A Theoretical Framework for End-to-End Video Quality Prediction of MPEG-based Sequences

Harilaos Koumaras, Anastasios Kourtis
Institute of Informatics and Telecommunications
NCSR Demokritos
Athens, Greece
{koumaras, kourtis}@iit.demokritos.gr

Cheng-Han Lin, Ce-Kuen Shieh
High Performance Parallel & Distributed Syst. Laboratory
National Cheng Kung University
Tainan, Taiwan
{jhlin5,shieh}@hpds.ee.ncku.edu.tw

Abstract— **This paper presents a novel theoretical framework for end-to-end video quality prediction of MPEG-based video sequences. 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 QoS-sensitive network parameters (i.e. packet loss) to video quality degradation. The efficiency of both the discrete models is experimentally validated, proving by this way the accuracy of the proposed framework.**

Keywords-Video Quality; Packet Loss;MPEG;H.264

I. INTRODUCTION

Today, the definition of the encoding parameters that satisfy a specific level of video quality is a matter of objective and subjective video quality assessments, each time taking place after the encoding process (post-encoding evaluation). Subjective quality evaluation processes of video streams require large amount of human resources, establishing it as a practically impossible procedure for a service provider, while repetitive use of objective metrics may be required on already encoded sequences for identifying the specified encoding parameters that satisfy a specific level of user satisfaction.

Thus, an open issue in video quality research community is the development of a method, which will be able to predict, according the content dynamics, the encoding parameters that satisfy a specific video quality level at a pre-encoding stage. By applying this technique, the Content Provider will be able to prepare/encode the available video signals quickly and easily at various quality levels, subject to the applied pricing policy and QoS framework.

Once the content provider has specified the appropriate encoding parameters that satisfy a specific level of user satisfaction, then the transmission of the content follows. A major challenge in wireless or congested wired communications is that the delivery channel may be time-varying and error prone. Digital video encoded services, in contrast with traditional network applications (i.e. ftp, telnet and World Wide Web), are highly sensitive to transmission problems (e.g. packet loss, network delay) and require high transmission reliability in order to maintain stream synchronization and initial quality between sender and receiver devices. Especially, in video transmission over wireless environments, each transmitted from one end video packet can be received at the other end either correctly or with errors or get totally lost. In the last two cases, the perceptual outcome is similar, since the decoder usually discards the packet with errors, causing decoding failure, which in turn results in perceived quality degradation.

Thus, another objective of the research community has been the development of methods that will model the transmission distortion of the video quality with the relative network QoS parameters, like packet loss ratio.

By providing a framework, which combines this network-to-video quality model with a video quality prediction method at a pre-encoding stage, the Content Provider will be able to predict the end-to-end delivered video quality, given the current network conditions and the selected encoding parameters. Such an end-to-end perceived QoS framework will not only play an essential role in performance analysis, control and optimization of video communication systems, but it will also contribute to a more efficient resource allocation and management in communication networks.

Towards this, the paper presents, describes and demonstrates a theoretical end-to-end video quality prediction framework for MPEG-based coded sequences. The rest of the paper is organized as follows: Section II discusses related work and Section III presents the proposed video quality prediction model. Afterwards, Section IV analyses the perceptual mapping of QoS sensitive parameters to video quality degradation. Then, Section V demonstrates the end-to-end video quality assessment framework, providing specific examples. Finally, Section VI concludes the paper.

II. RELATED WORK

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

The subjective test methods, proposed by ITU and VQEG, involve an audience of people, who watch a video sequence and score its quality as perceived by them, under specific and controlled watching conditions. Afterwards, the statistical analysis of the collected data is used for the evaluation of the

perceived quality. Objective evaluation methods, on the other hand, can provide PQoS evaluation results faster than subjective ones, but require large amount of machine resources and sophisticated apparatus configurations.

The objective methods can be classified depending on the requirement of the original/undistorted signal in the evaluation process. Thus, we have Full Reference (FR) Methods [1-3], which require the undistorted source video sequence as a reference entity. Reduced Reference Methods, which use only partial extracted structural features from the original signal [4] and No Reference Methods, which do not require any reference video signal in the evaluation process [5]. Finally, the development of methods for predicting the video quality at a pre-encoding stage is still a research challenge, where only few research papers have been published in the field [6-8].

