

VIDEO QUALITY PREDICTION BASED ON THE SPATIAL AND TEMPORAL CLASSIFICATION OF THE UNCOMPRESSED CONTENT

Harilaos Koumaras Anastasios Kourtis
{koumaras, kourtis}@iit.demokritos.gr
NCSR Demokritos
Athens, Greece

ABSTRACT

This paper deals with the notion of user satisfaction relative to the consumption of modern encoded video applications and services. Towards this, the concept of the Perceived Quality of Service (PQoS) is introduced and exploited for quantifying purposes. The objective of the paper is to research the relationship of the spatiotemporal dynamics of the content to the deduced perceptual quality of a MPEG-based encoded video. Once this relationship has been established, it is shown how it can be further exploited towards proposing a technique for video quality prediction at a pre-encoding state.

I. INTRODUCTION

Today, Quality of Service (QoS) is considered as the most essential issue in the field of mobile communications, having granted a quite diverse meaning depending on the provided service and system type. From an engineering aspect, QoS is strongly related with the appropriate traffic management and control, which offers efficient and flawless transmission eliminating Network QoS relative phenomena like delay, jitter and packet loss. From a customer perspective, the QoS concept coincides with the user satisfaction relative to the service consumption, which is mainly described as Perceived Quality of Service (PQoS). The concept of PQoS, although in general is customer-centric, practically it is significantly differentiated by the type of the delivered service.

Regarding video applications, the use of encoding techniques for the compression of the stream and the transmission of it over error-prone wireless environments cause visual quality degradation. Therefore, for video encoded services the issue of the PQoS is expressed in terms of encoding parameters (e.g. bit rate, frame rate etc.) and network-related impairments (e.g. packet loss, delay and delay variation for IP networks).

The existing literature of video quality assessment techniques focuses on models and techniques for evaluating and assessing the PQoS level of an already encoded and/or served video service. Thus, the aim of the current methods is the quantification of the user experience in terms of satisfaction. However, from a service provider aspect, there is a need for techniques, which will initially predict the PQoS level of a multimedia service according to the selected application parameters (i.e. codec type, bit rate etc.) and subsequently provide an estimation for the perceptual degradation of the delivered service due to the current network conditions.

Among the various encoding parameters that play significant role in the deduced perceived quality of service (PQoS) (e.g. bit rate, spatial and temporal resolution), the dynamics of the content (i.e. spatial and temporal activity of

the content) are critical for the final perceptual outcome. Although a lot of research has been focused on developing techniques and methods for estimating the video quality of a compressed/encoded video signal, the issue of studying the relation between the activity of the content and the deduced video quality has not been performed. The engineers generally agree that video with low spatiotemporal activity require less bit rate in order to achieve the same quality level in comparison with more competitive videos, featuring high spatiotemporal contents, but there is not any specific quantitative scheme proving this hypothesis.

This paper presents a study on the perceptual effectiveness of the spatiotemporal dynamics of the content in correlation to the encoding bit rate, considering that the rest encoding parameters (e.g. spatial and temporal resolution, encoding scheme, GOP pattern etc.) remain constant. Towards this, we provide results, depicting the actual perceived efficiency for various activity levels and not only the engineering effectiveness of each one, which may be measurable by simple error-based metrics, but as they are actually perceived by the human visual system through a respective objective assessment metric.

In this framework this paper uses five reference video clips, which are representative of different spatial and temporal activity levels, covering by this way all the range of the spatiotemporal scale. Afterwards, for each clip the relative PQoS vs. Bit rate curve for MPEG-4 encoding is drawn, showing how the differentiation in the content affects the deduced video quality.

