A NOVEL FULL-REFERENCE VIDEO QUALITY METRIC AND ITS APPLICATION TO WIRELESS VIDEO TRANSMISSION

Yang Peng and Eckehard Steinbach

Institute for Media Technology, Technische Universität München, Munich, Germany

ABSTRACT

In this paper, we present a novel objective video quality metric that captures the trade-off between the picture quality and the temporal resolution of a compressed video. The proposed metric is based on PSNR, frame rate as well as spatial and temporal activity measures that are obtained from the video. The content-independency of the metric makes it useful for the dynamic optimization of wireless video transmission. With the proposed metric, it is possible to adjust the trade-off between spatial and temporal qualities such that the user satisfaction is maximized. Our metric is very accurate, as verified by statistical analysis with data from subjective tests. We integrate the metric into a real-time wireless video transmission system and show in our experiments that the system, with our metric's ability to predict perceptual quality, can deliver significantly improved perceptual quality for arbitrary videos over a wide range of channel conditions.

1. INTRODUCTION

The transmission of video over wireless channels faces many challenges, including limited transmission capacity, time-varying channel conditions and stringent delay requirements of video applications. Extensive studies have been performed to improve the quality of video services, mostly by adapting video coding and scheduling parameters to channel characteristics. In order to achieve higher adaptivity, some schemes involve both spatial and temporal quality adjustment, such as [1] for scalable video streaming and our prior work [2] for low-delay live streaming, where one needs to decide how to trade off spatial and temporal quality to achieve the best possible overall perceptual quality. However, most of the studies considering joint control of spatio-temporal quality are based on PSNR as the quality metric, which has been shown to have poor correlation with quality evaluation results from subjective tests in such context [3]. Other objective quality metrics, such as those described in [4], although they show higher accuracy than PSNR for videos with spatial quality impairments, are not suitable either because they do not consider the impairments caused by frame rate changes.

Various prior works studied the impact of frame rate changes on perceptual quality. Some focus on modeling the temporal impact alone without any spatial quality impairment (e.g., [5]). The quality metric proposed in [3] considers both quantization and frame rate. It is a weighted sum of PSNR and frame rate reduction, where the weight is based on a motion measure. However, this metric ignores the content-dependency of the PSNR's impact on the overall quality, and therefore is content-dependent. Without including additional parameter(s), the metric can not provide accurate quality predictions when videos with different characteristics are involved. In addition, the motion measure, based on motion vector magnitude, depends strongly on the motion estimation (ME) scheme used in the video codec. Another PSNR-based metric is proposed in [6], which is the product of a spatial quality term based on PSNR and a temporal correction factor. There are two parameters in this metric that still need to be determined for each video individually, rendering the metric content-dependent and inapplicable in an automatic system. The metric in [6] is extended in [7] by estimating the metric parameters using content features, which makes the metric content-independent.

In this paper, we propose a novel PSNR-based video quality metric that considers both spatial and temporal quality impairments, and integrate the metric into the real-time wireless video transmission system described in [2]. All the metric parameters are either constants, or parameters that can be calculated directly from the video, which makes the metric content-independent and useful for automatically exploiting the trade-off between spatial and temporal quality. The proposed metric in this paper has some structural similarity with the metric in [7]. In comparison to [7], however, our metric has only two standard video activity measures that can be easily computed (four measures with significantly higher complexity need to be computed in [7]), and therefore is more suitable for real-time applications. In addition, the metric in [7] is dependent on the ME scheme in the video codec – a different ME configuration would require a new set of metric coefficients; our metric has no codec-dependent parameters. Our metric predicts the perceptual quality very well, significantly better than the metric in [3] and as well as the metric in [7] for our test dataset. With the integration of the proposed metric, our system in [2] operates with very high channel and content adaptivity, providing superior perceptual quality for arbitrary video content over a wide range of channel conditions.

2. SUBJECTIVE QUALITY EVALUATION

To find out how spatial and temporal quality affect the overall perceptual quality, we conducted a subjective test using three source videos (SRC) with various spatial and temporal characteristics: Mother&Daughter (MD), Foreman (FM) and Football (FB). The SRCs are in CIF (352x288) resolution and have an original frame rate (FR) of 30 fps. Each SRC is temporally downsampled to 15, 10 and 7.5 fps to generate four different temporal quality levels. Then for each temporal quality level, we encode the video using an MPEG-4 video codec to generate three different spatial quality levels, with average PSNR at about 38dB, 34dB and 31dB. Since this PSNR value only indicates the spatial quality level, it is referred to as SPSNR. Each video is encoded in IP...P structure with a constant quantization parameter (QP) that results in one of the spatial quality levels. After encoding, frame repetition is performed so that each processed video sequence (PVS) has the same duration (8 seconds).