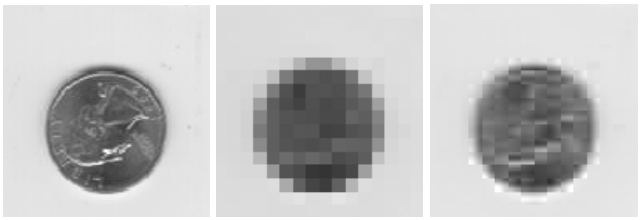
Regarding the mapping of the network QoS sensitive parameters (like delay, packet loss etc.) to perceived video quality, S. Kanumuri et al performed a very analytical statistical model of the packet-loss visual impact on the decoding video quality for MPEG-2 video sequences [9], specifying the various factors that affect perceived video quality and visibility (e.g. Maximum number of frames affected by the packet loss, on what frame type the packet loss occurs etc). Similarly, in [10] it is presented another transmission/distortion modeling for real-time video streaming over error-prone wireless networks. In this work, it is introduced a modeling of the impulse transmission distortion (i.e. the visual fading behavior of the transmission errors). However, the already proposed models are very codec and content specific, while they do not also provide any end-to-end video quality estimation, namely the degradation during the encoding process and the transmission/streaming procedure.

In this framework, this paper proposes a generic model for end-to-end video quality prediction which provides end-to-end video quality assessment, estimating the worst case degradation of the initial quality, regardless of the used video codec and the dynamics of the transmitted encoded sequence.

III. Modelling 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. The DCT operates on a X block of NxN image samples or residual values after prediction and creates Y, which is a NxN block of coefficients. The advantage of the DCT transform is that it is possible to reconstruct quite satisfactorily the original image, applying the reverse DCT on a subset of the DCT coefficients, without taking under consideration the rest coefficients with insignificant magnitudes (see Figure 1). Thus, with cost of some video quality degradation, the original frame can be satisfactorily reconstructed with a reduced number of coefficient values.

In this framework, a FR perceived quality metric, which provides very reliable assessment of the video quality is the *SSIM* metric [1], [3]. The *SSIM* is a FR metric for measuring the structural similarity between two image/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 concept of averaging the SSIM for the whole video duration can be exploited for deriving a single perceived quality measurement, which is more practical, especially for the service providers.



a. b. c.

Figure 1. Sub-figure a is the source image, while b is reconstructed using only 1 DCT coefficient. Image c uses 4 coefficients out of the total 64 DCT Coefficients for each block (i.e. 8x8).

For the experimental needs of this paper, five reference clips were used, which were transcoded from their original uncompressed format to ISO H.264 Baseline Profile, at various VBR bit rates (ranging from 50kbps up to 500kbps). For each corresponding bit rate, a different H.264 compliant file with CIF (Common Interface Format) resolution (352x288) was created. The frame rate was set at 25 frames per second (fps) during all the transcoding process.

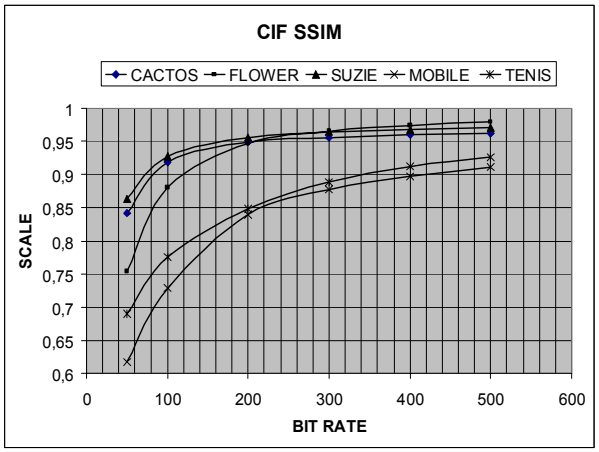


Figure 2. The $<PQoS>_{SSIM}$ vs. bit rate (kbps) curves for the test signals

Each H.264 video clip was then used as input in the *SSIM* estimation algorithm. From the resulting *SSIM* vs. time graph, the average <PQoS> value 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 2.

