Benchmark of State of the Art Objective Measures for 3D Stereoscopic Video Quality Assessment on the Nantes Database

Emil Dumic[1], Sonja Grgic[1], David Jiménez Bermejo[2], Luis Alberto da Silva Cruz[3]
[1] Department of Wireless Communications,
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
[2] Grupo de Aplicación de Telecomunicaciones Visuales,
Technical University of Madrid, Spain
[3] Instituto de Telecomunicações,
Department of Electrical and Computer Engineering, University of Coimbra, Portugal
emil.dumic@fer.hr

Abstract **- In this work we present a study on the performance of existing state of the art 2D objective image and video quality measures, tested on the new 3D stereoscopic video NAMA3DS1-COSPAD1 database. Different image and video quality measures have been tested on test material affected by different types of degradations. Results show that in some cases, image quality measures give better results than video quality measures. This paradoxical result means that more effort should be devoted to designing new 3D video quality assessment measures. The results presented and accompanying discussion can be used to motivate and guide future research directed towards definition of effective new 3D stereoscopic video quality measures.**

Keywords **- NAMA3DS1-COSPAD1 database, image quality measures, video quality measures, 3D video, stereoscopic video**

I. INTRODUCTION

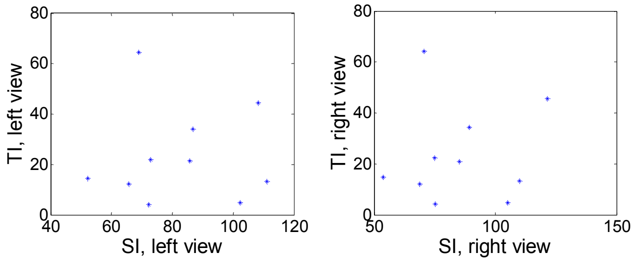
Although several image and 2D video quality measures with good performance have been developed, 'true' 3D video quality measures have been far less researched and proposed, especially for larger combinations of different degradation types. However, similarly to what happens with image quality measures which do not perform very well on video sequences because of missing correlation between successive frames in time [1], 2D stereo video quality measures applied to 3D video also suffer from performance problems related to the reduced correlation between left and right views which can occur in connection to some types of degradations. This study presents an empirical study on the performance of existing image and 2D video measures applied to stereoscopic video which has been subjected to specific types of degradations and subjectively evaluated, and assembled into a test dataset known as the Nantes *NAMA3DS1-COSPAD1* database. The cursory analysis of the results singles out some problems of the direct application of those measures, providing useful information about the applicability of each evaluated measure to each degradation type considered. This type of work is fundamental to identify problems deserving further work as well as quantifying the nature and magnitude of the non-adequacies that result from a direct extension of the measures studied to 3D contents.

II. RELATED WORK

Since 3D video became a foreseeable format of interest for visual information diffusion, the question of end-user perceived quality started attracting the attention of video coding and transmission experts in academia and industry. Works were published on the development and evaluation of stereo-video quality measures aimed at predicting the final quality after the effects of compression-related degradations and transmission impairments. Hewage et al. presented in [2] a study on the effectiveness of several measures derived from 2D video measures applied to 3D stereo video rendered from colour+depth 3D video. This study established that measures such as PSNR, SSIM and VQM when applied to the colour component or to the left and right views exhibited a reasonably high correlation with image quality and depth perception subjective scores, with room to improvement. In [3] Benoit et al. present and in-depth review of works dealing with the factors which impact perceived 3D image quality, also paying attention to the effects of display technology on final quality. Although a significant amount of the work is devoted to performance of objective 3D image quality measures and their correlations with subjective scores, this work did not extend into 3D video evaluation. In [4] Aflaki et al. look into the effects of asymmetric quality and resolution of 3D stereo video on final quality and after subjective evaluation of mixed resolution stereo video propose an logarithmic equation to model the relationship between subjective scores and downsampling ratios. More recently Le Callet and collaborators proposed in [5] a new 3D stereo video dataset which includes original content with varied characteristics and degradations together with subjective evaluation scores of the dataset components, pointing out future avenues for research. Some of these suggested research topics were explored in [6] and [7] based on experiments conducted by different research groups in a collaborative effort.

