2013 *International Workshop on Quality of Multimedia Experience*, Klagenfurt, Austria, July 2013

PERCEPTUAL EXPERIENCE OF TIME-VARYING VIDEO QUALITY

Abdul Rehman and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
Email: abdul.rehman@uwaterloo.ca, zhouwang@ieee.org

ABSTRACT

In real-world visual communications, it is a common experience that end-users receive video with significantly time-varying quality due to the variations in video content/complexity, codec configuration, and network conditions. How human visual quality-of-experience (QoE) changes with such time-varying video quality is not yet well-understood. To investigate this issue, we conduct subjective experiments designed to examine the quality predictability between individual video segment of relatively constant quality and combined video consisting of multiple segments that have significantly different quality. Our data analysis suggests that simple models that pool segment-level quality, such as linear averaging and weighted-averaging, nonlinear min- and median-filtering, and distortion-weighted averaging, are limited in predicting the overall human quality assessment of the combined video. We thus propose a quality adaptation model that is asymmetrically tuned to increasing and decreasing quality. The proposed asymmetric adaptation (AA) model leads to improved performance of both subjective and objective quality assessment approaches when using segment-level quality scores to predict multi-segment time-varying video quality. The video database together with the subjective data will be made available to the public.

Index Terms— visual quality-of-experience, time-varying video quality, temporal pooling, video quality assessment

1. INTRODUCTION

In practical network digital video communication systems, the source video content is subject to a series of distortions during the compression and transmission processes before being delivered to the end receivers. Very often, the quality of the received video varies over time. The source of such time-varying video quality may be at the sender side or within the communication network. At the sender side, video is compressed to meet the bandwidth constraints. Because of the large variations in the spatial/temporal/motion complexity in the video content, it is difficult to maintain constant video quality while making the best use of the communication channels, which often prefer approximately constant bit rate. In the communication network, packet loss and delay occur in somewhat random fashion, which, combined with the complexity of the coded video stream, often result in complicated distortions and quality variations when the video is decoded at the receiver side. Error correction and concealment techniques are commonly applied to partially recover the video but their performance varies as well.

Video quality assessment (VQA) has been an active subject of study in the past decades [1], but how human visual quality-of-experience (QoE) changes with time-varying video quality (in the scale of seconds or longer, rather than frames) is still an unresolved issue. Although quite many video quality databases have been built and subjective experiments conducted to study spatial and temporal video quality, they are not directly applicable in developing and validating computational models of time-varying video quality, because most video sequences in these databases consist of one scene or occasionally a few scenes of similar content and distorted in similar fashion, and thus in the scale of seconds or longer, they have fairly stable quality. Much less has been done in the area of predicting perceptual experience of time-varying video quality. Viewer response to time-varying video quality using a single stimulus continuous quality evaluation (SSCQE) in light of forgiveness, recency, and negative-peak and duration-neglect effects were studied in [2]. The findings of this study were applied in the form of an infinite impulse response (IIR) filter model for pooling in [3]. Asymmetric and smooth tracking of time-varying video quality by human subjects was observed and modeled in [4]. Temporal summation based on recursive formulations was used to model the low pass nature of the perceived continuous video quality [5] and hysteresis effect [6]. The historical experiences of the users' satisfaction while consuming a certain video streaming stimulus is modeled and quantified for web QoE in [7] and for VoIP in [8]. These models employ support vector machines and iterative exponential regeression to account for the memory effect. The difference in successive MOS values is exponentially weighted in a symmetric fashion as long as the difference is below a certain threshold. [9] investigates the human perception of variations in layer encoded video resulting in time-varying quality characteristics. Recently, the problem of video quality assessment with dynamically varying distortion on mobile devices was studied in [10].

