Analysis of the impact of temporal, spatial, and quantization variations on perceptual video quality

Andreas Rossholm, Muhammad Shahid, Benny Lövström

IEEE/IFIP Network Operations and Management Symposium: Management in a Software Defined World

2014 Krakow

Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

Andreas Rossholm, Muhammad Shahid, Benny Lövström
Department of Applied Signal Processing, Blekinge Institute of Technology, 37179 Karlskrona, Sweden
Email: benny.lovstrom@bth.se

***Abstract*—The growing consumer interest in video communication has increased the users' awareness in the visual quality of the delivered media. This in turn increases, at the service provider end, the need for intelligent methodologies of optimal techniques for adapting to varying network conditions. Recent studies show that constraints on the bandwidth of transmission media should not always be translated to an increase in compression ratio to lower the bitrate of the video. Instead, a suitable option for adaptive streaming is to scale down the video temporally or spatially before encoding to maintain a desirable level of perceptual quality, while the viewing resolution is constant. Most of the existing studies to examine these scenarios are either limited to low resolution videos or lack in provisioning of subjective assessment of quality. We present here the results of our campaign of subjective quality assessment experiments done on a range of spatial and temporal resolutions, up to VGA and 30 frames per second respectively, under a number of bitrate conditions. The analysis shows, among other things, that keeping the spatial resolution is perceptually preferred among the three parameters that have impact on the video quality, even in the case with high temporal activity.**

I. Introduction

As video communications constantly continues to grow both regarding its share of all data traffic and the amount of data in absolute terms, the consumers demand on perceived quality also increases. Also, new cellular wireless technology evolves and an increasing share of all data communication will be wireless. This results in many new scenarios with different services and requirements where the provider want to optimize the perceived quality or quality of experience (QoE). One new challenge with new mobile networks like 3G and 4G is that even if high peak link rates are possible the cellular wireless networks experience rapid link rate variation and occasional long delays in one or both direction. This requires either long receiver buffers, resulting in long end-to-end delay, or fast adaptation, resulting in need for the possibility to change used band width [1]. In this context the need of optimizing the delivered quality of experience by a service provider is raised. To this end, one significant issue to be resolved is finding the best trade-off among spatial resolution, temporal resolution, and quantization level, giving the optimal value of QoE in a given scenario. In practice, this includes applications such as adaptive streaming [2], [3], as well as different real time video communication services where maintaining the desired level of perceived quality is required in fluctuating network conditions. Also, there is a growing demand for objective quality measurement or monitoring techniques estimating perceived video quality in these scenarios, especially to be able to compare different spatial and temporal resolutions. An overview of various types of contemporary objective Video Quality Assessment (VQA) is presented in [4].

978-1-4799-0913-1/14/$31.00

The quest of discovering the optimal trade-off has been the subject of video scalability for assuring stipulated level of visual quality. For service providers, it is useful to ascertain the best QoE of a video at a given bandwidth capacity. In order to optimally address any fluctuations in the transmission network, it becomes pertinent to determine the parameter that can be scaled up or down with minimal deviation in the level of delivered visual quality. To serve this matter, a number of studies have been made that focus on examining the impacts of changes in the aforementioned three parameters of a video. Subjective quality assessment of low resolution, QCIF ($176x144$) and CIF ($352x288$), videos encoded using H.264/AVC has been reported in [5] for 150 test scenarios. Under low bitrate conditions, it was concluded that small frame size is mostly preferred. For CIF resolution or high temporal (30 fps) resolution at low bitrates, it was found that it was most efficient to reduce quantization except for video sequences containing very low spatial activity. It was also pointed out that a minimum threshold value of 0.1 Bits Per Pixel (BPP) is required to achieve good or excellent perceptual quality. Subjective experiments conducted using low resolution videos, CIF, in [6] show that frame rate can be compromised to maintain the perceptual quality by keeping the compression ratio at low value. Similar results can be observed in the study reported in reference [7]. Impact of encoding strategy on the quality of MPEG-2 encoded videos, QCIF and CIF, while transmitted over lossy network has been investigated in [8]. It has there been observed that videos with high spatial activity are perceptually preferred with higher spatial resolution, and videos with higher temporal activity are preferred in full frame-rates. The validity of these results needs to verified in the case of videos encoded by H.264/AVC.

