A Framework for End-to-End Video Quality Prediction of MPEG Video

Harilaos Koumaras, C-H Lin, C-K Shieh, Anastasios Kourtis

PII: S1047-3203(09)00094-7
DOI: 10.1016/j.jvcir.2009.07.005
Reference: YJVCI 836

To appear in: *J. Vis. Commun. Image R.*

Received Date: 6 January 2009
Revised Date: 16 April 2009
Accepted Date: 13 July 2009

Please cite this article as: H. Koumaras, C-H. Lin, C-K. Shieh, A. Kourtis, A Framework for End-to-End Video Quality Prediction of MPEG Video, *J. Vis. Commun. Image R.* (2009), doi: 10.1016/j.jvcir.2009.07.005

ELSEVIER

JOURNAL OF VISUAL Communication & IMAGE Representation

www.elsevier.com/locate/jvci

A Framework for End-to-End Video Quality Prediction of MPEG Video

Harilaos Koumaras[*α, c], C-H Lin[b], C-K Shieh[b], and Anastasios Kourtis[α]

"a NCSR DEMOKRITOS, Institute of Informatics and Telecommunications, Athens, Greece"

"b Cheng Kung University, Tainan, Taiwan"

"c Business College of Athens (BCA), Computer Science Department, Athens, Greece"

Abstract

This paper proposes, describes and evaluates a novel framework for video quality prediction of MPEG-based video services, considering the perceptual degradation that is introduced by the encoding process and the provision of the encoded signal over an error prone wireless or wire-line network. The concept of video quality prediction is considered in this work, according to which the encoding parameters of the video service and the network QoS conditions are used for performing an estimation/prediction of the video quality level at the user side, without further processing of the actual encoded and transmitted video content. The proposed prediction framework consists of two discrete models: i. A model for predicting the video quality of an encoded signal at a pre-encoding stage by correlating the spatiotemporal content dynamics to the bit rate that satisfies a specific level of user satisfaction, and ii. A model that predicts primarily the undecodable frames (and subsequently the perceived quality degradation caused by them) based on the monitored averaged packet loss ratio of the network. The proposed framework is experimentally tested and validated with video signals encoded according to MPEG-4 standard.

Keywords: Video Quality; MPEG; PQoS; Packet Loss; Prediction

1. Introduction

Today, the quality level of a digitally encoded video signal that is transmitted over a resource-constrained network depends largely on the performance capabilities of the encoder itself and the available bandwidth of the transmission network. Both of these factors may introduce to the video signal a respective perceptual loss, due to which the finally delivered video at the end-user is degraded in comparison to the original uncompressed one. This situation has set new research challenges for the assessment of video quality as a part of the encoding and network-resource management system, making Video Quality Assessment (VQA) an active research area during the last years.

VQA in general is the process of assessing the perceptual level of a video service, which has been come through a procedure of encoding (i.e. compressing), data loss or other processing. Particularly, VQA focuses on quantifying the degradation that is introduced by encoding systems and/or during the transmission of the encoded signal over error-prone or resource-constrained transmission channels. Thus, an original video signal may experience two phases of degradation that VQA must address:

i. Quality degradation due to encoding procedure,
ii. Quality degradation due to transmission errors.

Concerning the first phase, the degradation is strongly related on the selected encoding parameters and mainly on the

[*] Corresponding author. Tel.: +30-210-650-3107; +30-210-72-53-783; e-mail: koumaras@iit.demokritos.gr, koumaras@bca.edu.gr.

encoding bit rate. Currently, the determination of the encoding bit rate that satisfies a specific level of video quality is a matter of repetitive post-encoding subjective or objective video quality assessments, each time taking place after the encoding process [1]. Subjective evaluation of video signals requires large amount of human resources, establishing it as an impractical procedure for a service provider. Similarly, the repetitive use of objective metrics [1-6] on already encoded sequences may require numerous test encodings for identifying the appropriate encoding parameters for a specific quality level, which apart from time consuming is financially unaffordable from a business perspective as well.

Concerning the second phase (i.e. during the transmission of the service), it must be considered that encoded video services, due to their interdependent frame-structure, are highly sensitive to transmission errors (e.g. packet loss) and require high transmission reliability in order between sender and receiver devices to maintain flawless video transmission and stream synchronization. Especially, in video streaming, each transmitted from one end video packet, it can be received at the other end, either intact, with errors or get totally lost. In the last two cases, the perceptual outcome is similar, since the decoder at the end-user usually discards the packet with errors, causing visual artifacts not only on the frame that the specific dropped packet belongs to, but also to the subsequent decoded frames that are dependent on the dropped one as well. Currently, the assessment of the quality degradation of a video signal due to network QoS issues is performed either by applying subjective or objective assessment methods during the decoding process at the end-user side. From the network perspective, this procedure does not provide a mapping between the network QoS parameters (e.g. packet loss ratio) and the respective degradation of the video quality [7], mainly due to the stochastic nature of the phenomenon. Therefore, such a procedure is not practically applicable by a network operator, who wants to monitor how the QoS network conditions affect the video quality of a multimedia service, because it requires not only continuous monitoring (which can be dealt with the appropriate management systems), but also perceptual evaluation as well.

Therefore, the concept of video quality prediction (and not assessment) is necessary to be introduced, describing a novel category of methods that do not require the encoded signal in order to define its quality level. More specifically, these methods predict the video quality level that will have a specific video content after the encoding process based on the encoding parameters. Further processing of the encoding/encoded data is not required, minimizing by this way the respective complexity and resource consumption.

In this context, the paper proposes, describes and evaluates a novel framework for end-to-end video quality prediction of MPEG video services, which moves beyond the existing VQA methods both at the encoding and transmission processes, without requiring as input the encoded and/or streamed/decoded digital video. More specifically the proposed framework focuses on twofold procedures:

i. The prediction of the encoding parameters that satisfy a specific video quality level in terms of encoding bit rate and content dynamics.
ii. The mapping of the packet loss ratio of the transmission channel to the respective quality degradation.

Through the proposed end-to-end video quality prediction framework, the content provider will be able to predict the finally delivered video quality level at the end-user side, considering specific encoding parameters and transmission conditions, without being necessary to further process the actual encoded and/or streamed video content. So, the aim of this paper is to present a methodology that is deterministic, reasonably accurate and simple enough to support the ever-increasing applications of large-scale streaming media over resource-constrained packet networks. Such an end-to-end perceived QoS framework will not only play an essential role in performance analysis, control and optimization of multimedia systems, but it will also contribute towards a more efficient network/system resource allocation, utilization and management.

For clarity reasons, the frequently used term in this paper of the Perceived QoS (PQoS) is defined hereby as the video quality that the end-user experiences while consuming a multimedia service. Our proposed test bed is based on simulated network scenarios using a network simulator NS-2 [33], which provides a lot of flexibility for evaluating different configurations used in this paper. For demonstration purposes, the proposed framework is tested also on a real packet-based network.

After this introductory section, the rest of the paper is organized as follows: Section 2 performs a literature review on the relevant research works, focusing both on the video quality assessment and the estimation of the degradation due to the conditions of the transmission channel. Also, in this section are described the fundamental concepts of a MPEG signal, which will be later used for the description of the proposed framework. Section 3 presents the proposed model for the prediction and determination of the encoding bit rate value that satisfies a specific level of user satisfaction. Similarly, Section 4 presents the proposed model of mapping the packet loss rate of the network on the decoding performance of the video decoder. In Section 5, it is performed the experimental validation and calibration of the proposed models, proving the accuracy and efficiency of them. In Section 6 the proposed end-to-end video quality prediction framework is introduced, described and evaluated. Section 7 presents an experimental application of the proposed end-to-end framework at a real small scale packet network test-bed with controllable network conditions. Finally, Section 8 concludes the paper.

2. Background

2.1. Video Quality Assessment Methods

The advent in the field of video quality assessment is found in methods that use error-sensitive functions between the encoded and the original/uncompressed video sequence in order to peform an assessment of the quality. These primitive methods [1], although they initially provided a quantitative approach of the degradation caused by the encoding procedure, practically they do not reflect accurately the video quality as it is observed and perceived by human viewers.

Beyond these primal models, currently the evaluation of the video quality is a matter of subjective and objective procedures, which are applied 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, who watch a video sequence and score its quality as perceived by the participants, under specific and controlled watching conditions. Afterwards, usually the Mean Opinion Score (MOS) is exploited as evaluation metric, which provides a numerical indication of the perceived quality of the media after compression and/or transmission, for further statistical analysis and processing of the collected data.

Generally speaking, subjective video quality evaluation processes require large amount of human resources, making them time-consuming (e.g. large audiences are required for evaluating test sequences), limiting their use to experimental purposes only. On the other hand, objective evaluation methods provide faster quality assessment, exploiting multiple metrics that use mathematical models to quantify the perceptual impact of the encoding artifacts (e.g. blockiness, blurriness, error blocks, etc) on the video quality level.

The majority of the objective methods require the undistorted video source as a reference entity in the quality evaluation process. Due to this requirement, they are characterized as Full Reference (FR) Methods [1-3]. In this direction, some novel full reference metrics have been recently proposed based on the video structural distortion and content entropy [4-8]. On the other hand, the fact that these methods require the original video signal as reference deprives their use in broadcasting/streaming services, where the initial undistorted signal is not always available or not accessible at the user side.

Due to this reason, 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) [9-13] or do not require any reference video signal (No Reference Methods). The NR methods can be classified into two classes: The NR-visual based and the NR-coded based. The first methods must initially decode the bit stream and estimate the video quality at the pixel domain [14-19], while the second ones assess the perceived quality directly through the compressed bit stream, without requiring any decoding process [20-25].

Towards this direction, in [39] it is suggested the use of the frame rate as an objective metric for exploring its impact on the perceptual quality of a video. According to this framework, the authors propose the use of the product of i) a metric that assesses the quality of a quantized video at the highest frame rate and ii) a temporal correction factor, which reduces the respective quality of the first metric to the one that corresponds to the actual frame rate.

Similarly [40], it examines the simultaneous impact of various parameters on the video quality level and proposes a cross-dimensional perceptual quality assessment framework, focusing on low-bit rate videos. More specifically, the paper produces outcomes on how specific combinations of the encoding parameters affect the video quality level.

However, all the aforementioned works have focused on mapping the video signal to a specific perceptual quality level by quantifying the quality degradation that has been caused by the encoding process. So, all the previous works have been focused on the unidirectional mapping of encoding parameters to video quality. Due to the need for forecasting the encoding parameters that satisfy a specific level of user satisfaction, some alternative objective methods have been proposed lately, which perform the reverse mapping (i.e. video quality to encoding parameters) and move beyond the simple post-encoding quality assessment by introducing the concept of video quality prediction for given encoding parameters and content dynamics at a pre-encoding state [26-28]. Towards this direction will focus also the content of this paper and more specifically the paper addresses the determination of the quality degradation caused by the encoding process of a video signal by extending the concept of the encoding video quality prediction along with the respective degradation caused by the network.

2.2. Quality Degradation due to Transmission Errors

The issue of deterministically mapping the perceptual impact of transmission errors (like packet loss) on the delivered perceptual video quality at the end-user is a fresh topic in the field of video quality assessment since the relative literature appears to be limited with a small number of relative published works.

In this framework, S. Kanumuri *et al* [29] proposed a very analytical statistical model of the packet-loss visual impact on the decoding video quality of MPEG-2 video sequences, specifying the various factors that affect the video quality and

visibility (e.g. Maximum number of frames affected by the packet loss, on what frame type the packet loss occurs etc). However, this study focuses mainly on the pure study of MPEG-2 decoding capabilities, without considering parameters of the streaming or the latest encoding standards.

