Free-Viewpoint Video Sequences: a New Challenge for Objective Quality Metrics

Philippe Hanhart #1, Emilie Bosc *, Patrick Le Callet *2, Touradj Ebrahimi #3

# *Multimedia Signal Processing Group, EPFL*
Station 11, 1015 Lausanne, Switzerland
[1] philippe.hanhart@epfl.ch
[3] touradj.ebrahimi@epfl.ch

* *LUNAM Université, Université de Nantes*
IRCCyN UMR CNRS 6597, Nantes, France
[2] patrick.lecallet@univ-nantes.fr

Abstract—**Free-viewpoint television is expected to create a more natural and interactive viewing experience by providing the ability to interactively change the viewpoint to enjoy a 3D scene. To render new virtual viewpoints, free-viewpoint systems rely on view synthesis. However, it is known that most objective metrics fail at predicting perceived quality of synthesized views. Therefore, it is legitimate to question the reliability of commonly used objective metrics to assess the quality of free-viewpoint video (FVV) sequences. In this paper, we analyze the performance of several commonly used objective quality metrics on FVV sequences, which were synthesized from decompressed depth data, using subjective scores as ground truth. Statistical analyses showed that commonly used metrics were not reliable predictors of perceived image quality when different contents and distortions were considered. However, the correlation improved when considering individual conditions, which indicates that the artifacts produced by some view synthesis algorithms might not be correctly handled by current metrics.**

I. Introduction

Free-viewpoint systems are meant to provide the viewer with the ability to interactively change his/her viewpoint to enjoy a 3D scene. Among these, free-viewpoint television is one of the key technologies brought by the development of 3D video applications. It opens the door to new applications in entertainment, post-production, teleconferencing, security applications, etc. These applications are based on a limited number of cameras for recording the 3D scene. Many 3D scene representations have been proposed [1], among which is the multiview video plus depth (MVD) format. The MVD format consists of a set of texture views and associated depth maps acquired at different viewpoints. From color and depth information, new virtual viewpoints can be rendered through depth-image-based rendering (DIBR) techniques [2].

The perceived image quality of free-viewpoint content can be affected at many stages of the processing chain. In particular, the impact of both compression and DIBR algorithms on the quality of virtual viewpoints has been shown [3], [4]. Considering compression, the Video Coding Experts Group (VCEG) and Moving Picture Experts Group (MPEG) have joined their efforts to develop new 3D video coding standards for advanced 3D video applications. These algorithms consider the quality of the synthesized views to optimize compression.

As any technology, the performance evaluation of free-viewpoint systems, in terms of quality of user experience, is essential. Subjective evaluations are time consuming and expensive. Therefore, objective quality assessment tools are needed as well. Very few metrics have been proposed for FTV applications [5], [6] and common full reference 2D metrics, e.g., PSNR and SSIM, are still mostly used [3], [4].

As outlined in [4], [7], the 3D warping process involved in the DIBR techniques induces distortions mainly known as *"cracks"* or *"holes"*, which are due to the sampling rate and the discovering of areas not visible from the reference viewpoint, but visible in the new viewpoint, and *"ghosting"*, which is due to the edge resolution in the depth maps. These distortions are different from those commonly encountered in video compression. Moreover, video compression related artifacts are often scattered over the whole image, whereas DIBR related artifacts are mostly located around the disoccluded areas. Most of the commonly used objective quality metrics were initially designed to address video compression related artifacts and are not reliable predictors of perceived quality of monoscopic and stereoscopic video sequences formed from synthesized views [8], [9]. As free-viewpoint systems rely on view synthesis to render new virtual viewpoints, it is legitimate to question the reliability of these metrics to assess the quality of free-viewpoint video (FVV) sequences.