The rest of the paper is organised as follows: Section II presents a brief literature review regarding the past and present trends in the field of video quality assessment. In Section III, we present a method for classifying a video sequence according to the spatiotemporal dynamics of its content. The relationship of the video quality (i.e. PQoS) to the spatial and temporal level of the video content is discussed in Section IV and in Section V, it is shown how this dependency can be exploited as future work for predicting the video quality vs. Bit rate curve of an uncompressed sequence. Finally, Section VI concludes the paper discussing the perspectives of the current research outcomes.

II. LITERATURE REVIEW

The advent in video quality was the application of pure mathematical error sensitive frameworks between the encoding and the original/uncompressed video sequence. These primitive methods, although they provided a quantitative approach about the quality degradation of the encoded signal, did not provide reliable measurements of the

perceived quality, because they miss out the characteristics and sensitivities of the Human Visual System (HVS).

Currently, the evaluation of the PQoS is a matter of objective and subjective evaluation procedures, each time taking place after the encoding process (post-encoding evaluation).

The subjective test methods, which have mainly been proposed by International Telecommunications Union (ITU) and Video Quality Experts Group (VQEG), involve an audience of people, who watch a video sequence and score its quality as perceived by them, under specific and controlled watching conditions. Afterwards, the statistical analysis of the collected data is used for the evaluation of the perceived quality. The Mean Opinion Score (MOS) is regarded as the most reliable method of quality measurement and has been applied on the statistical analysis of almost all the known subjective techniques.

Subjective picture/audio quality evaluation processes require large amount of human resources, establishing it as a time-consuming process (e.g. large audiences evaluating video/audio sequences). Objective evaluation methods, on the other hand, can provide PQoS evaluation results faster, but require large amount of machine resources and sophisticated apparatus configurations. Towards this, objective evaluation methods are based and make use of multiple metrics, which are related to the content's artifacts (i.e. tiling, blurriness, error blocks, etc.) resulting during an encoding process.

The majority of the objective methods proposed in the literature requires the undistorted/uncompressed source video as a reference entity in the quality evaluation process, and due to this are characterized as Full Reference (FR) Methods. However it has been reported that these complicated FR methods do not provide more accurate results than the simple mathematical measures (such as PSNR). Due to this some new full reference metrics that are based on the video structural distortion, and not on error measurement, have been proposed [1-3].

On the other hand, the fact that these methods require the original video signal as reference deprives their use in commercial video service applications, where the initial undistorted clips are not accessible. Moreover, even if the reference clip is available, then synchronization predicaments between the undistorted and the distorted signal (which may have experienced frame loss) make the implementation of the FR Methods difficult and impractical.

Due to these reasons, the recent research has been focused on developing methods that can evaluate the PQoS level based on metrics, which use only some extracted structural features from the original signal (Reduced Reference Methods) [4] or do not require any reference video signal (No Reference Methods). The NR methods can be classified into two classes: The NR-pixel based and the N-bitstream based. The first methods must initially decode the bit stream and estimate the video quality at the pixel/visual layer [5], while the second ones can evaluate the perceived quality by receiving as input the compressed bitstream, without requiring any decoding.

However, the issue of developing methods for predicting the video quality at a pre-encoding stage is still a research

challenge, where some earlier works of the authors have been published in the field [6-8]. This paper will also contribute towards this direction.

III. CLASSIFICATION OF THE SPATIOTEMPORAL DYNAMICS

Perceptual quality of a video sequence for specific encoding parameters and settings may vary with the spatiotemporal dynamics of the content. It is well established from the fundamental principles of the video coding theory that action clips with high dynamic content are perceived as degraded in comparison to the sequences with slow-moving clips, considering identical encodings.

In order to classify a video clip according to the spatial and temporal complexity of its content, a spatiotemporal grid [9] is considered as it is depicted on Figure 1.