The SAMVIQ method [8], which is specifically designed for multimedia contents, is adopted to collect subjective quality scores. We choose SAMVIQ because it can provide more accurate and reliable subjective data, especially when various types of impairments are involved. SAMVIQ uses a continuous quality scale graded from

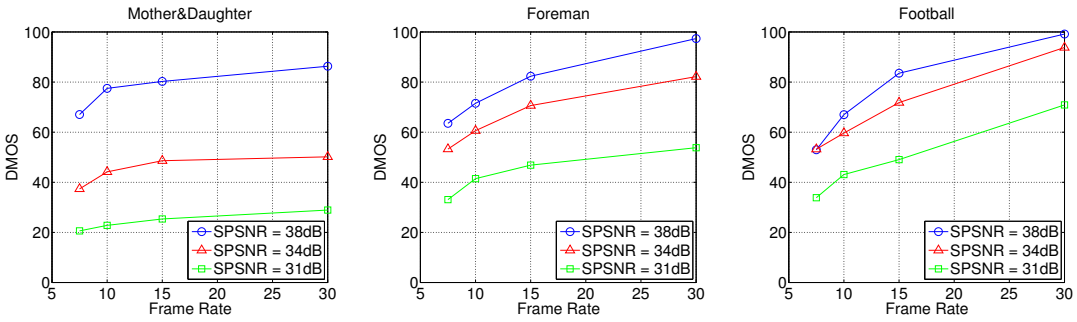

Fig. 1. DMOS versus frame rate for different test videos.

Table 1. Two-way ANOVA results for full FR videos

Source of Variation	*df*	*F-value*	*p-value*
SPSNR	2	364.1	<0.0001
Content	2	220.3	<0.0001
SPSNR : Content	4	20.5	<0.0001

0 to 100 to provide the subjective scores (MOS). We follow the instructions in [8] for designing the subjective test. Twenty seven test subjects (non-experts between the age of 20 and 26) participated in our test, which is conducted in a controlled room with identical LCD monitors. Data from 25 subjects are verified to be valid based on the screening process described in [8]. For each PVS, a DMOS value is computed by $DMOS = MOS(PVS) - MOS(SRC) + 100$, which is used in our data analysis as the subjective quality rating.

3. VIDEO QUALITY MODELING

3.1. Spatial Quality Modeling

First, we consider only the videos without temporal impairment, i.e., at 30fps, to study how the perceived spatial quality (DMOS) changes as a function of SPSNR at full frame rate. We perform a two-way analysis of variance (ANOVA) with repeated measures on the subjective data. The ANOVA results are reported in Table 1, which indicate that both SPSNR and content have significant impact ($p < 0.0001$) on the perceived quality. The interaction between them is also found to be significant ($p < 0.0001$), indicating that the impact of SPSNR is content-dependent.

Based on our observations and the ANOVA results, the spatial video quality (SVQM) is modeled with a logistic function (e.g., [9]):

$$SVQM = \frac{100}{1 + e^{-(SPSNR + w_s \cdot SA + w_t \cdot TA - \mu)/s}}, \quad (1)$$

where SA and TA measures the spatial and temporal activity of the video content, respectively. We adopt the spatial and temporal perceptual information measures in [8], slightly modified, in our metric as SA and TA, which are defined as:

$$SA = mean_{time}\{std_{space}[Sobel(F_n)]\}, \quad (2)$$

$$TA = mean_{time}\{std_{space}[F_n - F_{n-1}]\}. \quad (3)$$

The constants w_s, w_t, μ and s are determined by a least-square non-linear fitting using the subjective data of the considered videos, which leads to $w_s = 0.0356, w_t = 0.236, \mu = 36.9, s = 2.59$.

3.2. Spatio-Temporal Quality Modeling

We now examine the impact of temporal impairment on the overall perceived quality. In Fig. 1, we plot DMOS against frame rate at different SPSNR levels for different SRCs. As expected, when frame rate decreases, DMOS becomes lower. DMOS decreases slower for low-motion SRC like MD and faster for high-motion SRC like FB, which indicates that the impact of frame rate is content-dependent and reducing frame rate has stronger negative impact for videos with higher temporal activity. We can also see that the curves for each SRC at different SPSNR levels have different slopes, indicating that the impact of frame rate is dependent on the spatial quality level. A three-way ANOVA with repeated measures confirms our observations. From the ANOVA results given in Table 2, we can see that all three factors (i.e., SPSNR, content and FR) have significant impact ($p < 0.0001$) on the overall perceived quality. The interaction between FR and content is significant ($p < 0.0001$), indicating that the impact of frame rate is content-dependent. The significant ($p = 0.0016$) interaction between SPSNR and FR indicates that the impact of frame rate is also quality-dependent.