In this work, we attempt to investigate the problem in a more straightforward way. In particular, we carry out subjective test on both individual video segments (each with a single scene) and combined video consisting of multiple segments that have significantly different quality. The test is designed to study how subjects (and objective models) react when there are quality variations between the scenes. We then study different approaches that use the quality of the individual segments to predict that

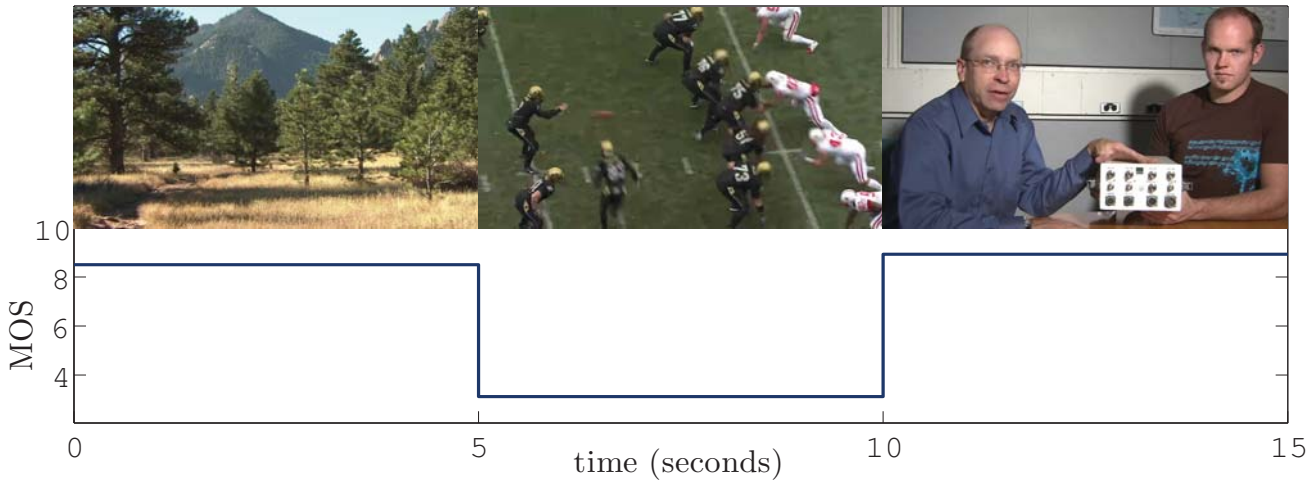


Fig. 1. A schematic example of a three-scene sequence with time-varying quality in the subjective test.

of the combined multi-segment video. This study is different from previous works, which typically focused on instantaneous video quality (often measured on a frame-by-frame basis) and its relationship to the aggregated quality of a video that contains one scene or multiple scenes. In reality, however, human subjects rarely judge video quality at such a high temporal resolution. Instead, based on our observation, they would rather give a single score to a segment of video, often of the same scene (regardless of the instantaneous quality variations between frames within the scene). Further, subjects tend to maintain their opinions until scene cut occurs, especially when adjacent scenes have very different content and quality. Eventually, the overall subjective opinion of the multi-scene video would be a result of pooling the segment-level quality. In this sense, our study better matches real-world scenarios, where a meaningful video content (such as a Youtube video) often contains multiple scenes with different levels of complexity and quality. The data collected from our subjective experiment allows us to study the quality predictability between individual video segments and combined multi-segment video. Our results show that none of the simple models such as linear averaging and weighted-averaging, nonlinear min- and median-filtering, and distortion-weighted averaging, produces impressive performance. We thus propose an asymmetric adaptation model to better account for the data. The model is useful in better understanding the psychological behavior of human subjects in evaluating time-varying video quality. It can also be directly applied to objective VQA algorithms to improve their performance, which is demonstrated using peak signal-to-noise-ratio (PSNR) and the multi-scale structural similarity index (MS-SSIM) [11] as examples.