III. Description of 3D Nantes Database

To be able to compare the different image and video quality measures under evaluation, we used the NAMA3DS1-COSPAD1 stereoscopic video database [5] comprising 10 original and 100 degraded stereoscopic sequences with 10 degradation types available from [8]. The sequences feature 1920×1080 progressive Full HD resolution per view and 25 frames per second. Two of the sequences have two scenes (with one scene cut), while the others have just one scene. It should also be noted that database has an overall size of around 333 GB. Three of the degradations represented in the database were made by applying H.264 compression to the original sequences, four by using J2K compression, an eighth degradation was introduced by downsampling, a ninth degradation was the result of an edge enhancement operation and the last one is a combination of downsampling and edge enhancement related degradations. The dynamic characteristics of the ten reference sequences (including scene cuts), measured by spatial and temporal activity indices, were computed for the left and right view according to the procedure defined in ITU-T recommendation P.910 [9]. The sequences' activity indices are plotted in Fig. 1 a) (left view) and Fig. 1 b) (right view) whose analysis shows that the sequences are very diverse in terms of their dynamic characteristics.



(a) (b)

Figure 1. Spatial versus temporal information: (a) left view; (b) right view

To obtain the subjective scores available as part of the database some evaluation sessions were conducted with the observers seated in a standardized room equipped with a Philips 46PFL9705H 46" stereoscopic display with shutter glasses was used to display the sequences. An ensemble of 29 observers, aged from 18 to 63 years old, evaluated the 110 different video sequences following an Absolute Category Rating-Hidden Reference (ACR-HR) methodology and quality scale. According to the study authors, more degradation types and respective scores are planned to be included in the database [10].

IV. Quality Measures Evaluated

So far several image and 2D video measures have been compared with the MOS (Mean Opinion Score) scores from the previously described 3D database: MS-SSIM (Multiscale Structural Similarity Index) [11], NQM (Noise Quality Measure) [12], PSNR (Peak Signal to Noise Ratio), SSIM (Structural Similarity Index) [13], VIF (Visual Information Fidelity) [14], VSNR (Visual Signal-to-Noise Ratio) [15], IW-SSIM (Information Weighted SSIM) [16], IW-PSNR (Information Weighted PSNR) [16], VQM (Video Quality Measure) [17], VQM_VFD (VQM with Variable Frame Delay) [18], IQM2 (Image Quality Measure 2) [19], RVQM (Reduced Video Quality Measure) [1] and PARMENIA [20]. Most of the measures can be obtained from [21], while VQM can be downloaded per request from [22]. VQM, RVQM and PARMENIA are 2D video quality measures, while others are image quality measures. Final grade of degraded 3D sequence was calculated from image quality measures as average grade over grades from all frames and both views, while grade from 2D video quality measures was calculated as average grade from grades from left and right view.

VQM was calculated using 'general' model and 'frcal' as calibration methods. Also, an experimental VQM measure based on neural network and with variable frame delay (VFD) processing was tested, in this case using 'vqm_vfd' and 'frcal' as calibration methods. However, due to computational limitations of the computer used in this work, neural network based VQM could be tested only on sequences with half resolution (960x540 pixels). These reduced resolution sequences were obtained from the full resolution original sources using nearest neighbor interpolation.

IQM2 was calculated using SPWT (Steerable Pyramid Wavelet Transform) with 2 orientations as recommended [19]. On each scale and orientation of SPWT transform modified SSIM (only structure and contrast component) was applied and final measure was calculated by multiplying modified SSIM grades over all subbands.

RVQM was calculated using 272 (width) x 272 (height) x 32 (frame) pixels, step size 16 pixels, 1st order and 3rd component of the Riesz transform. On each scale modified SSIM (only structure and contrast component) in 3 dimensions was applied and final measure was calculated by multiplying modified SSIM grades over all subbands. With increasing frame size (width x height), results for RVQM were just slightly better, while calculation time was much higher.

