A METRIC FOR NO-REFERENCE VIDEO QUALITY ASSESSMENT FOR HD TV DELIVERY BASED ON SALIENCY MAPS

H. Boujut*, J. Benois-Pineau*, T. Ahmed*, O. Hadar**, and P. Bonnet***

*LABRI UMR CNRS 5800, Universite Bordeaux 1/IPB-Matmeca-Enseirb
351 cours de la Liberation 33405
Talence cedex - France
{boujut, benois-p, tad}@labri.fr

**Communication Systems Engineering Dept.
Ben Gurion University of the Negev
Beer Sheva, Israel, 84105
hadar@cse.bgu.ac.il

***Audemat WorldCast Systems Group
20, av Neil Armstrong, Parc d'activite J.F. Kennedy
33700 Bordeaux-Merignac – France
bonnet@worldcastsystems.com

ABSTRACT

This paper contributes to objective video quality assessment of broadcasted HDTV content without reference. In this context we present a new No-Reference video quality metric taking into account the behavior of Human Visual System. This new metric called WMBER is based on macro-blocks error detection weighted by saliency maps computed at the decoder side. Moreover, both macro-block error detection and saliency maps processing require only partial decoding allowing real-time performance. A subjective experiment has been carried out to evaluate the performances of the proposed metric. The results are compared to the Full Reference metric MSE. The evaluation of the results shows that the proposed method provides a very good prediction of subjective measures.

Index Terms— No-Reference, Objective Video Quality Assessment, H.264/AVC, HDTV delivery, Saliency Maps

1. INTRODUCTION

With the introduction of HDTV Broadcast, specifically DVB-T/S [***1***], and the wide use of IPTV services, the quality assessment of broadcasted video services became an important research topic both for academia and industries. This is due to the necessity of optimization of bandwidth allocation, better system design and optimal geographical positioning of broadcast equipment for the delivery of HD content which would satisfy user requirements and enhance his quality of experience. Video quality assessment is motivated by introduction of lossy video coding standards at the beginning of 80th. The majority of techniques proposed were dedicated to the visual quality assessment due to the degradations induced by encoding process and adopted in IUT-T recommendation [***2***]. Today, the HD delivery raises a new challenge: how to objectively asses the quality of impaired video stream at the decoder side which may suffer from signal degradation and packet errors. Transmission errors yield strong visual degradations due to simple error resilience mechanisms. These mechanisms are implemented in typical industrial decoders of the actual HDTV standard H.264/AVC [***3***]. The delivered HD video is perceived by the humans vision system (HVS) and we believe that its quality assessment can be modeled based on user perception via an accurate definition of saliency maps in video scenes [***4***]. Furthermore, the objective quality metrics in our context have to be based on the loss of blocks in H.264 encoded HD stream during the delivery. In this paper we propose a new objective video quality assessment metrics for HD video delivery combining saliency maps and loss of blocks information. In the position paper [***5***], the authors propose to design the "Quality Estimator to achieve the required accuracy for its application over the set of input content and artifacts for which it was designed". In the present paper we design a No-Reference Video Quality Assessment (NR VQA) metric for a scenario of transmission over IP and DVB of HD video with two kinds of errors: packet loss and RF signal distortion. In this paper we focus on HD video. Comparative studies of HD video with SD video make no sense in our application and were essentially conducted in the context of mobile transmission and hand carried devices [***6***]. In section 2 we introduce visual saliency maps in order to model human gaze attraction both by spatial and temporal content of video scenes. In section 3 the new NR VQA metric WMBER is introduced for HD encoded video. In section 4 we briefly describe the prediction method of subjective quality metric MOS from proposed objective quality metric WMBER. The experiments are described in section 5 while results, conclusion and perspectives are presented in Section 6.

2. SALIENCY MAPS

The Human Visual System (HVS) has the property of focusing the attention on narrow areas in the visual scene

978-1-61284-350-6/11/$26.00 ©2011 IEEE

called salient areas. These salient areas send stimulus to the HVS. Inside video scenes, salient stimuli are characterized by high color contrasts, motion and edge orientation.