In our previous study [10], we performed a subjective quality evaluation to assess the quality of FVV sequences corresponding to a smooth camera motion during a time freeze, which were generated through DIBR from 3D content represented in the MVD format. Only depth maps compression was concerned (and not color view compression, as in a classical scenario) as it has been shown that depth compression has a critical impact on the quality of synthesized views. Six MVD contents were considered. The depth maps were compressed with seven algorithms using three quantization

2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), Sep. 22-24, 2014, Jakarta, Indonesia.
978-1-4799-5896-2/14/$31.00 ©2014 IEEE.

parameters. Two additional methods were also included to increase the variety of distortions. Two different modes for the synthesis process were considered, resulting in a total of 276 processed stimuli and 12 reference stimuli.

In this paper, we analyze the performance of several commonly used objective quality metrics (PSNR, SSIM, MS-SSIM, VIF, VIFP, UQI, and IFC) using the FVV sequences and corresponding ground truth subjective scores[1] obtained in [10]. For each metric, objective scores were fitted to subjective scores using cubic fitting. As compliant with the standard procedure for evaluating the performance of objective metrics [11], [12], the following properties of the estimation of subjective scores were considered in this study: accuracy, monotonicity, and consistency. Several performance indexes, such as Pearson and Spearman correlation coefficients, root-mean-square-error, and outlier ratio, were computed to compare the metrics estimation of subjective scores. Statistical tests were performed to determine if the difference between two metrics is statistically significant.

The remainder of the paper is organized as follows. The dataset and corresponding subjective scores used as ground truth are described in Sec. II. The different metrics benchmarked in this study are defined in Sec. III. In Sec. IV, the methodology used to evaluate the performance of the metrics is described. Section V provides a detailed analysis of the objective results and discusses the reliability of objective metrics. Finally, Sec. VI concludes the paper.

II. Dataset and Subjective Scores

The dataset was composed of six multiview video plus depth (MVD) contents, four real scenes with estimated depth maps (*Book Arrival*, *Newspaper*, *Kendo*, and *Balloons*) and two synthetic scenes with ground truth depth maps (*GT Fly* and *Undo Dancer*), with different visual characteristics, resolutions, and frame rates. The depth maps were compressed using seven algorithms, labeled from *C1* to *C7*. Three quantization parameters were selected for each depth map compression algorithm, according to the visual quality of the rendered views. Two additional methods were also included to increase the variety of distortions: low pass filtered depth maps (noted *F*) and depth maps with low-pass filtered applied on edges only (noted *FE*). Two different modes for the synthesis process, referred to as *VS1* and *VS2*, were considered:

- *VS1*: *Blended Mode* disabled. All pixels visible in the closer reference view are copied to the virtual view, and only hole areas are filled from the farther reference view.
- *VS2*: *Blended Mode* enabled. A weighted blending based on the baseline distance is used for hole filling. So pixels from the reference camera that are closer to the virtual view are assigned a higher weight.

The ACR-HR methodology was used to assess the FVV sequences using a five-level quality scale (1: *Bad*; 2: *Poor*; 3: *Fair*; 4: *Good*; 5: *Excellent*). The combination of contents, view synthesis modes, depth map compression algorithms, and bit rates resulted in a total of 276 processed stimuli and 12 reference stimuli to be assessed. The subjective evaluations were conducted in an ITU conforming test environment. The stimuli were displayed on a Panasonic BT-3DL2550 screen (1920×1080p), and according to [13]. Twenty-seven naïve observers participated in the subjective quality evaluation test into two 30-minutes sessions. All subjects underwent a screening to examine their visual acuity, color vision, and stereo vision. Four subjects were detected as outliers and all their scores were removed from the results. Then, the differential mean opinion scores score (DMOS) was computed between the mean opinion scores (MOS) of each test stimulus and its associated hidden reference, as well as associated 95% confidence interval. More details about dataset and subjective evaluations can be found in [10].