Considering the case of high resolution video conferencing applications, video scalability has been tested for high definition videos ($1920x1080$) in [9]. It was observed that the quality level can be maintained by decreasing frame rate and frame resolution to cater the constraints of the transmission bandwidth. Hence, high compression rates can be avoided. Moreover, as the bandwidth begins to grow, it is perceptually preferred to increase the frame rate up to a certain higher level first and the frame resolution can be increased afterwards. Unfortunately, these conclusions have been drawn only from the results of objective metrics, with no subjective assessment to support the results. A detailed discussion and a review of the studies performed on the video scalability for quality can

be found in [10] and the references therein.

By examining the existing drives to investigate the impacts of three basic parameters of video encoding, the requirement of a comprehensive study on a wider range of videos, in a highly interesting bandwidth range, supported by subjective assessment of quality becomes evident. Therefore, we present here the details of an extensive campaign of subjective quality assessment experiments of videos encoded using combinations of multiple levels of the bitrate, frame rate and resolution. This enables examinations of e.g. the perceptual trade off between spatial and temporal resolution at a certain bitrate. The rest of this paper is organized as follows. In Section II the video sequences used in the test are described, encoding configuration, as well as the subjective assessment setup. Also the pre-processing of the sequences before the assessment is described. In Section III the findings from the subjective tests are given, and finally in Section IV conclusions are drawn.

II. Test Stimuli and Subjective Video Quality Assessment

To perform a comprehensive subjective quality assessment that can be used to infer useful conclusions, it is imperative to select the SouRCe sequences (SRCs) carefully. Such SRCs should possess a variety of spatio-temporal characteristics to be representative of most commonly used videos. To this end, we followed the ITU recommendation P.910 [11] for the selection of SRCs based on spatial perceptual information (SI) and temporal perceptual information (TI). The SI and TI values are calculated in the luminance plane of a video. The five SRCs used in this study are Children, City, Elisa, Ice, and Soccer, all of 10 s duration. The starting frame of each of the sequences is shown in Fig. 1, and table I gives a short description of the content and lists the original frame rate of the sequences as well as their SI and TI characteristics.

TABLE I. *The original frame rate and a brief description of the SRCs used in the experiment*

Sequence	Frame rate [fps]	Description
Children	30	Two children sitting on the floor, slowly moving, low SI and low TI
City	25	Panning view over a city from an airplane, high SI and medium TI
Elisa	30	Head and shoulder of a talking woman, medium SI and low TI
Ice	25	Several persons skating on white ice, low SI and high TI
Soccer	25	Close up view of soccer game, panning, low SI and high TI

A. Encoding configuration

The SRCs have been encoded following the standard H.264/AVC using the JM reference software to produce Processed Video Sequences (PVSs). For the encoding of the PVSs a number of combinations of resolutions (Res), frame rates (FR) and bitrates (BR) have been used, based on several considerations. For the bitrates, the band width fluctuation and the built in limitations running realtime communication over cellular wireless network was taken into count. Based on this and the requirements of a realistic BPP value, and also de facto configurations from industry, the resolution and frame rate was limited, as shown below.

- BR: 50, 150, 300, 600, and 900 kbps
- Res: VGA = 640×480, HVGA (Half VGA) = 480×320, QVGA (Quarter VGA) = 320×240, and MVGA (mobile VGA) = 192×144
- FR: A: 30, 15, 10 fps, and B: 25, 12.5, 8.33 fps

All used combinations of bitrate, resolution, and frame rate are shown in Table II, where columns A and B shows the different combinations for the videos with original frame rates 30 fps and 25 fps, respectively. It can be seen in table II that it results in 38 combination for every SRC. To conduct a suitable subjective test the combination of resolution, frame rate and bitrate is based on realistic combinations used in practice, which means that combinations with too low or very high BPP are excluded.

