*), representing how often a typical viewer observes a noticeable visual error and a related instantaneous metric relating to the fraction of viewers that find a given video portion to be within acceptable quality levels. The value of *MTBF* is addressed in the context of video quality, and objective measurements that correlate well with subjective evaluations of *MTBF* are investigated for different video clips at bit rates in the range of 1.5 - 5 Mbps. The *full-reference* objective metrics of *PSNR* and *JND* are found to have correlation coefficients of 0.69 and 0.88 respectively. In contrast, a *reduced-reference* objective metric, Spatial Temporal Join Metric [1] (*STJM*) is found to have a correlation coefficient of 0.84. A method for estimating the *MTBF* of a video sequence from objective measurements is also described.

1. INTRODUCTION

In audiovisual communications, the need for perceptually meaningful objective metrics is broadly recognized. Such measures have the dual role of (a) understanding signal quality in completed algorithm designs and (b) providing an in-the-loop metric for real-time algorithm steering. For video, Subjective testing is the ideal approach, since it involves real viewers evaluating the end output. Objective testing for video is desirable, since subjective testing takes up a lot of time and effort. Current subjective testing methodologies involve viewers rating portions of videos on a continuous / discrete scale. Objective metrics perform some operations on the output video and compare these with, if available, some information about the input video. The output of the objective metric is correlated with the subjective scores to observe how effective the former is, and the degree of correlation is used as a confidence measure for using the objective metric to evaluate video.

This work involves the estimation of subjective metrics from users in an intuitive fashion for a few test clips and evaluating the correlation of objective scores of different *full-reference* and *reduced-reference* metrics with the subjective scores. A highly correlated objective metric can be used to estimate the subjectively meaningful metrics at the output for a wide range of video content and video quality. The *reduced-reference* qualification implies the use of primitives from the reference video that need very low overhead for communication to the place of quality measurement.

In current subjective testing techniques, the discrete-point scales of Mean Opinion Score (MOS) and Mean Impairment Score (MIS) are well understood and provide useful quality measurements under conditions in which there is adequate training of subjects, and if the mean scores are appropriately qualified by a standard deviation score reflecting inter-viewer differences. There are quite a few established methods for subjective testing [2], and they involve the viewers watching different video clips and giving each clip a score, or giving a continuous score using a user feedback device like a slider or throttle. Some of the desired characteristics of a testing scheme involve ease, intuitiveness, effectiveness, and giving the user real-time feedback about the current score. In this paper, we use testing strategies based on just asking the viewers whether they observe visual artifacts or not, and demonstrate how this can be used effectively to get global quality evaluation and continuous quality evaluation metrics. The subjective tests involve different video clips from the VQEG (Video Quality Experts Group) at bit rates in the range of 1.5 - 5 Mbps. Objective measurements that correlate well with subjective evaluations of 'Mean Time Between Failures' (*MTBF*) are investigated. One of the objective metrics that is considered is the so-called *reduced-reference* metric Spatial Temporal Join Metric (*STJM*) [1]. It is designed based on principles from the VQM metric developed by VQEG [3].

The paper is organized as follows: Section 2.1 explains the intuitively appealing subjective metrics, section 2.2 describes the test environment and section 2.3 describes the calculation of the subjective metrics. Section 3.1 explains some of the existing objective metrics. Section 3.2 illustrates how objective metrics and the subjective metrics can be compared. Section 3.3 explains how the subjectively meaningful metrics can be calculated from existing objective metrics. Section 4 concludes the paper and lists some future research possible.