2. SUBJECTIVE STUDY

2.1. Video Database

We start building our video database by selecting video segments, each of which contains a single scene, thus in the rest of the paper, the terms "scene" and "segment" are interchangeable. Four reference video segments are selected that contain indoor and outdoor scenes, flat areas and complex patterns, camera zooming/panning and object motion towards different directions. The video sequences are progressively scanned, with high definition (HD) resolution (1280×800), and in YUV 4:2:0 format. All the videos are five seconds long, with a frame rate of 30 frames/second. Every raw video scene is compressed at three quality levels using the recent high efficiency video coding (HEVC) reference software HM 8.0 [12]. The three quality levels are obtained by adjusting the quantization parameter (QP) of the encoder, for which a small-scale initial subjective test was conducted, such that each scene has three compressed versions at high-, medium- and low-quality levels (the distribution of quality levels will be discussed later). In the end, a total of 147 video sequences are included in the database, which are classified into three categories:

- 12 single-scene 5-second-long sequences, created by HEVC compression;
- 27 two-scene 10-second-long sequences, constructed by concatenating two of the single-scene sequences with combinations of varying quality;
- 108 three-scene 15-second-long sequences, constructed by concatenating three single-scene sequences with combinations of varying quality.

Figure 1 shows representative frames extracted from a three-scene test sequence, where the time-varying segment-level quality are indicated by the variations of the Difference of Mean Opinion Score (DMOS). A large number of combinations are included in the 2-scene and 3-scene categories to provide precise information necessary to study human behaviors in evaluating time-varying video quality. In addition, single-scene videos are used as prefixes of two-scene videos. Likewise, two-scene videos are used as prefixes of three-scene videos. As a result, by simply

asking each subject to score every sequence (1-scene, 2-scene, or 3-scene), we have the chance to monitor, track, and record the changes in quality scores along with the subject.

2.2. Subjective Test

Our subjective test generally follows the Absolute Category Rating (ACR) methodology, as suggested by ITU-T recommendation P.910 [13]. Although SSCQE [13] is designed for continuously tracking instantaneous video quality over time, it is not adopted in our experiment for the following reasons. First, as mentioned earlier, in practice human subjects often opt to judge video quality on per scene or segment basis, discounting the instantaneous quality variations between frames within a scene. Second, in our database, the same coding configuration and parameters are applied to the full duration of each scene, which is also roughly constant in terms of content and complexity. As a result, a single quality score is sufficient to summarize its quality. Third, in SSCQE, there is time delay between the recorded instantaneous quality and the video content, and such delay varies between subjects and is also a function of slider "stiffness". This is an unresolved issue of the general SSCQE methodology, but is avoided when only a single score is acquired. Fourth, we observe that humans tend to keep their opinions unless there is a significant change in video quality that attracts their attention. This is more realistically matched to real-world scenarios when subjects are watching a movie or online video. Compared with SSCQE, ACR is much simpler and provides more reliable and more realistic quality evaluations in our video database.

Thirty naïve subjects (17 males, 13 females) - all university undergraduate and graduate students - took part in the 40-minute subjective test. The first few video sequences were repeated at the end of the test to measure the fatigue factor. We found out that there were no bias or significant difference between the scores obtained, for the same set of video sequences, in the beginning and at the end of the test. The viewing distance is set to be four times of the picture height. Instructions were given to the subjects in both written and oral forms. A training session preceded the test where the subject was shown examples of distorted video sequences expected in the test. All the reference video sequences were also shown during the training session. During the main test, the 147 distorted video sequences were ordered randomly irrespective of their categories. Subjects scored the quality of each video sequence according to the eleven-grade $0-10$ numerical quality scale suggested in ITU-T recommendation P.910 [13].

After screening the data, 4 subjects were discovered to be outliers as they gave significantly different scores to the same video sequences when they were randomly repeated, and the scores given by the remaining 26 subjects were averaged to produce a mean opinion score (MOS) for each test sequence. Figure 2 plots the MOS scores versus video indices. Thanks to the initial subjective test before determining the Qp parameters used to create the compressed videos (as mentioned in Section 2.1), the resulting MOS values scatter in a wide range of the available scales $[0-10]$, which allows us to study different cases of quality transitions between the scenes.