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

$$< PQoS >_{SSIM} = C_1 \ln(BitRate) + C_2 \quad (1)$$

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

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>$_{SSIM}$ of the specific bit rate the constant C_1 and C_2, which is derived by the analytical logarithmic expression of Table I:

$$C_{S-T} \triangleq \{m : (\frac{1}{N}\sum_{i=1}^{N} SSIM(f_i), C_1, C_2)_n = F_n, n \in [BitRate_{\min}, BitRate_{\max}]\} \quad (2)$$

where the following notation has been followed:

-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 consist the movie m

Thus, deriving the sets C_n for various contents, ranging from static to very high spatial and temporal ones, we define a reference hyper-set RS, containing various C_n sets for specific spatiotemporal levels can be deduced:

$$RS = \{C_{S-T_{Low}}, ..., C_{S-T_{High}}\} \quad (3)$$

TABLE I
FITTING PARAMETERS AND R^2 FOR THE TEST SIGNALS

Test Signal	Logarithmic Function	R^2
Cactos	0.0490ln(x)+0.6719	0.8593
Mobile&Calendar	0.1295ln(x)+0.1274	0.9759
Flower Garden	0.0947ln(x)+0.4163	0.9979
Table Tennis	0.1033ln(x)+0.2940	0.9938
Suzie	0.0443ln(x)+0.7075	0.8901

For clarity reasons, an example follows: Consider a non-reference video clip of duration 25 seconds was encoded in H.264/CIF at 100 kbps. Then, the resulted instant *SSIM* curve for this clip was used to estimate the average *SSIM* value, which was estimated equal to 0.8. Afterwards, using this value as input in ADV, we define that the C_{S-T}, which contains the triplet element at 100kbps that minimizes the ADV belongs to Table Tennis reference clip. Thus, the equation that better describes the variation of the <PQoS>$_{SSIM}$ vs. the bit rate is <PQoS>$_{SSIM}$ = 0.1033ln(x)+0.2940. Thus, if the Content Provider wishes to offer this video clip at 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.

IV. MODELING PACKET LOSS IMPACT ON VIDEO QUALITY

From the hierarchical structure of MPEG encoding [11], each video frame may be classified as directly or indirectly undecoded. A directly undecoded frame is caused by a loss of packets, which are necessary for the decoding process of the specific frame. On the other hand, an indirectly undecoded video frame is considered when this frame depends on a directly undecoded frame. For simplicity, we do not consider any error concealment method in this paper, setting the Decoded Threshold (DT) [11] to 1.0. Therefore, our analysis provides the worst-case boundary of video quality degradation due to errors in transmission, since one packet loss leads to an undecoded frame.

In this paper for the video quality assessment we adopt a modified version of the Decoded Frame Rate (Q) metric [12] for video streams transmitted over packet-switched networks. Thus, the main modification is the use of the packet loss rate instead of the frame loss rate. The value of Q lies between 0 and 1.0, where 1.0 denotes no quality degradation and 0 is the worst theoretical case. The Q is defined as follows:

$$Q = \frac{N_{dec}}{(N_{total-I} + N_{total-P} + N_{total-B})} \quad (4)$$

where N_{dec-I}, N_{dec-P}, and N_{dec-B} is the total number of each type of frames and N_{dec} is the summation of the successfully decoded frames in the video flow.

Based on the probability theory, we next calculate for each frame type (i.e. I, P and B) the probability to be successfully decoded, considering a p packet loss rate, $C_I C_P C_B$ as the mean number of packets to transport the data of each frame type and N_{GOP} the total number of GOPs in the video flow.

