Visual quality assessment of synthesized views in the context of 3D-TV

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

Visual quality assessment of synthesized views in the context of 3DTV

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

Abstract DIBR is fundamental to 3DTV applications because the generation of new viewpoints is recurrent. As any tool, DIBR methods are subject to evaluations thanks to the assessment of the visual quality of the resulting generated views. This assessment task is peculiar because DIBR can be used for different 3DTV applications: either in a 2D context (free viewpoint video), or in a 3D context (3D displays reproducing stereoscopic vision). Depending on the context, the factors affecting the visual experience may differ. This chapter concerns the case of use of DIBR in the 2D context. It addresses two particular cases of use, in FTV: visualization of still images and visualization of video sequences, in the 2D context. Through these two cases, the main issues of DIBR are presented, in terms of visual quality assessment. Two experiments are proposed as case studies addressing the problematic of this chapter: the first one concerns the assessment of still images and the second one concerns the video sequences assessment. The two experiments question the reliability of subjective and objective usual tools when assessing the visual quality of synthesized view in a 2D context.

1.1. Introduction

3DTV technology has brought out new challenges such as the question of synthesized views evaluation. Indeed, the success of the two main applications referred to as "3D Video"- namely 3D Television (3DTV) that provides depth to the scene, and Free Viewpoint Video (FVV) that enables interactive navigation inside the scene ([1]) - relies on their ability to provide an added value (depth, or immersion) coupled with high-quality visual content. Depth-Image-Based-Rendering algorithms are used for virtual view generation, which is required in both applications. This process induces new types of artifacts. Consequently it impacts on the quality, which has to be identified considering various contexts of use. While many efforts have been dedicated to visual quality assessment in the last twenty years, some issues still remain unsolved in the context of 3DTV. Actually, DIBR opens new challenges because it mainly deals with geometric distortions, which have been barely addressed so far.

Virtual views synthesized either from decoded and distorted data or from original data, need to be assessed. The best assessment tool remains the human judgment as long as the right protocol is used. Subjective quality assessment is still de-

licate while addressing new type of conditions because one has to define the optimal way to get reliable data. Tests are time-consuming and consequently one should draw big lines on how to conduct such experiment to save time and observers. Since DIBR introduces new conditions, the right protocol to assess the visual quality with observers is still an unanswered question. The adequate assessment protocol might vary according to the expected answer that researchers investigate (impact of compression, DIBR techniques comparison …).

Objective metrics are meant to predict human judgment and their reliability is based on their correlation to subjective assessment results. As, the way to conduct the subjective quality assessment protocols is already questionable, the correlation between objective quality metrics, that is to say their reliability, in a DIBR context is also questionable.

Yet, trustworthy working groups base partially their future specifications, concerning new strategies for 3D video, on the outcome of objective metrics. Considering the test conditions may rely on usual subjective and objective protocols (because of their availability), the outcome of wrong choices could result to a poor quality of experience for users. Then, new tests should be carried on to determine the reliability of subjective and objective quality assessment tools in order to exploit their results for the best.

This chapter is organized as follows: first, Section 1.2 refers to the new challenges related to DIBR process. Section 1.3 gives an overview of two experiments we propose to evaluate the suitability of usual subjective assessment methods and the reliability of the usual objective metrics. Section 1.4 presents the results of the first experiment, concerning the evaluation of still images. Section 1.5 presents the results of the first experiment, concerning the evaluation of video sequences. Section 1.6 addresses the new trends regarding the assessment of synthesized views. Finally, Section 1.7 concludes the chapter.

1.2. New challenges in the DIBR context in terms of quality assessment

1.2.1. Sources of distortions

The major issue in DIBR consists in filling in the disoccluded regions of the novel viewpoint: when generating a novel viewpoint, regions that were not visible in the former viewpoint, become visible in the novel viewpoint [2]. However, the appropriate color information related to these discovered regions is often unknown. Inpainting methods that are either extrapolation or interpolation techniques, are meant to fill the disoccluded regions. However, distortions from inpainting are specific and dependant on a given hole-filling technique, as observed in [3].

Another noticeable problem refers to the rounding of pixel positions when projecting the color information in the target viewpoint (3D warping process): the pixels mapped in the target viewpoint may not locate at an integer position. In this case the position is either rounded to the nearest integer or interpolated.

Finally, another source of distortion relies on the depth map uncertainties. Errors in depth maps estimation cause visual distortion in the synthesized views because the color pixels are not correctly mapped. As well, the problem is similar when depth maps suffer important quantization from compression methods [4].

1.2.2. Examples of distortions

In this section, typical DIBR artifacts are described. As explained above, the sources of distortions are various and their visual effect on the synthesized views are perceptible as in the spatial domain as in the temporal domain. In most of the cases, these artifacts are located around large depth discontinuities, but they are more noticeable in case of high texture contrast between background and foreground.

Object shifting: a region may be slightly translated or resized, depending on the chosen extrapolation method (if the method chooses to assign the background values to the missing areas, object may be resized), or on the encoding method (blocking artifacts in depth data result in object shifting in synthesis). Figure 1**Erreur ! Source du renvoi introuvable.** depicts this type of artifact.

a b

Figure 1: Shifting/Resizing artifacts. The shape of the leaves, in this figure, is slightly modified (thinner or bigger). The vase is also moved to the right.

Blurry regions: This may be due to the inpainting method used to fill the disoccluded areas. It is obvious around the background/foreground transitions. These remarks are confirmed on Figure 2 around the disoccluded areas.

a b

Figure 2: Blurring artifacts (*Book Arrival*). a: original frame. b: synthesized frame.

Incorrect rendering of textured areas: inpainting methods can fail in filling complex textured areas. To overcome these limitations, a hole filling approach based on patch-based texture synthesis is proposed in [5].

Flickering: when errors occur randomly in depth data along the sequence, pixels are wrongly projected: some pixels suffer slight changes of depth, which appears as flickers in the resulting synthesized pixels. To avoid this methods such as [6] propose to acquire background knowledge along the sequence and to consequently improve the synthesis process.

Tiny distortions: in synthesized sequences, a large number of tiny geometric distortions and illumination differences are temporally constant and perceptually invisible. Due to the rounding decimal point problem mentioned in Section 1.2.1 and to depth inaccuracy, slight errors may occur when affecting a color value to a pixel in the target viewpoint. This leads to tiny illumination errors that may not be perceptible to human. However, pixel-based metrics may penalize these distorted zones.

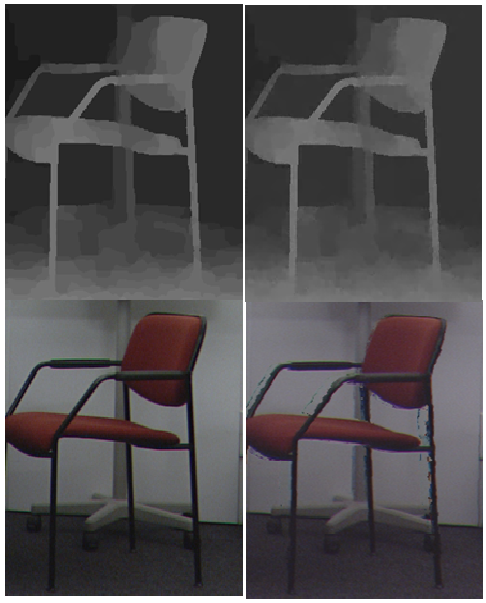
When encoding either depth data or color sequences before performing the synthesis, compression-related artifacts are combined with synthesis artifacts. Artifacts from data compression are generally spread within the whole image, while artifacts inherent to the synthesis process are mainly located around the disoccluded areas. The combination of both type of distortion, depending on the compression method, relatively affects the synthesized view. Indeed, most of the used compression methods are 2D video codecs inspired, and are thus optimized for human perception of color. As a result, artifacts occurring especially in depth data induce severe distortions in the synthesized views. In the following, a few examples of such distortions are presented.

Blocking artifacts: this occurs when the compression method induces blocking artifacts in depth data. In the synthesized views, whole blocks of color image seem to be translated. Figure 3 illustrates the distortion.

a b

Figure 3: Blocking artifacts from depth data compression result in distorted synthesized views (Breakdancers). a: Original depth frame (up) and color original frame (bottom). b: Distorted depth frame (up), synthesized view (bottom).

Ringing artifacts: when ringing artifacts occur in depth data around strong discontinuities, objects' edges appear distorted in the synthesized view. Figure 4 depicts this artifact.



a b

Figure 4: Ringing artifacts in depth data lead to distortions in the synthesized views. a: Original depth frame (up) and original color frame (bottom). b: Distorted depth frame (up) and synthesized frame (bottom).

