

## CAN 3D SYNTHESIZED VIEWS BE RELIABLY ASSESSED THROUGH USUAL SUBJECTIVE AND OBJECTIVE EVALUATION PROTOCOLS?

E. Bosc<sup>1</sup>, M. Köppel<sup>2</sup>, R. Pépion<sup>3</sup>, M. Pressigout<sup>1</sup>, L. Morin<sup>1</sup>, P. Ndjiki-Nya<sup>2</sup>, P. Le Callet<sup>3</sup>

<sup>1</sup> IETR UMR CNRS 6164, INSA de Rennes

20, avenue des Buttes de Coesmes, 35708 RENNES CEDEX 7, France

<sup>2</sup> Fraunhofer Institut for Telecommunications, HHI, Einsteinufer 37, 10587 Berlin, Germany

<sup>3</sup> IRCCyN, Université de Nantes, Rue Christian Pauc, 44306 Nantes, France

### ABSTRACT

This paper addresses the problem of evaluating virtual view synthesized images in the multi-view video context. As a matter of fact, view synthesis brings new types of distortion. The question refers to the ability of the traditional used objective metrics to assess synthesized views quality, considering the new types of artifacts. The experiments conducted to determine their reliability consist in assessing seven different view synthesis algorithms. Subjective and objective measurements have been performed. Results show that the most commonly used objective metrics can be far from human judgment depending on the artifact to deal with.

**Index Terms**— Virtual view synthesis, multi-view video, 3DTV, quality assessment, quality metrics

### 1. INTRODUCTION

Recently, multi-view video processing has gained a growing interest. 3D video refers to two main applications: 3D television (3DTV), that provides a depth feeling, and Free Viewpoint Video (FVV), that allows navigation inside the scene [1]. These emerging applications make the problem of evaluating the 3D visual experience a huge subject of investigation. As pointed out in [2], the assessed factors are more numerous than in traditional 2D video: image quality, visual comfort and depth are to be taken into consideration. This paper focuses on image quality.

Multi-view video plus depth (MVD) data [3] can be used to offer 3DTV or FTV. MVD data consist of two types of videos: a first set of conventional video sequences acquired from the same scene at slightly different viewpoints, referred as “texture data”; a second set of associated depth video sequences, referred as “depth data”. Depth data provide information on scene geometry and help in virtual intermediate view generation. When targeting either 3DTV or a FTV application, virtual view generation is very likely to be required. Indeed, the appreciation of a 3D content relies on the stereopsis phenomenon: an observer needs to be presented a pair of stereoscopic images with a strong binocular disparity. Human brain is then able to fuse the pair of images and to interpret the

3D scene. Thus, 3DTV displays should provide the appropriate stereoscopic pairs to ensure the immersion feeling. On the other hand, for FTV applications, a user may wish to navigate around the scene, which makes virtual view synthesis generation essential. Finally, depending on the available bandwidth or on the decoder, all the acquired video sequences may not be available. In this case, virtual view generation is also needed. Considering the users demand for acceptable image quality as a minimum, the quality of reconstruction of virtual views cannot be ignored.

Many new distortions have been listed in [4]. Among them, the *keystone effect* that makes the image look like a trapezoid; the *ghosting effect* that is a shadow-like artifact; the *cardboard effect* when depth is perceived as unnatural, as discrete incoherent planes. Synthesis errors can be added to this list as projection errors can occur. These new types of artifacts have to be taken into consideration when evaluating synthesized views.

However, up to now, there is no dedicated assessment framework, nor objective metric for 3D video quality evaluation. [5] addressed the problem of measuring the quality of a synthesized view from encoded color and depth video. When trying to determine the optimal ratio between color and depth data in the context of 3D video compression, the authors observed that PSNR (Peak Signal to Noise ratio) and VSSIM (Video Structural SIMilarity index) led to different conclusions, regarding the compression choices. PSNR seemed unstable depending on the direction of the targeted virtual viewpoint. Though, the subjective scores correlated the VSSIM scores. Recently, [6] reconsidered the synthesis quality evaluation framework. The authors showed the importance of the chosen reference for synthesis quality evaluation: they pointed out the fact that depending on the chosen reference (original view or control synthesis, i.e. image synthesized from uncompressed data), PSNR scores do not measure the same distortion. They showed that distortions from compression may be *masked* by distortion from synthesis process. Consequently, they recommend to use the control synthesis as a reference when assessing a codec performances. Peak

Signal to Perceptual Noise Ratio (PSPNR) [7] derived from PSNR, is the metric used by the 3D Video (3DV) group of MPEG. In [8], the authors proposed a new full reference metric that takes into consideration depth data and consequently the regions that are more likely to be distorted in the synthesis. The new framework is validated by its high correlation score with the perceptual-like metric Video Quality Metric (VQM) [9] results, and subjective assessments from 15 non-expert observers.