TABLE II. *The PVS combinations*

A			B		
Reso-lution	FR [fps]	BR [kbps]	Reso-lution	FR [fps]	BR [kbps]
MVGA	10	50	MVGA	8.33	50
MVGA	10	150	MVGA	8.33	150
MVGA	10	300	MVGA	8.33	300
MVGA	15	50	MVGA	12.5	50
MVGA	15	150	MVGA	12.5	150
MVGA	15	300	MVGA	12.5	300
MVGA	30	150	MVGA	25	150
MVGA	30	300	MVGA	25	300
QVGA	10	50	QVGA	8.33	50
QVGA	10	150	QVGA	8.33	150
QVGA	10	300	QVGA	8.33	300
QVGA	10	600	QVGA	8.33	600
QVGA	15	50	QVGA	12.5	50
QVGA	15	150	QVGA	12.5	150
QVGA	15	300	QVGA	12.5	300
QVGA	15	600	QVGA	12.5	600
QVGA	30	150	QVGA	25	150
QVGA	30	300	QVGA	25	300
QVGA	30	600	QVGA	25	600
HVGA	10	150	HVGA	8.33	150
HVGA	10	300	HVGA	8.33	300
HVGA	10	600	HVGA	8.33	600
HVGA	10	900	HVGA	8.33	900
HVGA	15	150	HVGA	12.5	150
HVGA	15	300	HVGA	12.5	300
HVGA	15	600	HVGA	12.5	600
HVGA	15	900	HVGA	12.5	900
HVGA	30	300	HVGA	25	300
HVGA	30	600	HVGA	25	600
HVGA	30	900	HVGA	25	900
VGA	10	300	VGA	8.33	300
VGA	10	600	VGA	8.33	600
VGA	10	900	VGA	8.33	900
VGA	15	300	VGA	12.5	300
VGA	15	600	VGA	12.5	600
VGA	15	900	VGA	12.5	900
VGA	30	600	VGA	25	600
VGA	30	900	VGA	25	900

B. Pre-processing the test sequences

Before executing the subjective assessment all the processed video sequences (PVSs) are pre-processed. The reason for this is to enable a more realistic test scenario as in streaming or realtime video applications, where the viewing resolution is usually fixed even if the source data is changed, e.g. down sampled, in the context of adapting to fluctuating bandwidth. Therefore all PVSs with spatial resolution MVGA, QVGA, and HVGA were up-scaled to VGA ($640x480$), performed with bicubic filtering as it produces sufficient quality and does not require too much of processing power. Also, all files with sub-sampled temporal resolution from the original

Children City Elisa Ice Soccer

Fig. 1. The SRCs used for generation of PVS

25fps or 30fps were up-sampled to the original frame rate by frame repetition. This was performed to limit difference in play out during the subjective assessment between the PVSs.

C. Subjective Video Quality Assessment Setup

The subjective quality assessment has been performed on 32 test subjects with video sequences described in the previous subsection. Since not all combinations of bitrates and frame rates are used, this results in a total of 190 sequences being used in the test. The setup of the subjective quality assessment follows ITU recommendations as given by ITU-R BT 500-12 [12] for the lab setup of our experiments. Particularly, the method followed was the single stimulus quality evaluation where a test video sequence is shown once without the presence of any explicit reference, corresponding to the reality where users see only the processed version of the video. Overall, the adopted methodology and lab setup has been summarized in [7]. The subjects who participated in the tests were of both genders, mainly students at the university and some staff members, and all of them were considered to be non-expert in the area of video quality assessment. In order to obtain reliable results out of the raw subjective scores on the quality scale of 1 to 100, a screening of the observers scores was employed to discard observers that are considered as outliers. The algorithmic details of these steps are reported in Annex 2 of [12]. After screening of our data no subject had to be rejected. Finally, the mean opinion score (MOS) was calculated and used in this work.