1.2.3. The peculiar task of assessing the synthesized view

The evaluation of DIBR systems is a difficult task because depending on the application (FTV or 3DTV), the type of evaluation differs. Not the same factors are involved in the two applications. The main difference between the two applications is the stereopsis phenomenon (fusion of left and right views in human visual system). This is used by 3DTV and this reproduces vision in relief. This includes psycho physiological mechanisms whose understanding is not complete so far. A FTV application does not have to be used in a 3D context. FTV can be applied in a 2D context. Consequently, the quality assessments protocols differ and address the quality of the synthesized view in two different contexts. It is obvious that stereoscopic impairments (such as cardboard effect, crosstalk, etc. as described in [7] and [8]), which occur in stereoscopic conditions, are not assessed in 2D conditions. As well, distortions detected in 2D conditions may not be perceptible in a 3D context.

Finally, artifacts, in DIBR, are mainly geometric distortions. These distortions are different from those commonly encountered in video compression, and assessed by usual evaluation methods: most video coding standards rely on DCT, and the resulting artifacts are specific (some of them are described in [9]). These artifacts are often scattered in the whole image, although DIBR related artifacts are mostly located around the disoccluded regions. Yet, most of the usual objective quality metrics were initially created to address usual specific distortions and may be unsuitable to the problem of DIBR evaluation. This will be discussed in Section 1.3.

Another aspect concerns the need for non-reference quality metrics. In particular cases of use, like FTV, references are unavailable because the generated viewpoint is virtual. In other words, there is no ground truth allowing a full comparison with the distorted view.

The next section addresses two case studies that question the validity of subjective and objective quality assessment methods for the evaluation of synthesized view in 2D conditions.

1.3. Two case studies to question the evaluation of synthesized view

In this section, we first present the aim of the studies, and the experimental material. Then we present the two subjective assessment methods whose suitability has been questioned in our experiments. We also justify the choice of these two methods. Finally we present a selection of the most commonly used metrics that also were included in our experiments.

1.3.1. Goal of the studies

We conducted two different studies. The first one addresses the evaluation of still images. An obviously important scenario to consider is the case in which the user switches the video to the “pause” mode. This case should be treated because it is likely to occur and may be subject to meticulous observation. The second study addresses the evaluation of video sequences.

The two studies question the reliability of subjective and objective assessment methods when evaluating the quality of the synthesized view. Most of the proposed metrics for assessing 3D media are inspired from 2D quality metrics. Previous studies ([10], [11]) already considered the reliability of usual objective metrics. However, often, experimental protocols 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 at the same time both compression distortion and synthesis distortion, without distinction. This is problematic because there

may be a combination of artifacts from various sources (compression and synthesis) whose effects are not clearly specified and assessed. The studies presented in this chapter concerns only synthesized views, observed in 2D conditions.

The rest of this section present the experimental material, the subjective methodologies and the objective quality metrics used in the studies.

1.3.2. Experimental material

Three different multiview plus depth (MVD) sequences are used in the two studies. The sequences are *Book Arrival* (1024x768, 16 cameras with 6.5cm spacing), *Lovebird1* (1024x768, 12 cameras with 3.5 cm spacing) and *Newspaper* (1024x768, 9 cameras with 5 cm spacing).

Seven DIBR algorithms processed the three sequences to generate, for each sequence, four different viewpoints.

These seven DIBR algorithms are labeled from A1 to A7:

- A1: based on Fehn [12], 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 [12]. Borders are inpainted by the method proposed by Telea [13].
- A3: Tanimoto et al. [14], it is the recently adopted reference software for the experiments in the 3D Video group of MPEG.
- A4: Müller et al. [15], proposed a hole filling method aided by depth information.
- A5: Ndjiki-Nya et al. [5], the hole filling method is a patch-based texture synthesis.
- A6: Köppel et al. [6], uses depth temporal information to improve the synthesis in the disoccluded areas.
- A7: corresponds to the unfilled sequences (i.e. with holes).

The test was conducted in an ITU conforming test environment.For the subjective assessments, the stimuli were displayed on a TVLogic LVM401W, and according to ITU-T BT.500 [16]. In the following, the subjective methodologies are first presented, and then the objective metrics are addressed.

Objective measurements were obtained by using MetriX MuX Visual Quality Assessment Package [17].

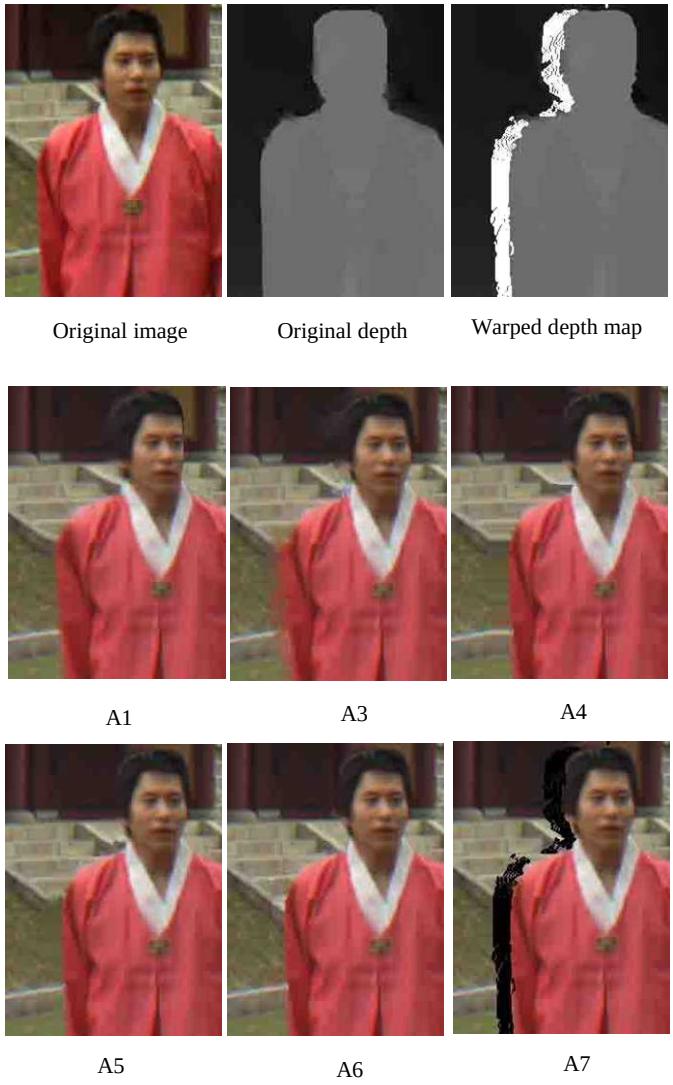


Figure 5: Synthesized frames ("*Lovebird1*" sequence)

1.3.3. Subjective assessment methodologies

Subjective tests are used to measure image or video quality. The International Telecommunications Union (ITU) [18] is in charge for the recommendations of the most commonly used subjective assessment methods. Several methods exist but there is no 3D-dedicated protocol. The available protocols both have their drawbacks and advantages and they are usually chosen according to the desired task. This depends on the distortion and on the type of evaluation [19]. They differ according to the type of pattern presentation (single-stimulus, double-stimulus, multi-stimulus), the type of voting (quality, impairment, or preference), the voting scale (discrete or continuous), the number of rating points or categories. **Erreur ! Source du renvoi introuvable.** depicts the proposed classification of subjective methods in [19]. The abbreviations of the methods classified in **Erreur ! Source du renvoi introuvable.** are referenced in Table 1.

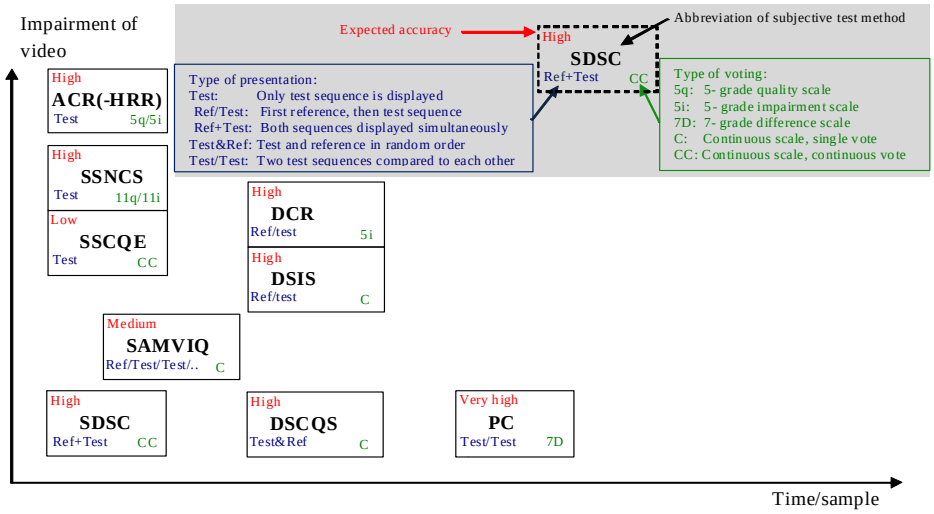


Figure 6: Commonly used subjective test methods, as depicted in [19]

Table 1: Overview of subjective test methods

Abbrev.	Full meaning	Ref.
DSIS	Double Stimulus Impairment Scale	[16]
DSQS	Double Stimulus Quality Scale	[16]
SSNCS	Single Stimulus Numerical Categorical Scale	[16]
SSCQE	Single Stimulus Continuous Quality Evaluation	[16]
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation	[16]
ACR	Absolute Category Rating	[18]
ACR-HR	Absolute Category Rating with Hidden Reference removal	[18]
DCR	Degradation Category Rating	[18]
PC	Pair Comparison	[18]
SAMVIQ	Subjective assessment Methodology for Video Quality	[18]

In the absence of any better 3D-adapted subjective quality assessment methodologies, the evaluation of synthesized views is mostly obtained through 2D validated assessment protocols. The aim of our two experiments is to question the suitability of a selection of subjective quality assessment methods. This selection is based on the comparisons of methods in the literature. Considering the aim of the two experiments that we proposed, the choice of a subjective quality assessment method should relies on consideration of reliability, accuracy, efficiency and easiness of implementation of the available methods.