III. Objective quality metrics

In this study, the performance of the following metrics in predicting image quality of FVV sequences was assessed:

1) PSNR: Peak Signal-to-Noise Ratio,
2) SSIM: Structural Similarity Index [14],
3) MS-SSIM: Multi-Scale Structural Similarity Index [15],
4) VIF: Visual Information Fidelity [16],
5) VIFP: VIF, pixel domain version [16],
6) UQI: Universal Quality Index [17],
7) IFC: Information Fidelity Criterion [18].

All above objective metrics were computed on the luma component of each frame of the FVV sequence and the resulting values were averaged across the frames to produce a global index for the entire FVV sequence. All objective metrics were computed using MeTriX MuX (v. 1.1)[2].

IV. Performance indexes

The results of the subjective tests can be used as ground truth to evaluate how well the objective metrics estimate perceived quality. The result of execution of a particular objective metric is a video quality rating (VQR), which is expected to be the estimation of the DMOS corresponding to a pair of video data. To be compliant with the standard procedure for evaluating the performance of objective metrics [11], [12], the following properties of the VQR estimation of DMOS were considered in this study: accuracy, monotonicity, and consistency.

First, a regression was fitted to each [VQR, DMOS] data set using cubic fitting, with the constraint that the function is monotonic on the interval of observed quality values:

$$DMOS_p(VQR) = a \cdot VQR^3 + b \cdot VQR^2 + c \cdot VQR + d$$

where a, b, c, and d are the parameters of the fitting function.

Then, the Pearson linear correlation coefficient (PCC) and the root-mean-square error (RMSE) were computed between $DMOS_p$ and $DMOS$ to estimate the accuracy of the VQR. To estimate monotonicity and consistency, the Spearman rank

[1] Available online at http://ivc.univ-nantes.fr/ivc/en/content/10-databases/13-free-viewpoint-synthesized-video-database

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

Table I Accuracy, consistency, and monotonicity indexes for the different metrics considering cubic fitting.

Metric	All contents				Average			
	PCC	SCC	RMSE	OR	PCC	SCC	RMSE	OR
PSNR	0.2671	0.2945	0.9072	0.5091	0.3284	0.4505	0.5663	0.3452
SSIM	0.0000*	0.0000*	0.9414	0.5641	0.2202	0.3670	0.6035	0.3741
MS-SSIM	0.0105	0.0611	0.9413	0.5604	0.1870	0.3942	0.6098	0.3960
VIF	0.0584	0.0948	0.9398	0.5714	0.2642	0.3415	0.5836	0.3853
VIFP	0.0798	0.1223	0.9384	0.5678	0.2624	0.3305	0.5847	0.3854
UQI	0.0000*	0.0000*	0.9414	0.5641	0.2395	0.3441	0.6007	0.3853
IFC	0.1289	0.0657	0.9335	0.5531	0.2808	0.3307	0.5799	0.3741

*The correlation for SSIM and UQI is null, which is due to the fact that the cubic fitting was constrained to be monotonic on the interval of observed quality values, whereas the non-fitted scores for these two metrics mostly showed a negative correlation with perceived quality (see Fig. 1(b), which shows the correlation between the obtained subjective scores and the corresponding non-fitted objective scores).

Table II Statistical analysis of the different metrics considering cubic fitting.

	PSNR	SSIM	MS-SSIM	VIF	VIFP	UQI	IFC
PSNR		$\neq\neq==$	$\neq\neq==$	$\neq\neq==$	$\neq\neq==$	$\neq\neq==$	$=\neq==$
SSIM	$\neq\neq==$		$=\neq==$	$=\neq==$	$=\neq==$	$=\neq==$	$=\neq==$
MS-SSIM	$\neq\neq==$	$=\neq==$		$====$	$====$	$=\neq==$	$====$
VIF	$\neq\neq==$	$=\neq==$	$====$		$====$	$=\neq==$	$====$
VIFP	$\neq\neq==$	$=\neq==$	$====$	$====$		$=\neq==$	$====$
UQI	$\neq\neq==$	$=\neq==$	$=\neq==$	$=\neq==$	$=\neq==$		$=\neq==$
IFC	$=\neq==$	$=\neq==$	$====$	$====$	$====$	$=\neq==$	