PARMENIA is a full-reference metric that is based on the pooling of different quality ratios that are obtained from three different approaches: Beucher's gradient, local contrast filtering and contrast and homogeneity Haralick's texture features.

V. Results

A. *Overall results*

The quality estimates produced by the previously described image and video quality measures have been compared to the database subjective scores using Spearman's rank order correlation. This correlation measure assesses how well an arbitrary monotonic function can describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Spearman's correlation coefficient is calculated like Pearson's correlation but over ranked variables. Overall results are presented in Fig. 2 from which observation it can be concluded that from all the tested measures, the best results were obtained using IW-SSIM and VIF. The VQM and IQM2 measures have correlation values above 0.8, with a peak value of 0.91 for VIF followed by 0.88 for IW-SSIM, but all other have lower correlation results, going as low as 0.57 for NQM. It should be noted that out of the 100 degraded sequences, 70 are degraded using compression, either H.264/AVC or J2K. Because all of the measures have lower correlation on the last three degradation types (downsampling, edge enhancement and their combination), overall correlation results may be misleading. This observation implies that the underlying quality measures produced by the estimators may be by either too pessimistic or to optimist, depending on the degradation mix of the "real-life" sequence which quality one wants to grade. This over-tuning of the image and 2D video quality measures to specific degradation types is a well-known problem that also affects measures applied to stereoscopic video contents, as shown in the following section.

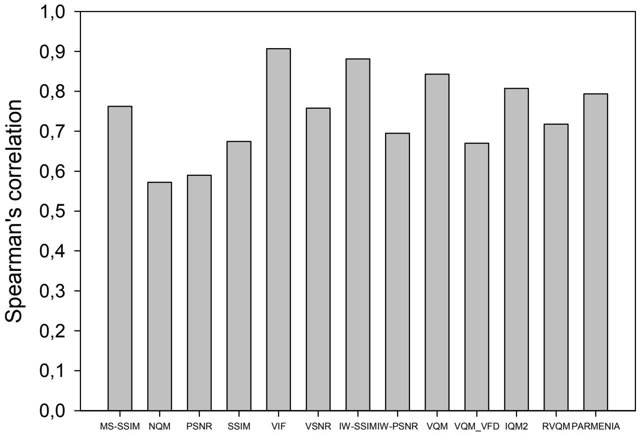


Figure 2. Spearman's correlation between different quality measures and MOS, 3D Nantes database

Also, some of the measures (e.g. VIF, IW-SSIM, NQM) had problems with calculation of some video sequences, which consist of 2 scenes: this means some preprocessing technique should be used firstly. Preprocessing technique could detect possible scene change and cut out black frames in between, which were problematic for calculation of some objective measures. In the presented results from all image quality measures, we manually omitted black frames from calculation.

B. *Results for separate degradation types*

In this section we present and discuss Spearman's correlation values calculated separately for the quality measures listed before and applied to the sequences affected only by the following degradation types: H.264 compression (30 sequences), JPEG 2000 compression (40 sequences) and reduction of resolution, sharpening and combination of resolution reduction and sharpening (30 sequences).

Results can be used to compare which of the objective measures will provide better correlation with MOS, for each specific degradation type, Tab. 1.

TABLE I. SPEARMAN'S CORRELATION FOR SEPARATE DEGRADATION TYPES

	H.264 (30 degradations)	JPEG 2000 (40 degradations)	Resolution reduction, sharpening and their combination (30 degradations)	Overall correlation (100 degradations)
MS-SSIM	0.6112	0.8448	0.6541	0.7624
NQM	0.6676	0.8557	0.1261	0.5719
PSNR	0.5064	0.7916	0.4111	0.5898
SSIM	0.7257	0.8854	0.2665	0.6745
VIF	0.8726	0.9043	0.7087	**0.9067**
VSNR	0.7041	0.9254	**0.7544**	0.7581
IW-SSIM	0.8396	0.9225	0.7464	0.8811
IW-PSNR	0.8347	**0.9408**	0.0189	0.6947
VQM	0.8804	0.9349	0.7513	0.8428
VQM_VFD*	**0.8906**	0.8904	0.1014	0.6700
IQM2	0.8033	0.8949	0.6811	0.8073
RVQM	0.6012	0.8208	0.6289	0.7176
PARMENIA	0.6464	0.7501	0.6015	0.7934