Generally, in the literature, the saliency of a visual scene is depicted by two saliency maps, the "spatial" and the "temporal" saliency maps [*7*]. The spatial saliency map S^{SP} is mainly based on color, contrast and luminance. The temporal saliency map S^{T} models the attraction of attention to motion singularities in a scene. Hence two sources of saliency can be modeled in transformed domain, such as Gabor decomposition of both: video frames and optical flow field [*8*] or in a baseband pixel domain, as we showed in [*9*]. Recently, temporal saliency maps were proposed on the basis of residual motion with respect to global model [*7*]. The latter is estimated using image signal on pixel basis. In our work, we take profit of motion information already present in video code-stream. Hence the primary motion features such as macro-block and sub macro-block motion vectors of H.264 are used to estimate the global model. Then the estimated motion is back projected to the smallest sub macro-blocks. The residual motion is computed as a difference of sub macro-block motion vectors and global motion vectors. Hence the temporal saliency map is obtained with lower computation burden. A spatiotemporal saliency map may be produced by combining the spatial and temporal saliency. Spatiotemporal saliency map fusion methods present in the literature remain simple like the sum or the multiplication of both saliency maps S_{mul}^{SP-T}. Typically, to obtain an integrated spatiotemporal saliency map three steps are required. The two first steps consist in extraction of both spatial and temporal saliency maps. The last step is the fusion. Several models which give good results already exist ([*7*], [*4*]) to predict the saliency of a video scene. Thus, we have used the algorithm presented in [*4*] to build the spatial and the temporal saliency maps. In [*9*] we proposed a new method for fusion of saliency maps in a log-space. In this paper we introduce a faster alternative by a squared sum of both spatial and temporal saliency maps. We will denote resulting saliency maps S_{LOG}^{SP-T} and S_{SQUARE}^{SP-T} respectively. The S_{LOG}^{SP-T} [*9*] is defined by (Eq. 1) with $\alpha = 0.5$. This fusion method has the same advantage S_{mul}^{SP-T} [*7*] that gives more importance to regions which have both high spatial and high temporal saliencies. Unlike S_{LOG}^{SP-T}, S_{mul}^{SP-T}provides null spatiotemporal saliency maps when the temporal saliency is very low.

$$S_{LOG}^{SP-T}(s_i) = \alpha \log(S^{SP}(s_i) + 1) + (1-\alpha)\log(S^{T}(s_i) + 1)$$

(Eq. 1)

The squared fusion method S_{SQUARE}^{SP-T} we propose in this paper is defined by (Eq. 2) which has similar fusion properties as S_{LOG}^{SP-T} when the temporal saliency is null. Its advantage is an obvious computational saving.

$$S_{SQUARE}^{SP-T}(s_i) = (S^{SP}(s_i) + S^{T}(s_i))^2 \qquad \text{(Eq. 2)}$$

3. NO-REFERENCE VIDEO QUALITY ASSESSMENT BASED ON SALIENCY MAPS

In this paper we propose a new No-Reference objective video quality metric called Weighted Macro-Block Error Rate (WMBER). This method is based on the identification of transmission errors in the H.264/AVC stream. Hence, each macro-block mb_i is labeled when it contains transmission errors, like wrong DC/AC or motion vector values (Figure 1). Then the error labels are propagated according to the H.264/AVC decoding process (Figure 2). So for each frame i of the video sequence, an error map is associated.

Figure 1 Original transmission error (error areas are tagged in red) Pedestrian Area sequence (TUM/Taurus Media Technik)

Figure 2 Propagation of transmission errors

Afterwards, the norm of the gradient $|\nabla_i|$ is computed and normalized between 0 and 1. For each labeled macro-block, the mean of the normalized norm of the gradient $\|\overline{\nabla_{mb_i}}\|$ is written in a matrix Err_i at the coordinates of the macro-block, thus forming a new matrix W_i. If the macro-block is not labeled, the value 0 is written in Err_i. The matrix Err_i is weighted by matrix W_i. Finally the WMBER is computed by equation (Eq. 3).

$$WMBER_i(W_i) = \frac{\sum_{mb=1}^{N} Err_i(mb) \times W_i(mb)}{\sum_{mb=1}^{N} W_i(mb)} \qquad \text{(Eq. 3)}$$

where N is the number of macro-blocks in the frame.

If the area is flat, the decoder error concealment does not produce an annoying artifact. Five weights matrix W_i are tested in this work: $2DGauss$ which is a simple symmetric Gaussian centered on the frame center, S_{LOG}^{SP-T}, S^{SP}, S_{mul}^{SP-T} and S_{SQUARE}^{SP-T}. Saliency map are computed from the H.264 stream received at the decoder side. The advantage is to take into account the saliency of visual artifacts when the error concealment mechanism has failed. Our assumption that strong visual artifacts are due to bad error concealment is directly confirmed by the work [***10***]. In this work they show the drop-off of subjective score due to the errors external to the region of interest in JPEG compressed images.