Similarly, in [30] is presented a transmission distortion model for real-time video streaming over error-prone wireless networks. In this work, an end-to-end video distortion study is performed, based on the modeling of the impulse propagation error (i.e. the visual fading behavior of the decoding artifact). The deduced model, although it is very accurate and robust, enabling the media service provider to predict the transmission distortion at the receiver side, is not a generic one. On the contrary, it is highly dependent on the video content dynamics and the selected encoder settings. More specifically, it requires an initial quantification of the spatial and temporal dynamics of the content, which will allow the appropriate calibration of the model. This prerequisite procedure (i.e. adapting the impulse transmission distortion curve based on the least mean square error criteria) is practically inapplicable by an actual content creator/provider. Moreover, due to the strong dependence of the proposed model on the spatiotemporal dynamics of the content its implementation is limited to sequences with very short duration (i.e. up to 10 sec), since the consideration of a unique impulse transmission distortion will not be accurate for longer video signals.

In [37], a mathematical approach of video quality assessment over a packet network is presented, which is based on mapping the jitter parameter of a packet network to packet loss based on the capacity of the playback buffer. Therefore a unified packet loss value is calculated, which comprises both the actual packet loss ratio and the deduced one by the respective network jitter parameter. This unified packet loss ratio is mapped primarily to frame loss and then to quality rating. However, for the model to be accurate an impairment degree is introduced, which sets the impact of each packet loss on the frame that it belongs to. The value of this parameter is calculated based on the content of the signal, making the proposed method to act as a reduced reference one. So, for the model to provide correct measurements, it must be fed with the characteristics of the encoded video. Also, the validation section presented in [37] is very limited, without testing the validity of the proposed model on an adequate number of experimental signals.

Finally, in [38] a theoretical analysis of the overall mean squared error in hybrid video coding is presented for the case of error prone transmission. This work has been validated with simulation on H.263 signals, but its aim has mainly focused on the effects of the INTRA coding and spatial loop filtering, while during the transmission a specific set of network parameters has been considered and tested.

Towards this direction will contribute also this paper, by proposing a mapping framework of network packet loss to video quality degradation in terms of subjective MOS.

2.3. Novelty of the Current Paper

In this context [26, 27, 28, 37, 38, 39, 40] our paper extends concepts and models of the existing literature to the case of MPEG-based encoded signals by proposing, testing and validating a generic model for end-to-end video quality prediction of to be encoded video services. Our proposed framework consists two discrete parts:

- A method for predicting and specifying for a given content the encoding parameters that satisfy a specific perceptual level
- A model for calculating the perceptual impact of the packet loss ratio on the delivered perceived quality of the transmitted service.

In comparison with the existing similar works [26-28, 39], the first part of the proposed framework in this paper moves beyond the existing works, because its aim is to provide for a specific video signal the decision on the encoding process that must be followed in order to satisfy a specific level of user satisfaction (i.e. quality level). Although, the proposed model relies on benefit functions like in [39], it differs and innovates because it provides quality prediction before the encoding process and not assessment as in [39] after the encoding process. Similarly in [40], the use of the multi-dimensional objective metrics is used for developing a decision assessment framework for deriving specific outcomes on the perceptual efficiency of the encoding process for low bit rate videos, while the first part of this paper provides a prediction framework for videos at any encoding bit rate.

Moreover, in contrast to any existing FR, RR or NR perceptual model that is applied on the decoded video signal, to the best of our knowledge, this work proposes one of the first models that extends the concept of assessment and provides end-to-end video quality *prediction* across all the lifecycle of the media content: From the service generation down to the content consumption at the viewer side. In this work, the concept of prediction is defined as the principle that the proposed framework is capable of calculating the delivered video quality of an encoded signal, without requiring as input the decoded video signal but only its encoding characteristics (i.e. encoding bit rate) and network statistics (i.e. packet loss).

The fact that the proposed method actually does not require any test encoding to be done or a video to be actually sent via the transport network to the end-user for performing its video quality estimation, it provides a content-aware aspect,

according to which at any time it can be predicted the video quality level that would have a video service if it was actually transporting over the specific network, subject to its encoding parameters and the current network conditions. This condition provides significant saving both in the bandwidth allocation and the service management. Concerning the bandwidth allocation the saving comes from the fact that specific video provisions may not be performed if the network conditions does allow an acceptable quality level. For the service management , the saving comes from the fact that adaptation actions may be taken in the requested video service in order to deal with the existing network conditions better and finally achieve optimized quality levels at the end-user.

This saving of resources that is performed by the proposed video quality prediction, outweighs the partial loss of accuracy that may be introduced by not considering the observed video quality at the user side, and instead replicating it theoretically. More specifically, the rest parts of this paper show that the accuracy of the proposed model is quite satisfactory, so by applying the proposed prediction model, many steps of the media delivery chain are saved in terms of resources and time, namely: The video encoding procedure, the video provision and the video consumption. The proposed model provides content-aware capabilities to the network that will be applied, which alleviates the network operators from using subjective/objective statistical methods for mapping network QoS to perceived QoS; a procedure that practically is not feasible because end-users must provide feedback for the perceptual level of the service that they consume. Moreover, the service provider wants to be able to avoid the provision of a service at an unacceptable quality level, a situation that currently cannot be avoided, since the delivered quality level can be assessed only when the service has been actually delivered (even in unacceptable quality). Therefore, the proposed prediction model can improve the efficiency of the media delivery in many aspects, providing also network/bandwidth resource saving.

In this paper both reference video signals and actual video trailers have been used as YUV source signals throughout the experimental section of the paper. The YUV files that are movie trailers have been generated by converting High-Definition or Standard-Definition trailers to YUV and are not released because they subject to copyright issues. The rest reference YUV test signals that have been used in this paper are available for downloading at [41].

2.4. MPEG Video Structure

The MPEG standard [31] defines three frame types for the compressed video streams, namely I (Intra-coded), P (Predictive-coded) and B (Bi-directionally predictive-coded) frames. The frame classification is mainly based on the procedure, according to which each frame type has been generated and encoded. The successive frames between two succeeding I frames define a Group Of Pictures (GOP). In the MPEG literature the GOP pattern is described by two parameters GOP (N, M), where N defines the GOP length (i.e. the total number of frames within each GOP) and the (M-1) is the number of B frames between I-P or P-P frames, as shown in Figure 1. The arrows indicate the encoding/decoding correlation between the frames and more specifically that the B and P frames depend on the respective preceding and succeeding I or P frames.

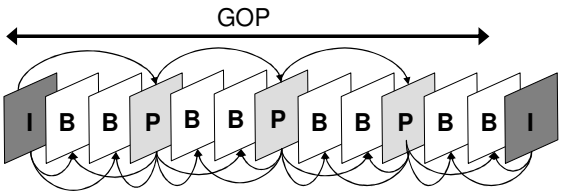

Fig. 1. A sample of MPEG GOP (N=12, M=3)

Therefore, from the hierarchical structure of MPEG encoding as it is depicted on Figure 1, a video frame may be considered as directly or indirectly undecodable. Direct undecodable is considered a video frame, which packets have been lost during their transmission. Thus in this paper we define as successful decoding of a frame, the case where all the packets that comprise the specific frame have been successfully received by the video decoder. On the other hand, indirect undecodable is considered a video frame when a reference frame is directly or indirectly undecodable. For simplicity, in this paper we consider the standard error concealment method i.e. the Zero Error Concealment (ZEC) and we set the Decodable Threshold (DT) [31-32] equal to 1.0, which means that any unsuccessfully decoded frame (i.e. a frame with at least one packet loss) is discarded at the decoder. Therefore, our analysis provides the worst-case scenario in terms of video quality degradation and decoding robustness.

3. Modeling and Predicting Video Quality

In digital video encoding the Block Discrete Cosine Transformation (BDCT) is exploited, since it exhibits very good energy compaction and de-correlation properties. In this paper, we use the following conventions for video sequences: Every real NxN frame f is treated as a N^2x1 vector in the space R^{N^2} by horizontal or vertical raster scanning.

The DCT is considered as a linear transform from $R^{N^2} \to R^{N^2}$. Thus, for a typical frame f, we can write:

$$F = Bf$$

Since B matrix is unitary, the inverse DCT can be expressed by B^t, where t denotes the transpose of a vector or matrix. Thus, the inverse transform can be described as:

$$f = B^t F$$

The elements of frame $F = Bf$ in the frequency domain can be expressed as the coefficients of the vector f, using the DCT basis in R^{N^2}. Thus

$$f = \sum_{n=1}^{N^2} F_n e_n$$

where e_n is the normalized DCT basis vector and F_n the DCT coefficients of f.

The high compression during the MPEG-related encoding process is (among other procedures) based on the quantization of the DCT coefficients, which in turn results in loss of high frequency coefficients. Within a MPEG block/macroblock, the luminance differences and discontinuities between any pair of adjacent pixels are reduced, by the encoding and compression process. On the contrary, for the pairs of adjacent pixels, which are located across and on both edge sides of the border of adjacent DCT blocks, the luminance discontinuities are increased by the encoding process, due to the independent quantization of the individual blocks. The blocking or blockiness effect is the most prominent visual distortion in a compressed sequence, due to the regularity of the pattern. Thus, after the quantization:

$$F_n^{'} = Q[F_n]$$

where $Q[\,]$ denotes the quantization process.

So, at the decoder side, the final reconstructed frame (after motion estimation and compensation modules) will be given by:

$$f^{'} = \sum_{n=1}^{N^2} F_n^{'} e_n$$

Thus, the perceived quality degradation per frame due to the encoding and quantization process can be expressed by an error based framework in the luminance domain Δf_Y between the original and the decoded frames.

$$\Delta f_Y \propto f_Y - f_Y^{'}$$

In this context, an average of the Perceived Quality of a video Service (PQoS) level for the whole encoding signal, consisting of *N* frames, can be derived by the following error-based equation:

$$< PQoS >_{video} \propto \sum_{i=1}^{N} \Delta f_{Y_i}$$

In this paper we deifne as Mean Perceived QoS (MPQoS) the averaged PQoS of a video signal for its whole duration. Based on this error-based framework, the *SSIM* objective quality metric [6] was selected for experimentally measuring and quantifying the degradation caused in the original uncompressed signal by the encoding process. The *SSIM* is a FR objective metric, which measures the structural similarity between two images/video sequences, exploiting the general principle that the main function of the human visual system is the extraction of structural information from the viewing field. The *SSIM* was preferred by the authors of this paper to be used for the experimental needs of this paper, because it has performed quite satisfactorily in the relative performance evaluation studies [6]. However, the selection of the specific metric does not limit the efficiency of the proposed model to *SSIM* metric only, since any other objective metric could be used instead of it for producing the experimental benefit functions of Fig. 3. Thus, considering that *f* and *f'* depicts the frames of the uncompressed and compressed signal respectively, then the *SSIM* is defined as:

$$SSIM(f, f') = \frac{(2\mu_f \mu_{f'} + C_1)(2\sigma_{ff'} + C_2)}{(\mu_f^2 + \mu_{f'}^2 + C_1)(\sigma_f^2 + \sigma_{f'}^2 + C_2)}$$

where μ_f, $\mu_{f'}$ are the mean of f and f', σ_f, $\sigma_{f'}$, $\sigma_{ff'}$ are the variances of f, f' and the covariance of f and f', respectively. The constants C_1 and C_2 are defined as:

$$C_1 = (K_1 L)^2 \quad C_2 = (K_2 L)^2$$

where L is the dynamic pixel range and $K_1 = 0.01$ and $K_2 = 0.03$, respectively.

Thus, *SSIM* metric can be considered as an appropriate metric for quantifying the video quality level of an encoded signal for the experimental and validation needs of this paper, without excluding the use of other metrics as well. The selection of the *SSIM* or any other metric does not alter the algorithm of the proposed model, since only the respective arithmetic values of the experimental data may differ among the various metrics, causing relative differences between them and not substantial ones.

