TEMPORAL HYSTERESIS MODEL OF TIME VARYING SUBJECTIVE VIDEO QUALITY

Kalpana Seshadrinathan

Intel Labs
Intel Corporation, Santa Clara, CA.

Alan C. Bovik

Lab. for Image and Video Engg. (LIVE)
University of Texas at Austin.

ABSTRACT

Video quality assessment (QA) continues to be an important area of research due to the overwhelming number of applications where videos are delivered to humans. In particular, the problem of temporal pooling of quality sores has received relatively little attention. We observe a hysteresis effect in the subjective judgment of time-varying video quality based on measured behavior in a subjective study. Based on our analysis of the subjective data, we propose a hysteresis temporal pooling strategy for QA algorithms. Applying this temporal strategy to pool scores from PSNR, SSIM [1] and MOVIE [2] produces markedly improved subjective quality prediction.

Index Terms— video quality, temporal pooling, MOVIE, quality assessment, LIVE Video Quality Database, hysteresis.

1. INTRODUCTION

Automated measurement of video quality has proven to be critical with the rising popularity of video applications that target human end users such as mobile video, video over the Internet, teleconferencing and video on demand. Video quality assessment algorithms (QA) attempt to measure the perceptual quality of a given video and usually generate spatially and temporally localized quality estimates, which are then combined to predict the overall quality of the video. Considerable work has been done in developing QA algorithms for video and several papers study pooling of spatially localized quality scores to produce frame level quality indices. Study of human behaviour while providing time-varying quality scores has received relatively little attention in the literature. Recency, forgiveness and negative peak duration neglect effects were reported based on data gathered using a single stimulus continuous quality evaluation (SSCQE) paradigm in [3] and a model that considers these effects was proposed in [4]. Smoothness of subjective time-varying quality scores was observed and modeled in [5]. A pooling method based on the hypothesis that subjective time-varying quality scores are smooth, asymmetric and saturating was proposed in [6].

Existing work is largely based on empirical observations of human perception of quality and do not directly attempt to study human behaviour in assessing instantaneous video quality and aggregating instantaneous quality to provide an overall impression of quality. We present the results of a study where we studied the relation between time-varying quality scores and the final quality score assigned by human subjects to a video using a subjective study that records both of these from human subjects. We have observed that there exists a hysteresis effect in the subjective judgment of video quality, whereby continuously recorded quality scores that trace a hight level of video quality rapidly transition to tracing a low level of quality following a distortion event, remaining there even after the event passes.

We reported the results of a human subjective study as the now publicly available LIVE Video Quality Database, where human subjects were asked to provide their quality judgment at the end of the presentation of a video sequence [7]. As part of this study, we also recorded time-varying quality scores from human subjects as the video was played out and this aspect of the study has not been reported yet. Based on our analysis of the subjective data, we propose a new hysteresis temporal pooling strategy for QA algorithms. We find that applying this temporal strategy to pool scores from such objective QA indices as PSNR, SSIM [1] and MOVIE [2] produces marked improvement as measured on the LIVE Video Quality Database.

2. SUBJECTIVE EXPERIMENT

We recently conducted a subjective study to assess the time-varying subjective quality of videos. The study deployed 10 uncompressed reference videos of natural scenes that span a wide range of content. All reference videos used in our study were progressively scanned, 768x432 pixels and in YUV 4:2:0 format. The frame rates of these videos were 25fps (7 videos) or 50 fps (3 videos). 9 videos were 10 seconds long, while one was 8.68 seconds long. We generated 150 distorted videos from the references using four different distortion types - MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP networks and through error-prone wireless networks. The distortion types included in our study were fairly diverse and included spatially and temporally uniform and transient distortions.

Each video in the LIVE Video Quality Database was assessed by 38 human subjects in a single stimulus study with hidden reference removal. All the videos in our study were viewed by each subject, which required one hour of the subject's time. To minimize the effects of viewer fatigue, we conducted the study in two sessions of thirty minutes each and the videos were played back to the subjects in a randomized order. The interface was designed to ensure precise playback of the video stimulus. The videos were viewed by the subjects on a calibrated Cathode Ray Tube (CRT) monitor. The monitor resolution was set to 100 Hz to avoid artifacts due to monitor flicker.

