RELIABILITY OF 2D QUALITY ASSESSMENT METHODS FOR SYNTHESIZED VIEWS EVALUATION IN STEREOSCOPIC VIEWING CONDITIONS

Emilie Bosc, Romuald Pépion, Patrick Le Callet, Muriel Pressigout, Luce Morin

RELIABILITY OF 2D QUALITY ASSESSMENT METHODS FOR SYNTHESIZED VIEWS EVALUATION IN STEREOSCOPIC VIEWING CONDITIONS

Emilie Bosc[1], *Romuald Pépion*[2], *Patrick Le Callet*[2], *Muriel Pressigout*[1], *Luce Morin*[1]

[1]Université Européenne de Bretagne, INSA de Rennes, IETR, UMR 6164, F-35708, Rennes, France

[2] LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597

Polytech Nantes, rue Christian Pauc BP 50609 44306 Nantes Cedex 3, France

ABSTRACT

This paper investigates the reliability of objective quality metrics commonly used for the quality assessment of 2D media, in the context of 3D Video. In the absence of any dedicated tool for the evaluation of synthesized views quality, we often rely on available 2D metrics for direct evaluation of 3D media quality, or with some adaptation to the 3D case. However, recent studies showed that the use of DIBR, depending on its in-painting strategy, can lead to downsides whose range in terms of quality, has never been experienced with 2D media. This paper questions the reliability of the objective quality metrics normally used for the quality assessment, when assessing stereopairs. Seven DIBR algorithms are used to generate novel viewpoints. A series of commonly used quality metrics then assess their quality. The results of our experiments showed that the metrics are not sufficient to faithfully predict human judgment. Moreover, we compared our results with an experiment run in monoscopic viewing condition and the differences are for the less unexpected since the preferred artifacts in monoscopic condition are the most rejected in stereoscopic condition. This paper proposes some explanations.

Index Terms **—** 3DTV, MVD, 3D video, DIBR, quality assessment.

1. INTRODUCTION

The ability to provide novel viewpoints is imperative in 3D Video applications such as 3D Television (3D-TV) that provides a depth feeling, and Free Viewpoint Video (FVV), that allows navigation inside the scene. Depth-Image-Based-Rendering [1] algorithms can be used in order to generate novel viewpoints of the same scene from Multi-view-Video-plus-Depth (MVD) data [2]. A recent study [3] showed that depending on the in-painting strategy, DIBR algorithms could induce specific artifacts whose annoyance can drop the overall media perceived quality. The success of 3D Video applications depends on the ability of the systems to provide high quality synthesized views and visually comfortable contents. These conditions should be controlled and measured through reliable tools. Yet, up to now, there is no standardized quality assessment framework for 3D media, despite the many efforts addressing this issue.

Several proposals for new objective or subjective quality evaluation methods rely on 2D existing tools [4, 5, 6]. However, the main criticism of the following examples in that they propose enhanced 2D objective evaluation tools for the assessment of stereoscopic contents while the reliability of these metrics have not been showed in the specific case of stereopairs containing a DIBR generated view. In numerous proposals, experimental protocols often involve depth and/or color compression, different 3D displays, and different 3D representations (2D+Z, stereoscopic video, MVD, etc...). In these cases, the quality scores obtained from subjective assessments are compared to the quality scores obtained through objective measurements, in order to find a correlation and validate the objective metric. The experimental protocols often assess both compression distortion and synthesis distortion, at the same time without distinction. This is problematic because there may be a combination of artifacts from various sources (compression and synthesis) whose effects are neither understood nor assessed.

In [4], Benoit *et al.* proposed a quality metric for the assessment of stereopairs using fusion of 2D quality metrics and depth information. Well-known 2D metrics, either Structural SIMilarity (SSIM)[7] or C4[8], are applied separately on each image (left and right view) and the scores are combined to obtain one overall score for the given stereopair. By taking into account the stereodisparity in their measure, Benoit *et al.* point out the fact that 2D metrics have limitations when assessing stereoscopic image quality, since SSIM is enhanced when adding the disparity distortion contribution. You *et al.*, in [9] reach the same conclusion regarding the use of disparity in the quality score of stereoscopic data. In this study [4], the proposed metric is not evaluated in the case of stereopairs containing a DIBR generated view.

