Objective quality prediction model for lost frames in 3D video over TS

Bruno Feitor, Pedro Assunção
Instituto de Telecomunicações
Instituto Politécnico de Leiria / ESTG
Email: bruno.ts.feitor@gmail.com

João Soares, Luís Cruz
Instituto de Telecomunicações
Universidade de Coimbra / DEEC
Email: lcruz@deec.uc.pt

Rui Marinheiro
Instituto Universitário de Lisboa / ISCTE-IUL
Instituto de Telecomunicações
Email: rui.marinheiro@iscte.pt

***Abstract*—This paper proposes an objective model to predict the quality of lost frames in 3D video streams. The model is based only on header information from three different packet-layer levels: Network Abstraction Layer (NAL), Packetised Elementary Streams (PES) and Transport Stream (TS). Transmission errors leading to undecodable TS packets are assumed to result in frame loss. The proposed method estimates the size of the lost frames, which is used as a model parameter to predict their objective quality measured as the Structural Similarity Index Metric (SSIM). The results show that SSIM of missing stereoscopic frames in 3D coded video can be predicted with Root Mean Square Error (RMSE) accuracy of about 0.1 and Pearson correlation coefficient of 0.8, taking the SSIM of uncorrupted frames as reference. It is concluded that the proposed model is capable of estimating the SSIM quite accurately using only the lost frames estimated sizes.**

I. Introduction

In recent years, quality evaluation of three-dimensional (3D) video has become an increasingly important issue, particularly when transmission over error prone networks is used, such as the near future 3DTV broadcast services over IP or DVB networks [1] and [2]. Since the users are the usual end consumers of multimedia content, they are also the most reliable evaluators of the actual quality experienced from the services provided by operators. However, collecting information about the users' quality of experience (QoE) is not a valid option in most real-time applications, especially in broadcast services. Thus, objective metrics capable of estimating the subjective video quality using simple parameters obtained from the coded stream and the transmission network, should be used in practical systems.

Depending on the type of information used objective quality methods can be classified into: Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) methods. FR methods like [3]–[5] determine the impaired video quality by comparison with the original signal. FR methods are impractical for video quality monitoring at locations remote from the video encoder, where the reference signal is not readily available. In RR methods, some kind of reduced information about the reference is sent through a side channel and used for computing objective quality scores [6]–[8]. The growing need to measure the quality of compressed video streams at any point of communication systems without recourse to suplementary information has been fostering increased research efforts on the development of NR quality methods [9]–[12].

3D video quality monitoring can be seen as an extension of similarly minded methods for 2D video [13], with anticipated applications to service planning and video quality monitoring. These two types of applications are compared in [14], which concludes that Packet Loss Rate (PLR), Burst Loss Frequency (BLF) and Invalid Frame Rate (IFR) can be successfully used for planning video systems, unlike video monitoring applications where they are not as effective. A model for monitoring 2D video quality is also proposed in [14], which grades video quality using measures of spatio-temporal complexity. The packet layer video quality model described in [15] only uses network level information (e.g., Packet Loss Rate) without access to video sequence information (e.g., spatio-temporal, error location in the video stream). This model estimates the Mean Square Error (MSE) from the PLR assuming a linear relationship, which is not very accurate, given the high variability of video content. The authors also point out that the variance of the actual MSE can be quite large, which implies that by simply measuring the PLR may not provide an accurate quality measure.

This paper presents a 3D video quality model that quantifies the objective video quality in the presence of frame losses without needing to decode the compressed stream. Figure 1 shows a potential application scenario for the proposed NR 3D video quality monitoring system. Since it only requires information that can be obtained from the transmission network such as the packets losses and the stream itself, this system is able to operate at any point of the communication path.

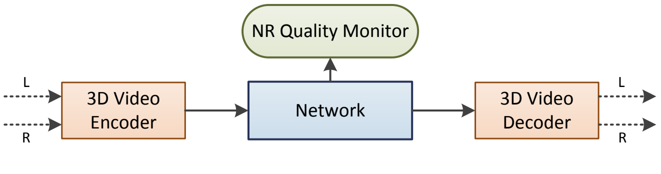


Figure 1. NR packet-layer application scenario.

Three models based on the dependency between the lost frame size and the video quality degradation are investigated. Next sections will demonstrate that a polynomial cubic fitting

This work was supported by grant IT/LA/P01131/2011 and Project FCT PTDC/EEA-TEL 114487/2009, PTDC/EEA-TEL/120666/2010

provides results accurate enough to allow practical use in real 3DTV packet based broadcast systems.