Each entry in the table corresponds to the results of the statistical tests performed on the following performance indexes (from left to right): PCC, SCC, RMSE, and OR. "$=$" means that there was no significant difference between the two metrics, whereas "$\neq$" means that the difference was significant. Reading: Line 2, column 4: SSIM and VIF are statistically different according to SCC, whereas they are similar according to the other performance indexes.

order correlation coefficient (SCC) and the outlier ratio (OR), were computed between $DMOS_p$ and $DMOS$, respectively. Finally, these four performance indexes were averaged across the different contents.

The root-mean-square error (RMSE) and the outlier ratio (OR) are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(DMOS_i - DMOS_{pi})^2}$$

$$\text{OR} = \frac{\text{total number of outliers}}{N}$$

$$\text{outlier: point for which } |DMOS_i - DMOS_{pi}| > CI_i$$

where N is the total number of points and CI_i is the 95% confidence interval corresponding to $DMOS_i$.

To determine whether the difference between two performance index values corresponding to two different metrics is statistically significant, a statistical test was performed according to [19].

A PCA was also applied between the DMOS and the objective scores to further investigate the correlation of the objective metrics with perceived quality. As the different metrics have different scales and PCA is sensitive to the relative scaling of the original variables, normalized variables with zero-mean and unit-variance were used.

V. Results

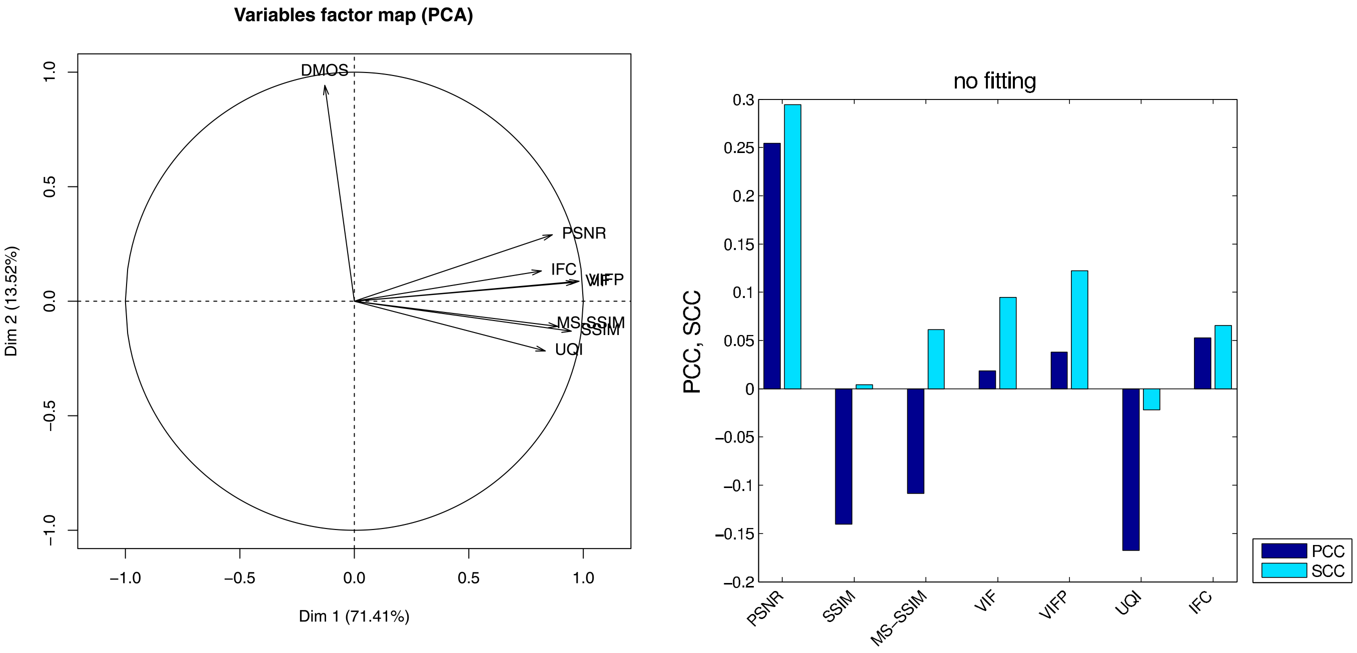
A. Correlation between objective and subjective scores