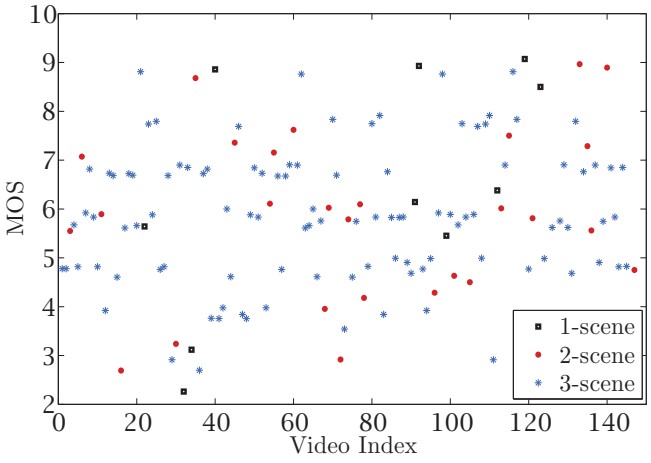


Fig. 2. MOS scores of all video sequences.

After each test session, we also discussed with each subject, inquiring about what strategy had been used by the subject to determine the scores. This step did not affect the data that had been collected, but helped us understand the data better, and also provided us with intuitive ideas that could be employed in the development of computational models that mimic human behaviors.

2.3. Observations

By investigating the subjective data collected and discussing with the subjects regarding their scoring strategies, we have a number of empirical observations. Although these observations are only qualitative, they provide useful insights in understanding the problem and in developing quantitative models that approximate human judgement. These observations are summarized as follows. Generally speaking, when watching a video with time-varying quality,

1. Subjects are *resistent* in updating their opinions. When there is a small quality variation between consecutive scenes, subjects tend to keep their opinions or change their opinions only slightly;
2. Subjects use *asymmetric* strategies in updating their opinions. A significant quality degradation between consecutive scenes results in a large penalty, as compared to the reward obtained by a similar quality improvement between consecutive scenes;
3. Subjects prefer *consistent* quality over time. Maintaining a "reasonable" quality for longer duration results in a small bias towards better subjective experience;
4. Subjects' judgments are not heavily influenced by the quality of the last (or the first) scene, which is in contrast to what was reported in [2]. This observation is also reflected in the numerical test results reported in Section 3.2.

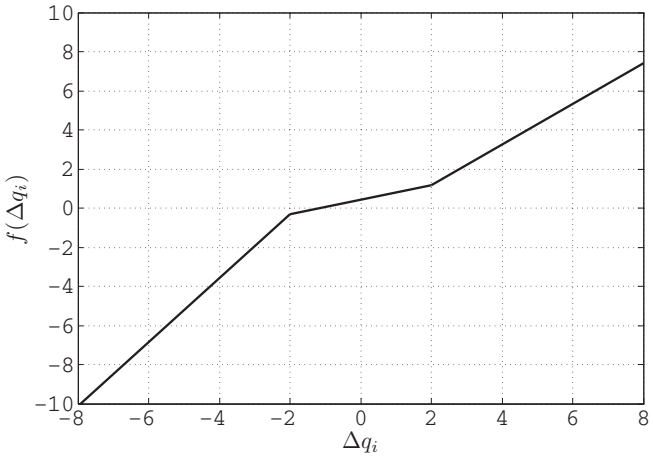


Fig. 3. Relationship between change in quality of successive scenes and change in perceptual quality experience.

3. OBJECTIVE MODEL

3.1. Asymmetric Adaptation (AA) Model

Based on the analysis of the subjective data and the observations described in Section 2.3, here we propose a model to better account for the perceptual experience of time-varying video quality. Assume that when subjects are watching a video, they maintain their overall opinions about the video quality until quality changes in consecutive scenes are observed. We can then focus on modeling the human strategy in updating their opinions.