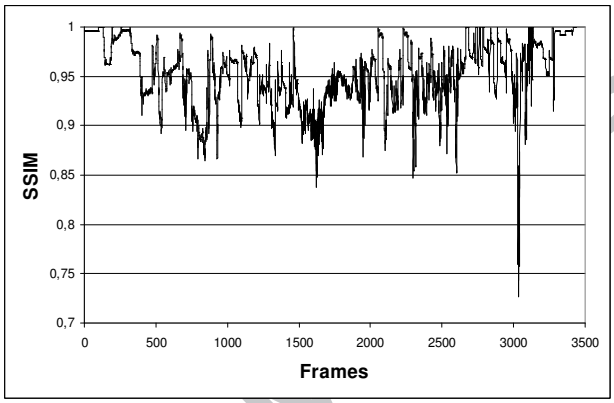

Fig. 2. The instant *SSIM* per frame of "16 Blocks" CIF 200kbps

Figure 2 depicts a typical example of the *SSIM* measurement per frame for the video trailer "16 Blocks", which was encoded using the MPEG-4/H.264 standard at 200 Kbps VBR with Common Intermediate Format (CIF) resolution and 25 frames per second (fps). The instant *SSIM* vs. time curve (where time is represented by the frame sequence) varies according to the spatiotemporal activity of each frame, which causes different quality degradation for the same quantization parameters. For frames with high complexity the instant *SSIM* level drops (i.e. <0.9), while for static frames the instant *SSIM* is higher (i.e. >0.9 or equal to 1, which denotes no degradation at all).

The concept of averaging the *SSIM* for the whole video duration can be exploited for deriving the Mean Perceived Quality of Service (MPQoS) as a representative perceptual parameter for the specific content. However, although the MPQoS provides a single perceived quality measurement, which is more practical especially for the service providers, for the case of long duration videos the use of just one representative measurement of the perceived quality may not be accurate, because the spatial and temporal activity level of the content may differ significantly, especially in the case of heterogeneous content. In such long sequences, the proposed average metric can be combined along with a shot boundary detection algorithm, which will lead to calculating partial MPQoS for various scenes. However, this case is not examined in the current paper, but since the proposed framework is tested and validated on videos with short duration then the successful extension to long duration videos in conjunction with a shot boundary detection method is considered as trivial. The paper aims at quality issues in short in duration signals, such as movie trailers, news or music clips with practically constant and homogeneous level of spatiotemporal activity.

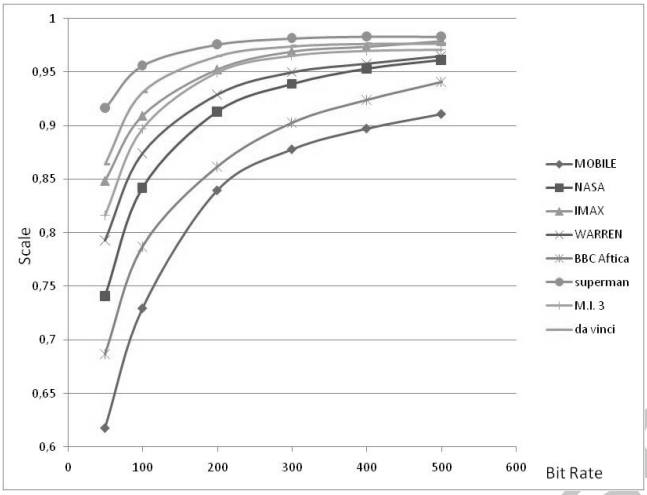

Fig. 3. The <PQoS>$_{SSIM}$ scale (0-1) vs. bit rate curves @25fps CIF VBR for various test signals

In this context, eight short in duration video clips were selected and used for the needs of this paper. The experimental set consisted trailer video clips with duration up to three minutes. Each trailer clip was encoded from its original H.264 format with Hi-Def resolution (i.e. 720p) to MPEG-4/H.264 Baseline Profile at diverse VBR values, applying exactly the same encoding procedure. For each bit rate, a different MPEG-4/H.264 compliant file with CIF resolution (352x288) was created. The frame rate was maintained constant at 25 frames per second (fps) during the encoding process of the test signals.

Each encoded video clip was then used as input in the SSIM estimation algorithm. From the resulting SSIM vs. time graph (like the one in Figure 2), the MPQoS value (denoted as <PQoS>$_{SSIM}$ for the rest of the paper) of each clip was calculated. This experimental procedure was repeated for each video clip in CIF resolution. The results of these experiments are depicted in Figure 3.

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

1. The minimum bit rate of the lowest <PQoS>$_{SSIM}$ value depends on the spatiotemporal activity level of the video clip.
2. The variation of the <PQoS>$_{SSIM}$ vs. bit rate is an increasing function, but non linear.
3. The quality improvement of an encoded video clip is not significant for bit rates higher than a specific threshold. This threshold depends on the spatiotemporal activity of the video content.

Moreover, each <PQoS>$_{SSIM}$ vs. bit rate curve can be successfully described by a logarithmic function of the general form

$$< PQoS >_{SSIM} = C_1 \ln(BitRate) + C_2$$

where C_1 and C_2 are constants strongly related to the spatial and temporal activity level of the content. Table 1 depicts the corresponding logarithmic functions for the test signals of Figure 3 along with their R^2 factor, which denotes the fitting efficiency of the theoretical logarithmic curve to the experimental one.

TABLE 1. FITTING PARAMETERS AND R^2 FOR DIFFERENT VIDEO

Test Signal	Logarithmic Function	R^2 factor
Mobile	0.1295ln(x)+0.1274	0.9759
Imax	0.0563ln(x)+0.6411	0.9514
M.I. 3	0.0668ln(x)+0.5747	0.9191
Da Vinci Code	0.0474ln(x)+0.6974	0.8833
Warren	0.0738ln(x)+0.5210	0.9528
Nasa	0.0950ln(x)+0.3892	0.9595
BBC – Africa	0.1098ln(x)+0.2702	0.9875
Superman	0.0282ln(x)+0.8167	0.8859

Based on the aforementioned analysis, we can describe the derived <PQoS>$_{SSIM}$ 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, where each element F_n is a triplet, consisting the <PQoS>$_{SSIM}$ at a specific bit rate and the constants C_1, C_2, which are derived by the analytical logarithmic expression of Table 1:

$$C_{S-T} = \{m : (\frac{1}{N}\sum_{i=1}^{N} SSIM(f_i), C_1, C_2)_n = F_n, n \in [BitRate_{\min}, BitRate_{\max}]\}$$

where
- *SSIM( )* is the function that calculates the perceived quality of each frame according to the *SSIM* metric
- *N* the total number of frames f_i *that* comprises the movie *m*

Thus, deriving the sets C_n for various contents, ranging from static to very high Spatial and Temporal (S-T) ones, a reference hyper set RS, containing a range of C_{S-T} sets for representative spatiotemporal levels can be deduced:

$$RS = \{C_{S-T_{Low}}, ..., C_{S-T_{High}}\}$$

Hence, considering an unknown video clip, which is uncompressed and the service provider wants to predict its corresponding C_{S-T} set that better describes its perceived quality vs. bit rate curve before the encoding process, then, it is defined for all the sets C_{S-T} the Absolute Difference Value (ADV) between the first C_{S-T} triplet element (i.e. the $<PQoS>_{SSIM}$ at a specific encoding $BitRate_n$) and the experimental measurement of the mean *SSIM* for the test signal at the same $BitRate_n$, for which all the reference sets C_{S-T} have been derived, utilizing the logarithmic equations of Table 1:

$$ADV = | F_{BitRate_n} : (\frac{1}{N}\sum_{i=1}^{N} SSIM(f_i)) - F'_{BitRate_n} : (\sum_{i=1}^{N} SSIM(f'_i)) |$$

Due to the fact that the additive property is valid, it is concluded that when the *ADV* between the average *SSIM* of the reference signal $F_{BitRate_i}$ and experimental signal $F'_{BitRate_i}$ is minimum, then the set C_{S-T}, which contains the triplet element that minimizes the *ADV*, describes better the $<PQoS>_{SSIM}$ vs. Bit Rate curve of the specific video. Thus, we have successfully approximated the PQoS vs. Bit rate curve of the specific video with actual cost only one test encoding and assessment at $BitRate_n$. Then the service provider can predict analytically through the estimated logarithmic expression all the bit rates that satisfy specific perceived quality levels, without requiring any other encoding tests. Thus, one only estimation of the $<PQoS>_{SSIM}$ at a specific encoding bit rate is adequate for the accurate determination of the $<PQoS>_{SSIM}$ vs. Bit Rate curve for a given signal/content.

Upon this presentation of the model that predicts the $<PQoS>_{SSIM}$ level of a signal during the encoding process, in the next section, it is examined the case of the quality degradation during the transmission process of the encoded video.

4. Modeling Packet Loss Impact on Video Quality

In this section, we discuss the impact of the packet loss during the transmission of a video over a lossy transmission channel on the percentage of the successfully decoded frames and afterwards we quantify the quality degradation at the end-user side. Due to the fact that the frames in a MPEG video sequence are interdependent, considering a packet loss, the visual distortion caused by a packet loss will be not limited only to the frame, to which the specific lost packet belongs to. On the contrary, spatial error propagation will take place, affecting all the frames that are dependent on the specific frame that the loss occurred. Thus, in order to calculate the error propagation due to a packet loss, the interdependencies of the coded frames must be taken under consideration.

At the user-side, the PQoS degradation induced by a packet loss depends on the error concealment strategy implemented by the decoder. A typical concealment strategy is the zero-motion method, in which a lost macroblock (or frame) is concealed by retaining the macroblock(s) located in the same spatial location of the previously successfully decoded frame. Thus, when packet loss occurs depending on the decoder structure there are two possible case scenarios: i. The decoder attempts to reconstruct the flawed video frames (with or without error concealment) causing the frames to exhibit spatial errors (block distortions and block errors), ii. The decoder completely discards a corrupted video frame and repeats the previous flawless frame until a new decoded frame (without errors) is available.

In this context, this section focuses on the second case scenario, while the first case will be considered into future work, and proposes a mathematical framework that models the percentage of the unsuccessfully decoded frames (i.e. lost frames) based on the frame drops that are caused by packet losses during the streaming process. An analytical model is introduced, which is used to predict the effect of the packet loss distribution on the delivered video quality.

4.1. The proposed model of packet loss to frame loss

For evaluation purposes of the packet loss impact on the PQoS level of a streaming service, it is adopted an objective evaluation metric, known as Decodable Frame Rate (Q) [34]. The value of Q lies between 0 and 1. The larger the value of Q,

the better the video quality received by the end user, since it means that the majority of the frames have been successfully received and decoded by the user's terminal. Therefore Q is defined as the fraction of successfully decoded frames to the total number of frames sent by a video source and contained in the video sequence:

$$Q = \frac{N_{dec}}{N_{total}}$$

Since the N_{dec} is the summation of the successfully decoded I, P and B frames (i.e. $N_{dec\text{-}I}$, $N_{dec\text{-}P}$, and $N_{dec\text{-}B}$) and the N_{total} is the summation of I, P and B frames (i.e. $N_{total\text{-}I}$, $N_{total\text{-}P}$, and $N_{total\text{-}B}$), the Q can be defined as:

$$Q = \frac{(N_{dec\text{-}I} + N_{dec\text{-}P} + N_{dec\text{-}B})}{(N_{total\text{-}I} + N_{total\text{-}P} + N_{total\text{-}B})}$$

Based on this Q metric, in the next sub-sections it is analytically calculated the expected numbers of theoretically expected successfully decodable frames per type (i.e. I, B, P) based on a structure GOP(12,3), which is a typical selection in MPEG-coded video applications due to its optimized trade-off between robustness and compression efficiency. However, the proposed model can be appropriately adapted to any possible GOP structure by reforming the relevant math equation, which will be presented in the next sections.