Each subject scored the quality of the video in two different ways. First, during the presentation of each video, a sliding bar scale for video quality was displayed on the screen. The quality scale had five labels marked on it to help the subject. The left end of the scale was marked "Bad" and the right end was marked "Excellent". Three equally spaced labels between these were marked "Poor", "Fair" and "Good", similar to the ITU-R Absolute Category Rating (ACR) scale. The cursor was set at the center of the quality scale at the beginning of playback of each video to avoid biasing the subject's quality percept. The subjects were asked to indicate the quality of the video they were viewing in a time varying fashion as

the video was played out by moving the mouse along the sliding bar. This procedure is similar to the SSCQE paradigm. A screenshot of the interface showing the video playback along with the quality scale is shown in Figure 1(a). Secondly, at the end of the playback of the video, a sliding bar scale for video quality was again displayed on the screen with the cursor at the center of the scale. Subjects were then asked to indicate the overall quality of the entire video that they just completed viewing as shown in Figure 1(b). This discrete score for each video was reported in the LIVE Video Quality Database and we refer the reader to [7] for further details.

3. ANALYSIS OF SUBJECTIVE RESULTS

We present the results of the time varying element of the subjective experiments in this section. We first studied the relationship between the continuous time scores and the final score assigned to each video by each subject. Let q_{ijk} denote the score assigned by subject i at the end of the presentation of video j in session $k = \{1, 2\}$. Let $f_{ijk}(t)$ represent the continuous time scores obtained from subject i during the playback of video j in session $k = \{1, 2\}$. The final scores obtained from the subjects were processed in a manner similar to that described in [7] to obtain Mean Opinion Scores (MOS) for each video. The main difference from the analysis conducted in [7] is that we did not compute difference scores, since our goal was to study the temporal pooling strategies used by humans. Scores obtained from each subject per session is first converted to Z-scores per session. The Z-scores were then converted to MOS scores by averaging them across subjects after subject rejection [7]:

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} q_{ijk}, \sigma_{ik}^2 = \frac{1}{N_{ik} - 1} \sum_{j=1}^{N_{ik}} (q_{ijk} - \mu_{ik})^2 \quad (1)$$

$$\zeta_{ij} = \frac{q_{ijk} - \mu_{ik}}{\sigma_{ik}}, \text{MOS}_j = \frac{1}{M} \sum_{i=1}^{M} \zeta_{ij} \quad (2)$$

where N_{ik} is the number of test videos seen by subject i in session k and $M = 32$ denotes the number of subjects (out of 38) whose scores were accepted.

A similar analysis was performed on the continuous time quality scores obtained from each subject to obtain continuous time MOS scores:

$$m_{ik} = \frac{1}{\sum_{j=1}^{N_{ik}} T_j} \sum_{j=1}^{N_{ik}} \sum_{t=1}^{T_j} f_{ijk}(t) \quad (3)$$

$$s_{ik}^2 = \frac{1}{\sum_{j=1}^{N_{ik}} T_j - 1} \sum_{j=1}^{N_{ik}} \sum_{t=1}^{T_j} (f_{ijk}(t) - m_{ik})^2 \quad (4)$$

$$z_{ij}(t) = \frac{f_{ijk}(t) - m_{ik}}{s_{ik}}, \text{MOS}_j^f(t) = \frac{1}{M} \sum_{i=1}^{M} z_{ij}(t) \quad (5)$$

Here, T_j denotes the duration of video j. We first studied the relation between the statistics of $\text{MOS}_j^f(t)$ and MOS_j. The cursor was always placed at the center of the slider when playback started and we found that subjects took on average about a second to respond to the quality of the video. We discarded the scores in $\text{MOS}_j^f(t)$ during the first second of playback of the video. We found that the mean of the continuous time quality scores was a very good indicator of the discrete score assigned by the subjects to the video at the end of the presentation. The linear correlation coefficient between the mean of the continuous time quality scores assigned by each subject (after ignoring the first second) and the final score assigned by the subject is 0.9620. A scatter plot is shown in Figure 2. We also performed an F-test on the mean of the continuous scores and the final scores and the hypothesis that the two data sets came from populations with the same mean was accepted at 95% confidence.