2. SUBJECTIVELY MEANINGFUL METRICS

2.1 Mean Time Between Failures (*MTBF*)

MTBF is a common term used in the measurement of quality of service. This is applied to subjective video quality evaluation as an intuitive global metric [4-6]. 'Probability of failure' (*PFAIL*) is a related instantaneous metric based on failure statistics, where failure corresponds to the occurrences of visual artifacts. *MTBF* represents how often visual artifacts are observed by a typical viewer; and 'Probability of Failure' reflects the fraction of viewers

that find a given video portion to be within acceptable quality levels. Whenever a perceptual artifact is observed, the viewer is asked to indicate this, say, using a buzzer. Common artifacts such as noise, blurriness, blockiness, etc. are explained and shown to the viewer prior to the testing. The viewer is allowed to keep the buzzer pressed if an entire stretch of video sequence looks bad.

The idea behind this methodology is that the viewer intuitively tends to give feedback intermittently, with a frequency correlating with how bad the video looks. Though the locations of the user responses are arbitrary for a particular viewer during a particular experiment, the results for a modest number of experiments with a sufficient number of viewers can be averaged out to generate a continuous score versus time, which is nothing but the probability that the average viewer would observe a visual artifact while watching the video. The user responses can be pooled over a period of time to determine the *MTBF* of the video. There are many advantages of this metric: it is highly intuitive, time invariant and the user need not have real-time feedback about the current score. This is intuitive, because the viewer just has to decide whether the video has any artifact or not. The metric is functional, being directly related to how consumers evaluate otherwise high-quality video. *MTBF* is not concerned with the physical categorization of an artifact, only that it is deemed visible. In this sense, it is non-diagnostic, but simple and universal.

It should be noted though, that observing the artifacts also depends on various factors such as the display type/size and viewing distance. The subjectively meaningful metrics represent the scores of a typical viewer for a typical output display environment.

2.2 Test environment

Some standard VQEG [3] sequences were used for video quality evaluation: 18 clips (*Tree*, *Barcelona*, *Music*, *EBU_Test*, *Rower*, *Race*, *Fries*, *Moving_Text*, *Rugby*, *Park*, *Building*, *Mobile*, *Cartoon*, *Waterfall*, *Football*, *Susie*, *Flower_Garden* and *BBC_Disk*) were concatenated to form a 140 second sequence. This was encoded at different bit rates (1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5 and 5.0 Mbps) using a publicly available MPEG-2 encoder, and shown to 8 different viewers. The sequences were shown on a 30-inch Television screen in a dark room, with the viewer sitting at a distance of 5 times the height. A simple MATLAB script was used to allow the user to indicate the occurrences of visual errors. Pilot experiments were conducted to estimate the delays between the occurrences of artifacts and the viewer's responses. The videos with different bit rate settings were shown in a random order to the viewers.

2.3 Determining subjective scores

Using the feedback from the user, one can determine the time locations corresponding to the artifacts that were perceived. This feedback vs. time plot for a viewer looks like an impulse train. This is low-passed filtered with respect to time using a Gaussian curve of width 1 second, to account for the viewers' inaccuracy in pinpointing the time of the error. The smoothed feedback vs. time plot for various viewers (in this case, 8) is averaged to get the *PFAIL* characteristics. Different temporal smoothening filters were used to verify that this subjective metric does not critically depend on the filter's width. The *MTBF* of the sequences is calculated as $MTBF = 1/(\text{mean}(PFAIL))$. For example, if the average probability of failure is 1/600, then an observable artifact occurs, on an average, once every 20 seconds (for a frame rate of 30 fps). Mathematically, it makes more sense to calculate *MTBF* in this manner rather than computing the mean time between viewers' responses. For example, if one had to compute the *MTBF* of a 10 second clip, and only 3 out of 8 viewers noticed some visual artifacts in the clip, then the mean times between the other viewers' responses are not available. However, computing the Probability of Failure leads to a better estimation of *MTBF*.