In the proposed model, the concept of theoretically expected decodable frames is subject to the following hypotheses:

- At the decoder it is not implemented any sophisticated error concealment method.
- The decoding threshold is considered equal to one (DT=1) independently of the content dynamics, meaning that one packet loss causes unsuccessful decoding of the respective frame, which practically equals to frame loss during the decoding process.
- The error propagation affects all the frames that are depended on the lost frame (where the packet loss took place). Considering that the previous hypothesis is valid (i.e. DT=1), the dependent frames are indirectly considered to fail during the decoding procedure.
- The packet loss rate is considered constant and uniform during the application period of the model.

Based on these hypotheses, it is clear that the proposed approach of modelling packet loss impact on the theoretically expected decodable frames of a video service follows a relative approach, where it is researched the degradation caused by the transmission packet loss ratio in relevance to the initial quality of the encoded video content.

A case that is not considered in the above hypotheses is the packet loss to happen on the control packets (i.e. ACK packets). Because he size of data packet (500~1000 Bytes) is much larger than the respective control packet (typically around 30 Bytes), the probability over a uniform packet loss ratio to lose a control packet in comparison to lose a data-packet is very low. So, we can practically ignore the loss effect of control packets and assume that all the packet-loss comes from data packets. Moreover, considering even that this not probable case of losing a control packet may occur, the fact that any video streamer implementation will have different reaction to this, makes the study of this case of low interest since it is not widely applicable but very dependent on the streamer implementation. So without loss of generality, this special codec-dependent case is not addressed by the paper.

Following this explanatory section, the proposed model is presented in the next sub-sections, considering constant packet loss ratio p for the whole service duration. In the appendix of the paper, for readability purposes is presented the notation explanation of all the used symbols.

Hereby is presented an analytical approach to the calculation of the expected number of successfully decodable frames, in order to analytically define the Q metric for GOP(12,3).

1) The expected number of successfully decodable I frames ($N_{dec\text{-}I}$)

In a GOP of an encoded sequence, the I frame (i.e. the first frame of the GOP) is successfully decodable only if all the packets that belong to the specific frame have been successfully received. Considering that $(1\text{-}p)$ is the probability for one packet to be successfully received, given a packet loss rate equal to p and taking under consideration that the I frame consists approximately C_I packets then the probability the first I frame of a sequence to be successfully decoded is

$$S(I) = (1 - p)^{C_I}$$

Which is the mathematical representation of the probability that all the packets carrying the data of the first I frame to have been successfully received at end-user side.

Consequently, the expected number of succesfully decodable I frames for the whole video sequence, considering that the total number of I frames is N_{GOP} (i.e. one I frame in each GOP) then the expression:

$$N_{dec\text{-}I} = (1 - p)^{C_I} * N_{GOP}$$

Represents the expected number of successfully decoded I frames of a streamed video sequence, which consists N_{GOP} GOPs, over a transmission network with *p* packet loss rate.

2) The expected number of successfully decodable P frames (N_{dec-P})

In a GOP, P frames can be successfully decoded only if the preceding I (only for the case of the first P frame in a GOP) or P frames (for all the rest P frames in a GOP) have been successfully decoded (see fig. 1) and all the packets that belong to the P frame under examination have been successfully received. Therefore, taking under consideration the previously proven equation for the successful decoding of the I frame within a GOP and the fact that successful reception of all the packets that belong to the P frame under examination and its preceding I frame is needed for successful decoding, then the probability of the first P frame to be decodable within a GOP is:

$$S(P_1) = (1-p)^{C_I} * (1-p)^{C_P} = (1-p)^{C_I+C_P}$$

Where C_P and C_I represent the average number of packets that construct a P and I frame respectively.

Respectively considering that the successful decoding of the second in turn P frame within a GOP is directly dependent on the successful reception of the packets that belong to it and indirectly dependent on the successful decoding of the previous P frame, then the probability of the second P frame within a GOP to be successfully decodable is:

$$S(P_2) = S(P_1) * (1-p)^{C_P} \Rightarrow$$

$$S(P_2) = (1-p)^{C_I} * (1-p)^{C_P} * (1-p)^{C_P} = (1-p)^{C_I+2C_P}$$

Extending this formulation to the rest P frames of a GOP, considering that there are totally N_p P frames in a GOP, and depending on their position, the probability of the 1st, 2nd, …, Np in turn P frame to be successfully decodable is provided by the following equations:

$$S(P_1) = (1-p)^{C_I} * (1-p)^{C_P} = (1-p)^{C_I+C_P}$$

$$S(P_2) = (1-p)^{C_I} * (1-p)^{C_P} * (1-p)^{C_P} = (1-p)^{C_I+2C_P}$$

……

$$S(P_{N_P}) = (1-p)^{C_I} * (1-p)^{N_P*C_P} = (1-p)^{C_I+N_P*C_P}$$

Thus, according to the above equations, the expected number of successfully decoded P frames within a GOP is provided by the equation

$$N_{dec-P} = (1-p)^{C_I} * \sum_{j=1}^{N_P} (1-p)^{jC_P}$$

Considering that the whole sequence contains N_{GOP} GOPs, then the total expected number of successfully decodable P frames for the whole video is given by the following expression:

$$N_{dec-P} = (1-p)^{C_I} * \sum_{j=1}^{N_P} (1-p)^{jC_P} * N_{GOP}$$

3) The expected number of successfully decodable B frames (N_{dec-B})

Within a GOP, B frames are successfully decodable only if the preceding and succeeding reference I or P frames are both decodable and all the respective packets that consists the specific B frame have been successfully received. Considering that consecutive B frames throughout the GOP structure have the same inter-coding dependencies on the preceding or succeeding I or P frames, we examine the consecutive B frames as a B frame group, except for the last B frame in a GOP, which is dependent only on the preceding P and succeeding I frame (making it indirectly dependent on two successive I frames).

Therefore, in a GOP the probability the first B frame/group to be successfully decodable depends on the successful decoding of the preceding I frame (i.e. $(1-p)^{C_I}$), the successful decoding of the succeeding P frame (i.e. $(1-p)^{C_P}$) and the successful reception of all the packets that consist the B frame/group under examination (i.e. $(1-p)^{C_B}$). Therefore the probability the first group of B frames in a GOP to be successfully decodable is described by the following expression:

$$S(B_1) = (1-p)^{C_I} * (1-p)^{C_P} * (1-p)^{C_B}$$

Where it has been taken under consideration that a I/P/B frame consists approximately C_I / C_P /C_B packets.

Similarly, considering that the second group of B frames depends on exactly the same I/P packets as the first one plus the

packets that consist the succeeding P frame, then the probability for successful decoding of the second B frame within a GOP is given by

$$S(B_2) = (1-p)^{C_I} * (1-p)^{2C_P} * (1-p)^{C_B}$$

It must be noted that the second group of B frames is independent on the successful reception of the packets that consist the first B group, since it is not related to it. Therefore it only depends on its own B packets and for this reason the term $(1-p)^{C_B}$ remains constant. Continuing the same process to the rest B frames of a GOP(N,M), the following expressions are derived that depict the probabilities each B frame in a GOP to be successfully decoded:

$$S(B_1) = (1-p)^{C_I} * (1-p)^{C_P} * (1-p)^{C_B}$$

$$S(B_2) = (1-p)^{C_I} * (1-p)^{2C_P} * (1-p)^{C_B}$$

$$\ldots\ldots .$$

$$S\left(B_{\frac{N}{M}-1}\right) = (1-p)^{C_I} * (1-p)^{\left(\frac{N}{M}-1\right)*C_P} * (1-p)^{C_B}$$

$$S\left(B_{\frac{N}{M}}\right) = (1-p)^{2C_I} * (1-p)^{\left(\frac{N}{M}-1\right)*C_P} * (1-p)^{C_B}$$

The last two expression refer to the penultimate and last B frame/group of a GOP structure, considering that the last B frame is also dependent on the I frame of the next GOP (for this reason the exponential $2C_I$ appears). According to MPEG structure that was presented on Section 2.4, within a GOP(N,M) there are M-1 subsequent B frames that in the above analysis have been considered as one B frame/group.

In order to calculate the total number of expected successfully decodable B frames in a video sequence that consists N_{GOP} GOPs, we sum all the respective probabilities of the B frames/groups, which have been calculated above for one GOP. Afterwards we multiply the result of the one GOP with (M-1) in order to measure all the B frames in the GOP and N_{GOP} in order to consider all the GOPs that formulate the sequence:

$$N_{dec\text{-}B} = (M-1)*\sum_{j=1}^{\frac{N}{M}} S(B_j) * N_{GOP}$$

Substituting the relative probabilities that have been already calculated, the expected number of successfully decodable B frames for the whole video sequence can be described as:

$$N_{dec\text{-}B} = (M-1)*\sum_{j=1}^{\frac{N}{M}} S(B_j) * N_{GOP}$$

$$= \left[(M-1) * (1-p)^{C_I} * \sum_{j=1}^{N_P} (1-p)^{jC_P} * (1-p)^{C_B} + (M-1) * (1-p)^{2C_I} * (1-p)^{N_P C_P} * (1-p)^{C_B} \right] * N_{GOP}$$

$$= \left[(1-p)^{C_I + N_P C_P} + \sum_{j=1}^{N_P} (1-p)^{jC_P} \right] * (M-1)*(1-p)^{C_I + C_B} * N_{GOP}$$

4) Expected Decodable Frame Rate Q

Based on the aforementioned proposed model of successfully decodable frames for each frame type within a MPEG video sequence, the Q metric becomes:

$$Q = \frac{N_{dec}}{(N_{total-I} + N_{total-P} + N_{total-B})} = \frac{N_{dec-I} + N_{dec-p} + N_{dec-B}}{(N_{total-I} + N_{total-P} + N_{total-B})} \Rightarrow$$

$$Q = \frac{(1-p)^{C_I} * N_{GOP} + (1-p)^{C_I} * \sum_{j=1}^{N_P} (1-p)^{jC_P} * N_{GOP} + \left[(1-p)^{C_I + N_P C_P} + \sum_{j=1}^{N_P} (1-p)^{jC_P} \right] * (M-1) * (1-p)^{C_I + C_B} * N_{GOP}}{(N_{total-I} + N_{total-P} + N_{total-B})}$$

According to this formulization, for typical parameterization the respective expected decodable frame rate for various packet sizes follows the form that is depicted on Figure 4. It must be noted that since the proposed model considers that a packet loss causes directly or indirectly undecodable frames, the model is independent on the spatiotemporal dynamics of the content and focuses on the frame loss of the test sequence.

Another significant notice it that the maximum value of packet loss under test is 0.1 (i.e. 10%) since at this packet loss rate approximately the 90% of the video signal has become undecodable. So, greater values do not add practical value to the paper. Therefore, for the rest of the paper, the experimental measurements will be focused on this range of packet loss ratio.

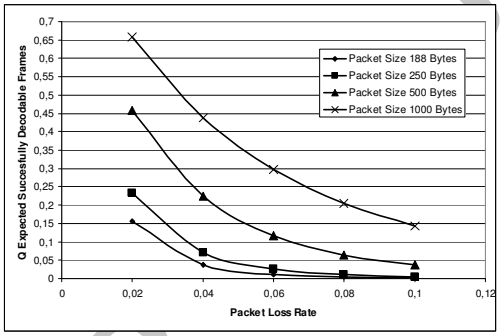

Fig. 4. Expected Decodable Frame Rate Q vs. Packet Loss Rate for various packet sizes

Hence, considering a transmission channel with uniform and constant packet loss ratio *p*, the respective Q rate of successfully decodable frames (i.e. frames without any direct or indirect packet loss) can be analytically predicted.

According to this formulization, for typical parameterization the respective expected decodable frame rate for various packet sizes follows the form that is depicted on Figure 4. It must be noted that since the proposed model considers that a packet loss causes directly or indirectly undecodable frames, the model is independent on the spatiotemporal dynamics of the content and focuses on the frame loss of the test sequence. Another significant notice it that the maximum value of packet loss under test is 0.1 (i.e. 10%) since at this packet loss rate approximately the 90% of the video signal has become undecodable. So, greater values do not add practical value to the paper. Therefore, for the rest of the paper, the experimental measurements will be focused on this range of packet loss ratio.