II. Model Overview

A broadcast network is assumed, which corresponds to a protocol stack with TS/PES/NAL. These are the packet layers used in the proposed model as depicted in Figure 2, where the data units from the Video Coding Layer (VCL) are wrapped into NAL packets which in turn make up the PES payload, and finally the whole PES packet is split into TS packets.

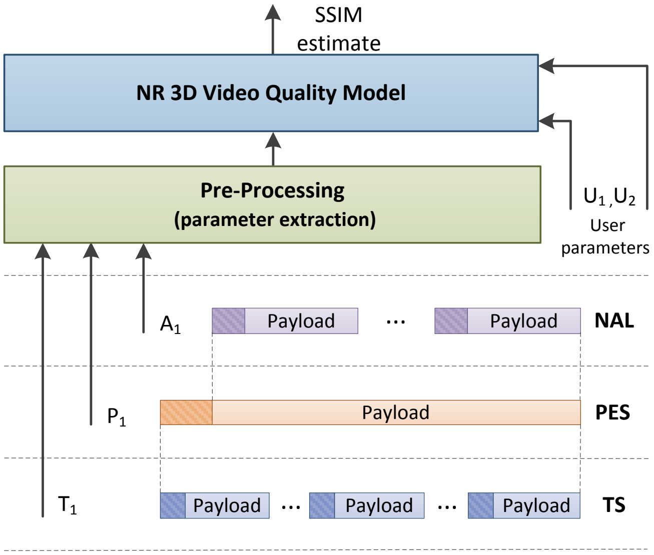


Figure 2. Packet layer model structure

The proposed model aims to estimate the quality degradation in isolated stereo frames due to errors in TS packets leading to frame loss. The degradation is measured by the difference between the SSIM of the error-free stereo frame (i.e., SSIM=1) and that of the displayed frame, assuming that frame-copy is used as the concealment method. Note that coding distortion is not included in this model, which means that SSIM=1 for any frame correctly received, thus decoupling the errors due to packet losses from the loss of fidelity due to lossy encoding. The proposed NR model uses an estimate of each lost frame size and different parameters for different frame types. The lost frame size is estimated from the average of the last frame sizes with the same type and view. The GOP structure is provided as input parameter, since this type of information is in general static. Nevertheless, estimation of different frame types can be done with reasonable accuracy even without knowing the GOP structure. The detection of lost frames and estimation of their sizes is based on the following parameters extracted from the packet headers:

- **T1: Continuity Counter (TS)**: enables detection of TS packet loss events.
- **P1: PES packet length**: provides the frame size.
- **A1: NAL_Ref_Idc**: identifies the frame type.
- **U1: GOP size**: GOP size information.
- **U2: GOP structure**: GOP structure information.

III. Proposed NR Model

Several experiments were performed to collect relevant statistical data on the impact of frame loss in 3D video streams quality. To conduct these experiments a 3D video sequence obtained by concatenating 5 individual sequences (Ballons, Champagne Tower, Kendo, Pantonime and Dog) was encoded with H.264/AVC Stereo High Profile using the reference software JM 18.2 . This sequence has spatial resolution of 1024x768 pixels, 30 fps frame rate, and was encoded using GOP size equal to 21 frames and IBPBP GOP structure. Three different datasets were created by encoding this sequence with QP ranging from 26 to 32, achieving different PSNR and bitrates. These datasets are described in Table I. The resulting video stream was encapsulated into a TS stream using the reference software FFMPEG. To emulate network conditions, packetized streams were then subject to packet loss events. These have to be created according to a error pattern, produced in a controlled manner, to better modeling individual types of impairments. Aiming this, at present a single frame is lost in each loss event, but other impairment models can be used [16].

Table I
QP'S, PSNR AND BITRATE.

	I-QP	P-QP	B-QP	PSNR (dB)	Bitrate (Kb/s)
Dataset-1	26	28	28	42	2847
Dataset-2	28	30	30	41	2225
Dataset-3	30	32	32	40	1727