3. OBJECTIVE METRICS

3.1 Existing objective metrics

Objective metrics can be broadly classified based on the amount of information available about the original video. Peak Signal-to-Noise Ratio (*PSNR*) and JND are *full-reference* metrics; meaning all the pixel values of the original and the processed video are used in computing the metric. *No-reference* metrics estimate the degradation in video just by looking at the end output and having some knowledge about the characteristics of encoding and channel errors in effect. For e.g., some metrics estimate the block distortions introduced in compression algorithms [7-9], while some metrics estimate the blurring/sharpness in the video. *Reduced-reference* metrics use some prior knowledge about the original video. For e.g., a watermark might be introduced in the original, and its distortion in the processed video could be used to estimate the video quality. Alternatively, a specific signature could be calculated over the original and transmitted along with the processed video. The same signature could be calculated over the processed video and compared with the original signature to determine the quality [1]. *No-reference* and *reduced-reference* metrics are considered to be more practical than *full-reference* metrics since the original video is in general not available at the receiving end. The *Spatial-Temporal "join" metric* [1] (*STJM*) is a *reduced-reference* metric. It works by comparing some features, like the amount of visual detail, and the presence of horizontal and vertical edges calculated over the input and output videos. In this paper, *PSNR*, JND [10] and *STJM* are compared with the subjective scores calculated.

3.2 Correlation between Subjective and Objective Metrics

The subjectively meaningful measure designed can be correlated with an existing objective metric to observe the latter's effectiveness, by pooling the data over different video clips and different bit rates (Fig. 1). *MTBF* characteristics seem to exhibit an exponential type of behavior to a certain extent. For example, as the *PSNR* of a sequence increases, the *MTBF* exponentially increases up to a point after which visual artifacts are practically not visible. This can be observed from the scatter plot of *log* (*MTBF*) vs. the objective metric. The knee of the exponential curve depends on the type of sequence, though. For example, while the "Music" sequence starts to look good (meaning, high values of *MTBF*) at *PSNR* = 31 dB, the "Susie" sequence starts to look good only at around *PSNR* = 38 dB. This happens because of the objective metrics' inefficiency in estimating video quality, and also because of the variations between users that results in errors in the subjective metric itself. The "Music" sequence has a lot of spatial detail in it, which masks many of the artifacts introduced,

and this is not captured by *PSNR,* which relies on a simple difference measure.

It should be noted though, that the objective metric behaves similarly for similar sequence. For example, the relationship between *MTBF* and *PSNR* for the "Rugby" and "Rower" sequences are very similar, because both the sequences have more or less the same amount of motion and noise information, and *PSNR* behaves similarly for both. The reliability of an objective metric can be estimated by plotting the line (or in general, polynomial) of best fit in the '*log* (*MTBF*) vs. metric' graph, and noting the correlation coefficient. In other words, the *MTBF* of a video should be predictable with good accuracy from the objective metric. In this sense, the *JND* metric has a tighter fit in the plot as compared to *PSNR*. This is understandable, as *JND* incorporates spatial masking and is expected to be a better metric than *PSNR*. The reduced-reference metric *STJM* incorporates some spatial masking, and performs better than *PSNR*. The *STJM* metric experimented in this work uses a block size of 8x8 pixels for every frame. With a frame size of 720x480 and a frame rate of 30 fps, the information overhead works out to a huge 2.6 Mbps. As anticipated in [1], we have observed that by calculating the features over selective time-varying regions of the video, the overhead can be reduced by about a factor of 50 without a significant change in the performance of the metric, as measured by the correlation of its scores with subjective scores of *MTBF*.

3.3 Estimating subjectively meaningful metrics from existing objective metrics

An ideal scenario for video quality evaluation would be to estimate the values of subjectively meaningful metrics from a no-reference metric or a reduced-reference objective metric like *STJM* [1]. By observing the scatter plot of the '*log* (*MTBF*) vs. *STJM*' graph (averaged over all viewers for different test clips at different bit rates), the exponential of best fit is determined to find the typical relationship between *STJM* and *MTBF* (Fig. 1). Since *MTBF* is inversely related to the instantaneous metric *PFAIL*, the typical relationship between *STJM* and *PFAIL* is also easily calculated and a lookup table can be generated. With this relationship, it is easy to calculate the *MTBF* of any arbitrary video: The *STJM* vs. time of the corrupted video is calculated, the 'probability of failure' vs. time is estimated using the lookup table, and then *MTBF* is calculated for the overall video sequence as 1/(mean (*PFAIL*)). Objective scores cannot directly predict the occurrences of visual artifacts. They can only predict the chances of an artifact being observed by a typical viewer. This is the reason why the instantaneous values of 'probability of failure' are calculated before calculating *MTBF*.