Table I reports the accuracy, consistency, and monotonicity indexes, as defined in Sec. IV, for the cubic fitting. The fitting, as defined in Sec. IV, was applied in two different ways:

a) on all contents at once,
b) on each content separately.

In the latter case, the performance indexes were computed separately on each content and then averaged across contents. When the fitting was applied on all contents at once, the correlation was lower than 0.15 for all metrics, except for PSNR, which showed a correlation around 0.3. The RMSE was around 0.9 for all metrics. The OR was higher than 55% on all metrics. These results show that there is a poor correlation between objective metrics and perceived quality. When the fitting was applied on each content separately, the obtained performance marginally improved, as the PCC and SCC scores are still in the range 0.18-0.33 and 0.33-0.45, respectively. The RMSE and OR decreased below 0.61 and 40%, respectively. However, these results still lead to the conclusion that there is a poor correlation between objective metrics and perceived quality.

When the fitting was applied on all contents at once, PSNR seems to outperform other metrics, even though the correlation was still very low. To determine if the difference between PSNR and the other metrics is significant, statistical tests were performed according to [19]. Table II reports the results of the statistical tests considering cubic fitting. Each entry



(a) PCA

(b) Correlation

Figure 1 Circle of correlations and Pearson and Spearman correlation scores between DMOS and objective scores.

in the table corresponds to the results of the statistical tests performed on the following performance indexes (from left to right): PCC, SCC, RMSE, and OR. The statistical tests were performed to determine whether the difference between two performance index values corresponding to two different metrics was statistically significant: "=" means that there was no significant difference between the two metrics, whereas "≠" means that the difference was significant. The results showed that PSNR was significantly different from the other metrics according to PCC and SCC values, except for IFC.

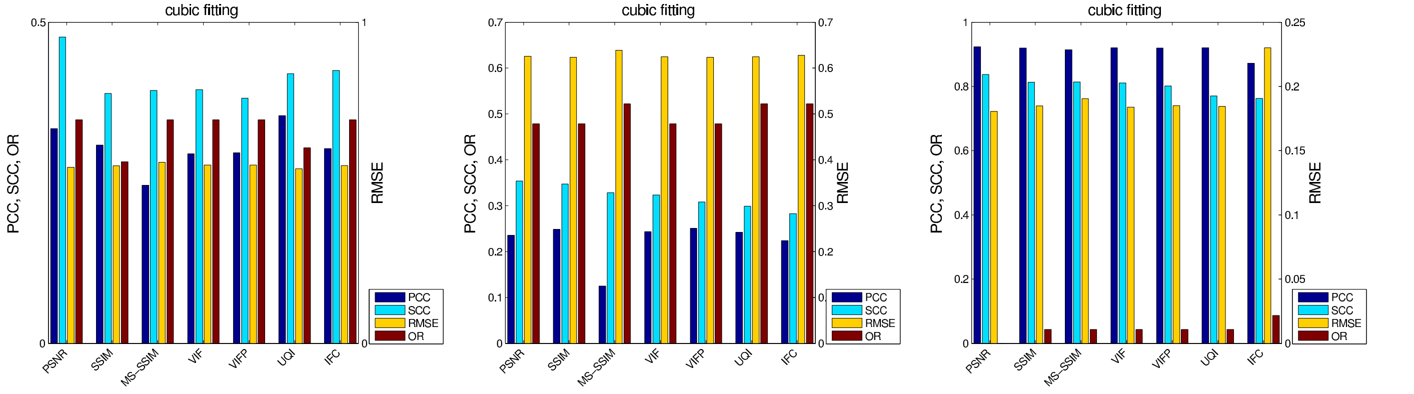
Figure 1 shows the correlation between the obtained subjective scores and the corresponding objective scores. Figure 1(a) depicts the circle of correlations derived from the PCA. Figure 1(b) depicts the Pearson and Spearman correlation coefficients between the DMOS and the objective scores. Only two components had an eigenvalue larger than 1 in the PCA. These two principal components explained 84% of the variance of the data.