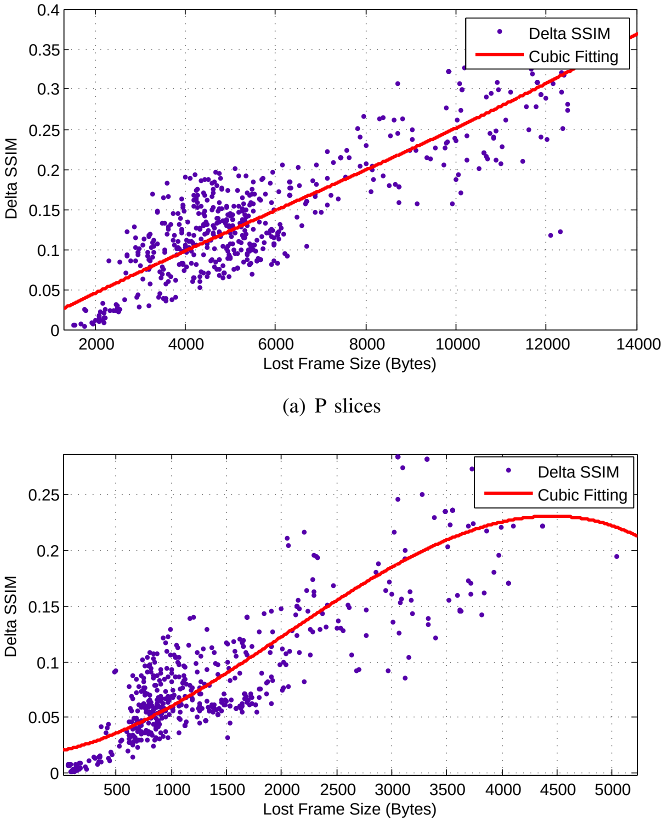
Then, the corrupted TS stream was decoded using frame-copy for concealment of lost frames. The decoded frames were used to compute the SSIM with reference to the corresponding uncorrupted decoded frames and the results stored. Linear, quadratic and cubic polynomials were then used as fitting models for the data obtained from the experiments with the three Datasets. The SSIM drop obtained for each lost frame ($Delta\_SSIM$) is plotted in Figure 3 as a function of the frame size for P and B slices using Dataset-3. The cubic fitting model is also shown. A similar behavior was found from the experiments carried out with the other Datasets used in this study. The generic equation that defines the proposed models is the polynomial in Eq. 1 where $dSSIM_n$ is the $Delta\_SSIM$, n is the polynomial degree, L_{fs} is the lost frame size in bytes and p_n are the polynomial coefficients.

$$dSSIM_n = p_n L_{fs}^n + ... + p_2 L_{fs}^2 + p_1 L_{fs} + p_0 \qquad (1)$$

After measuring the accuracy obtained from fitting models with polynomials of different degrees it was concluded that third order models are accurate enough to estimate the dependency between lost frame size and $dSSIM_n$. The polynomial coefficients pertaining to each fitted curve are shown in Table II for P-frames and Table III for B-frames.

IV. Simulation Results

To validate the proposed models a 3D video sequence was captured to simulate a real-time case analysis. Towards that



(a) P slices

(b) B slices

Figure 3. Delta SSIM vs frame size for P and B slices

Table II
CURVE FITTING COEFFICIENTS FOR P-FRAMES.

	Linear	Quadratic	Cubic

Dataset-1	p0=0.03596
	p1=3.66E-06
	-
	-
		p0=0.175
		p1=-2.11E-05
		p2=9.26E-10
		-
			p0=0.05365
			p1=9.29E-06
			p2=-1.19E-09
			p3=4.22E-14

Dataset-2	p0=-0.04488
	p1=2.61E-05
	-
	-
		p0=3.89E-02
		p1=6.11E-06
		p2=8.92E-10
		-
			p0=4.74E-03
			p1=1.78E-05
			p2=-1.87E-10
			p3=2.72E-14

Dataset-3	p0=-0.1276
	p1=9.01E-05
	-
	-
		p0=-0.2097
		p1=1.43E-04
		p2=-6.66E-09
		-
			p0=-0.03292
			p1=-2.92e-05
			p2=3.86E-08
			p3-3.28E-12

end the stereo video sequence was captured using a 3D Digital HD Video Camera Recorder (Sony HXR-NX3D1U NXCAM) with approximately 4000 stereoscopic frames. The following configurations were used for encoding: 1920x1080@30Hz, GOP size 21 frames, IBPBP GOP structure, I-frame QP 38, P-frame QP 40 and B-frame QP 40. To evaluate single frame quality degradation one frame was lost in each GOP.

Table III
CURVE FITTING COEFFICIENTS FOR B-FRAMES.