4. MOS PREDICTION BY SUPERVISED LEARNING

In our recent work [***9***] we proposed a supervised learning method for prediction of subjective score from objective quality metric. This prediction method requires a training data set of n known pairs (x_i, y_i) to be able to predict y from x. Here (x_i, y_i) pairs are objective metrics output values associated with MOS values from the subjective experiment. y is the predicted MOS from a given objective metric output value x. The prediction is performed using equation (Eq. 4) known as Similarity Weighted Average classifier (Eq. 5).

$$y = \frac{\sum_{i=1}^{n} s(x_i,x) y_i}{\sum_{i=1}^{n} s(x_i,x)} \quad \text{(Eq. 4)}$$

$$s(z,x) = exp[-|x - z|] \quad \text{(Eq. 5)}$$

In the original paper [***11***] the authors show good generalization properties due to the monotonicity of the exponential similarity measure (5), this was a reason for us to choose this prediction scheme. The other reason is that it does not require a heavy training as it is the case of many classifies such as Neuronal Networks and SVMs and proved of be more accurate than the polynomial fitting usually employed [***12***].

5. TESTS AND RESULTS

5.1 SUBJECTIVE EXPERIMENT

We carried out subjective experiments to measure the quality of HDTV video streams transmitted over lossy networks. To get more participants and more reliable results, the experiment was done in two research laboratories: LaBRI (University of Bordeaux) and Communication Systems Engineering Dept. (Ben Gurion University of Negev (BGU)). Twenty different video sequences of 10 seconds were selected to compose a representative sample of broadcasted HDTV programs. The selection of video sequences was done according to two features called spatial and temporal information, described in ITU-T Rec. P.910 [***13***]. Video sequences come from four different corpora: The Open Video Project [***14***], NTIA/ITS [***15***], TUM/Taurus Media Technik [***16***] and French HDTV. According to copyrights, video sequences from the French HDTV corpus are not available outside France.

Video sequences were encoded into the H.264/AVC format [***3***] using the x264 [***17***] software with a bit-rate of 6000Kb/s. Two models of transmission impairments were applied to each video sequence (Table 1). The first one, we called it IP model, simulates IP packet networks according to ITU-T Rec. G.1050 [***18***]. Hence, three kinds of networks: managed, semi-managed and unmanaged were simulated using five packet loss profiles. The second model, we called it RF model, simulates radio frequency transmission impairments by introducing bit corruption in Transport Stream (TS) packets. To simulate the RF model, three levels of bit corruption were chosen. After processing the 20 video sources (SRC) with the 8 impairment profiles, 160 processed video sequences (PVS) were generated. So, the total number of video sequences assessed by the participants of the experiment was 180.

Model	Profile	Loss	Burst
IP	0	0.05%	No
	1	1%	No
	2	1%	Yes
	3	5%	No
	4	5%	Yes
RF	5	0.01%	No
	6	0.1%	No
	7	1%	No

Table 1 Loss profiles

The experiment was carried out by following the ACR-HR experimental protocol described in the VQEG Report [***12***]. The experiment room and the lightning conditions were compliant with the ITU-R Rec. BT.500-11 [***2***]. The distance between the subject head and the screen was three times the height of the screen. The video sequences were displayed with a resolution of 1920x1080 pixels using a HDMI cable. In order to be compliant with ITU-R Rec. BT500.11, the experimentation time was reduced to 30 minutes by splitting the video dataset in two parts. Therefore, each participant has seen only 90 videos, i.e. 10 source (SRC) with the 8 related processed video streams (PVS). The experiment was done with the two video sub-datasets at LaBRI and one video sub-dataset at BGU. To avoid the "leaning effect" each participant has seen the video sequences in a unique order and a "warm-up" session of 5 minutes was done before starting the experiment. Hence for a total of 35 participants, 22 were gathered at LaBRI, i.e. 11 for each sub-set and 13 at BGU. The Mean Opinion Score (MOS) subjective metric was computed by using methods described in [***12***], [***2***] and involving test subjects.