In [5], the authors proposed a depth-based perceptual quality metric. It is a tool that can be applied to Peak-Signal-to-Noise-Ratio (PSNR) or SSIM. The method uses a weighting function based on depth data at the target viewpoint, and a temporal consistency function to take the motion activity into account. The final score includes a factor that considers non-moving background objects during view synthesis. Inputs of the method are the original depth map (uncompressed), the original color view (originally acquired, uncompressed) and the synthesized view. Validation of the performances is achieved by synthesizing different viewpoints from distorted data: color views suffer two levels of quantization distortion; depth data suffer four different types of distortion (quantization, low pass filtering, borders shifting, and artificial local spot errors in certain regions). The study [5] shows that the proposed method enhances the correlation of PSNR and SSIM to subjective scores.

Conze *et al.*[6] proposed a full-reference objective quality assessment metric that targets artifacts related to view synthesis. More precisely, their method relies on the observation that thin objects, object borders, transparency, variations of illumination or color differences between left and right views, periodic objects are the most critical elements to be rendered through DIBR. Their method

is known as the View Synthesis Quality Assessment (VSQA) and is defined as an extension of any existing 2D image quality. In [6], VSQA is used as an extension of SSIM. VSQA considers features of the spatial environment and the complexity in terms of textures, the diversity of gradient orientations and the presence of high contrast of the synthesized views.
Regarding the correlation scores with the subjective scores, these studies [4, 5, 6] showed improvements of 2D existing tools. Nevertheless, these studies come before the essential step evaluating the performances of these 2D existing tools in the presence of DIBR related artifacts. Yasakethu *et al.* [10] and You *et al.* [9] already proposed an evaluation of 2D usual image/video quality metrics for the assessment of stereoscopic contents. However, the tested material did not contain any DIBR related artifacts and only targeted compression related artifacts. In [11], Hanhart *et al.* investigated the correlation between PSNR-based quality metrics and the perceived quality of asymetric stereopairs (made of a decoded view and a synthesized view), in the context of the MPEG 3D video coding standardization effort. The results of the study showed that in this scenario, the PSNR of the decoded view and/or the average PSNR of both views were appropriate for predicting the perceived quality of the stereopairs. In a previous study [3], usual assessment methods have been evaluated but the targeted case of use included monoscopic viewing conditions only.
In this paper, we propose an evaluation of 2D usual objective quality metrics, and usual subjective quality assessment methods when rating stereoscopic contents in the context of DIBR, based on the database used in [3]. The paper is organized as follows. Sec. 2 presents the previous results of the study in monoscopic viewing conditions. Sec. 3 presents the experimental protocol of our study in stereoscopic viewing conditions. Sec. 4 gives the results of the study and Sec. 5 concludes the paper.

2. PREVIOUS WORK

In a previous study [3], the relevance of the use of 2D usual quality assessment methods has been questioned when addressing the quality of DIBR synthesized views. This study was motivated by the fact that the synthesis process and the inpainting strategies induce specific distortions that are different from the commonly encountered artifacts in 2D imaging system. Given this fact, the evaluation of usual quality assessment methods should be considered when dealing with synthesis distortions. DIBR related artifacts only were included in the tests. The three test MVD sequences are Book Arrival (1024×768, 16 cameras with 6.5cm spacing), Lovebird1 (1024×768, 12 cameras with 3.5 cm spacing) and Newspaper (1024×768, 9 cameras with 5 cm spacing). Seven DIBR algorithms processed the three sequences to generate four different viewpoints per sequence (84 synthesized views in total). These seven DIBR algorithms are labeled from A1 to A7:
-A1: based on Fehn [1], where the depth map is pre-processed by a low-pass filter. Borders are cropped, and then an interpolation is processed to reach the original size.
-A2: based on Fehn [1]. Borders are inpainted.
- A3: Tanimoto *et al.* [12]. It is the recently adopted reference software for the experiments in the 3D Video group of MPEG.
-A4: Müller *et al.*[13] proposed a hole-filling method aided by depth information.
- A5: Ndjiki-Nya *et al.* [14]. The hole-filling method is a patch-based texture synthesis.
-A6: Köppel *et al.*[15] uses depth temporal information to improve the synthesis in the disoccluded areas.
-A7: corresponds to the unfilled sequences (i.e. with holes).
Test images were "key" frames ("keys" were randomly chosen from the same set of synthesized views). That is to say that for each of the three reference sequences, only one frame was selected out of each synthesized viewpoint, for assessing still-images. ACR-HR (Absolute Categorical Rating with Hidden Reference Removal) [16] methodology was used: the quality of each image is rated independently on a category scale (5: excellent; 4: good; 3: fair; 2: poor; 1: bad), including the reference version of each image.
The subjective evaluations were conducted in an ITU conforming test environment. The stimuli were displayed on a TVLogic LVM401W, and according to ITU-T BT.500 [17]. Objective measurements were obtained by using MetriX MuX Visual Quality Assessment Package [18].
The results of this previous study in monoscopic viewing conditions showed that usual objective assessments do not correlate with subjective assessments. Rankings of algorithms from objective metrics are not reliable, considering the differences with the obtained subjective results. The presented experiments revealed that using only the objective metrics seems not sufficient for assessing virtual synthesized views, though they give information on the presence of errors.
In this paper, we extend the previous study by questioning the reliability of 2D usual image quality metrics when assessing stereoscopic pairs containing DIBR related artifacts. The next sectionsl present the new experimental protocol and the results of the study.