The circle of correlations allows the observation of correlations between variables and principal components. Each measured variable is represented as a vector. The vector length represents the combined strength of the relationships between measured variable and principal components. The vector direction indicates whether these relationships are positive or negative. Since the data is not perfectly represented by the two principal components, the variables are positioned inside the circle of correlations. The closer the variable is to the circle, the more important it is to the principal components. The smaller the angle between two measured variable's vector representations, the higher their correlation. In Fig. 1(a), it can be observed that the objective metrics are grouped, which shows that they are correlated with each other. However, the angle between most of the objective metrics and DMOS is close to $\frac{\pi}{2}$, which indicates that subjective scores are not correlated to objective metrics. This is confirmed by the analysis of Pearson and Spearman correlation scores in Fig. 1(b): these correlation scores are very low since they do not reach 0.3.

Another interesting observation concerns the contributions of the variables to the principal components in Fig. 1(a). The variables for which the contribution value is larger than the average contribution for the first component are VIFP, VIF, SSIM, MS-SSIM, and PSNR. The only variable for which the contribution value is larger than the average contribution for the second component is DMOS.

B. Scope of validity of the objective metrics

Huynh-Thu and Ghanbari [20] have shown that even PSNR can be a valid quality measure if the video content and the codec type are not changed. It is well known that objective metrics can better handle some types of degradations and often fail when different types of degradations are combined. In this study, different views synthesis modes, contents, depth map compression algorithms, and bit rates were considered. As it was shown in our previous study [10], the view synthesis mode had an impact on perceived quality and modified the behavior of a compression algorithms. Therefore, we benchmarked and analyzed the different metrics on sub-groups of stimuli, where only one view synthesis mode and one codec were considered. Figure 2 shows the minimum and maximum PCC values (across all metrics) for the different sub-groups. It can be

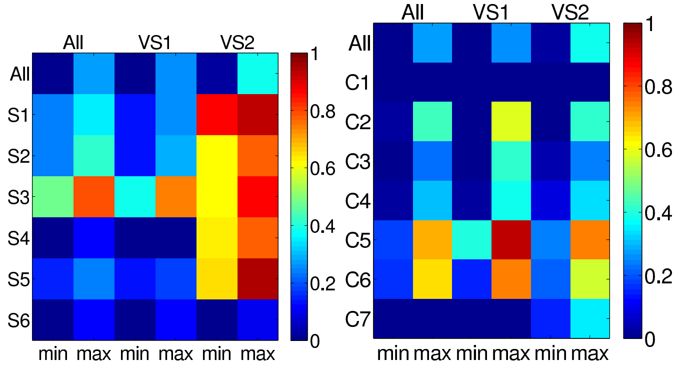


(a) *VS1* and *VS2*

(b) *VS1* only

(c) *VS2* only

Figure 3 Accuracy, consistency, and monotonicity indexes when considering only content *S1*.



(a) Contents

(b) Codecs

Figure 2 Minimum and maximum PCC values across all metrics for the different sub-groups.
Reading: Contents, line 4, columns 1 and 2: $\min(PCC) \approx 0.5$, $\max(PCC) \approx 0.8$ for content *S3* when all synthesis modes are considered.

observed that the correlation can be quite high when only *VS2* is considered and the analysis is performed for each content separately (except for content *S6*, where the correlation remains very low).