4. CONCLUSIONS

Innovative methods for estimating the quality of video are investigated. Video quality is described in units that are subjectively meaningful. 'Mean Time Between Failures' (*MTBF*) is a highly intuitive concept and represents how often visual artifacts are observed by a typical viewer; and 'Probability of Failure' reflects the fraction of viewers that find a given video portion to be within acceptable quality levels. These are functionally extremely pertinent for entertainment video and reflect measures that video solution providers tend to use. The testing method is potentially much easier than the current subjective testing procedures and due to the ease and intuitiveness of this method, extensive subjective test data can be collected with a large range of subjects from the naïve to the expert. Unlike the five-point methodology, the *MTBF* measure has a potential to be robust in the context of heterogeneous stimuli, such as video signals (and displays) of varying spatiotemporal resolution. The only important teaching at the beginning of the test is exposure to artifacts that may already exist in anchoring material. The correlation of *MTBF* with objective metrics produces meaningful results and a method for estimating the *MTBF* of a video sequence using an objective metric is explained. Estimation of this metric can be used for real-time monitoring of video quality, and also for tuning existing video coding algorithms

Future work will include the use greater rigor in experimental design and the observation of test results for different video content & formats like VGA, QVGA, QCIF for various encoding schemes like H.264 and MPEG4. Additional work on statistical methods for correlating candidate objective measurements with the subjective metric is also planned. The design of the objective metrics used for this purpose shall also be improved upon, in the areas of spatial/temporal masking and pooling. Predicting the subjectively meaningful metrics using reduced-reference objective metrics is a commercially attractive topic, and the role of the communication overhead and computational complexity involved in this system's effectiveness shall also be looked into.

5. REFERENCES

[1] S. Wolf, and M. H. Pinson, "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system", *Proc. of SPIE Int. Symp. on Voice, Video, and Data Commun.,* Boston, MA, Sept. 1999

[2] M. Pinson and S. Wolf, "Comparing Subjective Video Quality Testing Methodologies", *SPIE Video Communications and Image Processing Conference,* Lugano, Switzerland, Jul. 8-11 2003

[3] http://www.its.bldrdoc.gov/vqeg/

[4] N. Suresh and N. Jayant, "Mean Time Between Failures: A Functional Quality Metric for Consumer Video", *Invited Talk, First Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics,* Scottsdale, AZ, Jan. 23-25, 2005

[5] N. Suresh and N. Jayant, "Subjective Video Quality Metrics based on Failure Statistics", *IASTED Int. Conference on Circuits, Signals and Systems*, Marina del Rey, CA, USA, Oct. 24-26, 2005

[6] N. Suresh and N. Jayant, "Mean Time Between Failures: A Subjectively Meaningful Quality Metric for Consumer Video", *Invited Talk, Second Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics,* Scottsdale, AZ, Jan. 22-24, 2006

[7] T. Vlachos: "Detection of blocking artifacts in compressed video." *Electronics Letters 36(13)*: 1106–1108, 2000.

[8] Z. Wang, A. C. Bovik, B. L. Evans: "Blind measurement of blocking artifacts in images." in *Proc. ICIP*, *vol. 3*, pp. 981–984, Vancouver, Canada, 2000.

[9] H. R. Wu, M. Yuen: "A generalized block-edge impairment metric for video coding." *IEEE Signal Processing Letters 4(11):*317–320, 1997.