3. EXPERIMENTAL PROTOCOL

In this experiment, synthesized still image quality is evaluated in stereoscopic conditions. In these conditions, we have several objectives. First, we aim at determining whether ACR-HR methodology is appropriate for the assessment of different DIBR algorithms; Second, the required number of participants enabling a reliable subjective assessment test is questioned; Third, we investigate whether the results of the subjective assessments are consistent with the objective evaluations; Fourth, we need to compare the obtained results to the monoscopic conditions results.
The material comes from the same set of synthesized views as described in Section 2. The stereopairs consist of two stereo-compliant views. One view is the original acquired frame and the other is a synthesized frame. All the synthesized frames used in this experiment are exactly the same as those used in the previous study (in monoscopic viewing conditions, with still-images).
ACR-HR methodology was used with 25 naive observers. The stimuli were displayed on an Acer GD245HQ screen, with NVIDIA 3D Vision Controller.
The objective measurements were realized over 84 synthesized views of the 84 tested stereopairs by the means of MetriX MuX Visual Quality Assessment Package [18] software. The reference was the original acquired image.
The following section describes the results of the study. The first part addresses the results of the subjective assessments and the second part presents the results of the objective evaluations.

4. RESULTS AND DISCUSSION

4.1. Subjective tests

The seven DIBR algorithms are ranked according to the obtained ACR-HR DMOS scores, as depicted in Table 1. In each section

of the table (monoscopic or stereoscopic viewing), the first line gives the DMOS scores obtained through the MOS scores. The second line gives the ranking of the algorithms, obtained through the first line. The first comment regarding the results in Table 1 refers to the fact that the rankings of the algorithm, according to the subjective scores, are completely different from the rankings obtained in monoscopic conditions. In particular, A1 was ranked as the best algorithm in monoscopic conditions. It is ranked as 5^{th} in stereoscopic conditions. This can be explained by the discomfort produced by the proposed stereopairs. Indeed the stereopairs presented to the observers were made up of one original acquired view and its stereo-compliant synthesized view. Considering the interpolation strategy used in A1 (the borders of the image are cropped and then an interpolation allows to reach the original size of the image), we can observe that objects are shifted. This shift is assumed to be the cause of discomfort, since corresponding objects will have too large disparity values. On the other hand, algorithms that were not ranked as the best in monoscopic conditions are better ranked in stereoscopic conditions. For instance, A6 or A3 that were ranked 4^{th} and 6^{th} respectively by the subjective scores in monoscopic conditions, are ranked 2^{nd} and 3^{rd} in stereoscopic conditions. The assumption is that the artifacts generated by these algorithms are more easily masked through human vision in stereoscopic conditions. They do not induce difficult-to-deal-with artifacts (in stereovision) such as shifts of objects.

Table 2 give the results of the Student's test from the ACR-HR scores for stereoscopic still images. This provides knowledge on the statistical equivalence of the algorithms. The number in parentheses indicates the minimum required number of observers that allows statistical distinction (VQEG recommends 24 participants as a minimum in the Multimedia test Plan [19], values in bold are higher than 24 in the table). In most of the cases, less than 24 observers gave clear-cut decisions between the algorithms. This is very different from the results obtained in monoscopic viewing conditions since six cases were recorded as not statistically different. Our assumption is that artifacts are differently perceived in stereoscopic viewing and in monsocopic viewing conditions. Artifacts that are not annoying in monoscopic viewing conditions, may be disturbing in stereoscopic conditions which explains the clear-cut decisions.

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**
Monoscopic ACR-HR	3.572	3.308	3.145	3.401	3.496	3.32	2.277
Rank order	1	5	6	3	2	4	7
Stereoscopic ACR-HR	3.647	3.637	3.660	3.678	3.658	3.662	3.548
Rank order	5	6	3	1	4	2	7

Table 1. Rankings of algorithms according to subjective scores. First section: results in monoscopic viewing conditions. Second line: results in stereoscopic viewing conditions.