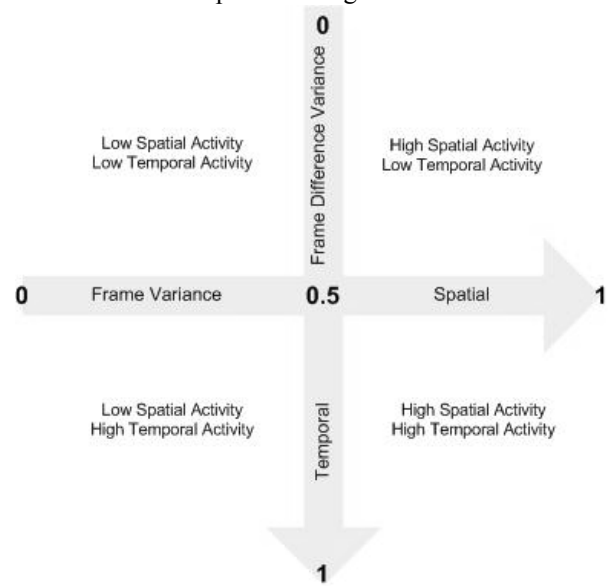


Figure 1: The Spatiotemporal grid used for classifying a video sequence according to its content dynamics

According to this approach, each video clip can be classified to four categories depending on its content dynamics, namely:

- Low Spatial Activity – Low Temporal Activity, which is defined as the upper left quarter in the grid.
- High Spatial Activity – Low Temporal Activity, which is defined as the upper right quarter in the grid.
- Low Spatial Activity – High Temporal Activity, which is defined as the lower left quarter in the grid.
- High Spatial Activity – High Temporal Activity, which is defined as the lower right quarter in the grid.

For the classification of each test signal, we use two discrete metrics for quantifying the spatial and temporal component of its content.

For the quantification of the spatial dynamics of a video sequence, the averaged frame variance is proposed as a metric. Considering that a frame consists N pixels x_i , then per frame its variance is defined as:

$$\sigma^2_{frame_y} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Therefore, based on the above equation, the averaged frame variance for the whole video duration, considering that the test signal numbers K total frames, is defined as

$$\frac{1}{K} \sum_{k=1}^K \sigma^2_{frame_y} = \frac{1}{K} \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^N (x_{k,i} - \bar{x}_k)^2$$

For the quantification of the temporal dynamics of a video sequence, the averaged variance of the successive frame luminance difference is proposed as a metric. Considering that a frame totally contains N pixels x_i and that the test signal numbers K total frames, then the averaged frame difference of the successive frame pairs is defined as:

$$\frac{1}{K-1} \sum_{k=2}^K \frac{1}{N} \sum_{i=1}^N (x_{k,i} - x_{k-1,i})$$

Therefore, the averaged variance for the overall duration of the test signal is defined as:

$$\frac{1}{K-1} \sum_{k=2}^K \left(\frac{1}{N} \sum_{i=1}^N (x_{k,i} - x_{k-1,i}) \right)^2 - \frac{1}{K-1} \sum_{k=2}^K \frac{1}{N} \sum_{i=1}^N (x_{k,i} - x_{k-1,i})$$

The scale in both axes refers to the normalized measurements (considering a scale from 0 up to 1) of the spatial and temporal component, according to the aforementioned metrics. The normalization procedure that has been followed in this paper, sets the test signal with the highest spatiotemporal content to the lower right quarter and specifically to the Cartesian (Spatial, Temporal) values (0.75, 0.75). This hypothesis, without any loss of generality, allows to our classification grid the possibility to consider also test signals that may have higher spatiotemporal content in comparison to the tested ones.






Suzie	
Cactus	
Flower Garden	
Table Tennis	
Mobile & Calendar	

Table 1: The five reference test signals

For the needs of this paper five short in duration reference sequences were used. These video sequences are depicted in table 1.

Applying the described spatial and temporal metrics on the reference signals of Table 1, their classification on the proposed spatiotemporal grid is depicted on Figure 2.