Regarding I frames in a GOP, they are successfully decoded only if all the packets that belong to an I frame received intact. Therefore, the probability that an I frame is decodable is $S(I) = (1-p)^{C_I}$. Consequently, the expected number of correctly decoded I frames for the whole sequence is

$$N_{dec-I} = (1-p)^{C_I} * N_{GOP} \quad (5)$$

A P frame is decodable only if the preceding I or P frames have been successfully decoded and all the P packets of this frame have also successfully received. In a GOP, there are N_p P frames, and the probability of the P frame to be decodable is

$$\begin{aligned} S(P_1) &= (1-p)^{C_I} * (1-p)^{C_P} = (1-p)^{C_I + C_P} \\ S(P_2) &= (1-p)^{C_I} * (1-p)^{C_P} * (1-p)^{C_P} = (1-p)^{C_I + 2C_P} \\ &\cdots\cdots \\ S(P_{N_P}) &= (1-p)^{C_I} * (1-p)^{N_P * C_P} = (1-p)^{C_I + N_P * C_P} \end{aligned} \quad (6)$$

Thus, the expected number of correctly decoded P frames for the whole sequence is

$$N_{dec-P} = (1-p)^{C_I} * \sum_{j=1}^{N_P} (1-p)^{jC_P} * N_{GOP} \quad (7)$$

Finally, the B frames in a GOP are decodable only if the preceding and succeeding I or P frames are both successfully decoded and all the B packets have successfully received. As consecutive B frames have the same dependency throughout the GOP structure, we consider the consecutive B frames as a B group. Especially, the last B frame in a GOP is encoded from the preceding P frame and succeeding I frame, so that it is influenced in the two I frames. In a GOP, the probability of the B frame to be decodable is

$$S(B_1) = (1-p)^{C_I} * (1-p)^{C_P} * (1-p)^{C_B}$$
$$S(B_2) = (1-p)^{C_I} * (1-p)^{2C_P} * (1-p)^{C_B}$$
$$\ldots\ldots . \quad (8)$$
$$S\left(B_{\frac{N}{M}-1}\right) = (1-p)^{C_I} * (1-p)^{\left(\frac{N}{M}-1\right)*C_P} * (1-p)^{C_B}$$
$$S\left(B_{\frac{N}{M}}\right) = (1-p)^{2C_I} * (1-p)^{\left(\frac{N}{M}-1\right)*C_P} * (1-p)^{C_B}$$

Where N defines the GOP length (i.e. the number of frames of each GOP) and the M-1 is the number of B frames between I-P or P-P frames. Thus, the expected number of correctly decoded B frames for the whole video is

$$N_{dec\text{-}B} = (M-1)*\sum_{j=1}^{\frac{N}{M}} S(B_j) * N_{GOP} \quad (9)$$
$$= \left[(M-1)*(1-p)^{C_I} * \sum_{j=1}^{N_P}(1-p)^{jC_P} * (1-p)^{C_B} + (M-1)*(1-p)^{2C_I} * (M-1)^{N_P C_P} * (1-p)^{C_B}\right] * N_{GOP}$$
$$= \left[(1-p)^{C_I+N_P C_P} + \sum_{j=1}^{N_P}(1-p)^{jC_P} * (1-p)^{C_B}\right] * (M-1)*(1-p)^{C_I+C_B} * N_{GOP}$$

In order to check experimentally the validity of the proposed model, the following procedure was performed: A video trace of the movie "*Aladdin*" was selected, which contains 89998 frames, and is encoded with MPEG-4 at 440kbps. This trace can be found online at [13]. Based on the selected encoding parameters, the video statistics are 7500 I frames, 22500 P frames and 59998 B frames with GOP structure IBBPBBPBBPBB (N=12, M=3). In order to perform the evaluation, we used the randomly uniform model (see Figure 3), which provides uniformly distributed losses with the mean loss rate (p) and corresponds to the worst case packet loss scenario, given that we have considered DT equal to 1.0.