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		**o(>25)**	**o(>25)**	↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A2$	**o(>25)**		↓(<24)	↓ (<24)	↓(<24)	↓(<24)	↑(<24)
$A3$	**o(>25)**	↑(<24)		↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A4$	↑(<24)	↑(<24)	↑(<24)		**o(>25)**	**o(>25)**	↑(<24)
$A5$	↑(<24)	↑(<24)	↑(<24)	**o(>25)**		**o(>25)**	↑(<24)
$A6$	↑(<24)	↑(<24)	↑(<24)	**o(>25)**	**o(>25)**		↑(<24)
$A7$	↓(<24)	↓(<24)	↓(<24)	↓ (<24)	↓(<24)	↓(<24)	

Table 2. Results of Student's t-test with ACR-HR results (stereo still images).

4.2. Objective measurements

This subsection addresses the objective quality measurements conducted over the selected "key" frames. The objective is to determine the consistency between the subjective assessments and the objective evaluations, and the most consistent objective metric.

The first step consists in comparing the objective scores with the subjective scores (previously presented). The consistency between objective and subjective measures is evaluated by calculating the Pearson linear correlation coefficients (PLCC) for the whole fitted measured points.

The coefficients are depicted in Table 3. Compared to in monoscopic viewing conditions, the metrics are still not highly correlated to human judgment. Fig. 1 illustrates the differences between the PLCC obtained in monoscopic conditions and those obtained in stereoscopic conditions and confirms the previous comment. Depending on the objective metric, we observe that the PLCC is slightly improved.

Fig 2 depicts the rankings of the algorithms obtained from the subjective and the objective scores through graphical tables. First line recalls the subjective results obtained in monoscopic viewing conditions. Second line gives the ACR-HR results in the extended study, i.e. in stereoscopic viewing conditions. The darker the blue the better ranked by the metric. The lighter the blue the worse ranked by the metric. In stereoscopic conditions, we observe that the rankings from the objective metrics are slightly closer to human judgment than in monoscopic conditions. If the assumption that a masking effect occurs in stereoscopic viewing conditions, explaining the differences between ACR-HR in monoscopic viewing conditions and ACR-HR in stereoscopic viewing conditions, then the fact that objective metrics are slightly closer to human judgment in stereoscopic conditions confirms that particular distortions related to DIBR algorithms are not taken into account by 2D usual objective measures.

However, this study also suggests that more reliable objective tools are required for assessing stereoscopic pairs including DIBR synthesized views.

	PSNR	**SSIM**	**MSSIM**	**VSNR**	**VIF**	**VIFP**
PLCC	46.98	45.06	**60.86**	26.44	38.46	42.96
	UQI	**IFC**	**NQM**	**WSN**	**PSNR**$_{HVSM}$	**PSNR**$_{HVS}$
PLCC	31.72	40.96	**52.66**	**51.58**	46.59	46.13

Table 3. Pearson correlation coefficients between DMOS and objective scores in percentage (stereoscopic still images).

5. CONCLUSION

In this paper, the reliability of usual image quality methods has been questioned in the case of stereoscopic viewing. Compared to the previous study in monoscopic viewing conditions, the obtained results showed noticeable differences. In particular, the subjective tests suggested that non-annoying artifacts in monoscopic viewing conditions may not be assessed with the same quality in stereoscopic viewing conditions especially in the case of shifting artifacts. ACR-HR methodology led to more clear-cut decisions in stereoscopic conditions which we assume to be related to visual discomfort induced by some DIBR distortions. Correlation of objective metrics with subjective scores is not proved in stereoscopic viewing conditions, though they are closer to subjective scores in stereoscopic viewing conditions than subjective scores in monoscopic viewing conditions. The results suggest that the tested objective metrics are not sufficient to predict the quality of stereoscopic images and that new objective assessment tools are required. This study also brought up the problem of non-matching

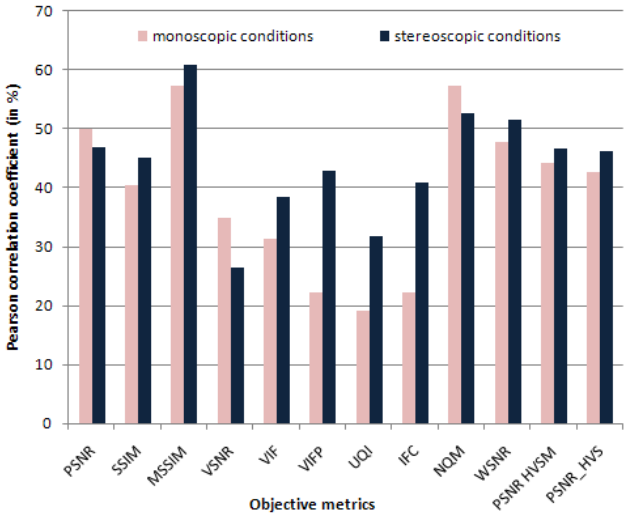

Figure 1. Comparison of Pearson linear correlation coefficients in monoscopic and stereoscopic conditions.