5. Experimental Validation and Calibration of the Proposed Prediction Models

5.1. Validation of the Encoding Quality Prediction Model

The proposed model of Section 3, for predicting the encoding bit rate that satisfies specific perceptual quality level was tested on a set of real captured video clips, containing content with duration spanning from 10 seconds up to 10 minutes. These video clips were captured in DV PAL format from common TV programs and then transcoded to MPEG-4/H.264. Applying the proposed model the predicted video quality was calculated for various bit rate values and then it was compared to the actually estimated one, which provided the data for drawing the relative performance plot that is depicted on Figure 5. According to this plot, the proposed model provides adequately accurate results by predicting successfully the respective video quality of short in duration video clips.

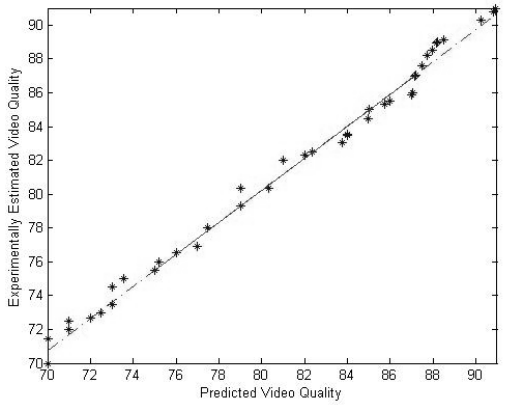

Fig. 5. Objective-estimated vs. predicted video quality

Moreover, the proposed prediction model was also validated for a set of media clips, which were captured from common television programs in DV PAL format and encoded at CIF resolution. The video clips had exclusively specific homogeneous content (i.e. talk show, football, swimming, speech etc.) with duration spanning from 15 seconds up to 60 seconds for the same content type.

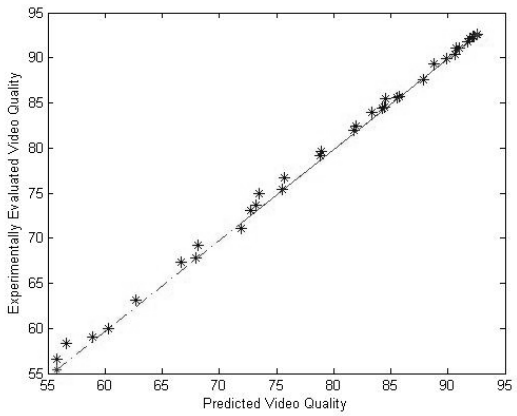

Fig. 6 Objective-estimated vs. predicted video quality for various in duration homogeneous contents

Applying the proposed prediction model and following the same validation procedure, the respective performance plot was derived and is depicted on Figure 6, showing satisfactory behavior of the model for homogeneous contents, independently of their short duration. This spatiotemporal homogeneity of the content can be met within short periods/shots of longer video sequences.

Therefore, the validity of the proposed prediction model for the encoding of the video sequences at various perceptual quality levels has been experimentally proved.

5.2. Validation and Calibration of the Proposed Frame Loss Prediction Model

1) Validation of the Proposed Frame Loss Prediction Model

The proposed model of packet loss impact on the percentage of successfully decodable frames of the transmitted video has been built based on some hypotheses, which have been stated in Section 4. Among these hypotheses, it has been considered that the packet rate remains constant and follows the uniform distribution, which means that the packet losses follow a specific periodic pattern like the one of Figure 7. Due to the uniform distribution of the packet losses and the hypothesis that the DT is equal to 1, the uniform packet loss scheme provides the theoretically worst case scenario of a specific packet loss rate impact on the expected number of successfully decodable frames.

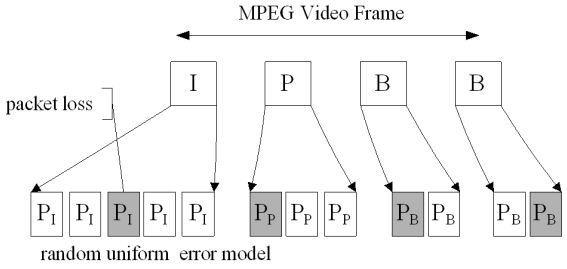

Fig. 7. Uniform Packet Loss Scheme

Therefore, in order to validate initially the accuracy and efficiency of the proposed model that maps the packet loss to frame loss considering DT=1.0, experimental measurements of the frame loss of MPEG sequences over a uniform packet loss scheme were performed.

The experiments were performed on NS-2 simulation environment, where three video traces were utilized namely "Fitzek", "Aladdin" and "Jurassic". These traces were encoded using the MPEG-4/H.264 format and GOP(12,3) and were specifically selected because they are representative of diverse spatiotemporal content. This diversity leads to different video statistics and properties, providing a wide range of video shot categories for the validation purposes of the proposed model. More specifically, Fitzek signal is characterized by short duration and high C_I values, while Aladdin and Jurassic signals are longer in duration videos with higher C_P and C_B values respectively. Table 2 contains the test signal statistics that were deduced for MPEG-4/H.264 encoding and GOP structure (12,3).

TABLE 2
STATISTICS OF TEST SIGNALS

Video File	Fitzek	Aladdin	Jurassic
Total Number of frames	67498	89998	89998
I frames	5625	7500	7500
P frames	16875	22500	22500
B frames	44998	59998	59998

TABLE 3
PACKETIZATION STATISTICS OF TEST SIGNALS

Packet Size	Video File	Fitzek	Aladdin	Jurassic
188/250 Bytes	Total Packets	568007	828604	1424102
	Packet I	207469	147600	219190
	Packet P	117676	244538	412709
	Packet B	242862	436466	792203
	C_I	36.88	19.68	29.23
	C_P	6.97	10.87	18.34
	C_B	5.39	7.27	13.20
500 Bytes	Total Packets	300831	436612	734785
	Packet I	105399	75651	111471
	Packet P	62700	127957	212016
	Packet B	132732	233004	411298
	C_I	18.74	10.09	14.86
	C_P	3.72	5.69	9.42
	C_B	2.95	3.89	6.86
1000 Bytes	Total Packets	173543	240723	389834
	Packet I	54423	39718	57632
	Packet P	35683	69459	111667
	Packet B	83437	131546	220535
	C_I	9.68	5.29	7.68
	C_P	2.11	3.09	4.96

	C_B	1.85	2.19	3.68

The validation process was performed for various packet sizes ranging from 188 bytes that correspond to Digital Video Broadcasting (DVB) services to 1000 bytes that is the typical size for video streaming application like Video On Demand (VoD) or IPTV. The statistics of the test signals for representative packet sizes are depicted on Table 3, where the distribution and concentration of the packets to the various frame types are displayed.

During the validation process, the test signals were streamed under the NS-2 simulation environment of a uniform packet loss scheme for the experimental estimation of their successfully decoded (i.e. flawless) frames. The experimentally derived results were compared to the corresponding predicted ones that were provided by the application of the proposed model. The two sets of experimentally measured and theoretically predicted results were used to form the respective performance plot that is depicted on figure 8, where it can be observed that the proposed model provides accurate prediction of the expected ratio of decodable frames over a uniformly packet loss scheme. The fact that it appears high density of experimental points in the beginning of the axes shows that for high packet loss ratios, the respective experimental/measured decodable number of frames is low, because the decoding process cannot be performed due to the high loss of data.

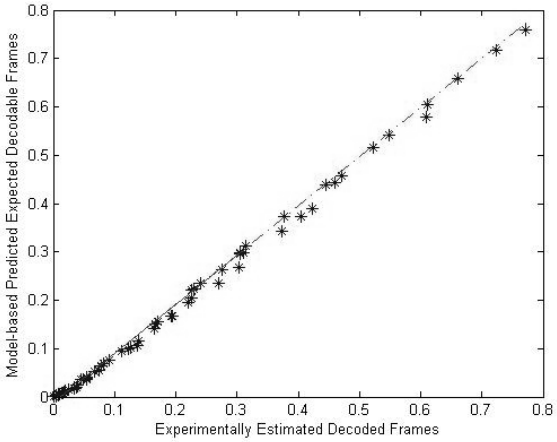

Fig. 8. Objective-estimated vs. predicted Decodable frames for uniform packet loss scheme

However, under real network transmission conditions, the packet loss scheme does not follow the uniform distribution, which is an unrealistic and theoretical case only. For this reason, the proposed model was also tested under bursty-based packet loss distribution scheme, which causes concentration of the lost packets under specific frames, similar to the case that it is depicted on Figure 9.

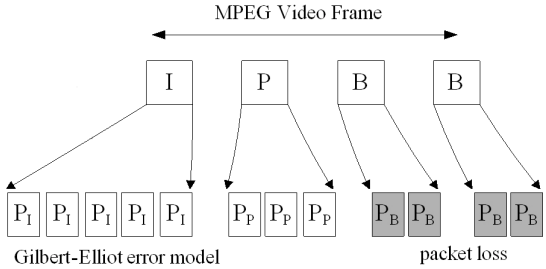

Fig. 9. Gilbert Elliot Packet Loss Scheme

For the validation purposes of the proposed model under bursty packet losses, the Gilbert Elliot (GE) model was selected since it has been deduced that provides efficient simulation conditions [35] and provides for the same packet loss rate of the uniform model, the packet losses grouped in bursts, approximating by this way the behavior of real error-prone transmission channels.

The experiments for the GE packet loss scheme were performed again on NS-2, using the same configuration as in the uniform case and the same test signals of Table 3. Following the same validation procedure as previously, the successfully decoded frames of the test signals were experimentally calculated. Afterwards, the corresponding expected number of decodable frames was also predicted by the application of the proposed model and the results were compared to the

experimental ones, in order to derive the respective performance plot, which is depicted on Figure 10. From the derived results, it can be observed that the proposed model does not predicted accurately the frame losses for burtsy packet loss environments and more specifically it over-estimates the impact of the packet loss on the relative frame loss, especially for the case of low/medium packet loss rates. This outcome was expected since it has been already discussed that the hypotheses of DT=1 and uniform packet loss scheme, on which the proposed model has been built, corresponds to the theoretically worst case scenario, providing the lowest threshold of frame loss.

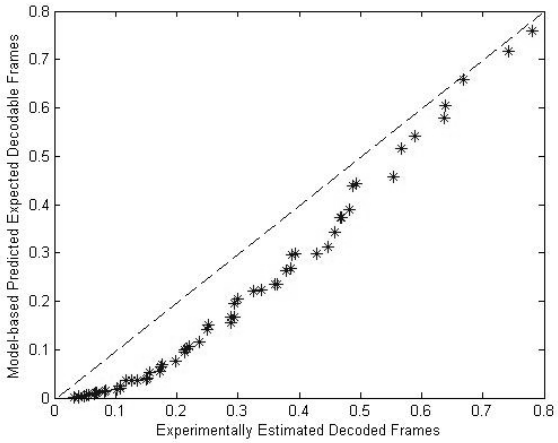

Fig. 10. Objective-estimated vs. predicted Decodable frames for Gilbert Elliot loss scheme

This overestimation of the packet loss impact by the proposed model can be also observed in Figure 11, where the experimentally derived curves for uniform and GE packet schemes are depicted in conjunction with the respective theoretically derived one from the proposed model.

