Quality evaluation of 3D video for QoE management in media networks

Pedro Rocha
Instituto de Telecomunicacoes
University of Coimbra
Dep. of Electrical and Computer Eng.
Coimbra, Portugal 3030-290
Email: pedro.rocha@student.uc.pt

Pedro Assuncao
Instituto de Telecomunicacoes
Polytechnic of Leiria
ESTG, Dep. of Electrical Eng.
Leiria, Portugal, 2401-911
Email: amado@co.it.pt

Luis A. da Silva Cruz
Instituto de Telecomunicacoes
University of Coimbra
Dep. of Electrical and Computer Eng.
Coimbra, Portugal 3030-290
Email: lcruz@deec.uc.pt

***Abstract*—The emergence of 3D video technologies and their foreseeable applications in Internet Protocol Television (IPTV) and video delivery systems using encoded 3D video contents over packet networks raises the question of how to manage the Quality of Experience (QoE) across diverse lossy channels. Of particular importance for QoE management is the problem of measuring the impact of data losses in packetized 3D video information and how it affects the quality experienced by end users, when the content is rendered and presented at their premises. Previous work by the authors showed that it is possible to model the perceived quality degradations through the use of artificial neural networks receiving as inputs several parameters describing the packet loss events. The first stage of such model was specifically developed for packetized 3D video in texture-plus-depth format where only the depth information was prone to transmission errors. This article presents an extension of the previous model by including the effect of texture information losses along with other specific aspects associated with the dual nature of this type of data loss. The validity of the model is verified through the use of extensive simulations and comparisons between real and estimated values of a recently proposed 3D video quality measure.**

I. Introduction

This paper addresses the problem of estimating the quality of 3D video represented in texture-plus-depth format, which after being encoded using HEVC and packetized is transmitted over packet networks. The quality estimation model proposed is based on Artificial Neural Networks (ANNs) and only requires information from bitstream and packet-level to predict the quality of 3D video reconstructed from packet streams transporting the texture and depth information which might have been subject to transmission losses. The full-reference 3D video objective measure described in [1] (3DSwIM) was chosen as the reference to be estimated due to the good performance reported by its authors. The neural network model follows from our previous work [2], consisting in a simple structure, using as input only 12 packet-layer-parameters (PLPs), 6 related with texture-carrying packets and 6 other with depth packets. This work extends the previous model to include loss information about the text stream, which is particularly useful when both streams are subject to the same type of network constraints. The simplicity of the model makes it usable in networked applications where a no-reference (NR) low-complexity real-time 3D video quality estimator is needed [3]. As will be explained the model was trained and validated using thousands of instances of vectors consisting of values of the 12 PLPs collected from simulated transmissions of encoded and packetized 3D video. Each of these vectors is complemented with the value of the 3DSwIM measure computed on the video recovered from the respective impaired packet stream. The set of these 13 component vectors is split and used in the supervised training of the model and its performance verification. The remaining of this paper is organised as follows. In Section 2 the proposed method is described. Section 3 presents the test conditions and Section 4 is divided into NN performance and accuracy, where results are analyzed. Section 5 concludes the paper.

II. Model, Inputs and Outputs

The proposed model is based on a two-layer feed-forward ANN with sigmoid hidden neurons and linear output neurons structurally similar to the one represented in Figure 1. The inputs to the model are a set of parameters describing the magnitude of the packet losses experienced by both the texture and depth coded streams.