	Linear	Quadratic	Cubic

Dataset-1	p0=0.01636
	p1=3.05E-05
	-
	-
		p0=-1.83E-02
		p1=5.69E-05
		p2=-3.48E-09
		-
			p0=-1.90E-02
			p1=5.78E-05
			p2=-3.77E-09
			p3=2.57E-14

Dataset-2	p0=0.006689
	p1=4.38E-05
	-
	-
		p0=-2.04E-03
		p1=5.34E-05
		p2=-1.80E-09
		-
			p0=1.30E-02
			p1=2.57E-05
			p2=1.07E-08
			p3=-1.50E-12

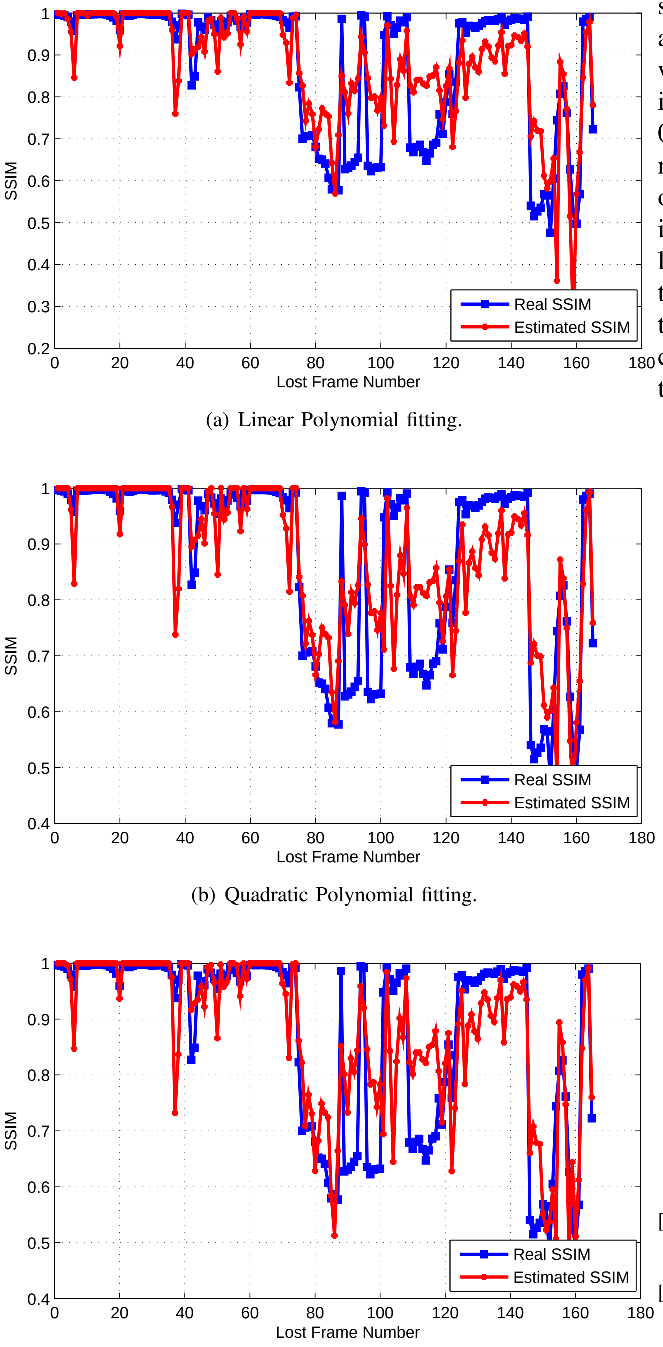
Dataset-3	p0=-0.006671
	p1=5.65E-05
	-
	-
		p0=2.07E-03
		p1=6.29E-05
		p2=-1.54E-09
		-
			p0=2.01E-02
			p1=2.13E-05
			p2=2.23E-08
			p3=-3.69E-12

In Figure 4 the SSIM for the lost frames estimated using linear, quadratic and cubic polynomial models is compared to the SSIM of the concealment frame that stands in for lost frame. It was found that the SSIMs estimated with the proposed models are quite similar to the real SSIM. The model performance was then evaluated with two different performance indicators of the SSIM estimators: Root Mean Square Error (RMSE) and the Pearson Correlation values for both P and B frames and the three model orders. Their correlation measurements are presented in Table IV. For P-frames, the RMSE decreases sightly with the increase of polynomial degree. As expected the Pearson Correlation increases with the polynomial degree. A similar behavior is observed for B-frames, though the model does not evidence such a strong correlation as that observed for the P-frames possibly due to the smaller size of B-frames and larger dispersion of the lost frame size vs. $dSSIM_n$ data around the fitted models.

Table IV
SIMULATION RESULTS.

Polynomial		P-frame	B-frame
Linear	RMSE	0.1045	0.1556
	Pearson Corr.	0.783	0.7099
Quadratic	RMSE	0.0965	0.1562
	Pearson Corr.	0.8197	0.7106
Cubic	RMSE	0.0956	0.1655
	Pearson Corr.	0.8224	0.7073

The simulation results show a strong correlation between the estimated and the real 3D video quality. Both RMSE and Pearson Correlation measurements show very promising values if one considers that a quality measure in the visual domain is being estimated using information obtained from the compressed packetized bitstream domain.