This paper investigates the reliability of different objective metrics on still synthesized images as still images can be a plausible case for FTV. The test objective metrics are pixel-based as well as perceptual-like metrics. The images are synthesized with seven different Depth Image Based Rendering (DIBR) algorithms. Subjective assessments allow to evaluate the correlation between human perception and objective measurements.

## 2. ALGORITHMS

In this section different depth-image-based rendering (DIBR) methods are presented. DIBR defines the process of synthesizing “virtual” views at a slightly different viewing perspective using an image or video and the associated per pixel depth information. A critical problem in DIBR is that regions occluded in the original view may become visible in the “virtual” view, an event also referred to as disocclusion. In the absence of original image data two extrapolation paradigms address this inherent problem: 1) One can preprocess the depth information in a manner that no disocclusion occur in the “virtual” view, or 2) replace the missing image areas (holes) with known suitable image information. In the following, a short overview will be given on relevant work in disocclusion handling in 3D video.

Fehn preprocesses the per pixel depth information with a 2D Gaussian low-pass filter [10]. This way large discontinuities are smoothed out in the depth map and disocclusions do not appear in the “virtual” image. However, this leads to geometric distortions in the virtual view. Larger baselines yielding to more disturbing artifacts. In a rectified camera setup this method fails to close holes on the left or right border image. Therefore, these areas are treated in two different ways. Either the border is cropped and the image is resized to the original size or the holes on the border are inpainted with [11]. These methods are referred to as *A1* and *A2* respectively in the rest of the paper. The cropping method is suitable for a stereo video where one view only is transmitted and the other one is rendered at the decoder side. In multi-view scenarios, this method is not applicable because all views, the original as well as the virtual views, have to be cropped to preserve the stereo impression. This would lead to image information losses in all views.

Tanimoto et al. [12] proposed a 3D view generation system. They are using an inpainting method [11] to fill missing parts in the “virtual” image. This algorithm is adopted as the

reference software for MPEG standardization experiments in the 3D Video group. This method is referred to as *A3* in the rest of the paper.

Müller et al. [13] proposed a hole filling method embedded in a 3D video system. Holes are filled linewise with neighboring background information. The corresponding depth values at the hole boundary are examined row-wise to find background color samples to copy into the hole. This color extrapolation of the suitable background pixel leads to better results than a simple linear interpolation. Generally, due to depth estimation, some boundary background pixels in fact belong to foreground objects. Thus their color information would lead to foreground color propagation into the hole. This method is referred to as *A4* in the rest of the paper.

In texture synthesis methods the unknown regions are synthesized by copying content from the known parts of the image to the missing regions. Ndjiki-Nya et al. [14] proposed a hole filling approach for DIBR systems based on patch-based texture synthesis. Holes with small spatial extend are closed by solving Laplacian equations. Larger holes are initialized by median filtering and then optimized via texture synthesis. This method is referred to as *A5* in the rest of the paper.

Köppel et al. [15] extended the *A5* approach by a background sprite. The sprite stores valuable background image information and is updated frame-wise. Using the original and synthesized image information from previous frames temporally consistency is achieved in a sequence. This method is referred to as *A6* in the rest of the paper. In the conducted subjective tests only images are analyzed. Thus, the capabilities of the approach to achieve temporal consistency in a sequence is not investigated. Algorithms *A2-A6* support multi-view scenarios.

Non-filled sequences (i.e. with holes) are referred to as *A7* in the rest of the paper.

## 3. EXPERIMENTAL PROTOCOL

The experiments have two main objectives: first to determine the tested algorithms performances and second, to assess the reliability of objective metrics for 3D images. Three test sequences have been used to generate four different viewpoints, that is to say twelve synthesized sequences for each test algorithm (84 synthesized sequences in total): *Book Arrival* (1024×768, 16 cameras with 6.5cm spacing), *Lovebird1* (1024×768, 12 cameras with 3.5cm spacing) and *Newspaper* (1024×768, 9 cameras with 5cm spacing). Altogether 43 naive observers participated in the subjective assessment session. The session was conducted in an ITU conforming test environment. Absolute categorical rating (ACR) [16] was used to collect perceived quality scores: stimuli are presented in a random order and are evaluated through a coarse resolution rating scale. Observers notes are then averaged, which is called MOS (Mean Opinion Score). The stimuli were displayed on a TVLogic LVM401W, and according to ITU-T BT.500 [17]. Considering the large size of the tested database,

only key frames of the rendered sequences were presented to the observers, as still images can also be a plausible scenario for FTV. Key frames were also evaluated through different objective metrics through MeTriX MuX Visual Quality Assessment Package [18]. For both objective metrics, the reference was the original acquired image.