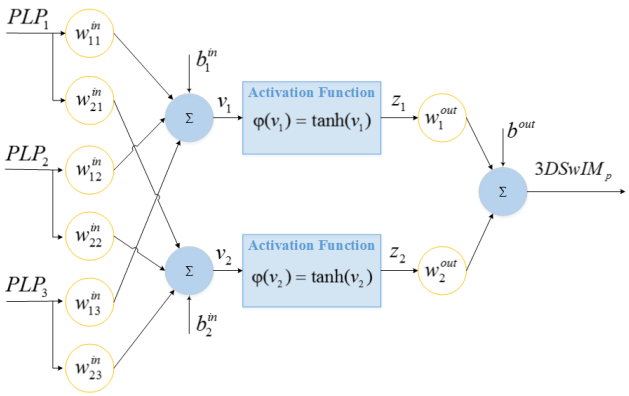

Fig. 1. Two-layer ANN with N=3 inputs and H=2 hidden nodes.

$$3DSwiM_p = \sum_{j=1}^{H}(w_j^{out} \cdot z_j) + b^{out} \quad (1)$$

$$z_j = tanh\left(\sum_{i=1}^{N}(w_{ji}^{in} \cdot PLP_i) + b_j^{in}\right) \quad (2)$$

The output of the model is an estimate of the objective quality measure ($3DSwIM_p$) of a video view synthesized from the decoded texture and depth and computed according to equations (1) and (2), where N is the number of input PLPs (PLP_i), H is the number of hidden nodes and w and b are the weights and biases of the ANN nodes.

III. Dataset and Model Training

The ANN model was trained and evaluated using a dataset of vectors constructed as described in the introduction section. The 3D video texture and depth components were encoded independently using HEVC reference software (HM v. 16.2), main profile, closed group of picture (GOP) with fixed structure IB..BP..., with a GOP size of 16 for the texture and 32 for the depth and 8 or 10 slices per frame for both components and the bitrates listed in Table I. Each slice was encapsulated in one network abstraction layer unit (NALU) and packetized in one RTP packet, with maximum transfer unit (MTU) set to 1500 bytes.

TABLE I
Video-plus-depth sequence videos used.

3D Video	Resolution	Frame Rate (fps)	V+D Bitrate (% Depth)
Balloons	1024x768	30	1.1 Mbps (26%)
Kendo	1024x768	30	1.1 Mbps (19%)
Newspaper	1024x768	30	1.1 Mbps (18%)
Champagne	1280x960	30	1.1 Mbps (13%)
Poz. CarPark	1920x1088	25	3.2 Mbps (29%)
Dancer	1920x1088	25	3.6 Mbps (5%)
GTFly	1920x1088	25	2.9 Mbps (13%)
Poz. Street	1920x1088	25	1.9 Mbps (10%)
Shark	1920x1088	25	3.5 Mbps (24%)

Using the Gilbert-Elliot model, 35 different trace files with 10000 events (lost or received packet) were generated. Different loss rates were considered (1%, 5%, 10%, 15% and 20%) with mean burst length varying from 3 to 6. With the different combinations of depth and texture, a total of 1023 impaired streams were generated and used to create hundreds of 3D videos with different degrees of degradations.

The following PLPs were defined and used in the model:

Packet Loss Rate (PLR): Rate of texture and depth slices lost during a 10 second transmission period.

Size of Lost Packets (SLP): The amount of information lost during a 10 second period.

Each PLP is computed for each slice type, I, P or B and for texture (t) and depth (d), totalling the 12 inputs of the NN: PLR_I^t, PLR_P^t, PLR_B^t, SLP_I^t, SLP_P^t, SLP_B^t, PLR_I^d, PLR_P^d, PLR_B^d, SLP_I^d, SLP_P^d, SLP_B^d,