(a) Linear Polynomial fitting.

(b) Quadratic Polynomial fitting.

(c) Cubic Polynomial fitting.

Figure 4. Estimated SSIM vs. Real SSIM of lost P-Frames.

V. Conclusion

A simulation study was undertaken with different 3D video streams. Frame sizes were found to be correlated with the quality degradation incurred as a result of loss of those frames. It was concluded that a polynomial fitting is able to model this dependency. The model coefficients for different polynomial degrees were determined using three different datasets.

The model was validated with a long real-life 3D video sequence. When applied to this sequence (after introduction of artificial packet losses) the quality estimator performed quite well with Pearson Correlation coefficients around 0.8 which indicate a strong positive correlation, and RMSE values around 0.1 evidencing a small estimation error. These results are quite remarkable as the proposed packet layer model is based only on the estimated lost frame size, which contains very little information about the characteristics and visual content of the lost frame. Further work will focus on adding the effects of temporal error propagation and propagation between views to the quality evaluation model. In addition, to emulate network conditions, different types of networking impairments will be tested individually and in a scenario-based combination.

References

[1] H. R. Wu, K. R. Rao, and Ashraf A. Kassim, "Digital video image quality and perceptual coding," *Journal of Electronic Imaging*, vol. 16, no. 3, pp. 039901, 2007.

[2] G. Ghinea, P. Muntean, F. Etoh, F.and Speranza, and H. Wu, "Special issue on quality issues on mobile multimedia broadcasting," 2008, vol. 54, pp. 424–727.

[3] ITU-T Recommendation J.144, *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*, Mar. 2004.

[4] Zhenghua Yu, Hong Ren Wu, S. Winkler, and Tao Chen, "Vision-model-based impairment metric to evaluate blocking artifacts in digital video," *Proceedings of the IEEE*, vol. 90, no. 1, pp. 154 –169, jan 2002.

[5] Pengwei Hao, Qingyun Shi, and Ying Chen, "Co-histogram and its application in remote sensing image compression evaluation," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, sept. 2003, vol. 3, pp. III – 177–80 vol.2.

[6] J. Baina, P. Bretillon, D. Masse, and A. Refik, "Quality of mpeg2 signal on a simulated digital terrestrial television," *Broadcasting, IEEE Transactions on*, vol. 44, no. 4, pp. 381 –391, dec 1998.

[7] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, sept. 2003, vol. 3, pp. III – 185–8 vol.2.

[8] S. Olsson, M. Stroppiana, and J. Baina, "Objective methods for assessment of video quality : state of the art," *Broadcasting, IEEE Transactions on*, vol. 43, no. 4, pp. 487 –495, dec 1997.

[9] S. Argyropoulos, A. Raake, M.-N. Garcia, and P. List, "No-reference bit stream model for video quality assessment of H.264/AVC video based on packet loss visibility," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 1169–1172.

[10] Toru Yamada, Yoshihiro Miyamoto, and Masahiro Serizawa, "No-reference video quality estimation based on error-concealment effectiveness," in *Packet Video 2007*, Nov. 2007, pp. 288 –293.

[11] Junghyun Han, Yo han Kim, Jangkeun Jeong, and Jitae Shin, "Video quality estimation for packet loss based on no-reference method," in *Advanced Communication Technology (ICACT), 2010 The 12th International Conference on*, Feb. 2010, vol. 1, pp. 418 –421.

[12] Dukgu Sung, Seungseok Hong, Yohan Kim, Yonggyoo Kim, and Taesung Parkand Jitae Shin, "No reference quality assessment over packet video network," in *IWAIT 2009*, 2009.

[13] Quan Huynh-Thu, P. Le Callet, and M. Barkowsky, "Video quality assessment: From 2D to 3D - challenges and future trends," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 2010, pp. 4025 –4028.

[14] Ning Liao and Zhibo Chen, "A packet-layer video quality assessment model with spatiotemporal complexity estimation," *EURASIP Journal on Image and Video Processing*, vol. 2011, no. 1, pp. 5, 2011.

[15] A.R. Reibman, V.A. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *Multimedia, IEEE Transactions on*, vol. 6, no. 2, pp. 327 – 334, 2004.

[16] ITU-T Recommendation G.1050, *Network model for evaluating multimedia transmission performance over Internet Protocol*, Mar. 2011.