Brotherton *et al*. [20] investigated the suitability of ACR and SAMVIQ methods when assessing 2D media. The study shown that ACR method allowed more test sequences (at least twice) to be presented for assessment compared to the SAMVIQ method. ACR method also proved to be reliable in the test conditions. Rouse *et al.* also studied the tradeoff of these two methods in [21], in the context of high definition still images and video sequences. They concluded that the suitability of the two methods may depend on specific applications.

A study was conducted by Huynh-Thu *et al.* in [22], and proposed to compare different methods according to their different voting scales (5-point discrete, 9-point discrete, 5-point continuous, 11-point continuous scales). The tests were carried in the context of high-definition video. The results shown that ACR method produced reliable subjective results, even across different scales.

Considering the classification of the methods, we selected the single-stimulus pattern presentation, ACR-HR (with 5 quality categories) and the double-stimulus pattern presentation PC for its accuracy. They are described and commented in the following.

Absolute categorical rating with Hidden Reference Removal (ACR-HR) methodology consists in presenting test objects (i.e. images or sequences) to observers one at a time. The objects are rated independently on category scale. The reference version of each object must be included in the test procedure and rated as any other stimulus. This explains the used term of ‘hidden reference”. From the scores obtained, a differential score (DMOS for Differential Mean Opinion Score) is computed between the mean opinion scores (MOS) of each test object and its

associated hidden reference. ITU recommends the 5-level quality scale depicted in Table 2.

Table 2 ACR-HR quality scale

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

ACR-HR requires many observers to minimize the contextual effects (previously presented stimuli influence the observer opinion, i.e. presentation order influences opinion ratings). Accuracy increases with the number of participants.

Paired comparisons (PC) methodology is an assessment protocol in which stimuli are presented by pairs to the observers: it is a double-stimulus method. The latter select the one out of the pair that best satisfies the specified judgment criterion, i.e. image quality.

The results of a paired comparisons test are recorded in a matrix: each element corresponds to the frequencies a stimulus is preferred over another stimulus. These data are then converted to scale values using Thurstone-Mosteller's or Bradley-Terry's model [23]. It leads to a hypothetical perceptual continuum.

The presented experiments follow Thurstone-Mosteller's model where naive observers were asked to choose the preferred item from one pair. Although the method is known to be highly accurate, it is time consuming.

The differences between ACR-HR and PC are of different types. First, with ACR-HR, even though they may be included in the stimuli, the reference sequences are not identified as such by the observers. Observers provide an absolute vote without any reference. In PC, observers only need to indicate their preference out of a pair of stimuli. Then the requested task is different: while observers assess the quality of the stimuli in ACR-HR, they just provide their preferences in PC.

The quality scale is another issue. ACR-HR scores provide knowledge on the perceived quality level of the stimuli. However the voting scale is coarse, and because of the single stimulus presentation, observers cannot remember previous stimuli and precisely evaluate small impairments. PC scores (i.e. “preference matrices”) are scaled to a hypothetical perceptual continuum. However, it does not provide knowledge on the quality level of the stimuli, but on the stimuli order of preferences. Moreover, PC is very well suited for small impairments, thanks to the fact that only two conditions are compared to each other. For these reasons, PC tests are often coupled with ACR-HR tests.

Another aspect concerns the complexity and the feasibility of the test: PC is simple because observers only need to provide preference in each double stimulus. However, when the number of stimuli increase, the test becomes hardly feasible as the number of comparisons grows as $\frac{N(N-1)}{2}$ with N, the number of stimuli. In the case of video sequences assessment, a double-stimulus method such as PC involves the use of either one split-screen environment (or two full screens), with the risk of distracting the observer (as explained in [24]), or one screen but sequences are displayed one after the other, which increases the length of the test. On the other hand, the simplicity of ACR-HR allows the assessment of a larger number of stimuli. However, the results of this assessment are reliable as long as the group of participants is large enough.

1.3.4. Objective quality metrics

The experiments that are proposed in this chapter require the use of objective quality metrics. The choice of the objective metrics used in these experiments is motivated by their availability. This section presents an overview of the objective metrics used in these experiments. Still-images and video sequences metrics are presented.

Objective metrics are meant to predict human perception of quality of images and thus avoid spending time in subjective quality assessment tests. They are then supposed to be highly correlated with human opinion. In the absence of approved metrics for assessing synthesized views, most of the studies rely on the use of 2D validated metrics, or on adaptations of such. There are different types of objective metrics, depending on their requirement for reference images. The objective metrics can be classified in three different categories according to the availability of the reference image: full reference methods (FR), reduced reference methods (RR), no-reference methods (NR). FR methods require references images. Most of the existing metrics rely on FR methods. RR methods require only elements of the reference images. NR methods do not require reference images. NR methods mostly rely on Human Visual System models to predict human opinion of the quality. Also, a prior knowledge on the expected artifacts highly improves the design of such methods.

As proposed in [25], we use a classification relying on tools used in the methods. Table 1Table 3 lists a selection of commonly used objective metrics and Figure 7 depicts the proposed classification.

Table 3 Overview of commonly used objective metrics

	Objective metric	**Abbrev.**
Signal-based	Peak Signal to Noise Ratio	PSNR
Perceptual-like	Universal Quality Index	UQI
	Information Fidelity Criterion	IFC
	Video Quality Metric	VQM
	Perceptual Video Quality Measure	PVQM

Structural-based	Single-scale Structural SIMilarity	SSIM
	Multi-scale SSIM	MSSIM
	Video Structural Similarity Measure	V-SSIM
	Motion-based Video Integrity Evaluation	MOVIE
HVS-based	PSNR- Human Visual System	PSNR-HVS
	PSNR-Human Visual System Masking model	PSNR-HVSM
	Visual Signal to Noise Ratio	VSNR
	Weighted Signal to Noise Ratio	WSNR
	Visual Information Fidelity	VIF
	Moving Pictures Quality Metric	MPQM

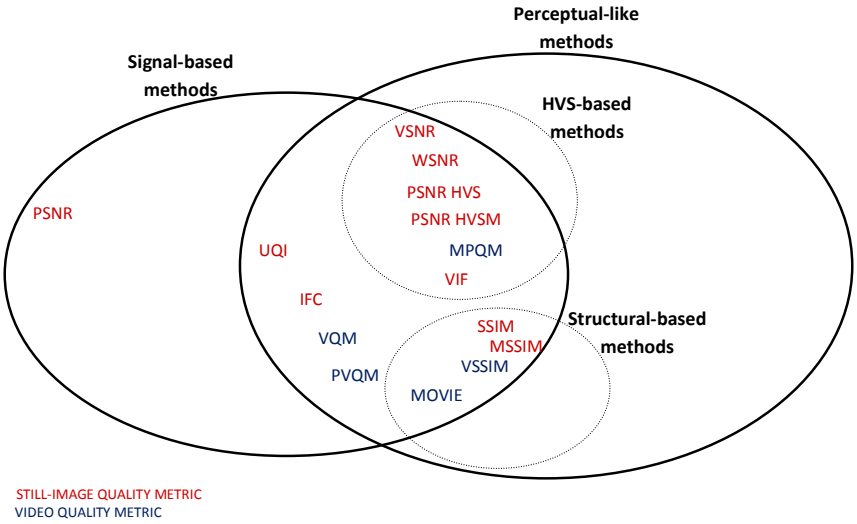


Figure 7: Overview of quality metrics

Signal-Based methods:

PSNR is a widely used method because of its simplicity. PSNR belongs to the signal-based methods category. It measures the signal fidelity of a distorted image compared to a reference. It is based on the measure of the Mean Squared Error (MSE). Because of the pixel-based approach of such a method, the amount of distorted pixels is depicted, but the perceptual quality is not: PSNR does not take into account the visual masking phenomenon. Thus, even if an error is not perceptible, it contributes to the decrease of the quality score. Indeed, studies (such as [26]) showed that in the case of synthesized views, PSNR is not reliable,

especially when comparing two images with low PSNR scores. PSNR cannot be used in very different scenario as explained in [27].