Table 2. Three-way ANOVA results for all videos

Source of Variation	*df*	*F-value*	*p-value*
SPSNR	2	783.0	<0.0001
Frame Rate (FR)	3	220.1	<0.0001
Content	2	158.5	<0.0001
SPSNR : Content	4	45.8	<0.0001
FR : Content	6	20.6	<0.0001
SPSNR : FR	6	3.6	0.0016
SPSNR : Content : FR	12	0.8	0.6367

Based on the observation that an interaction exists between spatial quality and temporal quality perception, the spatio-temporal quality (STVQM) is modeled as the product of the SVQM in (1) and a temporal quality term:

$$STVQM = SVQM \cdot \frac{1 + a \cdot TA^b}{1 + a \cdot TA^b \cdot \frac{30}{FR}}, \quad (4)$$

where FR denotes the frame rate and TA denotes the temporal activity measure in (3). A product form is also adopted in [6] and [7], but without statistically verifying the interaction. The two constants a and b are determined by a least-square non-linear fitting using the subjective data, which leads to $a = 0.028, b = 0.764$.

3.3. Performance Evaluation

Three statistical measures are used to evaluate the performance of our proposed metric STVQM: Pearson Correlation (PC), Residual Mean Square Error (RMSE) and Outlier Ratio (OR). We compare STVQM to three other metrics: STPSNR, the metric in [3] (referred to as Feghali) and the VQMTQ metric in [7]. STPSNR here is referred to as the PSNR averaged over all frames in a PVS, including the repeated frames in case of frame rate reduction. STPSNR considers both spatial and temporal impairment and is widely used in the

Table 3. Performance measures for the entire dataset

Metric	PC	RMSE	OR
STPSNR	0.484	19.1	0.83
Feghali	0.471	19.6	0.78
VQMTQ	0.995	2.82	0
STVQM	0.994	2.66	0.03

Table 4. Pearson Correlation for individual SRCs

Metric	MD	FM	FB	Average
STPSNR	0.978	0.863	0.929	0.923
Feghali	0.974	0.906	0.943	0.941
VQMTQ	0.999	0.997	0.985	0.994
STVQM	0.997	0.998	0.990	0.995

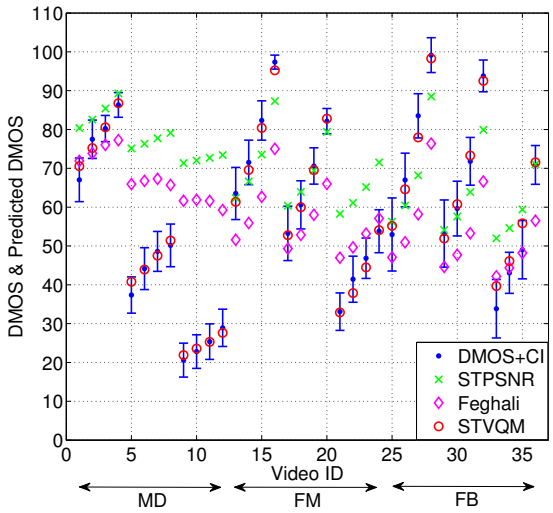

Fig. 2. Performance evaluation and comparison for STVQM.

literature when frame rate change is present. In our comparison, the metric coefficients in the Feghali metric are determined by testing many combinations and picking the one that leads to the best average correlation coefficient over the three SRCs, similar as done in [3]. For STPSNR and the Feghali metric, a least-square linear mapping is also performed to map the metrics to the DMOS scale. The metric coefficients in VQMTQ are determined by a least-square non-linear fitting using our subjective data. The statistical measures of all four metrics are summarized in Table 3. We can see that STPSNR and the Feghali metric fail drastically in predicting DMOS when several SRCs with different characteristics are considered. The main reason here is that none of them considers the content-dependency of the PSNR's impact. We also summarize the correlation coefficients for each SRC individually in Table 4. As expected, STPSNR and the Feghali metric have much better performance for individual SRCs. In comparison, our proposed metric STVQM achieves very high accuracy in predicting DMOS, significantly better than STPSNR and the Feghali metric, both for the entire dataset as well as for individual SRCs. Comparing our metric with VQMTQ, significance tests show that the difference between VQMTQ and our metric is not statistically significant – our metric performs as well as VQMTQ, with the advantages discussed in Section 1. A graphical presentation of the metric performance is provided in Fig. 2, where for 36 different test videos (12 quality levels for MD, FM and FB), DMOS with its 95% confidence interval (CI) and the metric predictions are plotted (for clarity, the results for VQMTQ are not plotted here).