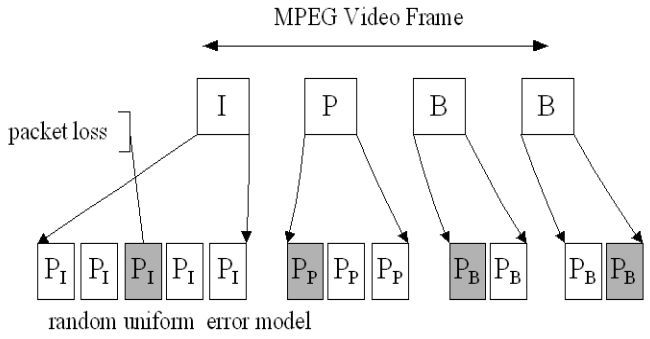


Figure 3. The randomly uniform packet loss scheme

We examined the packet loss rates from 2% up to 20% with step 2%. Furthermore, we assumed that the maximum transmitting packet size is 1000 bytes. The experiments were simulated on NS-2 [14]. The simulation topology was the same as in [11]. The results of this procedure are depicted on Figure 4, which shows the video quality degradation for various packet loss rates.

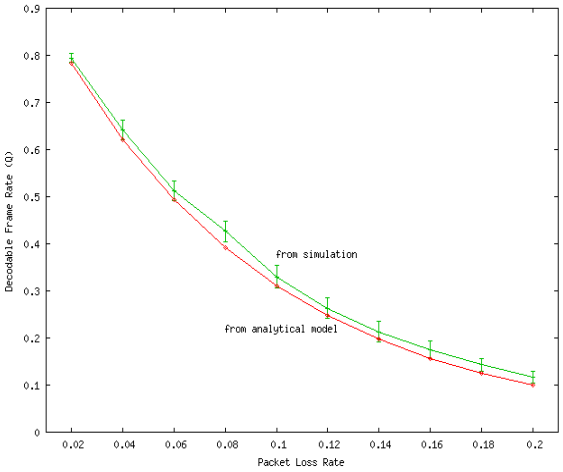


Figure 4. The effect of the packet loss models on the delivered video quality

As it can be observed, there is a significant good match between the theoretically expected video quality degradation and the corresponding experimentally derived, proving the validity of the proposed theoretical framework.

V. THE PROPOSED END-TO-END FRAMEWORK

Based on the aforementioned theoretical frameworks of video quality prediction at a pre-encoding stage and packet loss modelling, this section demonstrates an end-to-end video quality prediction model, which is based on the combination of the two frameworks.

Consider that a hypothetical Content Provider, which want to stream over the network a music video clip at various quality levels. We consider that the Content Provider possesses the reference hyper set RS, containing C_n sets from the test signals of Table I. Initially, the video clip is encoded at MPEG-4/CIF 100 kbps. Then, the resulted encoded clip is used as input to the *SSIM* algorithm and the resulted instant *SSIM* curve is used for the estimation of the <*SSIM*> value, which was estimated equal to 0.8. Afterwards, using this value as input in ADV, we define the C_{S-T}, which contains the triplet element that minimizes the ADV. For the derived <*SSIM*> value, the optimal C_{S-T} set belongs to Table Tennis reference clip. Thus, the equation that describes better the variation of the $<PQoS>_{SSIM}$ vs. the bit rate is

$$<PQoS>_{SSIM} = 0.1033 ln(x) + 0.2940$$

So, the Content Provider is now able to identify the bit rates that correspond to various quality levels (e.g. 0.70, 0.80 and 0.90), by simply using the above equation.

TABLE II
PREDICTED BIT RATE VALUES FOR SPECIFIC QUALITY LEVELS

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

Table II shows the corresponding encoding bit rate values that $<PQoS>_{SSIM}$ provides for the specific video clip. Thus, one only test measurement of the average *SSIM* at a specific encoding bit rate is enough for the accurate determination of the PQoS vs. Bit Rate curve for a given video clip.