5.2 EVALUATION

In this section, we compare two objective video quality metrics with the results of the subjective experiment described in the previous section. The first one is the Mean Squared Error (MSE) computed between the original non degraded video and a degraded version. It is a Full Reference metric. The second one is WMBER, the proposed method, which is a No-Reference metric. For the WMBER metric we have tested five different weight methods which are $2DGauss$, S_{LOG}^{SP-T}, S^{SP}, S_{mul}^{SP-T} and S_{SQUARE}^{SP-T}. For the 20 SRC and the 160 PVS, a MOS value is computed. The Similarity-Weighted method described in section 5 and the cubic polynomial function are used to predict the MOS. Therefore, to train and evaluate the prediction methods, a dataset of 180 data pairs objective metric/MOS is built for each metric. To validate the results of the metrics, the Kx2 cross-validation method is applied. This method randomly split the dataset into two equal parts, one part is used for training the prediction method and the other is used for the evaluation. Then, the evaluation set is used for training and the training set for evaluation. The process is run five times to validate each metric. The evaluation is performed by computing the Pearson Correlation Coefficient (PCC) (Eq. 6) denoted by R

$$R = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}} \qquad \text{(Eq. 6)}$$

where x_i is the MOS, y_i the predicted MOS and N the number of data pairs in the evaluation dataset. The final performance score is the mean of the 10 R values with the standard deviation.

6. RESULTS

In this section we compare our No-Reference metric WMBER with the Full Reference metric MSE. For both metrics, the MOS values are predicted using Similarity Weighted Average learning method. In [***9***] we have shown that the prediction of MOS from the MSE using Similarity Weighted Average gives very good results. The VQEG Report on the Validation of the Video Quality Models for High Definition Video Content proposes to use a cubic polynomial function to map the objective metrics output values to the MOS. However this method gives poor results compared to the Similarity Weighted Averaged method. In fact, the correlation between the MOS and the MSE is 0.99 when predicted by the Similarity Weighted Averaged and is 0.64 when predicted by the cubic polynomial function. Table 2 and Figure 3 give the comparative results of the 6 objectives metrics.

We compare the correlation of the predicted MOS with the MOS for each metric. MSE provide the best results. Nevertheless the use of a Full–reference metric MSE in the context of HD delivery for quality assessment is totally un-realistic. Hence the proposed No-reference metric has to approach the full-reference measure in the best way. From tests result we observe that the No-Reference metrics $WMBER(S^{SP})$ have the best results followed by $WMBER(S_{SQUARE}^{SP-T})$. To our opinion, the best performance of purely spatial WMBER is due to the error concealment mechanisms of H.264 industrial decoders which is efficient to recover motion vectors. In this case the artifacts due to the transmission are not perceptible as the lost blocks are compensated with a smooth motion field. We suppose that finer ponderation of matrix E with artificial contour map (as in [***10***]) instead of gradient energy will yield finer results. We note that $WMBER(S^{SP})$ and $WMBER(S_{SQUARE}^{SP-T})$ have good correlation with the MOS for No-Reference Quality Assessment protocol.

	IP Model		RF Model	
Metric	PCC	σ_{PCC}	PCC	σ_{PCC}
MSE	0.999	0.001	0.987	0.003
$WMBER(2DGauss)$	0.840	0.010	0.877	0.015
$WMBER(S_{LOG}^{SP-T})$	0.748	0.031	0.763	0.024
$WMBER(S^{SP})$	0.883	0.009	0.900	0.015
$WMBER(S_{mul}^{SP-T})$	0.714	0.024	0.786	0.048
$WMBER(S_{SQUARE}^{SP-T})$	0.860	0.015	0.895	0.013

Table 2 Metrics evaluation results

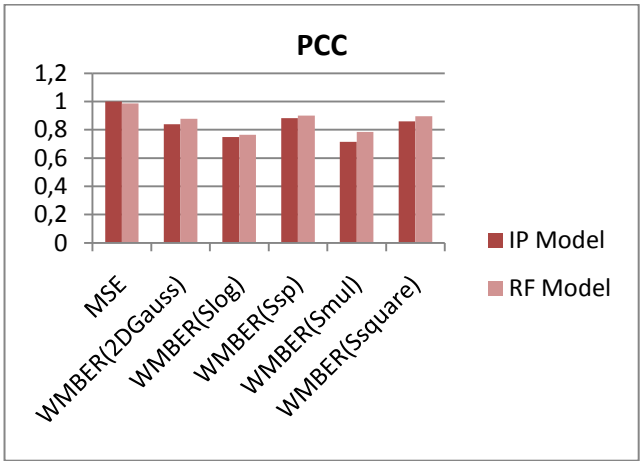

Figure 3 Metrics correlation with MOS

7. CONCLUSION

In this paper we were interested in the essential part of HD video content delivery which is the quality of perceived content. Taking into account modern distribution channels such as packetized networks, we studied the quality assessment for delivered HDTV content delivered in H.264 standard. The goal of this work was to qualify the HDTV broadcasting chain by only using the delivered video content, hence the No-Reference quality metrics. As the human is the target of the delivered HDTV content, we used