Let n be the number of scenes in a video sequence, q_i be the perceptual quality of the i-th scene in the sequence (i.e., the quality when the single scene is assessed), l_i be the time span of the i-th scene, and Q_i be the perceptual quality experience after the i-th scene (i.e., the quality opinion after the first i scenes are watched). The change in the quality of successive individual scenes can be calculated by

$$\Delta q_i = \begin{cases} q_i, & i = 1 \\ q_i - q_{i-1}, & i = 2, 3 \dots n \end{cases}. \quad (1)$$

We model the quality opinion update between watching the $(i-1)$-th and the i-th scenes as

$$Q_i = \begin{cases} q_i, & i = 1 \\ \alpha_i f(\Delta q_i) + (1-\alpha_i) Q_{i-1}, & i = 2, 3 \dots n \end{cases}, \quad (2)$$

where $\alpha_i = l_i / \sum_{k=1}^{i} l_k$ controls the scale of change that decreases as time progresses, and the function f determines how subjective opinion changes as a function of Δq_i. In a simple special case, when $f(x) = x$, the model corresponds to quality averaging over time. However, the observations discussed in Section 2.3 suggest that f should be nonlinear. In particular, based on Observation 1 in Section 2.3, f should change slowly when $|\Delta q_i|$ is small; By Observation 2, f needs to change faster with negative values of Δq_i and slower for positive values of Δq_i; By Observation 3, f should be slightly positive when Δq_i is close to 0. Combining all the desired properties, we use a piecewise linear function to approximate f, which is plotted in Figure 3, where the three linear pieces correspond to significantly decreasing Δq_i, small change of Δq_i, and significantly increasing Δq_i, respectively. Because of the asymmetric properties of f, we call our quality updating scheme the asymmetric adaptation (AA) model.