Afterwards, based on a network monitoring system (i.e. packet loss rate), we will examine how these initial video quality levels, will be further degraded. Considering that a monitoring systems provides us that the average packet loss rate at the transmission network is for example 10%, then it can be predicted from the packet loss model (see Figure 4) that the worst case is that the end-user (i.e. the content consumer) will experience video quality degradation (due to network parameters) for the 70% of the total duration of the sequence. For the rest 30%, the user will experience normal playback without any perceived artifacts. Thus, if the Content Provider would like to calculate a representative value of the Expected Delivered Video Quality (EDVQ) level at the content consumer, she/he would apply the following equation:

$$EDVQ = (Initial\_Video\_Quality) * (Percentage\_of\_Succesfully\_Decoded\_Frames) \quad (10)$$

Equation 10 provides a specific end-to-end video quality prediction for the worst case scenario, if error concealments methods are not taken under consideration (i.e. D.T. equal to 1.0).

VI. Conclusions

This paper has presented a theoretical framework for end-to-end video quality prediction for MPEG-based video sequences. 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 QoS-sensitive network parameters (i.e. packet loss) to video quality degradation. The efficiency of both the 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 prediction framework.

The advances of the proposed framework are that it is generic in nature, since it can be applied on MPEG-based encoded sequences, independently of the used encoding standard. Moreover, it also exploits the novel issue of predicting the video quality of an encoded service at a pre-encoding stage, which provides new facilities at the Content Provider side. Also, by applying the randomly uniform packet loss model, the proposed framework overpasses any stochastic predicaments in mapping the packet loss ratio to video quality degradation, since it calculates the worst case scenario.

As future work, the authors will extent the current framework by including more network parameters (e.g. packet size) in the network parameters to video quality degradation mapping. Moreover, the reference experimental hyper-set *RS* will be further extended with more test sequences and clips.

Acknowledgment

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

References

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

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

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

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

[5] Lu, L., et al. Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video. in IEEE International Conference on Multimedia. 2002.

[6] H. Koumaras, A. Kourtis, D. Martakos, "Evaluation of Video Quality Based on Objectively Estimated Metric", Journal of Communications and Networking, Korean Institute of Communications Sciences (KICS), Vol. 7, No.3, pp.235-242, Sep 2005 (available online http://citeseer.comp.nus.edu.sg/747634.html)

[7] H. Koumaras, A. Kourtis, D. Martakos, J. Lauterjung, "Quantified PQoS Assessment Based on Fast Estimation of the Spatial and Temporal Activity Level", Multimedia Tools and Applications, Springer Editions, accepted for publication.

[8] H. Koumaras, E. Pallis, G. Xilouris, A. Kourtis, D. Martakos, J. Lauterjung, "Pre-Encoding PQoS Assessment Method for Optimized Resource Utilization", 2nd Inter. Conference on Performance Modelling and Evaluation of Heterogeneous Networks, Het-NeTs04, Ilkley, United Kingdom, 2004.

[9] S. Kanumuri, P. C. Cosman, A.R. Reibman, V.A. Vaishampayan, "Modeling Packet-Loss Visibility in MPEG-2 Video", IEEE transactions on Multimedia, Vol.8, No.2, pp.341-355, April 2006.

[10] Z. He, H. Xong, "Transmission Distortion Analysis for Real-Time Video Encoding and Streaming over Wireless Networks", IEEE Transactions on Circuits and Systems for Video Technology, Vol.16, No.9, pp.1051-1062, September 2006

[11] Cheng-Han Lin, Chih-Heng Ke, Ce-Kuen Shieh, Naveen Chilamkurti, "The Packet Loss Effect on MPEG Video Transmission in Wireless Networks", The IEEE 20th International Conference on Advanced Information Networking and Applications (AINA'06), April 18-20, 2006, Vienna, Austria

[12] A. Ziviani, B. E. Wolfinger, J. F. Rezende, O. C. M. B. Duarte, and S. Fdida, "Joint Adoption of QoS Schemes for MPEG Streams," Multimedia Tools and Applications Journal, to appear.

[13] Video Trace Aladdin available online for downloading at http://trace.eas.asu.edu/TRACE/pics/FrameTrace/mp4/indexa4f6.html

[14] C.-H.-Ke, C.-H.-Lin, C.-K. Shieh, and W.-S. Hwang, "A Novel Realistic Simulation Tool for Video Transmission over Wireless Network," presented at The IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC2006), Taiwan, 2006.