4. APPLICATION TO WIRELESS VIDEO TRANSMISSION

To show the effectiveness of our proposed video quality metric in real systems where the trade-off between spatial and temporal quality can be exploited to achieve best possible overall perceptual quality, we integrate our metric in the real-time wireless video transmission system presented in our prior work [2]. One of the main ideas proposed in [2] is to integrate retransmission into a video transmission system without introducing any additional delay, which is achieved by dynamically adjusting the resource allocation between source coding and retransmission. Both quantization adjustment and frame skipping are considered in the resource allocation. The trade-off between spatial and temporal quality is exploited in two places where decisions need to be made. The first one is to decide (Decision I), before a video frame is encoded, whether we should use a coarser quantization or skip the next frame in order to make resources available for potentially necessary retransmissions. It is similar to a rate control scheme considering both quantization and frame skipping. The second place is after transmitting a frame in its allocated time slot(s). As some parts of the frame may still be lost (due to unpredictable channel variations), we need to decide (Decision II) whether we should just conceal the lost parts (sacrificing the spatial quality), or skip the next frame (sacrificing the temporal quality) and continue retransmitting the lost packets. In [2], the decisions are made in a heuristic manner based on thresholds that only involve channel statistics. Since the trade-off between spatial and temporal quality is content-dependent, the best decisions should be different for different contents. In order to make the best decisions automatically, we integrate our quality metric into the decision making and select the option that leads to the best predicted perceptual quality.

In our experiments, Decision I is made every second, intended to adapt to slow channel changes; Decision II is made every frame, intended to adapt to fast channel changes. Since we need to make Decision I before the encoding, SPSNR is estimated from the source coding rate using the rate-distortion model in [10], which can be written as:

$$D(R) = \sigma^2 \cdot e^{-\alpha \cdot R}, \tag{5}$$

where D is the MSE, R is the source coding rate, σ^2 is the variance of the source data and α is a content-dependent parameter. In this work, α is estimated from the video frames in the previous second, so are SA and TA in (1) and (4). For Decision II, SPSNR is calculated on the concealed frame, SA and TA are computed from the original frame. Here we assume concealment and quantization have the same impact on the perceptual quality if they introduce the same MSE, as in various previous works (e.g., [10]). More sophisticated combinations, such as a weighted one, could also be implemented.

5. EXPERIMENTAL RESULTS

The performance of the wireless video transmission system for Football and Mother&Daughter is shown in Fig. 3. The wireless channel is assumed to be a packet erasure channel with random packet losses. Four different system variations are compared: 1) SP: Decision I is quantization adjustment, Decision II is concealment; 2) TP: Decision I is frame skipping, Decision II is concealment; 3) ADP: Decision I is made adaptively based on our quality metric STVQM, Decision II is concealment; 4) ADP+: both Decision I and II are made adaptively based on STVQM. We note that adaptive decisions based on STPSNR would almost always avoid frame skipping, resulting in a performance similar to SP. From the results we can see that ADP is highly adaptive, both to the channel statistics as well