In the results VQM_VFD* means VQM_VFD tested on half of the Full HD resolution (as explained before). Based on these results and some reasoning we can speculate that VQM_VFD applied to the full resolution sequences would produce higher quality ratings. In light of these observations and of the results presented in Tab. 1, it can be concluded that the best results were obtained by VQM_VFD for the cases of degradation by H.264 compression, IW-PSNR for degradations by JPEG2000 compression and VSNR measure for non-compression related artifacts.

Nearly all measures show better correlation values for compression related degradations. When evaluating the correlation results for degradations related to resolution reduction, sharpening and their combination (30 degraded sequences), general results are much worse: some of the measures do not correlate at all with the subjective scores (NQM, SSIM, IW-PSNR and VQM_VFD have correlation below 0.3), while the best correlation is obtained by VSNR measure with a rather low value of 0.7544. Probably because of that, overall correlation is lower for some of the tested objective measures. A good example would be VQM_VFD which gives best correlation in H.264 compression, but does not correlate at all in non-compression related artifacts, thus achieving lower overall correlation results.

When comparing H.264 versus JPEG 2000 compression, it can be seen that nearly every measure provides better correlation with JPEG 2000, than with H.264 (VQM_VFD gives practically the same correlation in both degradations). This may be due to the higher complexity and magnitude of the H.264 compression related visual defects and resulting measure errors.

C. *Computation time*

Computation times for all above mentioned measures are given in Fig. 3. The computer used in the work was equipped an Intel i7-4770 (3.4 GHz) processor, 16 GB of RAM and running Windows 7 and Matlab 2012a. Timing information is presented for the first sequence (which consists of 400 frames) and first degradation type (H.264, QP=32). Image quality measures were calculated by first loading each video frame (left and right views) into memory and then the average score was calculated. Successive reading of the video file (frame by frame) may also need additional time which slows the computation of the measures.

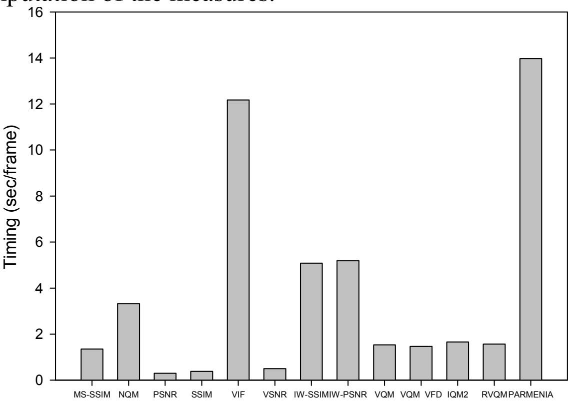


Figure 3. Timing (seconds per frame, both left and right) for all tested objective measures, 3D Nantes database

VI. Conclusions

In this paper we tested existing 2D objective image and video quality measures on new 3D Nantes (NAMA3DS1-COSPAD1) stereoscopic video database. From overall correlation results obtained and discussed before, we conclude readily that the best results were obtained using IW-SSIM and VIF measures, from among all the measures tested. When comparing each of the degradation types separately, best results were obtained for VQM_VFD in H.264 compression, IW-PSNR for JPEG2000 compression and VSNR measure for non-compression related artifacts. From the calculation time results, it can be seen that best performing measures were also the slowest, making them unusable in any real-time application. Overall results show that the performance of the measures is not very good, hinting that further study on the extension of existing image and 2D video quality measures to 3D stereoscopic video is needed. A second conclusion is that current measures do not perform well on all types of degradations (best Spearman's correlation for non-compression related artifacts was about 0.75) thus preventing their use in a non-controlled way agnostic to the degradation mix.

Therefore we are of the opinion that in the future, new 3D objective quality measures should be researched and proposed, designed with particular care on what concerns their computation complexity and performance on stereoscopic video affected by different types of degradation.