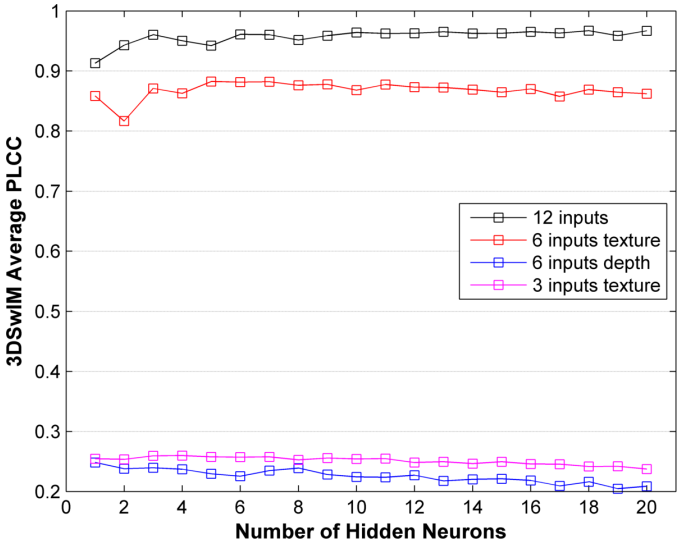

Fig. 2. PLCC between estimated 3DSwiM and real 3DSwIM scores.

Since the 3DSwIM measure evaluates the quality of 3D video views, a number of views were generated from the impaired decoded texture and depth streams. Several views were synthesized with the View Synthesis Reference Software (VSRS) 3.5 and the 3DSwIM measure was used to compute the reference (ground-truth quality values). The reference sequence input for the 3DSwIM was in all cases the same synthesized video but without degradations. Finally the Levenberg-Marquardt algorithm was used to train the ANN. Different ANN topologies were trained and validated varying the number of hidden nodes from 1 to 20. The performance of the 3D video quality estimator was evaluated by the average Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE).

IV. Model Performance

A. Effect of hidden neurons and PLPs subsets

The experimental results obtained during the ANN training showed that increasing the number of hidden neurons increased the performance of the model with a saturation effect visible after 10 hidden neurons. Also using the full set of 12 PLPs increased significantly the performance relative to the case of using only 6 texture or 6 depth related PLPs, from about 0.86 average PLCC obtained using only 6 texture PLPs to about 0.96 when using all 12 PLPs.

These behaviors are illustrated in Figure 2 which shows the correlation coefficient between estimated and real 3DSwiM as a function of the number of hidden neurons and PLPs, and in Figure 3 that shows a close up of the results with 12 and 6 texture inputs.

It is clear the impact the number and type of PLPs used have in the NN's performance, as the number of PLPs and hidden neurons increase, the correlation values increase. Due

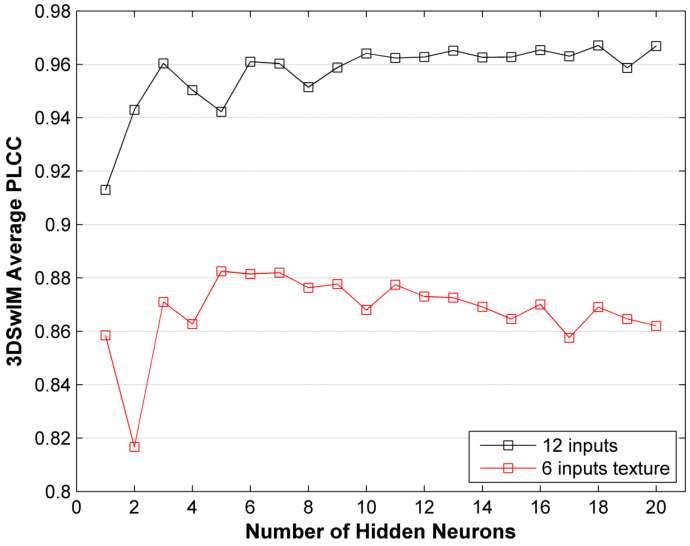

Fig. 3. PLCC between estimated 3DSwiM and real 3DSwIM scores. Close up on the 12 input and 6-texture input plots.

to the presence of texture degradations, the NN training won't be accurate enough with only 6 depth inputs. When 3 texture inputs are used, the model that has SSIM as a reference has a much better performance than the model that uses 3DSwIM algorithm, even though they are not good enough to be considered. only knows the size of the lost packet, assuming they are all of equal importance. With 6 texture PLPs it is possible to have a good correlation with the best results being provided by the ANN using 12 PLPs.