Perceptual-like methods:

Considering that signal-based methods are unable to correctly predict the perceptual quality, perceptual-like metrics have been introduced. They make use of perceptual criterion such as luminance or contrast distortion.
UQI [28] is a perceptual-like metric. The quality score is the product of the correlation between the original and the degraded image, a term defining the luminance distortion, a term defining the contrast distortion. The quality score is computed within a sliding window and the final score is defined as the average of all local scores.
IFC [29] uses a distortion model to evaluate the information shared between the reference image and the degraded image. IFC indicates the image fidelity rather than the distortion. IFC is based on the hypothesis that, given a source channel and a distortion channel, an image is made of multiple independently distorted subbands. The quality score is the sum of the mutual information between the source and the distorted for all the subbands.
VQM was proposed by Pinson and Wolf in [30]. It is a FR video metric that measures perceptual effects of numerous video distortions. It includes a calibration step (to correct spatial/temporal shift, contrast, and brightness according to the reference video sequence), an analysis of perceptual features. VQM score combines all the perceptual calculated parameters. VQM method is complex but the correlation to subjective scores is good according to [31]. The method is validated in typical video processing conditions.
Perceptual Video Quality Measure (PVQM) [32] is meant to detect perceptible distortions in video sequences. Different indicators are used. First, an edge-based indicator allows the detection of distorted edges in the images. Second, a motion-based indicator analyses two successive frames. Third, a color-based indicator detects non-saturated colors. Each indicator is pooled separately across the video and incorporated in a weighting function to obtain the final score. This method was not available so it was not tested in our experiments.

Structural-based methods:

Structural-based methods are also included in the perceptual-like metrics. They are based on the assumption that human perception is based on the extraction of structural information. Thus, they measure the structural information degradation. SSIM [33] was the first method of this category. It is considered as an extension of UQI. It combines image structural information: mean, variance, covariance of pixels, for a single local patch. The blocksize depends on the viewer distance to the screen. However, a low variation of the SSIM measure, can lead to an important error of MOS prediction.

Then, many improvements to SSIM were proposed, and adaptations to video assessment were introduced. MSSIM is the average SSIM scores of all patches of the image. V-SSIM [34] is a FR video quality metric which uses structural distortion as an estimate of perceived visual distortion. At the patch level, SSIM score is a weighted function of SSIM of the different component of the image (i.e. luminance, and chromas). At the frame level, SSIM score is a weighted function of patches' SSIM scores (based on the darkness of the patch). Finally at the sequence level, VSSIM score is a weighted function of frames' SSIM scores (based on the motion). The choice of the weights relies on the assumption that dark regions are less salient. However, this is questionable because the darkness may depend on the used screen.
MOVIE [35] is a FR video metric that uses several steps before computing the quality score. It includes the decomposition of both reference and distorted video by using a multi-scale spatio-temporal Gabor filter-bank. A SSIM-like method is used for the spatial quality analysis. An optical flow calculation is used for the motion analysis. Spatial and temporal quality indicators determine the final score.

Human-Visual-System (HVS)-based methods:

HVS-based methods rely on human visual system modelling from psychophysics experiments. Due to the complexity of the human vision, studies are still in progress. HVS-based models are the result of tradeoffs between computational feasibility and accuracy of the model. HVS-based models can be classified into two categories: neurobiological models and models based on psychophysical properties of human vision.
The models based on neurobiology estimate the actual low-level process in human visual system including the eye and optical nerve. However, these models are not widely used, because of their complexity [36].
Psychophysical HVS-based models are implemented in a sequential process that includes luminance masking, color perception analysis, frequency selection, contrast sensitivity implementation (based on the contrast sensitivity function CSF [37]) and modeling of masking and facilitation effects [38].
PSNR-HVS [39], based on PSNR and UQI, takes into account the Human Visual System (HVS) properties such as its sensitivity to contrast change and to low frequency distortions. In [39], the method proved to be correlated to subjective scores, but the performances of PSNR-HVS method are tested on a variety of distortions specific to 2D image compression which are different from distortions related to DIBR.
PSNR-HVSM [40] is based on PSNR but takes into account Contrast Sensitivity Function (CSF) and between-coefficient contrast masking of DCT basis functions. The performances of the method are validated considering a set of images containing Gaussian noise or spatially correlated additive Gaussian noise, at different locations (uniformly through entire image, mostly in regions possessing a high masking effect or, mostly in regions possessing a low masking effect).

VSNR[41] is also a perceptual-like metric: it is based on a visual detection of distortion criterion, helped by CSF. VSNR metric is sensitive to geometric distortions such as spatial shifting and rotations, transformations which are typical in DIBR applications.
WSNR that uses a weighting function adapted to HVS denotes a weighted Signal to Noise Ratio, as applied in [42] . It is an improvement of PSNR that that uses a CSF-based weighting function. However, although SNR is more accurate by taking into account perceptual properties, as with PSNR method, the problem remains the accumulation of degradations errors even in non-perceptible areas.
IFC has been improved by the introduction of a HVS model. The method is called VIF[43]. VIFP is a pixel-based version of VIF. It uses wavelet decomposition and computes the parameters of the distortion models, which enhance the computational complexity. In [43], five distortion types are used to validate the performances of the method (JPEG and JPEG 200 related distortions, white and Gaussian noise over the entire image), which are quite different from the DIBR related artefacts.
MPQM [44] uses a HVS model. In particular it takes into account the masking phenomenon and the contrast sensitivity. It has high complexity and its correlation to subjective scores is varying according to [31]. Since, the method is not available it is not tested in our experiments.

Only a few commonly used algorithms (in the 2D context) have been described above. Since they are all dedicated to 2D applications, they are optimized to detect and penalize specific distortions of 2D image and video compression. As explained in 1.2, distortions related to DIBR are very different from 2D known artefacts. There exist many other algorithms for visual quality assessment that are not covered here.

1.3.5. Experimental protocols

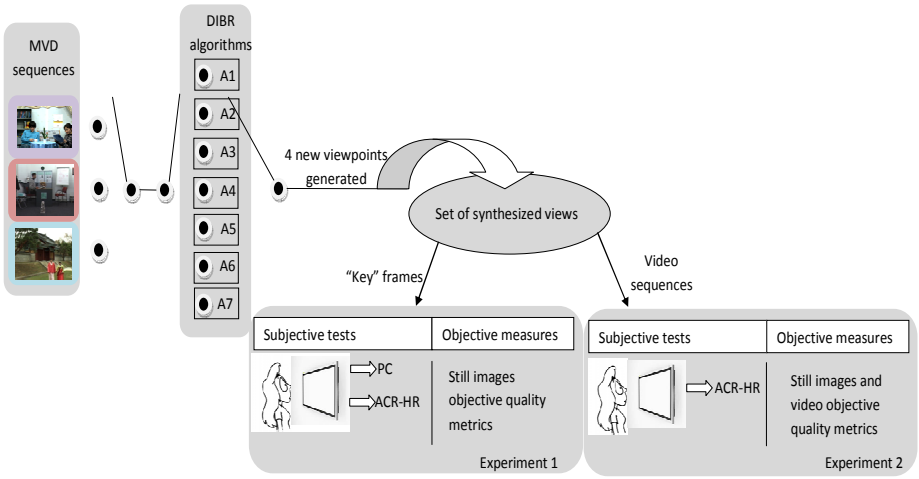


Figure 8: Experimental protocols

Two experiments were conducted. The first one addresses the evaluation of still images. The second study addresses the evaluation of video sequences. Figure 8 depicts the overview of the two experiments.

The material for both experiments comes from the same set of synthesized views as described in Section 1.3.2. However, in the case of the first experiments, on still-images, the test images are "key" frames ("keys" were randomly chosen) from the same set of synthesized views, due to the complexity of PC tests when number of items increases. That is to say that for each of the three reference sequences, only one frame was selected out of each synthesized view viewpoint.

In both experiments, the suitability of subjective quality assessment methods and the reliability of objective metrics are addressed.

Concerning the subjective tests, two sessions were conducted. The first one addressed the assessment of the still images. Forty-three naive observers participated in this test. The second session addressed the assessment of the video sequences. Thirty-two naïve observers participated in this test.

In the case of video sequences, only ACR-HR test was conducted, but both ACR-HR and PC were carried for the still-images context. PC test with video sequences would have required either two screens, or switching between items. In the case of the use of two screens, it involves the risk of missing frames of the tested sequences, because one cannot watch simultaneously two different video sequences. In the case of the switch, it would have increased considerably the length of the test.