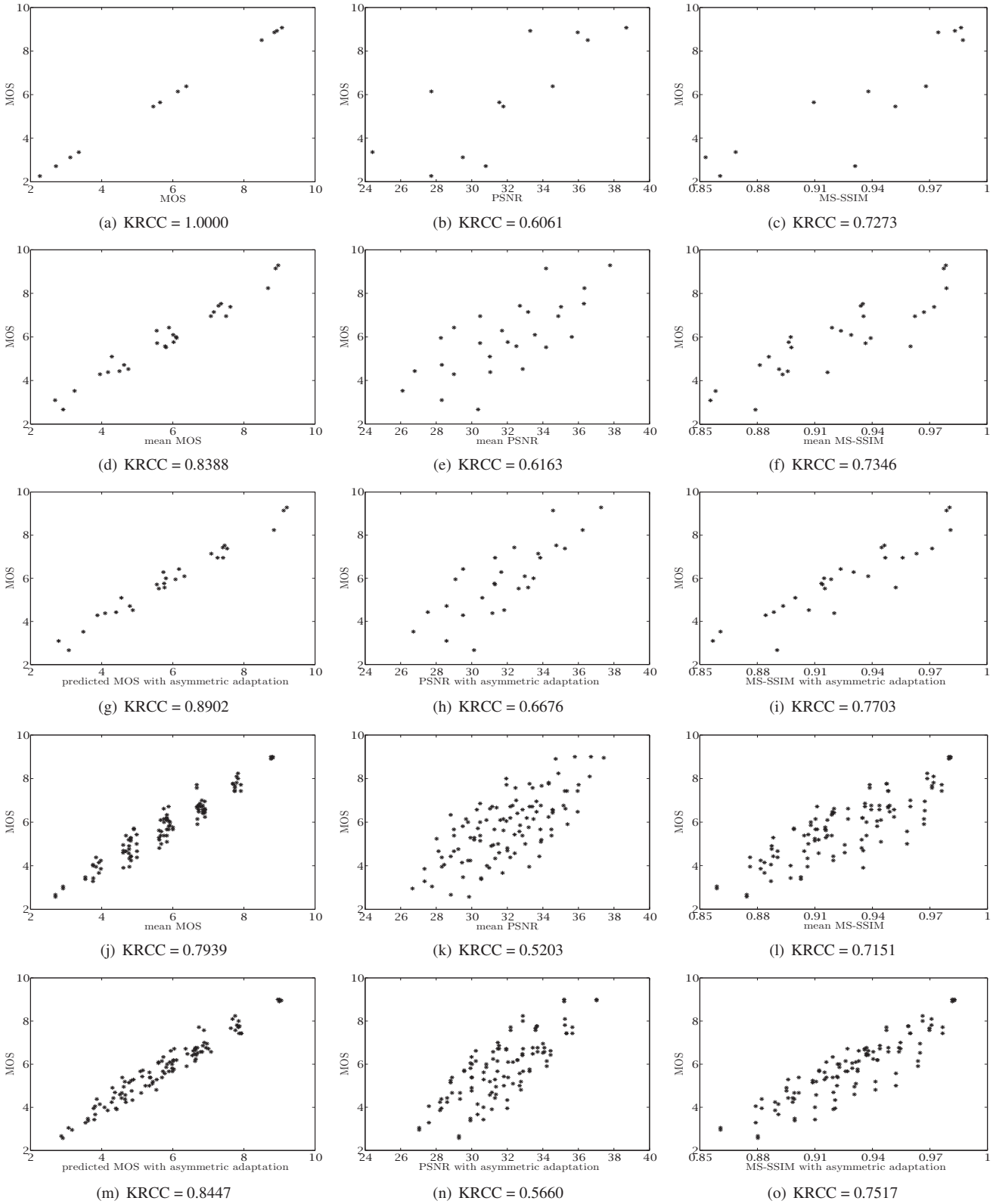
3.2. Validation

We test the proposed AA model by using it to predict the MOS value of a sequence from the MOS values of individual scenes that compose the sequence. All the MOS values are available in the subjective database described in Section 2. In addition to the proposed AA model, a series of other predictive models are also included for comparison. These include the Mean, Min, Max, and Median MOS values of all scenes, the MOS value of the first scene (FS) and the last scene (LS), weighted average MOS with increasing weights (W+), where $w = [\frac{1}{6}\ \frac{2}{6}\ \frac{3}{6}]$ for 3 scenes; decreasing weights (W-), where $w = [\frac{3}{6}\ \frac{2}{6}\ \frac{1}{6}]$, and distortion-based weights (DW), where $w = 1/\mathrm{MOS}$. Correlation between the predicted and actual sequence-level MOS scores is then calculated to provide quantitative evaluation of the performance. The results are reported in Table 1, where due to space limit, only Kendall's rank-order correlation coefficient (KRCC) results are given, but other measures give similar results. Furthermore, Figs. 4(d) and 4(g), and Figs. 4(j) and 4(m) compare the scatter plots of the actual MOS values versus Mean- and AA-predicted MOS values for 2-scene and 3-scene sequences, respectively. It can be observed that AA provides better predictions then Mean-MOS, which is one of the best in Table 1 among all other pooling methods being compared.

If a pooling scheme is effective at predicting sequence-level quality using the quality of each segment, then it should also be useful in improving objective VQA models in the pooling stage. We use the well-known PSNR and MS-SSIM [11] as examples to verify this. Note that the purpose here is not to find the best objective VQA approach, but to demonstrate the usefulness of the proposed model. The PSNR and MS-SSIM values are computed for each frame and then averaged within each scene, resulting the scene-level PSNR and MS-SSIM measures, which are used as the basis to predict the sequence-level MOS. The quantitative results are shown in Table 1 and the corresponding scatter plots for Mean- and AA-prediction are given in Fig. 4. It can be seen that the pooling schemes being tested generally behave consistently when using MOS, PSNR and MS-SSIM as the basis for scene-level quality measurement, and the proposed AA model generally outperforms the other approaches.

4. CONCLUSION

The major contributions of this work are twofold. First, we created a video database and carried out subjective test that are designed to directly examine the perceptual experience of time-varying video quality. The database, together with the subjec-



(a) KRCC = 1.0000 (b) KRCC = 0.6061 (c) KRCC = 0.7273

(d) KRCC = 0.8388 (e) KRCC = 0.6163 (f) KRCC = 0.7346

(g) KRCC = 0.8902 (h) KRCC = 0.6676 (i) KRCC = 0.7703

(j) KRCC = 0.7939 (k) KRCC = 0.5203 (l) KRCC = 0.7151

(m) KRCC = 0.8447 (n) KRCC = 0.5660 (o) KRCC = 0.7517

Fig. 4. Scatter plots of sequence-level actual MOS (vertical axis) versus predicted MOS (horizontal axis) using different scene-level base quality measures and different pooling strategies. Column 1: predicted by scene-level MOS; Column 2: predicted by scene-level PSNR; Column 3: predicted by scene-level MS-SSIM. Row 1: 1-scene sequence; Rows 2 and 3: 2-scene sequence; Rows 4 and 5: 3-scene sequence; Rows 2 and 4: Mean prediction; Rows 3 and 5: AA prediction.