Figure 3 depicts the accuracy, consistency, and monotonicity indexes when considering only content *S1*. The results show that there is a poor correlation between objective metrics and perceived quality when the views are not blended (*VS1*, see Fig. 3(b)) as the PCC value is lower than 0.25 on all metrics, whereas the correlation is high when the views are blended (*VS2*, see Fig. 3(c)) as the PCC value is higher than 0.8 on all metrics. These results show that the objective metrics can achieve a good correlation with perceived quality if content characteristics are considered, but cannot handle the artifacts produced by some view synthesis algorithms.

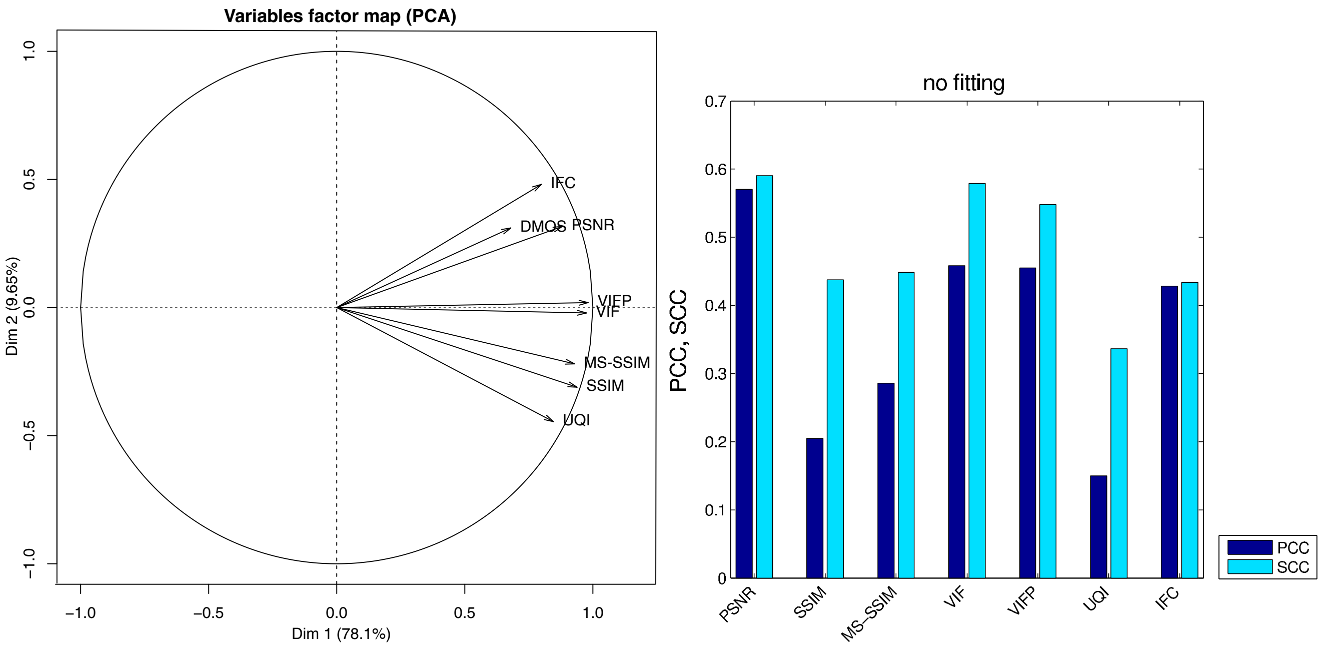
C. A specific case: C6

Figures 4(a) and (b) show the circle of correlations derived from PCA and the Pearson and Spearman correlations coefficients with subjective and objective scores of *C6* related stimuli only. In Fig. 4(a), the two principal components explained 87.7% of the variance of the data. The variables for which the contribution value is larger than the average contribution for the first component are VIFP, VIF, SSIM, and MS-SSIM, according to the principal component coefficients. These objective metrics are known to be perception-oriented. The variables for which the contribution value is larger than the average contribution for the second component are IFC, UQI, DMOS, PSNR, and SSIM. In addition, Fig. 4(a) shows that the angle between the vectors representing DMOS, IFC, and PSNR are very small, which indicates a large correlation between these variables. These results are in line with the obtained correlation scores in Fig. 4(b) regarding PSNR: according to Pearson and Spearman coefficients, PSNR is the most correlated objective metric. Our observation of *C6* related depth maps shows that this coding method distorts only slightly small pixel blocks around the edges. So the quality of the resulting synthesized views is close to that of the reference stimuli, which explains the higher objective scores.