The objective measurements were realized over the 84 synthesized views by the means of MetriX MuX Visual Quality Assessment Package [17] software except for two metrics: VQM and VSSIM. VQM were available at [45]; VSSIM was

implemented by the authors, according to [34]. The reference was the original acquired image. It should be noted that still image quality metrics used in the study with still images, are also used to assess the visual video sequences quality by applying these metrics on each frame separately and averaging the frames scores.

Table 4 summarizes the experimental framework. The next sections present the results of the first experiment assessing the quality of still-images, and then the results of the second experiment assessing the quality of video sequences.

Table 4 Overview of the experiments

		Experiment 1(still-images)	**Experiment 2 (video sequences)**
Data		Key frames of each synthesized view	Synthesized video sequences
Subjective tests	**Nb. of participants**	43	32
	Methods	ACR-HR, PC	ACR-HR
Objective measures		All available metrics of MetriX MuX	VQM, VSSIM, Still-image metrics

1.4. Results on still images (experiment 1)

1.4.1. Subjective tests

The seven DIBR algorithms are ranked according to the obtained ACR-HR and PC scores, as depicted in Table 5. This table indicates that the rankings obtained by both testing method are consistent. For ACR-HR test, the first line gives the DMOS scores obtained through the MOS scores. For PC test, the first line gives the hypothetical MOS scores obtained through the comparisons. For both tests, the second line gives the rankings of the algorithms, obtained through the first line.

Table 5 Rankings of algorithms according to subjective scores

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**
ACR-HR	3.539	3.386	3.145	3.40	3.496	3.32	2.278
Rank order	1	4	6	3	2	5	7
PC	1.038	0.508	0.207	0.531	0.936	0.454	-2.055
Rank order	1	4	6	3	2	5	7

In Table 5, although the algorithms can be ranked from the scaled scores, there is no information concerning the statistical significance of the quality difference of two stimuli (one more preferred than another one). Then statistical analyses have been conducted over the subjective measurements: a student's t-test has been performed over ACR-HR scores, and over PC scores for each algorithm. This provides knowledge on the statistical equivalence of the algorithms. Table 6 and Ta-

ble 7 show the results of the statistical tests over ACR-HR and PC values respectively. In both tables, the number in parentheses indicates the minimum required number of observers that allows statistical distinction (VQEG recommends 24 participants as a minimum [46], values in bold are higher than 24 in the table).

A first analysis of these two tables indicates that PC method leads to clear-cut decisions, compared to ACR-HR method: indeed, the distributions of the algorithms are statistically distinguished with less than 24 participants in 17 cases with PC (only 11 cases with ACR-HR). In one case (between A2 and A5), less than 24 participants are required with PC, and more than 43 participants are required to establish the statistical difference with ACR-HR. The latter case can be explained by the fact that the visual quality of the synthesized images (and thus, some distortions) may seem very similar for non-expert observers. This makes the task more delicate for observers. These results indicate that it seems more difficult to assess the quality of synthesized views than in other contexts (for instance, quality assessment of images distorted through compression). Indeed, the results with ACR-HR method, in Table 6, confirm this idea: in most of the cases, more than 24 participants (or even more than 43) are required to distinguish the classes (Remember that A7 is the synthesis with holes around the disoccluded areas).

However, as seen with rankings results above, methodologies give consistent results: when the distinctions between algorithms are stable, they are the same with both methodologies.

Finally, these experiments show that fewer participants are required for a PC test than for an ACR-HR test. However, as stated before, PC tests, while efficient, are feasible only with a limited number of items to be compared. Another problem, pointed out by these experiments, concerns the assessment of similar items: with both methods, 43 participants were not always sufficient to obtain a stable and reliable decision. Results suggest that observers had difficulties assessing the different types of artefacts.

Table 6 Results of Student's t-test with ACR-HR results Legend: ↑: superior, ↓: inferior, ^{O}: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "32" observers participate.

	A1	A2	A3	A4	A5	A6	A7
A1		**↑(32)**	↑(<24)	**↑(32)**	**^{O} (>43)**	**↑(30)**	↑(<24)
A2	**↓(32)**		↑(<24)	**^{O} (>43)**	**^{O} (>43)**	**^{O} (>43)**	↑(<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 7 Results of Student's t-test with PC results. Legend: ↑: superior, ↓: inferior, ^{O}: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "less than 24" observers participate.

	A1	A2	A3	A4	A5	A6	A7
A1		↑ (<24)	↑(<24)	↑(<24)	↑(<24)	↑(<24)	↑(<24)
A2	↓(<24)		**↑(28)**	**^{O} (<43)**	↓(<24)	**^{O} (>43)**	↑(<24)
A3	↓(<24)	**↓(28)**		↓(<24)	↓(<24)	↓(<24)	↑(<24)
A4	↓(<24)	**^{O} (>43)**	↑(<24)		↓(<24)	**↑ (>43)**	↑(<24)
A5	↓(<24)	↑(<24)	↑(<24)	↑(<24)		↑(<24)	↑(<24)
A6	↓(<24)	**^{O} (>43)**	↑(<24)	**↓(>43)**	↓(<24)		↑(<24)
A7	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	

As a conclusion, this first analysis, involving still images quality assessment, reveals that more than 24 participants may be necessary for these types of test.

PC gives clear-cut decisions, due to the mode of assessment (preference) while algorithm's statistical distinctions with ACR-HR are slightly less accurate. With ACR-HR, the task is not easy for the observers because, although each DIBR induces specific artifacts, the impairments among the tested images are small. Thus, when evaluating the performances of different DIBR algorithms with this methodology, this aspect should be taken into account.

However, ACR-HR and PC are complementary: when assessing similar items, like in this case study, PC can provide a ranking, while ACR-HR gives the overall perceptual quality of the items.

1.4.2. Objective measurements

The results of this subsection concerns the measurements conducted over the same selected "key" frames.

The whole set of objective metrics give the same trends. Table 8 provides correlation coefficients between obtained objective scores. It reveals that they are highly correlated. This table shows that the behavior of the tested metrics was the same when assessing images containing DIBR related artifacts. Thus, they have the same response when assessing DIBR related artifacts. Note the high correlation scores between pixel-based and more perceptual-like metrics such as PSNR and SSIM (83.9%).

The first step consists in comparing the objective scores with the subjective scores (in section 1.4.1). The consistency between objective and subjective measures is evaluated by calculating the correlation coefficients for the whole fitted measured points. The coefficients are presented in Table 9. In the results of our test, none of the tested metric reaches 50% of human judgment. This reveals that contrary to the received opinion, the objective tested metrics, whose efficiency has been proved for the quality assessment of 2D conventional media, do not reliably predict human appreciation in the case of synthesized views.

Since it is argued in [47] that correlation is different from agreement (as illustrated in Figure 9), we check the agreement of the tested metrics by comparing the ranks affected to the algorithms. Table 10 presents the rankings of the algorithms, obtained from the objective scores. Rankings from subjective scores are mentioned for comparison. They present a noticeable difference concerning the ranking order of A1: judged as the best algorithm out of the seven by the subjective scores, it is ranked as the worst by the whole set of objective metrics. Another comment refers to the assessment of A6: often judged as the best algorithm, it is judged as one of the worst algorithms through the subjective tests. The ensuing assumption is that objective metrics detect and penalize non-annoying artifacts.

Table 8 Correlation coefficients between objective scores in percentage

	PSNR	**SSIM**	**MSSIM**	**VSNR**	**VIF**	**VIFP**	**UQI**	**IFC**	**NQM**	**WSNR**	**PSNR $_{HSVM}$**	**PSNR $_{HSV}$**
PSNR		83.9	79.6	87.3	77.0	70.6	53.6	71.6	95.2	98.2	99.2	99.0
SSIM	83.9		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 HSVM	99.2	83.2	77.9	86.1	79.4	72.2	50.9	73.5	92.3	97.4		99.9
PSNR HSV	99.0	83.5	78.3	85.8	80.2	72.9	50.8	74.4	91.8	97.1	99.9	

Table 9 Correlation coefficients between objective and subjective scores in percentage

	PSNR	**SSIM**	**MSSIM**	**VSNR**	**VIF**	**VIFP**	**UQI**	**IFC**	**NQM**	**WSNR**	**PSNR HVSM**	**PSNR HVS**
ACR-HR	31.1	19.9	11.4	22.9	19.6	21.5	18.4	21.0	29.5	37.6	31.7	31.0
PC	40.0	23.8	34.9	19.7	16.2	22.0	32.9	20.1	37.8	36.9	42.2	41.9

Table 10 Rankings according to measurements

	A1	A2	A3	A4	A5	A6	A7
ACR-HR	2.388	2.234	1.994	2.250	2.345	2.169	1.126
Rank order	1	4	6	3	2	5	7
PC	1.038	0.508	0.207	0.531	0.936	0.454	-2.055
Rank order	1	4	6	3	2	5	7
PSNR	18.75	24.998	23.18	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
VSNR	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	1	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
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

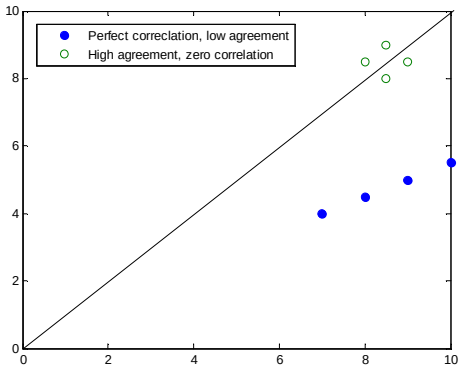


Figure 9: Difference between correlation and agreement [47]

1.5. Results on video sequences (experiment 2)

1.5.1. Subjective tests

In the case of video sequences, only ACR-HR test was conducted, as mentioned before.