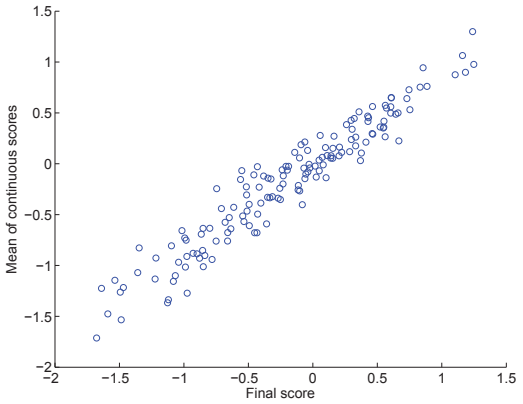


Fig. 2. Scatter plot showing final scores assigned by subjects against the mean of the continuous scores.

It has been reported that the final quality judgment from the subjects is influenced heavily by the quality of the last segment of the video [3]. However, we found that schemes where continuous time scores from later segments of the video were weighed more heavily did not correlate as well with the final score as the overall mean on our database. We believe that this is possibly due to the short duration of the videos used in our study (8-10 seconds) and the fact that we used natural videos where the temporal nature of the distortions varied considerably. This is also inline with the results of [4], where this effect was found to be negligible for 8 second long videos.

Our analysis suggests that computing the average of frame level quality scores from QA algorithms can serve as a good indicator of the overall quality of the video, provided that frame level quality scores from the QA algorithm match continuous time quality scores from human subjects. However, we found that the continuous time scores provided by human subjects follow a smoother trajectory over time. Further, subjects react sharply to drops in video quality and provide poor quality scores for such regions and do not react as sharply to improvements in quality thereon. We refer to this as a hysteresis effect, since the memory of poor quality elements in the past causes subjects to provide lower quality scores immediately afterward. Since this memory is retained even after the time varying video quality returns to acceptable levels, it is a form of subjective hysteresis.

Objective algorithms, on the other hand, tend to be less smooth, react quickly to improvements in quality and do not consider the effects of prior quality scores. This is illustrated in Figure 3, which shows subjective continuous time quality scores against objective scores from the Temporal MOVIE algorithm for 2 videos in the database [2]. The Temporal MOVIE index is computed every 8 frames and the resulting values are scaled for visibility and interpolated using straight lines in the figure. In the figure shown on the left, notice that Temporal MOVIE reacts to the sudden drop in video quality indicated by the subjects but recovers quickly as the quality improves over time, while human subjects do not react to the improvement as sharply. In the figure on the right, Temporal MOVIE scores show multiple frames where video quality drops sharply, which re-

(a) (b)

Fig. 1. (a) Screenshot from the subjective study interface displaying the video to the subject with the time-varying quality scale. (b) Screenshot from the subjective study interface that prompts the subject to enter a quality score for the video they completed viewing.

sults in a steadily decreasing subjective percept of quality. Note that the human response to quality lags the waveform obtained from the Temporal MOVIE index, which accounts for the delay between the subject viewing a change in quality and indicating this change using the interface. In the next section, we model these effects in the pooling stages of objective video QA algorithms to improve their performance in matching human perception.