[10] JNDmetrix http://www.sarnoff.com/products_services/video_vision/jndmetrix/visual_quality.asp

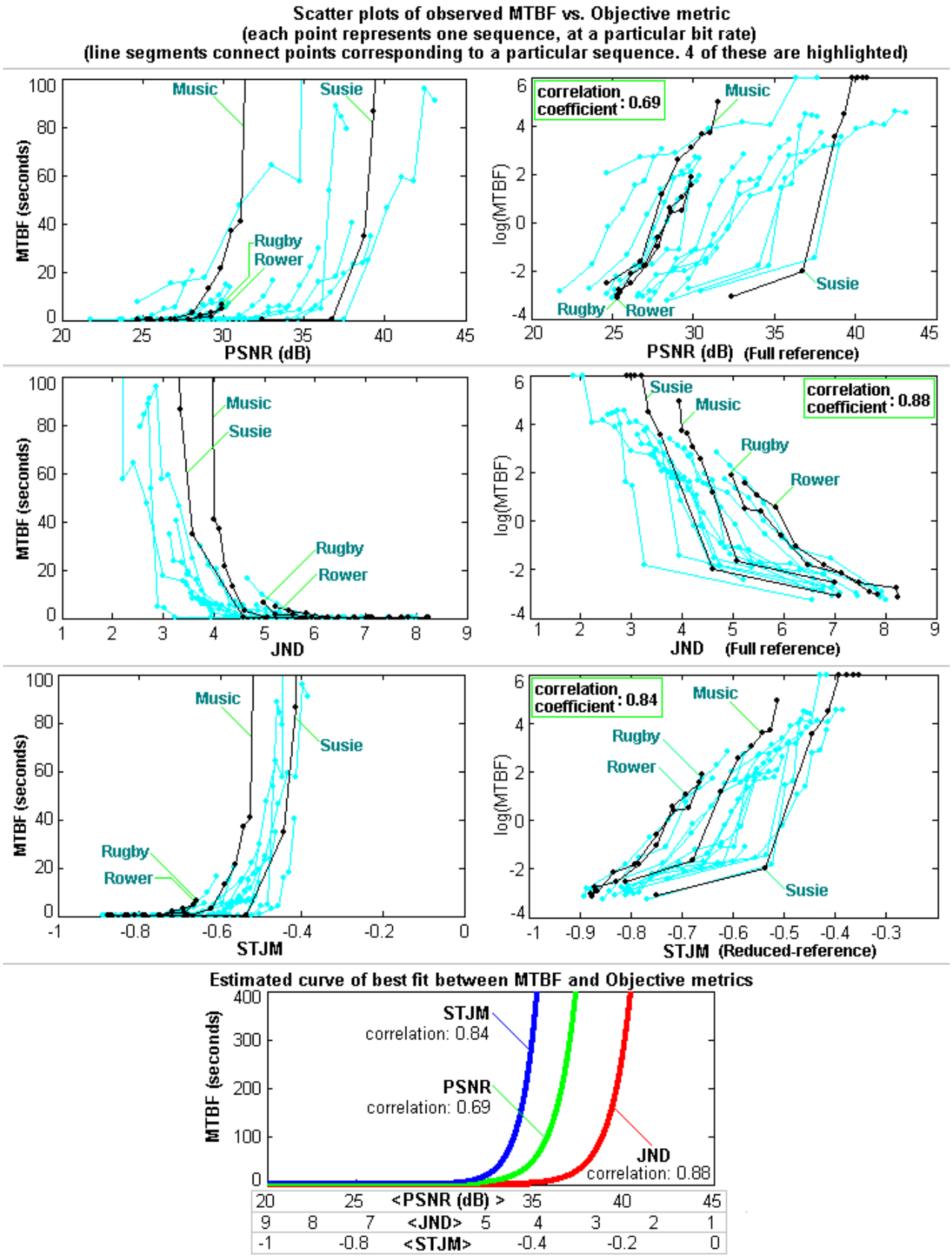

Fig. 1: Relationship between *MTBF* and objective metrics