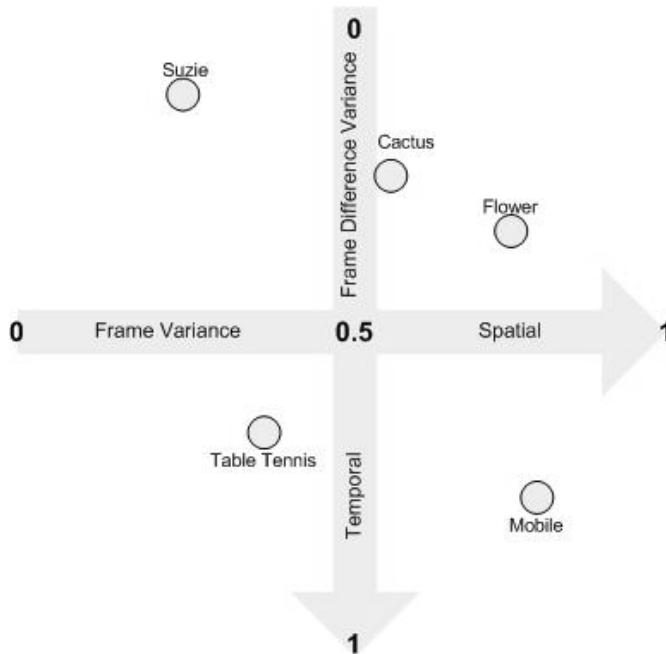


Figure 2: The Spatiotemporal classification of the test signals.

According to Figure 2, it can be observed that the spatiotemporal dynamics of the selected reference signals are distributed to all the four quarters of the spatiotemporal grid, indicating their diverse nature of the content dynamics. Moreover, the validity of the proposed metrics is certified by these experimental results, showing that they provide adequate differentiation among the dynamics of the signals under test.

In the next Section, we discuss the correlation between the proposed spatiotemporal classification and the mean PQoS of the encoded clip, derived for the whole video duration.

IV. SPATIOTEMPORAL ACTIVITY AND VIDEO QUALITY

The evaluation of the video quality is a subject of subjective or objective evaluation methods, which take place after the encoding process. From a business aspect, the application of subjective methods is not practically possible and for this reason objective methods are usually exploited, which provide in a fast and economically affordable way, a quality evaluation for any frame of the video sequence.

In this section, it is exploited the concept of averaging the objective quality evaluation per single frame over the whole video duration, deriving by this way a single indicative PQoS level for each video service, which is essential from a marketing perspective, especially for the service providers.

It must be noted that the used sequences in this paper are reference signals with limited duration and therefore with practically homogeneous content (i.e. constant spatial and

temporal activity level). The case of longest in duration videos is out of the scope of this paper and therefore is not examined.

For the experimental section, each test video clip of Table 1, was encoded from its original uncompressed format to ISO MPEG-4 (Simple Profile) format, at different constant bit rates (spanning a range from 50kbps to 1.5Mbps for CIF (Common Intermediate Format) with key-frame period equal to 100 frames in both cases). For each corresponding bit rate, a different ISO MPEG-4 compliant file was created. The frame rate was set at 25 frames per second (fps) for the whole encoding process.

Each ISO MPEG-4 video clip was then used as input in a no-reference objective quality measurement tool [10]. From the resulting PQoS per frame measurements, the Mean PQoS (MPQoS) value of each clip was calculated. This experimental procedure was repeated for each video clip under test. The results of these experiments are depicted in Figure 3, where PQ_L denotes the lowest acceptable MPQoS level (corresponding to 70 in the scale from 1 to 100 for CIF resolution) and PQ_H denotes the best MPQoS level that each video can reach.