4. TEMPORAL POOLING STRATEGY FOR QA

We propose a new temporal pooling strategy to account for the subjective effects seen in Section 3. We hypothesize that the average of frame level quality scores obtained from objective QA algorithms can serve as a good indicator of the overall quality of the video, provided that modeling of the memory effect and the sharp reaction of subjects to drops in video quality is accounted for. Let $g(t)$ represent time varying scores obtained from an objective QA algorithm for a video. First, we recursively define a memory component to quality at each time instant t_i using quality scores obtained from the QA algorithm over the previous $t = \tau$ seconds:

$$x(1) = g(1) \tag{6}$$

$$x(t_i) = \min[x(t)], t = \{\max(1, t_i - \tau), t_i - 1\} \tag{7}$$

The minimum of the quality scores over the previous τ seconds accounts for the fact that subjects are intolerant of poor quality video events. The recursive computation ensures that the memory element x is smoothly varying, similar to the behavior of human subjects as shown in Section 3. While a model that linearly combines quality estimates for the current frame with that of the previous frame has been proposed [4], this model is not based on subjective experiments of human performance and is very different from the memory model that we propose here.

We also construct a current quality element at each time instant t_i using quality scores obtained from the QA algorithm in the next $t = \tau$ seconds. To account for the fact that subjects respond strongly to drops in quality, we sort the quality scores in ascending order and combine them using a Gaussian weighting function [8]. Let $\mathbf{v} = \{v_1, v_2, \ldots, v_k\}$ denote the sorted elements and $\mathbf{w} = \{w_1, w_2, \ldots, w_k\}$ represent the descending half of a Gaussian weighting function that sums to 1 ($1/K \sum_{k=1}^{K} w_k = 1$). The standard deviation of the Gaussian window was chosen to be $(2K - 1)/12$.

$$\mathbf{v} = \text{sort}[g(t)], t = \{t_i, \min(t_i + \tau, T_j)\} \tag{8}$$

$$y(t_i) = \sum_{k=1}^{K} v_k w_k, k = \{1, 2, \ldots, k\} \tag{9}$$

We then linearly combine the memory and current elements of quality to produce time varying quality scores that account for the hysteresis effect and approximate the continuous time quality judgments from human subjects. The overall video quality is computed as the mean of the time varying scores, accounting for the finding that the overall subjective quality assigned by humans is well approximated by the mean of the continuous time quality scores.

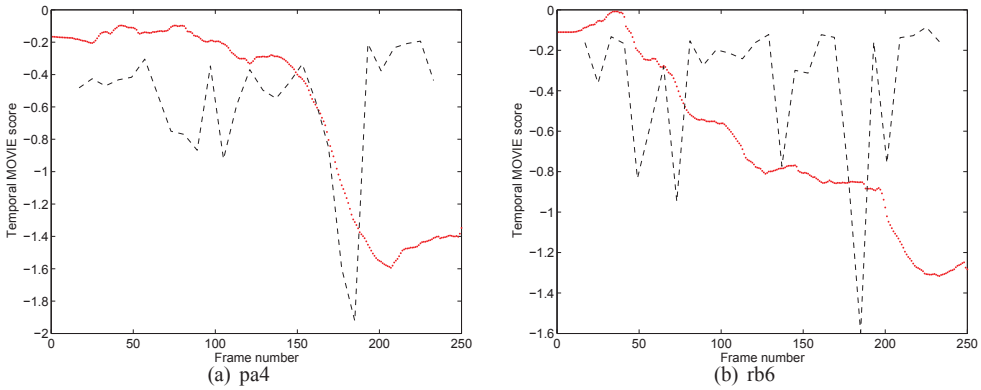
$$g'(t_i) = \alpha y(t_i) + (1 - \alpha) x(t_i) \tag{10}$$

$$G = \frac{1}{T} \sum_{t} g'(t) \tag{11}$$

5. RESULTS AND CONCLUSION

We tested the performance of PSNR, SSIM [1] and MOVIE [2] using the hysteresis pooling strategy on the LIVE Video Quality Database [7]. The original implementations of PSNR and SSIM [1] use the mean of the quality scores computed at each frame as the frame level quality score. However, it has been pointed out that humans are sensitive to small regions of poor quality in a video and the mean often overestimates frame level qualities. Pooling using the coefficient of variation (CoV) have been proposed to improve the spatial pooling of quality scores [7]. We found that the performance of both PSNR and SSIM improves considerably by using the CoV to pool the quality scores at each frame.