These observations show that objective metrics are strongly content dependent, as previously shown in [20]. Therefore, content characteristics should be considered by objective metrics or the benchmarking of objective metrics should be made on a per content basis for fair comparison.

VI. Conclusion

In this paper, we analyzed the performance of several commonly used objective quality metrics on free-viewpoint video sequences using subjective scores as ground truth. The considered free-viewpoint video sequences were generated from decompressed data and simulating a smooth camera motion during a time freeze. The results showed that objective metrics achieved low correlation with subjective scores when various conditions were considered. However, the correlation with perceived quality improved when content characteristics were considered. In addition, the artifacts produced by some view synthesis algorithms might not be correctly handled by the objective quality metrics. These results motivate the need to design better objective metrics that can accurately assess the specific artifacts generated by the view synthesis process.



(a) PCA

(b) Correlation

Figure 4 Circle of correlations and Pearson and Spearman correlation scores between DMOS and objective scores when considering subjective and objective scores for compression algorithm *C6*.

Acknowledgment

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1), French ANR-PERSEE project (ANR-09-BLAN-0170), and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET.

References

[1] A. Smolic, K. Mueller, P. Merkle, P. Kauff, and T. Wiegand, "An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution," in *Picture Coding Symposium (PCS)*, May 2009, pp. 1–4.

[2] C. Fehn, "3D-TV using depth-image-based rendering (DIBR)," in *Picture Coding Symposium (PCS)*, 2004, pp. 307–312.

[3] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Muller, and T. Wiegand, "The effects of multiview depth video compression on multiview rendering," *Singal Processing: Image Communication*, vol. 24, no. 1-2, pp. 73–88, 2009.

[4] L. Do, S. Zinger, Y. Morvan, and P. de With, "Quality improving techniques in DIBR for free-viewpoint video," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, May 2009, pp. 1–4.

[5] J. Kilner, J. Starck, J. Y. Guillemaut, and A. Hilton, "Objective quality assessment in free-viewpoint video production," *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 3–16, January 2009.

[6] K.-J. Oh, S. Yea, A. Vetro, and Y.-S. Ho, "Virtual view synthesis method and self-evaluation metrics for free viewpoint television and 3d video," *International Journal of Imaging Systems and Technology*, vol. 20, no. 4, pp. 378–390, December 2010.

[7] Y. Liu, S. Ma, Q. Huang, D. Zhao, W. Gao, and N. Zhang, "Compression-Induced Rendering Distortion Analysis for Texture/Depth Rate Allocation in 3D Video Compression," in *Data Compression Conference (DCC)*, March 2009, pp. 352–361.

[8] 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, Nov. 2011.

[9] P. Hanhart and T. Ebrahimi, "Quality Assessment of a Stereo Pair Formed From Two Synthesized Views Using Objective Metrics," in *Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2013.

[10] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, "A quality assessment protocol for free-viewpoint video sequences synthesized from decompressed depth data," in *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, July 2013, pp. 100–105.

[11] ITU-T Tutorial, "Objective perceptual assessment of video quality: Full reference television," International Telecommunication Union, 2004.

[12] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, July 2012.

[13] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, January 2012.

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

[15] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, vol. 2, November 2003, pp. 1398–1402.

[16] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, February 2006.

[17] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[18] H. Sheikh, A. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[19] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," in *18th International Conference on Digital Signal Processing (DSP)*, July 2013.

[20] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, June 2008.