Rankings in stereoscopic viewing conditions

	A1	A5	A4	A2	A6	A3	A7
ACR-HR monoscopic	1	2	3	4	5	6	7
ACR-HR stereoscopic	5	1	3	6	2	4	7
PSNR	7	2	3	4	1	5	6
SSIM	7	1	1	4	3	6	5
MSSIM	7	2	1	4	2	6	5
VSNR	7	1	3	5	2	6	4
VIF	7	2	2	5	1	6	4
VIFP	7	1	1	5	1	6	4
UQI	7	3	1	5	1	6	4
IFC	7	3	2	5	1	6	4
NQM	7	2	3	4	1	5	6
WSNR	7	2	3	4	1	5	6
PSNR HSVM	7	2	3	4	1	5	6
PSNR HSV	7	2	3	4	1	5	6

BEST

WORST

Figure 2. Rankings according to measurements (stereoscopic still images).

features that may occur when views of stereopairs are not coherent: in our study, depending on the DIBR distortions, we assume that binocular rivalry or binocular suppression may have occurred, explaining the differences between subjective scores obtained in monoscopic viewing and those obtained in stereoscopic viewing conditions.

6. ACKNOWLEDGMENTS

This work is supported by the French National Research Agency as part of PERSEE project (ANR-09-BLAN-0170). We would like to acknowledge the Fraunoffer HHI for providing the synthesized views.

7. REFERENCES

[1] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in *Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XI*, 2004, vol. 5291, pp. 93–104.

[2] A. Smolic, K. Mueller, P. Merkle, N. Atzpadin, C. Fehn, M. Mueller, O. Schreer, R. Tanger, P. Kauff, and T. Wiegand, "Multi-view video plus depth (MVD) format for advanced 3D video systems," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q. 6, JVT- W*, pp. 21–27, 2007.

[3] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-D synthesized view assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1332–1343, 2011.

[4] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP Journal on Image and Video Processing*, vol. 2008, 2009.

[5] E. Ekmekcioglu, S. T. Worrall, D. De Silva, W. A. C. Fernando, and A. M. Kondoz, "Depth based perceptual quality assessment for synthesized camera viewpoints," in *Proc. of Second International Conference on User Centric Media, UCMedia 2010*, 2010.

[6] P.-H.i Conze, P. Robert, and L. Morin, "Objective view synthesis quality assessment," in *Stereoscopic Displays and Applications*, SPIE, Ed. 2012, vol. 8288 of *Proc. SPIE*, pp. 8288–56.

[7] Z. Wang, A. C Bovik, H. R Sheikh, and E. P Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[8] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *2003 International Conference on Image Processing*, 2003, pp. 14–17.

[9] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in *Proc. Int. Workshop Video Processing and Quality Metrics*, 2010.

[10] S. L. P. Yasakethu, C. Hewage, W. Fernando, and A. Kondoz, "Quality analysis for 3D video using 2D video quality models," *Consumer Electronics, IEEE Transactions on*, vol. 54, no. 4, pp. 1969–1976, 2008.

[11] P. Hanhart, F. De Simone, and T. Ebrahimi, "Quality assessment of asymmetric stereo pair formed from decoded and synthesized views," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, 2012, p. 236241.

[12] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3-D warping using depth information for FTV," *Elsevier Signal Processing: Image Communication*, vol. 24, pp. 65–72, 2009.

[13] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3-D video systems," *EURASIP Journal on Image and Video Processing*, 2008, Article ID 438148, 11 pages.

[14] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Depth image based rendering with advanced texture synthesis," in *Proc. IEEE International Conference on Multimedia & Expo (ICME)*, 2010.

[15] M. Köppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2010.

[16] ITU-T, "Subjective video quality assessment methods for multimedia applications," Tech. Rep. Rec. P910,2008.

[17] ITU-R BT., *500, Methodology for the subjective assessment of the quality of television pictures*, 1993.

[18] Metrix Mux, "Metrix mux home page," http://foulard.ece.cornell.edu/gaubatz/metrix_mux/.

[19] VQEG, "VQEG 3DTV group," ftp://vqeg.its.bldrdoc.gov/Documents/Projects/multimedia/.