III. Results

In the context of adaptive streaming an estimation of the variable bandwidth over a channel is used as a restriction for available bitrates to use for the video codec. With this in mind, the MOS results for the five test sequences with their 38 combinations is presented in Fig. 2-6 where the MOS scores are plotted versus the bitrate. To be able to identify different resolutions and frame rates in the figures the resolution is color coded and the frame rate is marked by different symbols.

A. Observations from the MOS results

Some observations can be made directly by studying the MOS results. It can be seen that the sequences with lowest temporal information (TI), Elisa, City, and Children, have clear differentiation between the different resolutions, indicating that the resolution has high significance. There is though a difference regarding City versus Elisa and Children, where the later have the highest spatial information, that for City it clearly differentiates between resolutions even for VGA and HVGA which is not the case for Elisa and Children even if highest resolution is always preferred. It can also be seen for these

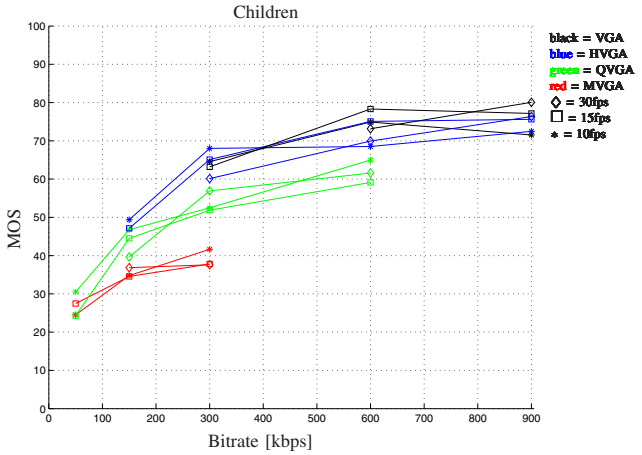

Fig. 2. MOS vs. bitrate for different frame rate and resolutions where Children is characterised to have low SI and low TI.

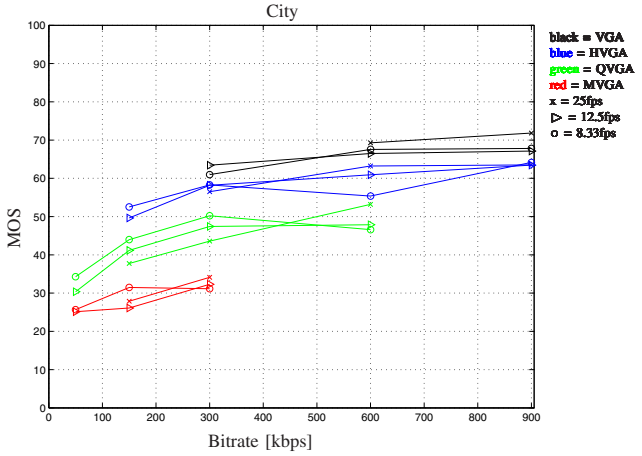

Fig. 3. MOS vs. bitrate for different frame rate and resolutions where City is characterised to have high SI and medium TI.

sequences that higher resolution over increased frame rate is always preferred. For the two sequences with highest temporal information (TI), Soccer and Ice, the tendency is the same but not to the same extent. It can be seen that for increased spatial resolution at lower frame rate is preferred over increase of frame rate but keeping the spatial resolution.