#### 4. RESULTS

Subjective ratings are illustrated on Figure 1. Algorithms are ordered by MOS ratings. For a given algorithm its rank varies depending on the data set. This suggests that the algorithms performances depend on the inner sequences properties (i.e. the depth range, the camera acquisition parameters).

On Figure 2, subjective scores are plotted over objective scores in order to find a correlation. Let  $d$  be the camera baseline of a sequence. Top graph shows the performances when synthesizing with large baseline between reference and target view ( $2 \times d$ ), and bottom graph corresponds to a shorter baseline ( $d$ ). As expected, it is observed that the shorter the baseline, the higher the scores (for objective as subjective measurements). For a view synthesized with short baseline, it can be observed that two PSNR scores varying from 20dB to 28dB (A1 and A4), MOS score remain nearly constant (from 2.6 to 2.5). As well, a variation of 2dB (A1 and A7) corresponds to about 1 MOS point. This is significant considering the coarse scale of MOS scores (from 1 to 5).

Statistical analyses have been conducted over the objective and subjective measurements. In order to determine whether classes of algorithms could emerge, a Student's t-test has been performed over the MOS scores for each test algorithm: on Table 1, statistically dependent pairs can be distinguished. It clearly indicates the statistical divergence of three algorithms: A7, A3 and A1 distributions differ from the other algorithms'. A7 and A3 count no statistically dependent pair, and A1 counts only one. In addition, Table 1 also indicates the required minimum number of observers that allows the statistical distinction (values in bold are higher than 24). It shows that the final ranking is obtained when 32 observers participate (VQEG recommends 24 observers).

The test with metrics other than PSNR led to nearly the same observations regarding the algorithms performances and their correlation with MOS scores. Besides, Table 2 confirms that all metrics are very correlated, even pixel-based ones compared to perceptual-based ones. Note that perceptual-like SSIM is very correlated to pixel-based PSNR (83.9%). Table 3 expresses the correlation coefficients between objective metrics and MOS scores, for the whole fitted measured points. It can be observed that the metrics closest to human judgment are WSNR (Weighted Signal-to-Noise Ratio), PSNR and NQM (Noise Quality Measure), (42.3%, 38.6% and 38.6% respectively). WSNR is a CSF-based weighting function and PSNR and NQM are pixel-based metrics. These metrics are also highly correlated according to Table 2. However, Figure 2 reveals the inconsistency between MOS and

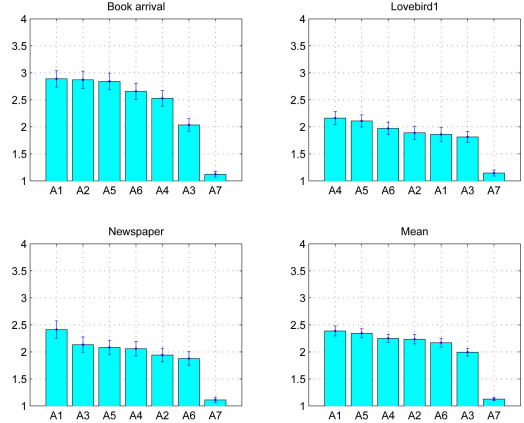


Fig. 1. MOS scores for the different sequences.