We measured the efficacy of the hysteresis temporal pooling strategy in terms of the Spearman rank order correlation coefficient (SROCC) and the linear correlation coefficient (LCC) on the LIVE Video Quality Database in Tables 1 and 2. MOVIE uses the CoV for pooling and entries for pooling using the mean in Tables 1 and 2 are hence left blank. We have two hysteresis parameters in our model: the duration of the memory effect modeling τ and the linear factor α. We found that $\tau = 2$ seconds and $\alpha = 0.8$ produced good results.



(a) pa4 (b) rb6

Fig. 3. Continuous time scores (solid line) from human subjects and objective scores from Temporal MOVIE (dashed line) for two videos.

Algorithm	M1	M2	M3
PSNR	0.5398	0.6058	0.6256
SSIM	0.5257	0.6974	0.7330
Spatial MOVIE	-	0.7270	0.8009
Temporal MOVIE	-	0.8055	0.8150
MOVIE	-	0.7890	**0.8394**

Table 1. SROCC of VQA algorithms for different pooling strategies. M1: spatial & temporal means, M2: spatial CoV & temporal mean. M3: spatial CoV and hysteresis based temporal pooling. The best performing algorithm is highlighted in bold font.

Algorithm	M1	M2	M3
PSNR	0.5621	0.6224	0.6272
SSIM	0.5444	0.7166	0.7460
Spatial MOVIE	-	0.7451	0.8176
Temporal MOVIE	-	0.8217	0.8262
MOVIE	-	0.8116	**0.8524**

Table 2. LCC of VQA algorithms for different pooling strategies. See caption of Table 1 for definitions of M1,M2,M3. The best performing algorithm is highlighted in bold font.

Since τ corresponds to the duration of the memory effect, a value of 2 seconds seems reasonable based on the human data illustrated in Figure 3. We found that the results don't vary significantly as τ varies from 1-3 seconds. α controls the contribution of the memory effect to the linear weighting and we found that the results are similar for values of α above 0.5. This shows that the subjective quality estimate is weighted more by the current element of quality, as compared to the memory element. It is seen that performing spatial pooling using CoV improves the performance of PSNR and SSIM considerably when compared to spatial pooling using the mean. The performance of all algorithms is improved by incorporating hysteresis based temporal pooling. These results are quite promising since the underlying QA algorithms are unchanged and the gains are solely due to better temporal pooling of quality scores.

In conclusion, we proposed a new hysteresis based temporal pooling strategy for video QA algorithms based on the results of a subjective study. We showed that scores assigned by humans to a video can be approximated quite well using the mean of their continuous time quality scores. We also demonstrated differences between the temporal evolution of quality scores obtained from human subjects and scores from objective QA algorithms. We proposed a hysteresis model for temporal pooling of quality scores and demonstrated that it performs quite well and results in improvements in the performance of three full reference video QA algorithms.

6. REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[2] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[3] D. E. Pearson, "Viewer response to time-varying video quality," in *Proc. SPIE*, vol. 3299, no. 1, 1998.

[4] M. Barkowsky, B. Eskofier, R. Bitto, J. Bialkowski, and A. Kaup, "Perceptually motivated spatial and temporal integration of pixel based video quality measures," in *Welcome to Mobile Content Quality of Experience*. Vancouver, Canada: ACM, 2007, pp. 1–7.

[5] M. A. Masry and S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 133–146, Feb. 2004.

[6] K. T. Tan, M. Ghanbari, and D. E. Pearson, "An objective measurement tool for mpeg video quality," *Signal Processing*, vol. 70, no. 3, pp. 279–294, Nov. 1998.

[7] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, 2010.

[8] H. G. Longbotham and A. C. Bovik, "Theory of order statistic filters and their relationship to linear fir filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 2, pp. 275–287, 1989.