Therefore it is recommended to use 10 hidden neurons and the full PLPs set.

B. Validation

To fully evaluate the performance of the subjective quality estimator a thorough set of tests were conducted using a *leave-one-out* methodology, with test data sets different from those used for training. The ANNs were retrained leaving the samples of one of the 9 sequences out of the training stage and using them only for testing, repeating the procedure 100 times with random starting points. The results are presented in Table II and Figure 4.

TABLE II
ACCURACY RESULTS, AVERAGE PLCC AND RMSE.

Left Out Video	Avg. PLCC	Avg. RMSE
Balloons	0.9753	0.0607
Kendo	0.9693	0.0645
Newspaper	0.9651	0.0910
PoznanCarPark	0.8837	0.1270
Champagne	0.9657	0.0659
Dancer	0.9351	0.0865
GTFly	0.9419	0.0753
PoznanStreet	0.9410	0.0726
Shark	0.9345	0.0894

Results show that the proposed estimator predicts the 3D quality scores with good accuracy as measured by the high PLCC and therefore can be used as quality predictor for 3D video represented in texture plus depth format, with components independently encoded in HEVC and subject to packet losses.

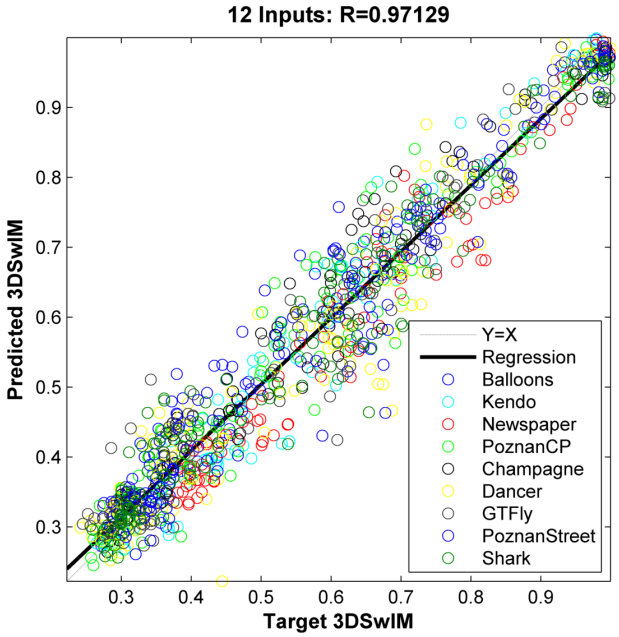

Fig. 4. 3DSwIM$_p$ vs $3DSwIM$- PLCC=0.971

V. CONCLUSION

This paper addressed the problem of estimating the quality of compressed 3D video (in video-plus-depth format) subjected to packet-losses in both the depth and texture streams. The performance results of the model proposed showed a very high accuracy when using the low-complexity 3DSwIM algorithm as reference for comparison. These characteristics make this method a good choice for QoE management in networked real-time 3D video services where NR quality monitors are essential system elements.

ACKNOWLEDGEMENTS

We wish to thank Federica Battisti for providing us with modified versions of the 3DSwIM script.

REFERENCES

[1] F. Battisti, E. Bosc, M. Carli, P. L. Callet, and S. Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Processing: Image Communication*, vol. 30, pp. 78 – 88, 2015.

[2] J. Soares, L. da Silva Cruz, P. Assuncao, and R. Marinheiro, "No-reference lightweight estimation of 3D video objective quality," in *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 763–767, Oct 2014.

[3] L. A. da Silva Cruz, M. Cordina, C. J. Debono, and P. A. A. Assuncao, "Quality monitor for 3-D video over hybrid broadcast networks," *IEEE Transactions on Broadcasting*, vol. 62, pp. 785–799, Dec 2016.