Table 11 shows the algorithms' ranking from the obtained subjective scores. The ranking order differs from the one obtained with ACR-HR test in the still image context slightly vary.

Table 11 Ranking of algorithms according to subjective scores

	A1	A2	A3	A4	A5	A6	A7
ACR-HR	3.523	3.237	2.966	2.865	2.789	2.956	2.104
Rank order	1	2	3	5	6	4	7

And, still, although the values allow the ranking of the algorithms, they do not directly provide knowledge on the statistical equivalence of the results. Table 12 depicts the results of the Student's t-test processed with the values. Compared to ACR-HR test with still images detailed in section 1.4.1, distinctions between algorithms seem to be more obvious. Statistical significance of the difference between the algorithms, based on the ACR-HR scores, exists and seems clearer in the case of the video sequences than in the case of still images. This can be explained by the exhibition time of the video sequences: watching the whole video, observers can refine their judgment, compared to still images. Note that the same algorithms were not statistically differentiated: A4, A3, A5 and A6.

Table 12 Results of Student's t-test with ACR-HR results. Legend: ↑: superior, ↓: inferior, ^{O}: statistically equivalent.

	A1	A2	A3	A4	A5	A6	A7
A1		↑(7)	↑(3)	↑(3)	↑(2)	↑(3)	↑(1)
A2	↓(7)		↑(2)	↑(2)	↑(1)	↑(2)	↑(1)
A3	↓(3)	↓(2)		^{O}**(>32)**	↑(9)	^{O}**(>32)**	↑(1)
A4	↓(3)	↓(2)	^{O}**(>32)**		^{O}**(>32)**	^{O}**(>32)**	↑(1)
A5	↓(2)	↓(1)	↓(9)	^{O}**(>32)**		↓(15)	↑(1)
A6	↓(3)	↓(2)	^{O}**(>32)**	^{O}**(>32)**	↑(15)		↑(1)
A7	↓(1)	↓(1)	↓(1)	↓(1)	↓(1)	↓(1)	

As a conclusion, ACR-HR test with video sequences gives clearer statistical differences between the algorithms than ACR-HR test with still images. This suggests that new elements allow the observers to make a decision: existence of flickering, exhibition time, etc.

1.5.2. Objective measurements

The results of this subsection concern the measurements conducted over the entire synthesized sequences.

As in the case of still images studied in the previous section, the rankings of the objective metrics (Table 13) are consistent with each other: the correlation coefficients between objective metrics are very close from the figures depicted in Table 8, and so they are not presented here. As with still images, the difference be-

tween the subjective-test-based ranking and the ranking from the objective scores is noticeable. Again, the algorithm judged as the worst (A1) by the objective measurements, is the one preferred by the observers. This can be explained by the fact that A1 performs the synthesis on a cropped image, and then enlarges it to reach the original size. Consequently, signal-based metrics penalize it while it gives good perceptual results.

2. Table 13 Rankings according to measurements

	A1	A2	A3	A4	A5	A6	A7
ACR-HR	2.70	2.41	2.14	2.03	1.96	2.13	1.28
Rank order	1	2	3	5	6	4	7
PSNR	19.02	24.99	23.227	25.994	26.035	26.04	20.89
Rank order	7	4	5	3	2	1	6
SSIM	0.648	0.844	0.786	0.859	0.859	0.859	0.824
Rank order	7	4	6	1	1	1	5
MSSIMM	0.664	0.932	0.825	0.948	0.948	0.948	0.888
Rank order	7	4	6	1	1	1	5
VSNR	13.14	20.41	18.75	21.786	21.965	21.968	20.73
Rank order	7	5	6	3	2	1	4
VIF	0.129	0.393	0.313	0.423	0.423	0.424	0.396
Rank order	7	5	6	2	2	1	4
VIFP	0.153	0.415	0.342	0.446	0.446	0.446	0.419
Rank order	7	5	6	1	1	1	4
UQI	0.359	0.664	0.58	0.598	0.598	0.598	0.667
Rank order	7	5	6	3	3	3	1
IFC	0.779	2.399	1.926	2.562	2.562	2.564	2.404
Rank order	7	5	6	2	2	1	4
NQM	8.66	15.933	13.415	16.635	16.739	16.739	10.63
Rank order	7	4	5	3	1	1	6
WSNR	14.41	20.85	18.853	21.76	21.839	21.844	16.46
Rank order	7	4	5	3	2	1	6
PSNR HSVM	13.99	19.37	18.361	21.278	21.318	21.326	16.23
Rank order	7	4	5	3	2	1	6
PSNR HSV	13.74	19.52	17.958	20.795	20.823	20.833	15.91
Rank order	7	4	5	3	2	1	6
VSSIM	0.662	0.879	0.809	0.899	0.898	0.893	0.854
Rank	7	4	6	1	2	3	5
VQM	0.888	0.623	0.581	0.572	0.556	0.557	0.652
Rank order	7	5	4	3	1	2	6

Table 14 presents the correlation coefficients between objective scores and subjective scores, based on the whole set of measured points. None of the tested objective metric reaches 50% of subjective scores. The metric obtaining the higher correlation coefficient is VSNR, with 47.3%. Figure 10 shows the same obtained correlation scores, with resulting ranking of tested metrics. It is easily observed

that the top metrics are perceptual-like metrics (they include psychophysical approaches).

Table 14 Correlation coefficients between objective and subjective scores in percentage

	PSNR	SSIM	MSSIM	VSNR	VIF	VIFP	UQI	IFC	NQM	WSNR	PSNR HVSM	PSNR HVS	VSSIM	VQM
ACR -HR	34.5	45.2	27	47.3	43.9	46.9	20.2	45.6	36.6	32.9	34.5	33.9	33	33.6

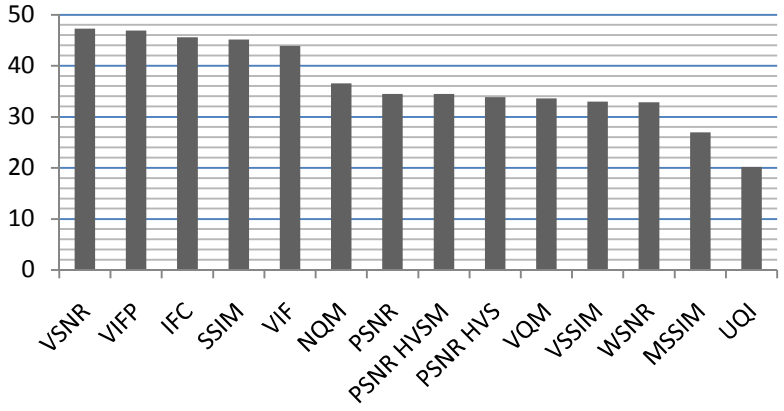


Figure 10: Ranking of used metrics according to their correlation to human judgment.

To conclude, performances of objective metrics, with respect to subjective scores, are different in the case of video sequences than in the case of still images. Correlation coefficients between objective and subjective scores were higher in the case of video sequences, by comparing Table 14 with Table 9. However, human opinion also differed in the case of video sequences. In the case of video sequences, perceptual-like metrics were the most correlated to subjective scores (also in video conditions). However, in both conditions, none of the tested metrics reached 50% of human judgment.

1.6. Discussion and future trends

This section discusses the future directions regarding the quality assessment of views synthesized with DIBR systems. The results presented in the previous sections proved the need for new subjective quality assessment protocols and improved objective metrics. This section addresses the issues related to the conception of a new subjective quality assessment method and the new trends for the objective metrics.

1.6.1. Subjective protocols

ACR-HR and PC are known for their efficiency in 2D conditions, though they showed their limitations in the two case studies presented in 1.3. Moreover, one may need to assess the quality of 3D media in 3D conditions. Defining a new subjective video quality assessment framework is a tough task, knowing the new complexity involved in 3D media. The difficulty of 3D-image quality evaluation, compared to 2D conventional images, is now more considered. Seuntiens [48] introduced new parameters to assess in addition to image quality, which are naturalness, presence and visual experience. Thus, a multi-dimensional quality indicator may allow a reliable assessment of 3DTV media. However, it may be difficult to define such terms in the context of a subjective quality assessment protocol, and there is no standardized protocol considering these aspects yet. ITU-R BT. 1438 recommendation [49] describes subjective assessment of stereoscopic television pictures and the methods are described in [16].