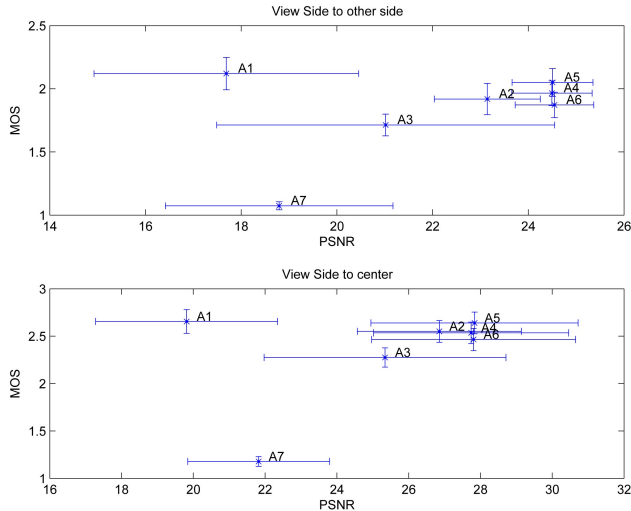
	A1	A2	A3	A4	A5	A6	A7
A1		↑(32)	↑(<24)	↑(32)	o(>43)	↑(30)	↑(<24)
A2	↓(32)		↑(<24)	o(>43)	o(>43)	↑(<24)	↑(<24)
A3	↓(<24)	↓(<24)		↓(<24)	↓(<24)	↓(<24)	↑(<24)
A4	↓(32)	o(>43)	↑(<24)		o(>43)	o(>43)	↑(<24)
A5	o(>43)	o(>43)	↑(<24)	o(>43)		↑(28)	↑(<24)
A6	↓(30)	o(>43)	↑(<24)	o(>43)	↓(28)		↑(<24)
A7	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	

Table 1. Results of Student's t-test. Legend: ↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "32" observers participate.

PSNR. And indeed, the algorithms rankings according to each metric, listed on Table 4, show a very important difference between human scores and metrics for A1 algorithm. It is ranked as the best of this set by humans but worst by the metrics. A6 generates the best objective results but subjective evaluations assign its quality as not as good. A5 generates coherent objective and subjective results. For A5, A2, A3 and A4 the results of objective metrics correspond with human scores. This suggests that the reliability of the objective metrics differ depending on the rendering algorithm used, i.e. on the induced artifact. Algorithms can induce non-perceptible or non-annoying artifacts. Then, this implies that commonly used metrics are not suited for assessing virtual synthesized views as they inflict serious costs to relatively acceptable degradations. These results point out the need for a new 3D-adapted metric.

	PSNR	SSIM	MSSIM	VSNR	VIF	VIFP	UQI	IFC	NQM	WSNR	PSNR <sub>hsvm</sub>	PSNR <sub>hsv</sub>
PSNR		83.9	79.6	87.3	77.0	70.6	53.6	71.6	95.2	98.2	99.2	99.0
SSIM			96.7	93.9	93.4	92.4	81.5	92.9	84.9	83.7	83.2	83.5
MSSIM	79.6	96.7		89.7	88.8	90.2	86.3	89.4	85.6	81.1	77.9	78.3
VSNR	87.3	93.9	89.7		87.9	83.3	71.9	84.0	85.3	85.5	86.1	85.8
VIF	77.0	93.4	88.8	87.9		97.5	75.2	98.7	74.4	78.1	79.4	80.2
VIFP	70.6	92.4	90.2	83.3	97.5		85.9	99.2	73.6	75.0	72.2	72.9
UQI	53.6	81.5	86.3	71.9	75.2	85.9		81.9	70.2	61.8	50.9	50.8
IFC	71.6	92.9	89.4	84.0	98.7	99.2	81.9		72.8	74.4	73.5	74.4
NQM	95.2	84.9	85.6	85.3	74.4	73.6	70.2	72.8		97.1	92.3	91.8
WSNR	98.2	83.7	81.1	85.5	78.1	75.0	61.8	74.4	97.1		97.4	97.1
PSNR <sub>hsvm</sub>	99.2	83.2	77.9	86.1	79.4	72.2	50.9	73.5	92.3	97.4		99.9
PSNR <sub>hsv</sub>	99.0	83.5	78.3	85.8	80.2	72.9	50.8	74.4	91.8	97.1	99.9	

Table 2. Correlation coefficients between objective metrics in percentage.



**Fig. 2.** Correlation between MOS and PSNR according to the baseline distance between reference and target view.

	PSNR	SSIM	MSSIM	WSNR	VIF	VIFP	UQI	IFC	NQM	WSNR	PSNR	HVS	PSNR
CC	38.6	21.9	16.1	25.8	19.3	19.2	20.2	19.0	38.6	42.3	38.1	37.3	