B. Analysis of Variance Based Comparison

To further evaluate the MOS scores ANalysis Of VAriance (ANOVA) [13] was used. ANOVA analysis is used to determine whether or not different factors or variables are statistically significant. We considered Res, FR, and BPP for this analysis to see their statistical significance on the

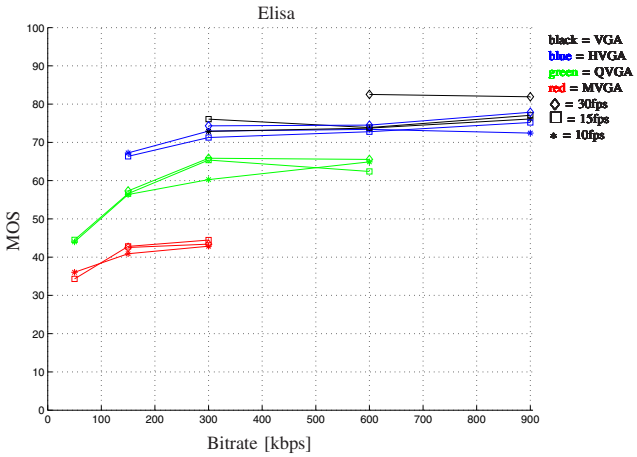

Fig. 4. MOS vs. bitrate for different frame rate and resolutions where Elisa is characterised to have medium SI and low TI.

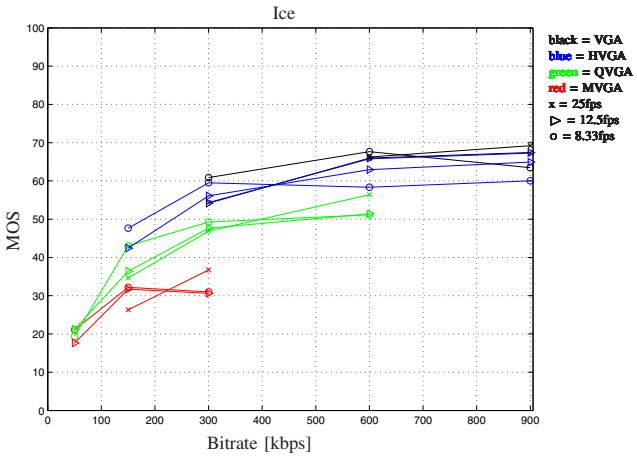

Fig. 5. MOS vs. bitrate for different frame rate and resolutions where Ice is characterised to have low SI and high TI.

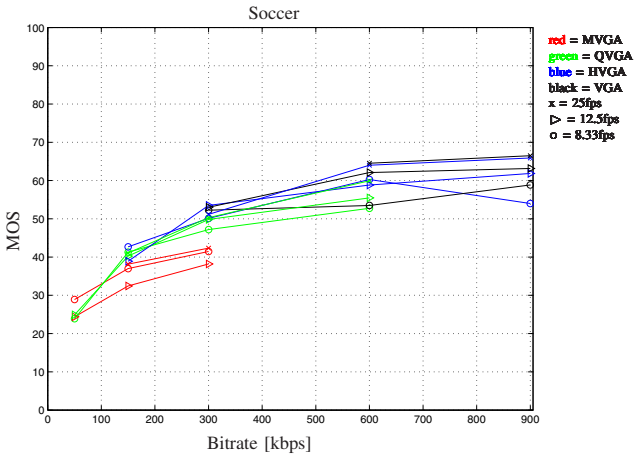

Fig. 6. MOS vs. bitrate for different frame rate and resolutions where Soccer is characterised to have low SI and high TI.

MOS scores, where BPP can be seen as an indicator of level of compression. To resolve the relative importance of these variables the multiway ANOVA technique was used and the variables are stated significant if the p-value was below 0.05. We used the Matlab function `anovan` for this purpose. The result from the ANOVA comparison is shown in Table III. It

TABLE III. *ANOVA applied to the sequences. The "*" marks the most significant variable.*

Sequences	Variable	Prob>F
All	FR	7.55e-11
	Res	5.81e-33*
	BPP	1.062e-19
Children	FR	4.22e-06
	Res	1.51e-09*
	BPP	9.75e-07
City	FR	0.0026
	Res	0*
	BPP	0.0035
Elisa	FR	0.0013
	Res	0*
	BPP	0.0308
Ice	FR	0.0001
	Res	0*
	BPP	0.0001
Soccer	FR	0.0002
	Res	0*
	BPP	0.0008

can be seen in Table III that for all the cases resolution (Res) has the highest significance which was also confirmed in the evaluations of the MOS scores illustrated in Fig. 2-6. Further the result indicates that BPP is the second most important variable, i.e. the compression level, except for Soccer and Ice which are the two sequences with highest temporal information where the frame rate (FR) has the same or higher significance.