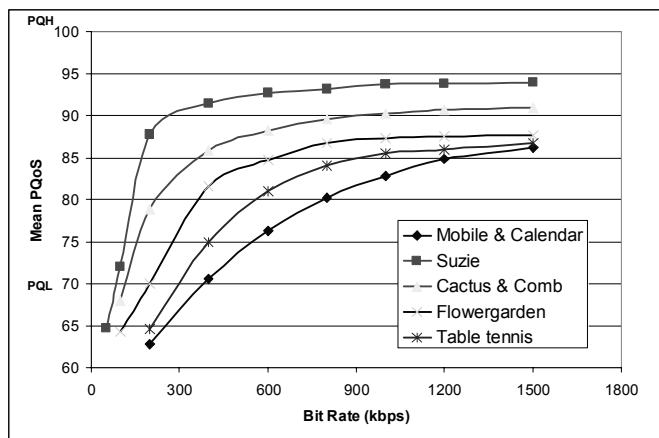


Figure 3: The Mean PQoS vs. Bit Rate curves

Referring to the curves of Figure 3, the following remarks can be made:

1. The minimum bit rate BR_L , which corresponds to the lowest MPQoS value depends on the spatiotemporal activity of the video content.
2. The variation of the MPQoS vs. bit rate is an increasing function, but non linear. Moreover, the quality improvement of an encoded video clip is not significant for bit rates higher than a specific perceptual threshold (PQ_H). This threshold depends on the S-T activity of the video content.
3. The shape of the MPQoS vs. bit rate curve is strongly related to the spatiotemporal dynamics of the encoded content.

Practically, the transposition of the curve to the upper-left area means that content with low spatiotemporal activity (i.e. a clip from the upper left quarter of the spatial-temporal grid) reaches a higher PQoS level at relatively lower bit rate in comparison to a sequence with higher spatiotemporal content (i.e. a clip except from the first quarter in the spatial-temporal

grid). In addition, when the encoding bit rate decreases below a threshold, which depends on the video content, the PQoS practically “collapses”.

On the other hand, the transposition of the curve to the lower-right area means that content with high spatiotemporal activity (i.e. a clip from the down right quarter) requires higher bit rate in order to reach a satisfactory PQoS level. Nevertheless, it reaches its maximum PQoS value more smoothly than in the low S-T activity case.

Moreover, as we have shown in earlier works [6-7], each MPQoS vs. bit rate curve can be successfully described by an exponential function of the general form

$$MPQoS = [PQ_H - PQ_L] (1 - e^{-\alpha [BR - BR_L]}) + PQ_L, \quad \alpha > 0 \text{ and } BR > BR_L$$

where PQ_L is the lowest acceptable perceived quality (i.e. 70) and BR_L the encoding bit rate that corresponds to the specific level. Parameter α defines the shape of the curve.

Therefore, based on the spatiotemporal classification of our test signals, which is depicted on Figure 2 and given the form of the Mean PQoS vs. bit rate curves, shown on Figure 3, the following relationships can be derived between the spatiotemporal quarter, to which belongs a specific video content and the respective BR_L , PQ_H and parameter α of its respective curve.

Therefore, we provide hereby in Table 2 our experimental results for the four quarters of the spatiotemporal grid.

S-T Quarter	BR_L	PQ_H	parameter α
Upper Left	95	93.91	0.0083
Upper Right	110	90.89	0.0063
Lower Left	250	86.72	0.0053
Lower Right	400	86.20	0.0045

Table 2: Dependencies between spatiotemporal quarters and the parameters of the proposed exponential approximation

V. FUTURE WORK

Based on the aforementioned analysis, we present here how these results can be further exploited towards developing a video quality prediction method based on the spatiotemporal quantification of the uncompressed signal.

Therefore, we can describe the derived MPQoS vs. bit rate curve of each test signal with N total frames, which is encoded at bit rate n from $BitRate_{min}$ to $BitRate_{max}$ as a set C_{S-T} , where each element F_n is a triplet, consisting the parameter α , the BR_L and the PQ_H .

$$C_{S-T} \triangleq \{m : (\alpha, BR_L, PQ_H)_n = F_n, n \in [BitRate_{min}, BitRate_{max}]\}$$

Hence, considering an unknown video clip, which is uncompressed and we want to predict its corresponding C_{S-T} set that better describes its perceived quality vs. bit rate curve before the encoding process, we define for all the sets C_{S-T} the Absolute Difference Value (ADV) between the C_{S-T} triplet elements and the theoretically derived ones from the spatiotemporal classification of the uncompressed content and the mapping rules of Table 2.