Table 1. KRCC Comparison between actual MOS and predicted MOS using different base quality measures (scene-level MOS, PSNR and MS-SSIM) and different pooling strategies (Mean, Min, Max, Median, FS, LS, W+, W-, DW and AA).

Base measure	MOS			PSNR			MS-SSIM		
Sequence type	1-scene	2-scene	3-scene	1-scene	2-scene	3-scene	1-scene	2-scene	3-scene
Mean	1.0000	0.8388	0.7939	0.6061	0.6163	0.5203	0.7273	0.7346	0.7151
Min	1.0000	0.7274	0.6245	0.6061	0.5722	0.4752	0.7273	0.6477	0.5214
Max	1.0000	0.6546	0.4973	0.6061	0.5477	0.4468	0.7273	0.5928	0.4639
Median	1.0000	0.8388	0.7033	0.6061	0.6163	0.6133	0.7273	0.7346	0.6601
FS	1.0000	0.5553	0.3574	0.6061	0.4365	0.3156	0.7273	0.5078	0.3452
LS	1.0000	0.5292	0.4390	0.6061	0.4763	0.3828	0.7273	0.5075	0.4113
W+	1.0000	0.7475	0.7299	0.6061	0.6562	0.5288	0.7273	0.6733	0.6657
W-	1.0000	0.8103	0.6553	0.6061	0.5307	0.4784	0.7273	0.7247	0.6136
DW	1.0000	0.8445	0.7808	0.6061	0.6220	0.5380	0.7273	0.7232	0.7133
AA	1.0000	0.8902	0.8447	0.6061	0.6676	0.5660	0.7273	0.7703	0.7517

tive data, will be made available to the public. Second, we have a number of useful observations from the subjective test, based on which we proposed an asymmetric adaptation (AA) model to mimic the human strategies in updating quality opinions when watching video with time-varying quality. The proposed AA model was found to be effective in predicting sequence-level MOS values using scene-level MOS scores. It also leads to improved quality prediction performance when adopted in the pooling stages of objective VQA methods. The results of the current study may help us better understand perceptual experience of time-varying video quality in more realistic scenarios. They also have the potentials to be employed in the optimization of modern video compression technologies and in the optimal allocation of network resources for improving the visual QoE of end-users.

5. REFERENCES

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, Morgan & Claypool Publishers, Mar. 2006.

[2] D. E. Pearson, "Viewer response to time-varying video quality," *Proc. SPIE Human Vision and Electronic Imaging*, vol. 3299, no. 1, pp. 16–25, 1998.

[3] 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*, 2007, pp. 1–7.

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

[5] M. A. Masry and S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Signal Processing: Image Comm*, vol. 19, pp. 133–146, 2004.

[6] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 1153–1156.

[7] T. Hosfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, "The memory effect and its implications on web qoe modeling," in *23rd International Teletraffic Congress (ITC)*, Sep. 2011, pp. 103–110.

[8] A. Raake, "Short- and long-term packet loss behavior: Towards speech quality prediction for arbitrary loss distributions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1957–1968, 2006.

[9] M. Zink, O. Künzel, J. Schmitt, and R. Steinmetz, "Subjective impression of variations in layer encoded videos," in *Proceedings of the 11th international conference on Quality of service*, 2003, pp. 137–154.

[10] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. Veciana, "Video quality assessment on mobile devices: Subjective,behavioral and objective studies," *Journal Selected Topics Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[11] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," Pacific Grove, CA, Nov. 2003, pp. 1398–1402.

[12] G. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[13] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Tech. Rep., International Telecommunication Union, Geneva, Switzerland, Apr. 2008.