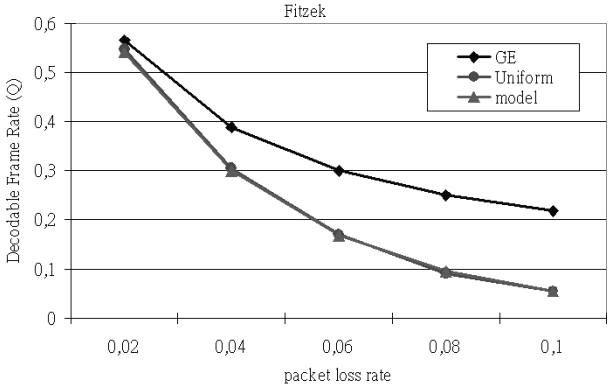

Fig. 11. Experimentally derived Successfully Decodable Frame Rate Q for bursty (Gilbert Elliot –GE) and Uniform packet loss scheme compared to the predicted one

Figure 11 depicts the case of the test signal Fitzek with packet size 500 Bytes as a representative example of the efficiency of the proposed model in comparison to the experimentally derived results of uniform and Gilbert Elliot packet loss schemes. As it can be observed, the proposed model provides excellent match between the uniform and the predicted case, which correspond also to the theoretically worst case. But for the case of bursty packet-loss scheme, the relative experimental curve appears an offset from the respective predicted one. Therefore, for better match between the predicted values of the proposed model and the experimentally measured under bursty loss conditions, which correspond to a realistic condition of a network, a calibration of the proposed model is required by the application of an appropriate offset multiplier.

2) Calibration of the Proposed Frame Loss Prediction Model

The results from the study of bursty schemes of the previous sub-section showed that the theoretically expected decodable frames for the worst case of the uniform packet loss scheme must be multiplied by an offset parameter in order to be more accurate and emulate the behavior of the Gilbert Elliot packet loss schemes. The offset multiplier for this purpose has been experimentally defined by performing repeating test-matches and calculated to be equal to the following expression:

$$Offset\ Multipier = \begin{cases} \frac{1}{-3.9204p + 1.0315} + \frac{0.05}{Q}, & 0.01 < p < 0.05 \\ 1, & 0.05 < p < 0.1 \end{cases}$$

Where

$$Q = \frac{N_{dec}}{(N_{total-I} + N_{total-P} + N_{total-B})} = \frac{N_{dec-I} + N_{dec-p} + N_{dec-B}}{(N_{total-I} + N_{total-P} + N_{total-B})} \Rightarrow$$

$$Q = \frac{(1-p)^{C_I} * N_{GOP} + (1-p)^{C_I} * \sum_{j=1}^{N_p} (1-p)^{jC_p} * N_{GOP} + \left[(1-p)^{C_I + N_p C_p} + \sum_{j=1}^{N_p} (1-p)^{jC_p} \right] * (M-1) * (1-p)^{C_I + C_B} * N_{GOP}}{(N_{total-I} + N_{total-P} + N_{total-B})}$$

Therefore, considering a packet loss ratio equal to p, the Calibrated Predicted Decodable Frames (CPDF) for bursty and non-uniform packet loss schemes is provided by the following formula:

$$CPDF = \begin{cases} \frac{Q}{-3.9204p + 1.0315} + 0.05, & 0.01 < p < 0.05 \\ Q, & 0.05 < p < 0.1 \end{cases}$$

Using again the NS-2 simulation environment, the experiment for the case of bursty packet loss scheme was repeated and the proposed calibrated model compared to experimentally measured values of successfully decoded frames. From this comparison the performance plot of Figure 12 was derived and the respective Fig. 13 (which can be compared to the respective Fig. 11).

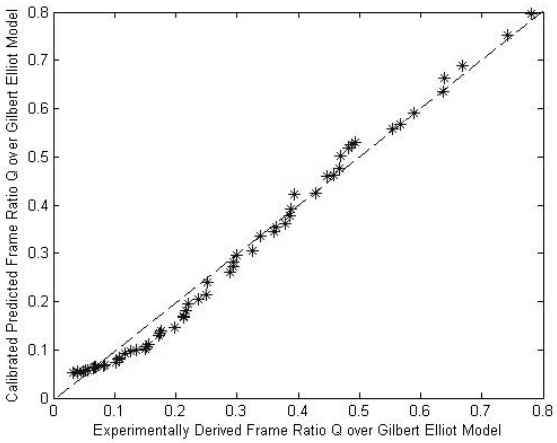

Fig. 12. Predicted by the calibrated model vs. objective Decoded frames for G-E packet loss scheme

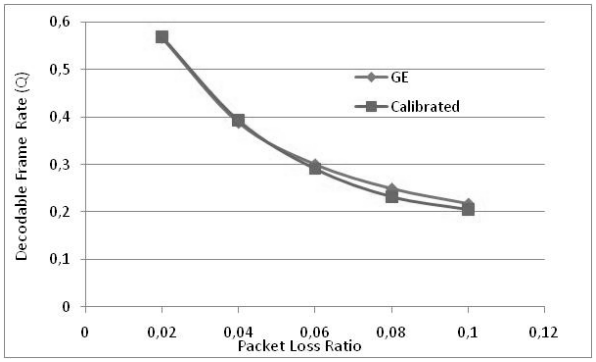

Fig. 13. Experimentally derived Successfully Decodable Frame Rate Q for bursty (Gilbert Elliot –GE) packet loss scheme compared to the calibrated model

Comparing the performance plots of the calibrated and non calibrated model, which are depicted on Fig. 12 and 10

respectively, it can be derived that the proposed offset multiplier provides satisfactory mapping of the predicted number of successfully decodable frames to the respective experimentally derived. Therefore, the proposed model has been successfully validated for the cases of both uniform and bursty packet loss schemes as well, providing satisfactory approximations and predicted measurements.

6. The proposed end-to-end framework

6.1. Mapping of CPDF to MOS

Based on the aforementioned proposed theoretical models of video quality prediction at a pre-encoding state and packet loss-to-frame loss, this section proposes an end-to-end video quality prediction framework of MPEG-based signals, which is based on the combination and exploitation of the two proposed models.

More specifically, the first model will be exploited as a metric for predicting the video quality degradation that is caused by the encoding process of the uncompressed video signal. Considering an MPEG-based compression, the degradation is primarily dependent on the selected encoding bit rate (i.e. the deduced perceptual quality will be lower in comparison to the initial content).

Therefore, the $<PQoS>_{SSIM}$ derived from the first proposed model of Section 3 will be used as a degradation multiplier to the initial quality level of the uncompressed content, which is considered as perceptually excellent (i.e. equal to 100 in the perceptual scale), indicating the relative degradation of the encoded content due to the MPEG compression.

Considering the degradation that may be induced due to transmission problems during the streaming of the service, the proposed model of Section 4 will be used, which maps the packet loss rate of the transmission channel to the expected number of decodable frames (or equivalently to the frame loss percentage). However, the proposed model maps the impact of the packet loss ratio to the respective frame loss, but the perceptual degradation caused by this frame loss, it is not estimated.

In the proposed model, we have considered the zero-motion concealment method for recovering the frame losses. This means that the transmitted degraded signal suffers from motion discontinuities, because the last successfully decoded frame is retained till the next successfully decoded one. Towards this, based on the literature, a relative mapping between the frame losses of an MPEG video and the respective perceptual degradation, has been already proposed in [36]. More specifically, due to the fact that the video signal may suffer several degradations at any stage of the transmission chain, resulting in severe motion discontinuities in video streaming, the end-user may perceive a jerky motion and instantaneous fluidity breaks. Packet losses in the transmitted networks are the main cause of this perceived jerkiness/break. Packet losses cause a sporadic frame discarding at the decoding process because of the limited buffering time, making the last successfully decoded frame to be displayed in the position of the dropped frame until the next successfully decoded frame follows. Therefore, the end-user will perceive a frozen playback followed by an abrupt displacement of the depicted objects.

In [36] the term temporal discontinuity is used as a perceptual synonym of the frozen playback. More specifically, the perceptual impact of a single burst of dropped frames on the perceptual degradation has been estimated for short video sequences of 10 sec duration, a period that is adequate to avoid the forgiveness effect and consider the spatiotemporal content dynamics practically constant.

Moreover, since the PQoS degradation due to frame losses and the respective temporal discontinuity is highly dependent on the spatiotemporal activity of the content, the selection of a short period of monitoring and assessment minimizes this dependence and makes possible the provision of a generic prediction model regardless of the spatiotemporal dynamics of the content. Furthermore, according to the statistical analysis of [36], this short duration minimizes the probabilities to have multiple bursty packet losses within this period, making the study of single packet loss bursts schemes statistically quite accurate and satisfactory.

Based on the model, which has been proposed, presented, tested and subjectively evaluated in [36] the mapping of the dropped frames to perceptual quality level with regard to the Mean Opinion Scores (MOS) [1] over various spatiotemporal contents of 10 sec duration and 25fps temporal rate is analytically described by the following expression:

$$MOS = \begin{cases} 85.8, & x=0 \\ 85.8 - \dfrac{53.03}{1+(562/x)^{1.01}}, & x>0 \end{cases}$$

where x is the discontinuity duration computed over all contents in msec. Therefore, this equation maps the objective metric of the discontinuity durations to the respective PQoS level degradation in terms of the subjective metric of MOS.

The fact that the 85.8 is used as maximum estimated quality is derived from the validation process of the model in [36] and it is based on the fact that statistically there is always a variance in the measured assessments by the viewers, even if they

characterize the sample as excellent. Therefore, the mean excellent value that it is subjectively achieved is around 85.8 out of 100.

In this point the authors would like to point out that the specific representation of the MOS metric in the range of 0-100 (and not in the usual one from 1 to 5) is used as a more detailed one in comparison to 1-5. This elaboration of the five scale of the MOS, it is a common tactic in the research field of video quality evaluation and a direct mapping is performed every 20 units to the respective 1-5 MOS unit (i.e. 0-20:1, 21-40:2, 41-60:3, etc.) [12]. So, the selection of this scale is not a proprietary approach for the need of this paper, but a common tactic followed for more accurate results.

Based on the 10sec duration restriction of the described model in [36], the estimation of the dropped frames due to the monitored packet loss ratio, which is the complimentary of the CPDF (i.e. the $CPDF'_{10\text{sec}}$) must be performed over a period of 10 seconds. Therefore:

$$CPDF'_{10\text{sec}} = 1 - \left(\frac{\text{Q}}{\text{-3.9204}p + 1.0315} + 0.05\right)$$

Based on this estimation of the frame loss provided by $CPDF'_{10\text{sec}}$ for single bursty packet loss schemes, the derived result will be used as input to the MOS prediction scheme for the final mapping of frame loss to PQoS.

Considering that the duration period of applying the proposed model in [36] to all the test signals is 10 sec, then the variable of the discontinuity duration can be also expressed in terms of duration percentage of the discontinuities over the 10 seconds. Thus, describing the variable x of $MOS = 85.8 - \frac{53.03}{1 + (562/x)^{1.01}}$ as duration percentage of the discontinuities over a period of ten seconds, this means that it can be further described as the percentage of the dropped frames over the total frames of a 10 sec signal. Moreover, this mapping is 1-to-1 without requiring any further sophisticated adoption. So, the variable x can be substituted by the ratio of the dropped frames (i.e. the $CPDF'_{10\text{sec}}$), which has been mathematically modeled and experimentally validated in the previous section. Upon this, the above equation can be further formulated as:

$$MOS = 85.8 - \frac{53.03}{1 + (562/CPDF'_{10\text{sec}})^{1.01}}$$

And substituting the $CPDF'_{10\text{sec}}$ expression, it is derived that the predicted MOS after the degradation caused by the packet losses is two-tailed depending on the packet loss ratio as follows:

- For $0.01 < p < 0.05$, MOS becomes:

$$MOS_{Predicted} = 85.8 - \frac{53.03}{1 + \left(\frac{562}{1 - \left(\frac{\text{Q}}{\text{-3.9204}p + 1.0315} + 0.05\right)}\right)^{1.01}}$$

- For $0.05 < p < 0.1$, MOS becomes:

$$MOS_{Predicted} = 85.8 - \frac{53.03}{1 + \left(\frac{562}{1 - \text{Q}}\right)^{1.01}}$$

Therefore by applying the proposed model over a period of 10 sec at the streaming channel of service, the respective perceptual level can be estimated.