Due to the fact that the additive property is valid, it is concluded that when the ADV of the quality between reference $F_{BitRate_i}$ and experimental $F_{BitRate_e}$ is minimum, then this C_{S-T} set contains the triplet element that minimizes the ADV and therefore describes better the specific video.

Thus, this algorithm can successfully predict the Mean PQoS vs. Bit rate curve of the specific video exploiting the proposed spatiotemporal classification. Consequently, the service provider can predict analytically the bit rates that satisfy specific perceived quality levels at a pre-encoding state.

VI. CONCLUSIONS

This paper has presented a spatiotemporal classification method, which provides a quantification of the video content dynamics through the application of objective metrics on the uncompressed domain. By developing mapping rules between the deduced content dynamics measurements and the parameters that specify the MPQoS vs. bit rate curves, it is demonstrated how the video quality is affected by the spatial and temporal activity level. Finally, it is discussed as a future work how the outcomes of the present paper can be further exploited in developing a video quality prediction model.

VII. ACKNOWLEDGEMENT

The work in this paper was carried out in the framework of the Information Society Technologies (IST) project ENTHRONE phase II/ FP6-38463.

REFERENCES

- [1] Wang, Z., H.R. Sheikh, and A.C. Bovik, "Objective video quality assessment, in *The Handbook of Video Databases: Design and Applications*", B. Furht and O. Marqure, Editors. 2003, CRC Press. p. 1041-1078.
- [2] VQEG. "Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment". 2000.
- [3] Wang, Z., A.C. Bovik, and L. Lu. "Why is image quality assessment so difficult?" in IEEE International Conference on Acoustics, Speech, and Signal Processing. 2002.
- [4] Gunawan, I.P. and M. Ghanbari. "Reduced-Reference Picture Quality Estimation by Using Local Harmonic Amplitude Information". in London Communications Symposium 2003. 2003.
- [5] Lu, L., et al. "Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video". in IEEE International Conference on Multimedia. 2002.
- [6] H. Koumaras, A. Kourtis, D. Martakos, "Evaluation of Video Quality Based on Objectively Estimated Metric", Journal of Communications and Networking, Korean Institute of Communications Sciences (KICS),

Vol. 7, No.3, pp.235-242, Sep 2005 (available online for download at <http://citeseer.comp.nus.edu.sg/747634.html>)

- [7] H. Koumaras, A. Kourtis, D. Martakos, J. Lauterjung, "Quantified PQoS Assessment Based on Fast Estimation of the Spatial and Temporal Activity Level", Multimedia Tools and Applications, Springer Editions, Published Online (available online for download at <http://www.springerlink.com/content/f8v2p4r852266415/>).
- [8] H. Koumaras, E. Pallis, G. Xilouris, A. Kourtis, D. Martakos, J. Lauterjung, "Pre-Encoding PQoS Assessment Method for Optimized Resource Utilization", 2nd Inter. Conference on Performance Modelling and Evaluation of Heterogeneous Networks, Het-NeTs04, Ilkley, United Kingdom, 2004.
- [9] N. Cranley and L. Murphy, "Incorporating User Perception in Adaptive Video Streaming Systems", in Digital Multimedia Perception and Design (Eds. G. Ghinea and S. Chen), published by Idea Group, Inc., May 2006. ISBN: 1-59140-860-1/1-59140-861-X
- [10] J. Lauterjung, "Picture Quality Measurement", Proceedings of the International Broadcasting Convention (IBC), Amsterdam, 1998, pp. 413-417.