saliency maps in our metrics to model the human visual attention. We thus proposed a new no-reference video quality metric based on spatiotemporal saliency maps. The interest of our contribution resides in the fact that both visual saliency maps and no-reference metrics are obtained without the full decoding of compressed video stream. The experiments conducted according to the VQEG evaluation protocol show that the proposed No-Reference metric WMBER is competitive with the Full-Reference classical MSE. Obviously the broadcasting network models we considered are not complete. Only packet loss and RF signal distortion has been taken into account. Those models might be improved by considering jitter and fading. This is an opening for the future. The No-Reference quality assessment is a challenging problem and we are only at the beginning of the road to the success. In the perspective of this work, we will refine the spatial weighting of WMBER. We will be also interested in a “semantic” saliency of content to incorporate it into no-reference visual quality assessment of delivered HDTV content.

REFERENCES

[1] European Broadcasting Union, "Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for digital terrestrial television," ETSI European Standard ETSI EN 300 744 V1.6.1, 2009.

[2] International Telecommunication Union, "ITU-R BT.500-11 Methodology for the subjective assessment of the quality of television pictures," Recommendation, 2002.

[3] ISO/IEC, "H.264 Advanced Video Coding," in Information technology - Coding of audio-visual objects, 2004, ch. Part 10.

[4] O. Brouard, V. Ricordel, and D. Barba, "Cartes de Saillance Spatio-Temporelle basées Contrastes de Couleur et Mouvement Relatif," CORESA, Feb. 2009.

[5] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," Signal Processing: Image Communication, vol. 25, pp. 469-481, Aug. 2010.

[6] Y. Pitrey, M. Barkowsky, P. Le Callet, and R. Pepion, "Subjective Quality Assessment of MPEG-4 Scalable Video Coding in a Mobile Scenario," Visual Information Processing (EUVIP), pp. 86-91, 2010.

[7] S. Marat, et al., "Modelling Spatio-Temporal Saliency To Predict Gaze Direction For Short Videos," IJCV, no. 82, pp. 231-243, Mar. 2009.

[8] K. Seshadrinathan and A. C. Bovik, "Motion Tuned Spatio-temporal Quality Assessment of Natural Videos," IEEE Transactions on Image Processing, vol. 19, no. 2, pp. 335-350, Feb. 2010.

[9] H. Boujut, O. Hadar, J. Benois-Pineau, T. Ahmed, and P. Bonnet, "Weighted-MSE based on Saliency map for assessing video quality of H.264 video streams," IS&T/SPIE Electronic Imaging, Jan. 2011.

[10] T. Shoham, D. Gill, and C. Sharon, "A novel perceptual image quality measure for block based image compression," IS&T/SPIE Electronic Imaging, vol. 7867, Jan. 2011.

[11] A. Billot, I. Gilboa, and D. Schmeidler, "Axiomatization of an exponential similarity function," Mathematical Social Sciences, no. 55, pp. 107-115, 2008.

[12] VQEG (Video Quality Experts Group), "Report on the Validation of Video Quality Models for High Definition Video Content," Report, 2010.

[13] International Telecommunication Union, "ITU-T Rec. P.910 Subjective video quality assessment methods for multimedia applications," Recommendation, 1999.

[14] The Open Video Project. (2010, Nov.) LABRI-ANR ICOS-HD. [Online]. http://www.open-video.org/collection_detail.php?cid=23

[15] NTIA/ITS. (2010, Nov.) VQEG FTP - NTIA source. [Online]. ftp://vqeg.its.bldrdoc.gov/HDTV/NTIA_source/HDTV_Readme.doc

[16] TUM / Taurus Media Technik. (2010, Nov.) HD test sequences Taurus Media Technik. [Online]. ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/test_sequences/1080p/ReadMe_1080p.txt

[17] Videolan. (2010, Nov.) x264 - a free h264/avc encoder. [Online]. http://www.videolan.org/developers/x264.html

[18] International Telecommunication Union, "ITU-T Rec. G.1050 Network model for evaluating multimedia transmission performance over Internet Protocol," Recommendation, 2007.