6.2. Correlating the proposed discrete models into an end-to-end quality metric

Combining the two discrete proposed models along with the aforementioned MOS extension of the second one, can be further exploited for providing a prediction of the Expected Delivered Video Quality (EDVQ) level at the end-user. In this direction, the video service provision can be considered as two events: i. the encoding event (A event) and ii. the transmission

event (B Event). Each of these two events can introduce quality degradation to the initial video signal through different mechanisms that have been already discussed and modelled by the proposed models of this paper.
Moreover, both the events are absolutely necessary in the video delivery process, which means that the intersection of A and B cannot be the empty set (i.e. encoding without transmission and transmission without encoding is not possible for an end-to-end video delivery framework). Therefore, the EDVQ can be considered as the intersection of the A and B events and described by the following conditional equation, which maps the EDVQ to the degradation that is caused to the initial video signal by both the A and B events (i.e. the encoding and the transmission process):

$$EDVQ(A \bigcap B) = Degradation(A \bigcap B)$$

Therefore, this relation –based on the fact that the intersection of A and B are not the empty set, can be further analyzed as following:

$$EDVQ(A \bigcap B) = Degradation(A) \bullet Degradation(B \mid A)$$

This relation means that the EDVQ can be dscribed as the product of the degradation that is occurred by the encoding event A and the transmission event B given the encoding event A. Considering this approach, the $Degradation(A)$ parameter can be successfully mapped to the degradation that is caused by the encoding process, and respectively the $Degradation(B \mid A)$ can be mapped to the degradation caused due to packet loss during the video delivery subject to the applied encoding parameters. It must be noted that the impact of the packet loss on the quality degradation is strongly related to the encoding parameters that have been applied on the specific video (i.e. event A), since the mean frame size plays a key role to the packetization scheme of the stream (as it has been already discussed in the respective model of packet loss-to-frame loss of this paper).
Therefore, the above formula can be re-written as follows:

$$EDVQ = (Degradation\_due\_to\_encoding) * (Degradation\_due\_to\_packet\_loss)$$

Substituting the above multipliers with the predicted quality degradation metrics that are provided by the two proposed discrete models of this paper, the expression becomes:

- $Event\_A = Degradation\_due\_to\_encoding = < PQoS >_{SSIM}$

  and

- $Event\_B \mid A = Degradation\_due\_to\_packet\_loss = MOS_{Predicted}$

The proposed prediction model of the $MOS_{Predicted}$ during the transmission phase is obvious that considers the encoding process for providing a forecast, since requires as input statistical information of the encoding process. Thus, it can successfully describe the event A|B of the above equation.

Therefore:

$$EDVQ = (< PQoS >_{SSIM}) * (MOS_{Predicted})$$

$$0 << PQoS >_{SSIM} < 1 \; and \; 0 < MOS_{Predicted} < 100$$

Where the $<PQoS>_{SSIM}$ can be considered as the long term multiplier, since it provides an average prediction of the video quality degradation caused by the encoding process over a video signal and the MOS is the short term multiplier, given that it provides prediction of the video quality degradation that it is caused by the packet losses of the transmission medium every ten seconds.

7. Experimental Demonstration of the proposed end-to-end framework

In order to be able to demonstrate the validity of our framework and verify the proper operation, we applied the proposed framework on a real small scale packet-network testbed, which is able of applying specific network conditions (i.e. packet loss ratio) by Linux-based scripts. The chosen architecture is illustrated in Fig.14, where the testbed is comprised of one autonomous domain, consisting of three PCs operated by Linux OS (kernel 2.6.xx) and a Measurement PC. The ingress PC is the content provider and service generator (i.e. the media streamer), the second PC in the domain is equipped with NistNet [42] for applying specific network conditions and the third PC is the media receiver (i.e. the end-user). For the validation purposes of this paper, specific packet loss pattern was applied (as it is illustrated on Fig 15) by appropriate configuration of NistNet. The measurement PC of the testbed has twofold responsibilities: Firstly was used for validating the accuracy of the loss pattern that NistNet applies and secondly for applying the proposed framework and deriving the experimental graphs that

are depicted on Fig. 15.

The experimental application of the proposed end-to-end framework is divided to two phases: a. The application of the prediction model for estimating the bit rate that corresponds to a specific quality level; b. The application of the model that predicts the impact of the packet loss during the content provision on its initial quality level.

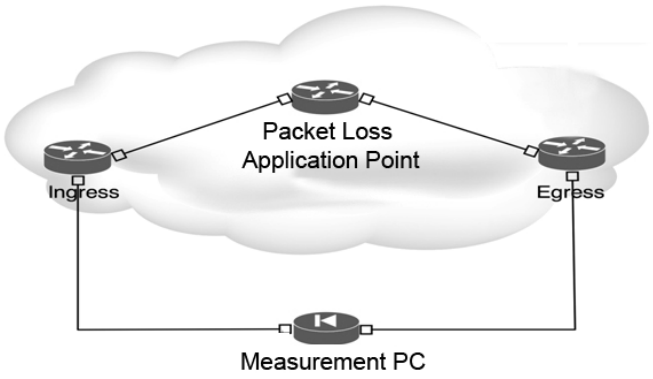

Figure 14. Testbed Architecture

For the needs of this experiment a music video trailer was selected as a test signal, which was initially encoded at MPEG-4/H.264 CIF 100 kbps. During the first phase of the experiment, it is also considered that the Content Provider possesses the reference hyper-set RS , containing the C_{S-T} sets derived from the test signals of Table 1. Thus, the encoded clip @100kbps is used as input to the *SSIM* algorithm and the resulted instant *SSIM* curve is used for the estimation of the <*SSIM*> value, which is estimated equal to 0.8. Afterwards, using this value as input in the ADV equation, it is defined the C_{S-T} that minimizes the ADV and therefore contains the optimal triplet element for the analytical description of the signal under test. More specifically, for the derived <*SSIM*> value, the optimal C_{S-T} set belongs to *BBC Africa* reference clip. Thus, the equation that describes better the variation of the $<PQoS>_{SSIM}$ vs. the bit rate is

$$<PQoS>_{SSIM} = 0.1098\ln(Bit\ Rate)+0.2702$$

Consequently, if the content provider wishes to offer this video clip at the perceptual qualities 0.70, 0.80 and 0.90, then by using the above equation is able to estimate the corresponding bit rates in a pre-encoding process. Table 5 shows the corresponding encoding bit rate values for the specific video clip.

TABLE 5
PREDICTED BIT RATE VALUES FOR SPECIFIC QUALITY LEVELS

$<PQoS>_{SSIM}$	BR (Kbps)
0.7	50.12
0.8	124.60
0.9	309.79

For the purposes of the specific experiment the $<PQoS>_{SSIM}$ =0.9 was selected, which means that the long term of the initial degradation due to encoding in EDVQ is 0.9 (i.e. 10% video quality degradation).

During the second phase, we selected a specific and periodic pattern of packet losses with an average value of 20% packets drop from the video stream. By this selection, the authors want to demonstrate that the proposed framework predicts statistically correctly the stochastic impact of a packet loss on the subsequent frames of a video sequence (i.e. the fact that the same packet loss ratio will diversely result in loss of I, B or P-frame packets, which in turn will cause different error propagation and therefore quality degradation). Afterwards, using the network testbed, the streaming of the video was performed with packet size equal to 1000 Bytes. Then the frame loss was experimentally calculated and then it was compared to the predicted one that it was derived by the proposed model, showing satisfactory match, in a similar way like the presented validated results in section 5c. Afterwards, the respective MOS estimations were estimated based on the proposed adaptation of the model described in [36]. The results of this demonstration are depicted in Figure 15, where the predicted

MOS values are depicted in parallel to the packet loss scheme that was used during this experimental procedure.

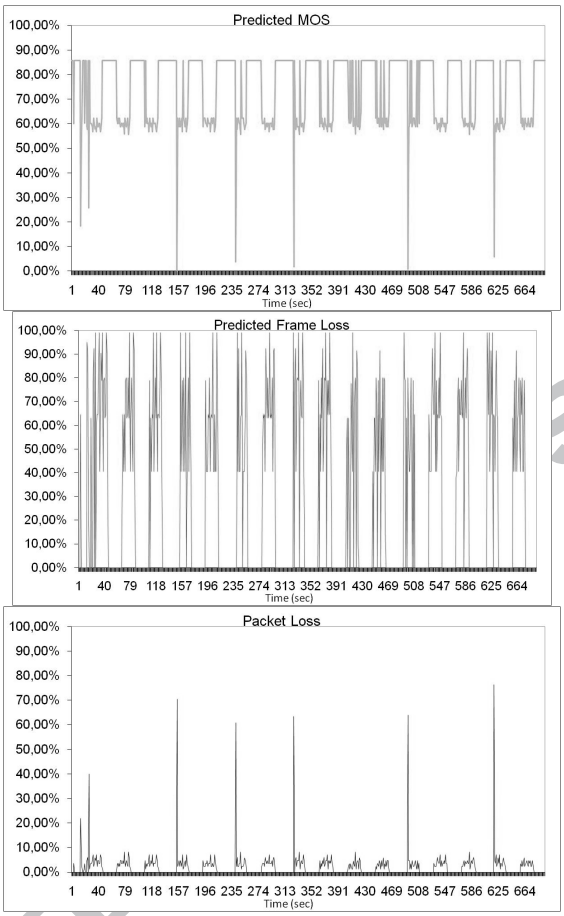

Fig. 15. Predicted MOS for the music video clip

More specifically, in Figure 15 the predicted MOS for the degradation of the perceptual quality has been calculated. For reference purposes the respective frame loss ratio is depicted in parallel to the packet loss scheme. The EDVQ can be finally calculated by the multiplication of the respective predicted MOS values (short term) with the $<PQoS>_{SSIM}$ =0.9 (long term), as they were estimated by the proposed end-to-end video quality prediction framework

As it can be observed, when packet loss spikes occurs (significant instant packet loss) then the predicted MOS value is instantly degraded. Moreover, for the same packet loss scheme, it can be deduced that its impact on the number of successfully decoded frames and the calculated MOS quality varies based on the type of the frames that the packet loss happened. For example see the degradation around the 400 sec of Figure 15 in comparison to the other ones. It can be clearly observed that the respective frame loss pattern is more severe than the rest ones, showing that the specific packet loss occurred on frame types that are more significant in the decoding process. Thus the model deterministically predicts the impact of the packet loss on the delivered video quality, without leaving out the stochastic nature of the quality degradation caused by packet losses during the service provision.

8. Conclusions

This paper presents a theoretical framework for end-to-end video quality prediction for MPEG-based services. The proposed framework encloses two discrete models: i) A model for predicting the video quality of an encoded signal at a pre-encoding stage and ii) A model for mapping packet loss ratio of the transmitting channel to video quality degradation. The efficiency of both discrete models has been experimentally validated, proving by this way the accuracy of the proposed framework, which combines the discrete models into a common end-to-end video quality assessment framework.

The advance of the proposed framework is that it can be applied on any MPEG-based encoded sequences, subject to

specific GOP structure. Moreover, it is also introduced the novel issue of predicting the video quality of an encoded service at a pre-encoding state, which provides new facilities at the content provider side. However, the validity of the proposed framework has been successfully experimentally tested both on uniform and bursty packet loss schemes.

9. ACKONWLEDGEMENT

Part of the work in this paper has been performed within the research framework of FP7 ICT-214751 ADAMANTIUM Project (www.ict-adamantium.eu). The authors would like also to thank the anonymous reviewers for their constructive comments, which help a lot for improving the quality of the paper.