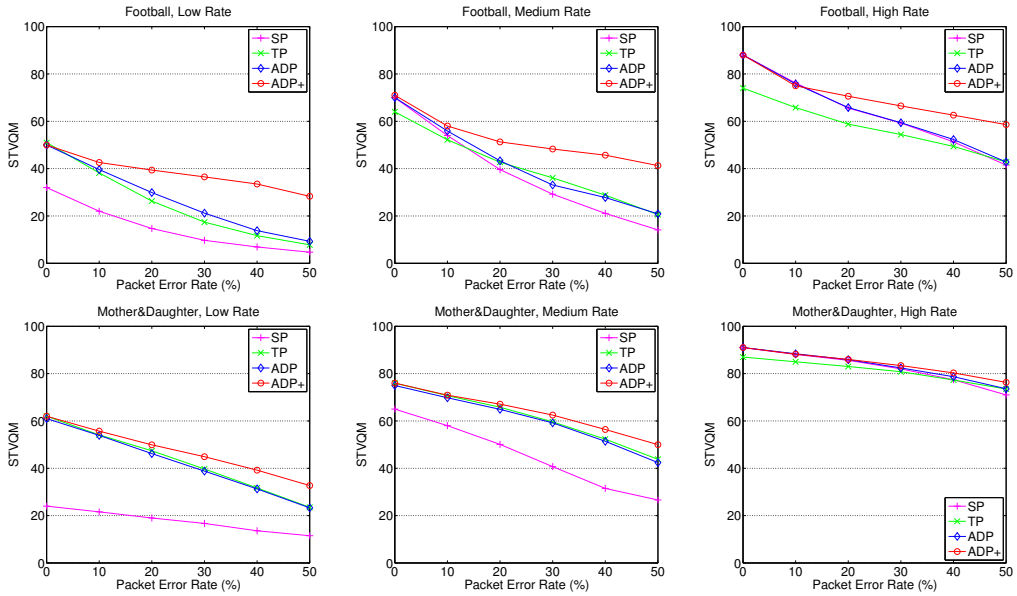

Fig. 3. System performance with the proposed perceptual quality metric STVQM. All results are averaged over 10 channel realizations.

as to the video content, as the curves of ADP always follow the option with better perceptual quality (predicted by STVQM), no matter at what rate, at what packet error rate or for which content. ADP+ adds another level of adaptability to channel variation and achieves higher performance gains for channels with higher loss rate and for videos with larger motion. The system runs automatically in real-time; there is no parameter that needs to be determined manually to achieve the best performance for a certain condition.

For the high motion video Football at low rate, TP (frame skipping) delivers better perceptual quality than SP (quantization), indicating that for high motion content at low rate/spatial quality level, frame skipping is preferred over quantization adjustment. As the rate/spatial quality level increases, the two curves move towards each other, meet at medium rate, and then separate at high rate, where SP delivers better perceptual quality than TP. This indicates that for high motion content at high spatial quality level, quantization adjustment is preferred over frame skipping. On the other hand, for the low motion video Mother&Daughter, frame skipping always leads to a better perceptual quality. The gain of choosing frame skipping over quantization adjustment becomes smaller as the rate/spatial quality level increases, until it saturates at high spatial quality level. We note that the case with zero packet error rate is equivalent to a rate control scheme that jointly adjusts quantization and frame rate. So from our results, we can also see how effective it would be to adopt our quality metric in applications involving such a rate control scheme.

6. CONCLUSIONS

In this paper, we proposed a perceptual quality metric that considers both spatial and temporal quality. Statistical analysis with subjective data has shown that the proposed metric is very accurate. Unlike most existing metrics, our metric is content-independent, and therefore can be used to automatically trade off spatial and temporal quality. We have integrated our metric in a real-time video transmission system proposed in our prior work. Experimental results have shown that with our metric's ability to accurately predict perceptual quality, the system can deliver improved perceptual quality with high adaptability to both video content and channel statistics.

7. REFERENCES

[1] T. Schierl, T. Stockhammer, and T. Wiegand, "Mobile video transmission using scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1204–1217, Sept. 2007.

[2] Y. Peng, F. Zhang, and E. Steinbach, "Error-resilient video transmission for short-range point-to-point wireless communication," in *Proc. of ICCCN'10*, Zurich, Switzerland, Sept. 2010.

[3] R. Feghali, D. Wang, F. Speranza, and A. Vincent, "Video quality metric for bit rate control via joint adjustment of quantization and frame rate," *IEEE Trans. on Broadcasting*, vol. 53, no. 1, pp. 441–446, Mar. 2007.

[4] ITU-R Rec. BT.1683, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," 2004.

[5] K. C. Yang, C. C. Guest, K. El-Maleh, and P. K. Das, "Perceptual temporal quality metric for compressed video," *IEEE Trans. on Broadcasting*, vol. 53, no. 1, pp. 441–446, Mar. 2007.

[6] Y. Ou, Z. Ma, and Y. Wang, "A novel quality metric for compressed video considering both frame rate and quantization artifacts," in *Proc. of VPQM'09*, Scottsdale, USA, Jan. 2009.

[7] Y. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 286–298, Mar. 2011.

[8] ITU-R Rec. BT.1788, "Methodology for the subjective assessment of video quality in multimedia applications," 2007.

[9] ITU-R Rec. BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.

[10] Z. He, J. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 511–523, June 2002.