IV. Conclusion

In this paper we have addressed the increasing interest of video communication and its attempt to maximize the perceptual quality during fluctuating bandwidth conditions. In many scenarios of streaming, realtime video communication, or other video applications, adaptive streaming is used to handle fluctuating network bandwidths. A suitable option for adaptive streaming is to scale down the video temporally or spatially before encoding to maintain a desirable level of perceptual quality while viewing resolution is constant. In a subjective assessment with five original sequences, 38 different combination of bitrate, frame rate, and resolution, 32 subjects were used. Both direct and statistical evaluation was made of the MOS scores, where MOS scores were plotted versus the bitrate, and ANOVA was used for statistical analysis. The result shows that preserving the spatial resolution throughout the process has the highest significance even in the scenarios with high temporal information. In comparison, in most studies when increasing the bitrate for a sequence with high SI this results in a preference for increased resolution or decreased quantization, while for sequences with high TI it results in a preference for increased frame rate. One of the reasons to this could be that all sequences were assessed at the same or limited number of different spatial resolutions. In our study, however, four different spatial resolutions were used, and all sequences were assessed at a fixed spatial viewing resolution. Future work planned includes using the presented results to develop a bit-stream based no-reference quality metric, as well as conducting a subjective study using higher resolutions to investigate the same parameters of video coding in other user scenarios.

REFERENCES

[1] K. Winstein, A. Sivaraman, and H. Balakrishnan, "Stochastic forecasts achieve high throughput and low delay over cellular networks," *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2013)*, April 2-5 2013.

[2] D. Robinson, Y. Jutras, and V. Craciun, "Subjective video quality assessment of HTTP adaptive streaming technologies," *Bell Labs Technical Journal*, vol. 16, no. 4, pp. 5–23, 2012.

[3] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 20–27, 2012.

[4] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, June 2011.

[5] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1316 –1324, nov. 2008.

[6] J. Korhonen, U. Reiter, and J. You, "Subjective comparison of temporal and quality scalability," in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*, 2011, pp. 161–166.

[7] M. Shahid, A. K. Singam, A. Rossholm, and B. Lovstrom, "Subjective quality assessment of H.264/AVC encoded low resolution videos," in *5th International Congress on Image and Signal Processing*, Oct. 2012, pp. 63 –67.

[8] R. Shmueli, O. Hadar, R. Huber, M. Maltz, and M. Huber, "Effects of an encoding scheme on perceived video quality transmitted over lossy internet protocol networks," in *IEEE Transactions on Broadcasting*, vol. 54, Sept 2008, pp. 628–640.

[9] A. Ciancio, J. F. L. De Oliveira, C. D. Estrada, and E. A. B. da Silva, "Impact of encoding configurations on the perceived quality of high definition videoconference sequences," in *IEEE International Symposium Circuits and Systems (ISCAS)*, May 2012, pp. 1716–1719.

[10] J.-S. Lee, F. De Simone, T. Ebrahimi, N. Ramzan, and E. Izquierdo, "Quality assessment of multidimensional video scalability," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 38–46, 2012.

[11] "Subjective video quality assessment methods for multimedia applications," September 1999, iTU-T, Recommendation ITU-R P910.

[12] "ITU-R Radio communication Sector of ITU, Recommendation ITU-R BT.500-12," 2009.

[13] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th ed. Ames, IA: Iowa State Univ. Press, 1989.