APPENDIX

NOTATIONS USED IN THE PAPER

$N_{total\text{-}I}\ N_{total\text{-}P}\ N_{total\text{-}B}$	The total number of I, P and B frames.
$N_{dec\text{-}I}\ N_{dec\text{-}P}\ N_{dec\text{-}B}$	The number of successfully decoded (i.e. flawless) I, P and B frames.
N_{dec}	The aggregate number of successfully decoded (i.e. flawless) frames at the end-user terminal from the sequence
N_{Total}	The total number of frames in the sequences
N_{GOP}	The total number of GOPs in the sequence.
$C_I\ C_P\ C_B$	The mean number of packets that transport the data of I, P and B frames respectively
p	Packet loss rate
Q	The fraction of successfully decoded frames to the total number of frames of a sequence
$S(I)$	The probability the I frame of a sequence to be successfully decoded
$S(Pn)$	The probability the n^{th} P frame within a GOP to be successfully decoded
$S(Bn)$	The probability the n^{th} B frame within a GOP to be successfully decoded
$N_P,\ N_B$	Total number of P and B frames within a GOP
EDVQ	Expected Delivered Video Quality

REFERENCES

[1] Z. Wang, H.R. Sheikh, A.C. Bovik, Objective video quality assessment, in The Handbook of Video Databases: Design and Applications, B. Furht, O. Marqure, Editors. CRC Press. (2003) 1041-1078.

[2] VQEG. Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment. 2000. Available: http://www.vqeg.org.

[3] Z. Wang, A.C. Bovik, L. Lu, Why is image quality assessment so difficult? Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing. 2002.

[4] U. Engelke, H.-J. Zepernick, Perceptual-based Quality Metrics for Image and Video Services: A Survey, 3rd EuroNGI Conference on Next Generation Internet Networks, Trondheim, Norway, May 2007

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Proc. 13 (4) (2004) 600-612.

[6] Z. Wang, L. Lu, A. C. Bovik, Video quality assessment based on structural distortion measurement, Signal Processing: Image Comm. 19 (2) (2004) 121-132.

[7] M. Ries, C. Crespi, O. Nemethova, M. Rupp, Content Based Video Quality Estimation for H.264/AVC Video Streaming, Proceedings, IEEE Wireless and Communications & Networking Conference, Hong Kong, March, 2007.

[8] E. A. Silva, K. Panetta, S. S. Agaian, Quantifying image similarity using measure of enhancement by entropy, Mobile Multimedia/Image Processing for Military and Security Applications, Proceedings of SPIE. 6579 (2007)

[9] I. P. Gunawan, M. Ghanbari, Reduced-Reference Picture Quality Estimation by Using Local Harmonic Amplitude Information, London Communications Symposium, 2003.

[10] M. Montenovo, A. Perot, M. Carli, P. Cicchetti, A. Neri, Objective evaluation of video services. Proceedings, 2nd Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2006.

[11] S. S. Hemami, M. A. Masry, A scalable video quality metric and applications. Proceedings, 1st Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2005.

[12] O. A. Lotfallah, M. Reisslein, S. Panchanathan, A framework for advanced video traces: Evaluating visual quality for video transmission over lossy networks. EURASIP Journal on Applied Signal Processing. (2006)

[13] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, A. C. Bovik, Quality-Aware Images. IEEE Trans. Image Processing. 16 (6) (2006) 1680-1689.

[14] H. R. Wu, M. Yuen, A generalized block-edge impairment metric for video coding. IEEE Signal Processing Letters. 11 (4) (1997) 317-320.

[15] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahim, A no-reference perceptual blur metric, Proceedings, IEEE Int. Conf. on Image Processing, 2002, pp. 57-60.

[16] J. Caviedes, S. Gurbuz, No-reference sharpness metric based on local edge kurtosis, Proceedings, IEEE Int. Conf. on Image Processing, 2002, pp. 53-56.

[17] A. Cavallaro, S. Winkler, Segmentation-driven perceptual quality metrics, Proceedings, IEEE Int. Conf. on Image Processing, 2004, pp. 3543-3546.

[18] R. R. Pastrana-Vidal, J. C. Gicquel, Automatic quality assessment of video fluidity impairments using a no-reference metric, Proceedings, 2nd Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2006.

[19] M. C. Q. Farias, S. K. Mitra, No-reference video quality metric based on artifact measurements, Proceedings, IEEE Int. Conf. on Image Processing, 2002, pp. 141-144.

[20] X. Marichal, W. Y. Ma, H. J. Zhang, Blur determination in the compressed domain using DCT information, Proceedings, IEEE Int. Conf. on Image Processing, 2002, pp. 386-390.

[21] R. Ferzli, L. J. Karam, J. Caviedes, A robust image sharpness metric based on kurtosis measurement of wavelet coefficients, Proceedings, 1st Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2005.

[22] E.P. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, F. Moschetti, Low bit rate quality assessment based on perceptual characteristics, Proceedings, Int. Conf. on Image Processing , 2003, pp. 182-192.

[23] S. Liu, A. C. Bovik, Effcient dct-domain blind measurement and reduction of blocking artifacts, IEEE Trans. on Circuits and Systems for Video Technology. 12 (12) (2002) 1139-1149.

[24] M. Ries, O. Nemethova, M. Rupp, Reference-free video quality metric for mobile streaming applications, Proceedings, 8th Int. Symp. on DSP and Communication Systems & 4th Workshop on the Internet, Telecommunications and Signal Processing, 2005, pp. 98-103.

[25] L. Lu, Z. Wang, A. C. Bovik, J. Kouloheris, Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video, Proceedings, IEEE International Conference on Multimedia, 2002.

[26] H. Koumaras, A. Kourtis, D. Martakos, Evaluation of Video Quality Based on Objectively Estimated Metric, Journal of Comm. and Netw., Korean Institute of Communications Sciences (KICS).7 (3) (20005) 235-242.

[27] 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. 34 (3) (2007) 355-374

[28] H. Koumaras, E. Pallis, G. Xilouris, A. Kourtis, D. Martakos, J. Lauterjung, Pre-Encoding PQoS Assessment Method for Optimized Resource Utilization, Proceedings, 2nd Inter. Conference on Performance Modeling and Evaluation of Heterogeneous Networks (Het-NeTs04), Ilkley, U. K., 2004.

[29] S. Kanumuri, P. C. Cosman, A.R. Reibman, V.A. Vaishampayan, Modeling Packet-Loss Visibility in MPEG-2 Video, IEEE Trans. Multimedia. 8 (2) (2006) 341-355.

[30] Z. He, H. Xong, Transmission Distortion Analysis for Real-Time Video Encoding and Streaming over Wireless Networks, IEEE Trans. Circuits and Systems for Video Technology. 16 (9) (2006) 1051-1062

[31] J. Mitchell, W. Pennebaker, MPEG Video: Compression Standard. Chapman and Hall, 1996. ISBN 0412087715

[32] C.-H. Lin, C.-H. Ke, C.-K. Shieh, N. Chilamkurti, The Packet Loss Effect on MPEG Video Transmission in Wireless Networks, Proceedings, IEEE 20th International Conference on Advanced Information Networking and Applications (AINA'06), Vienna, Austria, 2006.

[33] NS-2 simulator, http://hpds.ee.ncku.edu.tw/~smallko/ns2/ns2.htm

[34] A. Ziviani, B. E. Wolfinger, J. F. Rezende, O. C. M. B. Duarte, S. Fdida, Joint Adoption of QoS Schemes for MPEG Streams, Multimedia Tools and Applications Journal. 26 (1) (2005) 59-80.

[35] J. P. Ebert, A. Willig, A Gilbert-Elliot Bit Error Model and the Efficient Use in Packet Level Simulation, Technical Report, TKN-99-002, Technical University of Berlin, March 1999.

[36] R. R. Pastrana-Vidal, J. C. Gicquel, C. Colomes, C. Hocine, Sporadic frame dropping impact on quality perception, Proceedings, SPIE Electronic Imaging, Human Vision and Electronic Imaging IX, 2004, pp. 182-193.

[37] L. Liuming, L. Xiaoyuan, Quality Assessing of Video Over a Packet Network, Proceedings, 2nd Workshop on Digital Media and its Application in Museum & Heritage, 2007, pp.365-369

[38] K. Stuhlmuller, N. Farber, M Link, B. Girod, Analysis of Video Transmission over Lossy Channels, IEEE Journal on Selected Areas in Communications. 18 (6) (2000) 1012-1032.
[39] Y-F Ou, Z. Ma, Y. Wang, "A Novel Quality Metric for Compressed Video Considering both Frame Rate and Quantization Artifacts", VPQM, Scottsdale, AZ, 2009
[40] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang; M. Etoh, "Cross-Dimensional Perceptual Quality Assessment for Low Bit-Rate Videos" IEEE Transactions on Multimedia, Volume 10, Issue 7, Nov. 2008 Page(s):1316 – 1324
[41] YUV Video Sequences, http://trace.eas.asu.edu/yuv/index.html
[42] NistNet http://snad.ncsl.nist.gov/nistnet/

Harilaos Koumaras was born in Athens, Greece. He received his BSc degree in Physics in 2002 from the University of Athens, Physics Department, his MSc in Electronic Automation and Information Systems in 2004, being scholar of the non-profit organization Alexander S Onassis, from the University of Athens, Computer Science Department and his PhD in 2007 at Computer Science from the University of Athens, Computer Science Department, having granted the four-year scholarship of National Centre of Scientific Research "Demokritos". He has received twice the Greek State Foundations (IKY) scholarship during the academic years 2000-01 and 2003-04. He has also granted with honors the classical piano and harmony degrees from the classical music department of Attiko Conservatory. Since 2004 he is a principal lecturer at the Business College of Athens (BCA) teaching modules related to Information Technology and Mathematics, Data Networks and Local Area Networks. From 2009, he has been elected as the Head of the Computer Science Department of BCA and Course leader of the respective franchised course of London Metropolitan University. At the same time, since 2003 is a research associate of the Digital Communications Lab at the National Centre of Scientific Research "Demokritos" by participating in numerous EC-funded and national funded projects. His research interests include objective/subjective evaluation of the perceived quality of multimedia services, video quality and picture quality evaluation, video traffic modeling, digital terrestrial television and video compression techniques. Currently, he is the author or co-author of more than 30 scientific papers in international journals, technical books and book chapters, numbering 55 non-self citations. He is an active editorial board member of Telecommunications Systems Journal and he has served many conferences as TPC member and has acted as reviewer for various journals, such as IEEE Transactions on Broadcasting and IEEE Transactions on Image Processing. Dr. Koumaras is a member of IEEE, SPIE and National Geographic Society.

Cheng-Han Lin is currently a Ph.D. candidate studying in the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan. Lin received his MS and BS degree from the Electrical Engineering Department of National Chung Cheng University in 2002 and 2004. His current research interests include wireless MAC protocols, multimedia communications, and QoS network.

Ce-Kuen Shieh is currently a professor teaching in the Department of Electrical Engineering, National Cheng Kung University. He received his PhD, MS, and BS degrees from the Electrical Engineering Department of National Cheng Kung University, Tainan, Taiwan. His current research areas include distributed and parallel processing systems, computer networking, and operating systems.

Anastasios Kourtis received his B.S. degree in Physics in 1978 and his Ph.D. in Telecommunications in 1984, both from the University of Athens. Since 1986, he has been a researcher in the Institute of Informatics and Telecommunications of the National Centre for Scientific Research "Demokritos", currently ranking as Senior Researcher. His current research activities include, digital terrestrial interactive television, broadband wireless networks, Perceived Quality of video services, end to end QoS and real time bandwidth management in satellite communications. He is author or co-author of more than 80 scientific publications in international scientific journals, edited books and conference proceedings. Dr. Kourtis has a leading participation in many European Union funded research projects in the frame of IST/FP5/FP6 (MAMBO, SOQUET, CREDO, WIN, LIAISON,

ENTHRONE). He has also coordinated three European funded Specific Targeted Research Projects (REPOSIT, ATHENA, IMOSAN). Currently, he is the project coordinator of the ICT-214751 EC-funded ADAMANTIUM project.