**Table 3.** Correlation coefficients between subjective and objective scores in percentage.

	A1	A2	A3	A4	A5	A6	A7
MOS	2.388	2.234	1.994	2.250	2.345	2.169	1.126
Rank order	1	4	6	3	2	5	7
PSNR	18.752	24.998	23.180	26.117	26.171	26.177	20.307
Rank order	7	4	5	3	2	1	6
SSIM	0.638	0.843	0.786	0.859	0.859	0.858	0.821
Rank order	7	4	6	1	1	3	5
MSSIM	0.648	0.932	0.826	0.950	0.949	0.949	0.883
Rank order	7	4	6	1	2	2	5
WSNR	13.135	20.530	18.901	22.004	22.247	22.195	21.055
Rank order	7	5	6	3	1	2	4
VIF	0.124	0.394	0.314	0.425	0.425	0.426	0.397
Rank order	7	5	6	2	2	1	4
VIFP	0.147	0.416	0.344	0.448	0.448	0.448	0.420
Rank order	7	5	6	1	1	1	4
UQI	0.237	0.556	0.474	0.577	0.576	0.577	0.558
Rank order	7	5	6	1	3	1	4
IFC	0.757	2.420	1.959	2.587	2.586	2.591	2.423
Rank order	7	5	6	2	3	2	4
NQM	8.713	16.334	13.645	17.074	17.198	17.201	10.291
Rank order	7	4	5	3	2	1	6
WSNR	13.817	20.593	18.517	21.597	21.697	21.716	15.588
Rank order	7	4	5	3	2	1	6
SNR	12.848	19.094	17.276	20.213	20.267	20.274	14.403
Rank order	7	4	5	3	2	1	6
PSNR hsvm	13.772	19.959	18.362	21.428	21.458	21.491	15.714
Rank order	7	4	5	3	2	1	6
PSNR hsv	13.530	19.512	17.953	20.938	20.958	20.987	15.407
Rank order	7	4	5	3	2	1	6

**Table 4.** Rankings according to measurements.

## 5. CONCLUSION

This paper addresses the issue of evaluating virtual synthesized views with the traditional objective metrics. The assessments of the seven test algorithms by objective measurements and subjective ratings show that among all tested objective metrics, WSNR and pixel-based PSNR and NQM are the most correlated with perceptual evaluation provided by MOS scores. However, the results also show PSNR's inability to predict human experience. New methods are then required for assessing virtual synthesized views as pixel-based

and perceptual-based metrics fail. Depth should be taken into account in such a metric as recently proposed in [8], because view synthesis produces geometric distortions. Registration process according to the original view coupled with weighted critical areas could be investigated in future work to build a new metric. In addition, paired comparisons experiments should be held on still images and video sequences in the future to refine the presented results.

## 6. ACKNOWLEDGMENTS

This work is partly supported by the French ANR-PERSEE project n° ANR-09-BLAN-0170, and ANR-CAIMAN project n° ANR-08-VERS-002.

## 7. REFERENCES

- [1] A. Smolic, K. Müller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3D video and free viewpoint VideoTechnologies, applications and MPEG standards," in *Proc. of IEEE ICME*, Toronto, Canada, Jul. 2006.
- [2] Pieter Seuntjens, *Visual Experience of 3D TV*, Ph.D. thesis, 2006.
- [3] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. of IEEE ICIP*, San Antonio, USA, Sep-Oct. 2007.
- [4] M Meesters, W Ijsselstein, and P Seuntjens, "A survey of perceptual evaluations and requirements of three dimensional TV," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 381–391, Mar. 2004.
- [5] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Müller, "Quality assessment of 3D video in rate allocation experiments," in *Proc. of IEEE ISCE*, Algarve, Portugal, Apr. 2008.
- [6] N. A. El-Yamany, K. Ugur, M. M. Hannuksela, and M. Gabbouj, "Evaluation of depth compression and view synthesis distortions in multiview-video-plus-depth coding systems," in *Proc. of 3DTV-Con*, Tampere, Finland, Jun. 2010.
- [7] Yin Zhao and Lu Yu, "Perceptual measurement for evaluating quality of view synthesis," Apr. 2009, MPEG Doc. M16407.
- [8] 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 UCMedia*, Palma de Mallorca, Spain, Sep. 2010.
- [9] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.
- [10] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in *Proc. of SPIE Conf. Stereoscopic Displays and Virtual Reality Systems X*, San Jose, USA, Jan. 2004.
- [11] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of Graphics, GPU, and Game Tools*, vol. 9, no. 1, pp. 23–34, 2004.
- [12] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3D 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 3D 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. of IEEE ICME*, Singapore, Jul. 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. of IEEE ICIP*, Hong Kong, China, Sep. 2010.
- [16] ITU-T Study Group 12, "ITU-T p.910 subjective video quality assessment methods for multimedia applications," 1997.
- [17] ITU-R BT., *500, Methodology for the subjective assessment of the quality of television pictures*, November, 1993.
- [18] "MetriX MuX page," [http://foulard.ece.cornell.edu/gaubatz/metrix\\_mux/](http://foulard.ece.cornell.edu/gaubatz/metrix_mux/).