Chen et al. [50] revisited the question of subjective video quality assessment protocols for 3DTV. This work points out the complexity of 3D media quality assessment. Chen *et al.* proposed to reconsider several conditions in this context, such as the viewing conditions (viewing distance, monitor resolution), the test material (depth rendering according to the chosen 3D display), viewing duration, etc. In the following, some of the requirements proposed by Chen *et al.* in [50] are mentioned:

- General viewing conditions: First the luminance and contrast ratio is considered, because of the crosstalk involves by 3DTV screens, and because of the used glasses (as active as polarized glasses cause reduction of luminance). Second, the resolution of depth as to be defined. Third, the viewing distance recommended by ITU standards may differ according to the used 3D display. Moreover, as the authors of the study claim it, depth perception should be considered as a new parameter to evaluate the Preferred Viewing Distance, such as human visual acuity or picture resolution.
- Source signals: the video format issue is mentioned. It refers to the numerous 3D representations (namely “Layer Depth Video” (LDV), “Multi-view Video-plus-Depth” (MVD), or “video plus depth” (2D+Z)) whose reconstruction or conversion lead to different types of artifacts.
- Test methods: as mentioned previously, new aspects have to be considered (naturalness, presence, visual experience), and visual comfort as well. The latter refers to the visual fatigue that should be measured to help in a standardization process.
- Observers: an adapted protocol should involve the measurement of viewers’ stereopsis ability, first. Second, the authors of [50] mention that the required number of participants may differ from 2D. Then further experiments should define this number.
- Test duration and results analysis: the duration of the test is still to be determined, taking into account the visual comfort. The analysis of the re-

sults refers to the definition of a criterion for incoherent viewer rejection and to the analysis of the assessed parameter (depth, image quality, etc.)

1.6.2. Objective quality assessment metrics

The experiments presented in this chapter shown the need for more adapted tools to correctly assess the quality of synthesized views. The most recent proposed 3D quality metrics propose to take into account the new modes brought by 3D. Among the proposed metrics, numerous target stereoscopic video, for instance, but not views synthesized from DIBR. Then they will not be referred to in this section.

Most of the proposed metrics for assessing 3D media are inspired from 2D quality metrics. It should be noted that, often, experimental protocols validating the proposed metrics, involve depth and/or color compression, different 3D displays, and different 3D representations (2D+Z, stereoscopic video, MVD, etc...). The experimental protocols often assess at the same time both compression distortion and synthesis distortion, without distinction. This is problematic because there may be a combination of artefacts from various sources (compression and synthesis) whose effects are not clearly specified and assessed.

In the following, we present the new trends, regarding new objective metrics for 3D media assessment, by distinguishing whether they make use of depth data in the quality score computation or not.

2D-like metrics

Perceptual Quality Metric (PQM) [51] is proposed by Joveluro *et al.* Although the authors assess the quality of decoded 3D data (2D+Z), the metric is applied on left and right views synthesized with a DIBR algorithm (namely [12]). Thus, the method can be cited in this section. The quality score is a weighted function of the contrast distortion and the luminance differences between both reference and distorted color view. So, the method can be classified as HVS-based. The method is sensitive to slight changes in image degradation and error quantification. In [51]PQM method performances are validated by evaluating views synthesized from compressed data (both color and depth data are encoded at different bitrates). Subjective scores are obtained by a SAMVIQ test, on a 3D 42-inch Philips multi-view auto-stereoscopic display. Note that compression, synthesis and factors inherent to the display are assessed at the same time without distinction in the experiments.

Zhao and Yu [52] proposed a FR metric, *Peak Signal to Perceptible Temporal Noise Ratio*. The metric evaluates quality of synthesized sequences by measuring the perceptible temporal noise within these impaired sequences.

Depth-aided methods

Ekmekcioglu *et al.* [53] proposed a depth-based perceptual quality metric. It is a tool that can be applied to 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. The inputs of the method are the original depth map (uncompressed), the original color view (originally acquired, uncompressed), the synthesized view. The 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 [53] shows that the proposed method enhances the correlation of PSNR and SSIM to subjective scores.

Yasakethu *et al.* [54] proposed an adapted VQM for measuring 3D Video quality. It combines 2D color information quality and depth information quality. Depth quality measurement includes an analysis of the depth planes. The final depth quality measures combines 1) the measure of distortion of the relative distance within each depth plane, 2) the measure of the consistency of each depth plane and 3) the structural error of the depth. The color quality is based on the VQM score. In [54], the metric is evaluated through left and right view (rendered from 2D+Z encoded data), and compared to subjective scores obtained by using an autostereoscopic display. Results show higher correlation than simple VQM.

Solh *et al.* [55] introduced the 3D Video Quality Measure (3VQM) predict the quality of views synthesized from DIBR algorithms. The method analyses the quality of the depth map against an ideal depth map. Three different analyses lead to three distortions measures: spatial outliers, temporal outliers, and temporal inconsistencies. These measures are combined to provide the final quality score. To validate the method, subjective tests were run in stereoscopic conditions. The stereoscopic pairs included views synthesized from depth map and colored video compression, depth from stereo matching, depth from 2D to 3D conversion. Results shown accurate and consistent scores compared to subjective assessments.

1.7. Conclusion

This chapter proposed a reflection considering both subjective quality assessment protocols and objective quality assessment methods reliability in the context of DIBR-based media.

Typical distortions related to DIBR were introduced. They are geometric distortions and mainly located around the disoccluded areas. When compression-related distortions and synthesis-related distortions are combined, the errors are generally spread in the whole image, increasing visual annoyance.

Two case studies were presented answering the two questions relating, first to the suitability of two efficient subjective protocols (in 2D), and second, to the reliability of commonly used objective metrics. Experiments considered commonly used methods for assessing conventional images, as subjectively or objectively, to assess DIBR-based synthesized images, from seven different algorithms.

Concerning the suitability of the tested subjective protocols, the number of participants required for establishing a statistical difference between the algorithms was almost the double of the number required by VQEG (24), which reinforce Chen et al. requirements [50]. Both methodologies agreed on the performances ranking of the view synthesis algorithms. Experiments also showed that the observers' opinion was not as stable when assessing still images as when assessing video sequences, with ACR-HR. PC gave stable results with fewer participants than ACR-HR, in the case of still images. Both methodologies have their advantages and drawbacks and they are complementary: assigning an absolute rating to distortions such as synthesized views' ones seemed a tough task to observers, although it provides knowledge on the perceived quality of the set. Small impairements are better evaluated with PC.

Concerning the reliability of the tested objective metrics, the results showed that objective metrics did not correlate the observers' opinion. Objective measures did not reach 50% of human judgment and they were all correlated with each other. The results suggest that tiny distortions are penalized by the objective metrics when not perceptible by humans. Then, objective metrics inform on the existence of distortions but not on their visible annoyance. Using the tested metrics is not sufficient for assessing virtual synthesized views.

The simple experiments that have been presented in this chapter reveal that the reliability of the tested objective metrics is uncertain when assessing intermediate synthesized views, in the tested conditions. Yet, reckoned organizations plan to base partially their future decisions, concerning new strategies for 3D video, on the outcome of such objective metrics. New standards have to been developed considering the new aspects brought by DIBR: location and type of artifacts, degree of annoyance of artifacts.

1.8. Acknowledgements

We would like to thank the experts who provided the synthesized sequences of the presented experiments, as well as the algorithms specifications: Martin Köppel and Patrick Ndjiki-Nya, from the Fraunhofer Institut for Telecommunications, HHI (Berlin).

We would like to acknowledge the Interactive Visual Media Group of Microsoft Research for providing the Breakdancers data set, the MPEG Korea Forum for providing the Lovebird1 data set, the GIST for providing the Newspaper data set, and HHI for providing Book Arrival.

References

[1] A. Smolic et al., « 3D Video and Free Viewpoint Video–Technologies, Applications and MPEG Standards », in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'06)*, 2006, p. 2161–2164.

[2] 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, p. 93–104.

[3] E. Bosc et al., « Towards a new quality metric for 3D synthesized view assessment », *IEEE Journal of Selected Topics*, 2011.

[4] D. V. S. . De Silva, W. A. . Fernando, et S. T. Worrall, « Intra mode selection method for depth maps of 3D video based on rendering distortion modeling », *IEEE Transactions on Consumer Electronics*, vol. 56, n°. 4, p. 2735-2740, nov. 2010.