Acknowledgment

We would like to acknowledge the financial support provided by FCT and Instituto de Telecomunicações (Portugal) to research project IT/LA/P01131/2011.

This work was supported in part by the project HORFI, TEC2012-38402-C04-01, which is partially funded by the Spanish Ministry of Science and Innovation.

References

[1] E. Dumic, S. Grgic, "Reduced Video Quality Measure Based on 3D Steerable Wavelet Transform and Modified Structural Similarity Index", Proc. of the 55th International Symposium ELMAR-2013, pp. 65-69, 2013.

[2] C.T.E.R. Hewage, S.T. Worrall, S. Dogan and A.M. Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-D video", Electronic Letters, Vol. 44, No. 16, pp. 963 – 965, 2008.

[3] A. Benoit, P. Le Callet, P. Campisi and R. Cousseau, "Quality Assessment of Stereoscopic Images", EURASIP Journal on Image and Video Processing, Volume 2008, Article ID 659024, 13 pages, 2008.

[4] P. Aflaki, M. M. Hannuksela, J. Hakala, J. Häkkinenb and M. Gabbouj, "Estimation of subjective quality for mixed-resoultion stereoscopic video", 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2011, pp. 1-4, 2011.

[5] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutierrez, N. Garcia, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences", Quality of Multimedia Experience (QoMEX), pp. 109-114, 2012.

[6] E. Bosc, R. Pepion, Patrick Le Callet, M. Pressigout, L. Morin, "Reliability of 2D quality assessment methods for synthesized views evaluation in stereoscopic viewing conditions", 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2012, pp. 1 – 4, 2012.

[7] K. Wang, M. Barkowsky, K. Brunnström, M. Sjöström, R. Cousseau and P. Le Callet, "Perceived 3D TV Transmission Quality Assessment: Multi-Laboratory Results Using Absolute Category Rating on Quality of Experience Scale", IEEE Trans. on Broadcasting, Vol. 58, No. 4, 2012.

[8] ftp://ftp.ivc.polytech.univ-nantes.fr/NAMA3DS1_COSPAD1/

[9] ITU-T, "Recommendation P.910, Subjective video quality assessment methods for multimedia applications", 2008.

[10] ftp://vqeg.its.bldrdoc.gov/Documents/Projects/3dtv/VQEG_3DTV_2012_113_grotruqoe3d1_draft_100_v1R1.docx

[11] Z. Wang, E.P. Simoncelli, A.C. Bovik, "Multiscale structural similarity for image quality assessment", 37th Proc. IEEE Asilomar Conf. on Signals, Systems and Computers, Vol. 2, pp. 1398-1402., 2003.

[12] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans and A. Bovik, "Image Quality Assessment Based on a Degradation Model", IEEE Trans. on Image Processing, Vol. 9, No. 4, pp. 636-650., 2000.

[13] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", IEEE Trans. on Image Proc., Vol. 13, No. 4, pp. 600-612., 2004.

[14] H.R. Sheikh and A.C. Bovik, "Image information and visual quality", IEEE Trans. Image Processing, Vol. 15, No. 2, pp. 430-444., 2006.

[15] D.M. Chandler and S.S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images", IEEE Transactions on Image Processing, Vol. 16, No. 9, pp. 2284-2298., 2007.

[16] Z. Wang, Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment", IEEE Trans. on Image Processing, Vol. 20, No. 5, pp. 1185-1198., 2011.

[17] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality", IEEE Trans. Broadcast, Vol. 50, No. 3, pp. 312–322, 2004.

[18] M. H. Pinson and S. Wolf, "Video Quality Model for Variable Frame Delay (VQM_VFD)", NTIA Technical Memorandum TM-11-482

[19] E. Dumic, S. Grgic and M. Grgic, "IQM2 - New image quality measure based on steerable pyramid wavelet transform and structural similarity index", Signal, Image and Video Processing, DOI: 10.1007/s11760-014-0654-3, 2014.

[20] D. Jiménez, "High definition video quality assessment metric built upon full reference ratios" PhD. Thesis. Available: http://oa.upm.es/14712/

[21] http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[22] http://www.its.bldrdoc.gov/resources/video-quality-research/request-software