[5] P. Ndjiki-Nya et al., « Depth image based rendering with advanced texture synthesis », in *Proc. IEEE International Conference on Multimedia & Expo (ICME)*, Singapore, 2010.

[6] M. Köppel et al., « Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering », in *Proc. IEEE International Conference on Image Processing (ICIP)*, Hong Kong, China, 2010.

[7] M. Meesters, W. Ijsselsteijn, et P. Seuntiens, « A survey of perceptual evaluations and requirements of three dimensional TV », *IEEE Transactions on Circuits And Systems for Video Technology*, vol. 14, n°. 3, p. 381-391, mars 2004.

[8] A. Boev, D. Hollosi, et A. Gotchev, « Classification of stereoscopic artefacts », *Mobile3DTV Project report, available online at http://mobile3dtv. eu/results*.

[9] M. Yuen et H. R. Wu, « A survey of hybrid MC/DPCM/DCT video coding distortions », *Signal Processing*, vol. 70, n°. 3, p. 247–278, 1998.

[10] S. L. P. Yasakethu, C. Hewage, W. Fernando, et A. Kondoz, « Quality analysis for 3D video using 2D video quality models », *Consumer Electronics, IEEE Transactions on*, vol. 54, n°. 4, p. 1969-1976, 2008.

[11] A. Tikanmaki, A. Gotchev, A. Smolic, et K. Mueller, « Quality assessment of 3D video in rate allocation experiments », in *IEEE Int. Symposium on Consumer Electronics (14-16 April, Algarve, Portugal)*, 2008.

[12] C. Fehn, « Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV », in *Proc. SPIE Conf. Stereoscopic Displays and Virtual Reality Systems X*, San Jose, USA, 2004.

[13] A. Telea, « An Image Inpainting Technique Based on the Fast Marching Method », *Journal of Graphics, GPU, and Game Tools*, vol. 9, n°. 1, p. 23-34, 2004.

[14] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, et Y. Mori, « Reference Softwares for Depth Estimationand View Synthesis », presented at the ISO/IEC JTC1/SC29/WG11 MPEG 2008/M15377, 2008.

[15] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, et T. Wiegand, « View synthesis for advanced 3D video systems », *EURASIP Journal on Image and Video Processing*, 2008.

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

[17] « MetriX MuX Home Page ». [Online]. Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/. [Accessed: 18-janv-2011].

[18] ITU-T, « Subjective video quality assessment methods for multimedia applications », Geneva, Rec. P910, 2008.

[19] M. Barkowsky, *Subjective and Objective Video Quality Measurement in Low-Bitrate Multimedia Scenarios*. Citeseer, 2009.

[20] M. D. Brotherton, Q. Huynh-Thu, D. S. Hands, et K. Brunnstrom, « Subjective multimedia quality assessment », *IEICE Transactions on Fundamentals of Electronics Communications and Computer ScienceES E SERIES A*, vol. 89, n°. 11, p. 2920, 2006.

[21] D. M. Rouse, R. Pépion, P. Le Callet, et S. S. Hemami, « Tradeoffs in subjective testing methods for image and video quality assessment », *Human Vision and Electronic Imaging XV*, vol. 7527, p. 75270F.

[22] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, et A. Raake, « Study of rating scales for subjective quality assessment of high-definition video », *IEEE Transactions on Broadcasting*, p. 1-14, mars-2011.

[23] J. C. Handley, « Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment », in *ISand TS PICS Conference*, 2001, p. 108-112.

[24] M. Pinson et S. Wolf, « Comparing subjective video quality testing methodologies », in *SPIE Video Communications and Image Processing Conference, Lugano, Switzerland*, 2003.

[25] S. Péchard, « Qualité d'usage en télévision haute définition: évaluations subjectives et métriques objectives », 2008.

[26] E. Bosc, M. Pressigout, et L. Morin, « Focus on visual rendering quality through content-based depth map coding », in *Proceedings of Picture Coding Symposium (PCS)*, Nagoya, Japan, 2010.

[27] F. Ebrahimi, M. Chamik, et S. Winkler, « JPEG vs. JPEG2000: An objective comparison of image encoding quality », in *Proc. of SPIE*, 2004, vol. 5558, p. 300–308.

[28] Z. Wang et A. C. Bovik, « A universal image quality index », *Signal Processing Letters, IEEE*, vol. 9, n°. 3, p. 81-84, 2002.

[29] H. R. Sheikh, A. C. Bovik, et G. de Veciana, « An information fidelity criterion for image quality assessment using natural scene statistics », *Image Processing, IEEE Transactions on*, vol. 14, n°. 12, p. 2117-2128, 2005.

[30] M. H. Pinson et S. Wolf, « A new standardized method for objectively measuring video quality », *IEEE Transactions on broadcasting*, vol. 50, n°. 3, p. 312–322, 2004.

[31] Y. Wang, « Survey of objective video quality measurements », *EMC Corporation Hopkinton, MA*, vol. 1748.

[32] A. P. Hekstra et al., « PVQM-A perceptual video quality measure », *Signal processing: Image communication*, vol. 17, n°. 10, p. 781–798, 2002.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, et E. P. Simoncelli, « Image quality assessment: From error visibility to structural simi-

larity », *Image Processing, IEEE Transactions on*, vol. 13, n°. 4, p. 600-612, 2004.

[34] Z. Wang, L. Lu, et A. Bovik, « Video quality assessment based on structural distortion measurement », *Signal processing: Image communication*, vol. 19, n°. 2, p. 121-132, févr. 2004.

[35] K. Seshadrinathan et A. C. Bovik, « Motion tuned spatio-temporal quality assessment of natural videos », *Image Processing, IEEE Transactions on*, vol. 19, n°. 2, p. 335–350, 2010.

[36] A. Boev, M. Poikela, A. Gotchev, et A. Aksay, « Modelling of the stereoscopic HVS ».

[37] J. Yang et W. Makous, « Spatiotemporal separability in contrast sensitivity », *Vision Research*, vol. 34, n°. 19, p. 2569–2576, 1994.

[38] S. Winkler, *Digital video quality: vision models and metrics*. Wiley, 2005.

[39] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, et M. Carli, « New full-reference quality metrics based on HVS », in *CD-ROM Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, Scottsdale, USA, 2006.

[40] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, et V. Lukin, « On between-coefficient contrast masking of DCT basis functions », in *CD-ROM Proc. of the Third International Workshop on Video Processing and Quality Metrics*, 2007, vol. 4.

[41] D. M. Chandler et S. S. Hemami, « VSNR: A wavelet-based visual signal-to-noise ratio for natural images », *Image Processing, IEEE Transactions on*, vol. 16, n°. 9, p. 2284-2298, 2007.

[42] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, et A. C. Bovik, « Image quality assessment based on a degradation model », *Image Processing, IEEE Transactions on*, vol. 9, n°. 4, p. 636-650, 2002.

[43] H. R. Sheikh et A. C. Bovik, « Image information and visual quality », *Image Processing, IEEE Transactions on*, vol. 15, n°. 2, p. 430–444, 2006.

[44] C. Van, E. Lambrecht, et O. Verscheure, « Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System », 1996.

[45] « Video Quality Research ». [Online]. Available:

http://www.its.bldrdoc.gov/vqm/. [Accessed: 19-juill-2011].

[46] VQEG 3DTV Group, « VQEG 3DTV Test Plan for Cross-talk Influences on user Quality of Experience ». 21-oct-2010.

[47] M. Haber et H. X. Barnhart, « Coefficients of agreement for fixed observers », *Statistical methods in medical research*, vol. 15, n°. 3, p. 255, 2006.

[48] P. Seuntiens, « Visual Experience of 3D TV », Doctoral thesis, Eindhoven University of Technology, 2006.

[49] ITU, « Subjective Assessment of Stereoscopic Television Pictures », in *Recommandation ITU-R BT. 1438*, 2000.

[50] W. Chen, J. Fournier, M. Barkowsky, et P. Le Callet, « New requirements of subjective video quality assessment methodologies for 3DTV », in *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM 2010*, Scottsdale, Arizona, U.S.A., 2010.

[51] P. Joveluro, H. Malekmohamadi, W. A. Fernando, et A. M. Kondoz, « Perceptual Video Quality Metric for 3D video quality assessment », in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2010, p. 1–4.

[52] Y. Zhao et L. Yu, « A perceptual metric for evaluating quality of synthesized sequences in 3DV system », in *Proceedings of SPIE*, 2010, vol. 7744, p. 77440X.

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

[54] S. L. P. Yasakethu, S. T. Worrall, D. De Silva, W. A. C. Fernando, et A. M. Kondoz, « A compound depth and image quality metric for measuring the effects of packet loss on 3D Video », in *Proc. of 17th International Conference on Digital Signal Processing*, Corfu, Greece, 2011.

[55] M. Solh, G. AlRegib, et J. M. Bauza, « 3VQM: A vision-based quality measure for DIBR-based 3D videos », in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, 2011, p. 1–6.