

# Objective Image Quality Assessment of 3D Synthesized Views

Federica Battisti, *Member, IEEE*, Emilie Bosc, *Member, IEEE*,  
Marco Carli, *Senior Member, IEEE*, Patrick Le Callet, *Member, IEEE*,  
and Simone Perugia.

## Abstract

Depth-Image-Based-Rendering (DIBR) techniques are essential for three dimensional (3D) video applications such as 3D Television (3DTV) and Free-Viewpoint Video. However, this process is based on 3D warping and can induce serious distortions whose impact on the perceived quality is far different from the one experienced in the 2D imaging processes. Since quality evaluation of DIBR-synthesized views is fundamental for the design of perceptually-friendly 3D video systems, an appropriate objective quality metric targeting the assessment of DIBR-synthesized views is momentous. Most of 2D objective quality metrics fail in assessing the visual quality of DIBR-synthesized views because they have not been conceived for addressing the specificities of DIBR-related distortions. In this paper, a new full-reference objective quality metric, 3DSwIM (3D Synthesized view Image quality Metric), dedicated to artifacts detection in DIBR-synthesized view-points is presented. The proposed scheme relies on a comparison of statistical features of wavelet subbands of two input images: the original image and the DIBR-based synthesized image. A registration step is included before the comparison step so that best matching blocks are always compared to ensure "shifting-resilience". In addition, a skin detection step weights the final quality score in order to penalize distorted blocks containing "skin-pixels" based on the assumption that a human observer is most sensitive to impairments affecting human subjects. Experimental tests show that the proposed method outperforms the conventional 2D and DIBR-dedicated quality metrics under test.

F. Battisti, M. Carli and S. Perugia are with the Department of Engineering, Università degli Studi Roma TRE, Rome, Italy.  
E. Bosc and P. Le Callet are with Polytech' Nantes, Nantes, France.

## I. INTRODUCTION

Three dimensional (3D) imaging still soundly arouses the public interest and curiosity as the wide range of proposed applications in the 3D market shows. Together with the increase in the popularity of this technology, issues linked to the need for delivery of 3D multimedia content raised. This has pushed the research field to face many challenges due to the large amount of data to be delivered and to the limitations of the transmission channel.

In many applications the creation of 3D content is performed through the use of multiple cameras that acquire the same scene at slightly different viewpoints, but may also require the generation of additional virtual views, i.e., non-acquired. Indeed, the recorded sequences are post-processed by means of computer graphics techniques for rendering the 3D effect. To this aim, the 2D video sequences can be associated to depth sequences that provide information on the scene geometry. The set of 2D video sequences and corresponding depth data is called Multi-view-Video-plus-Depth (MVD) [1] data. This 3D scene representation can be exploited for generating virtual viewpoints of the scene by using view synthesis. This allows to render a virtual scene as if it is recorded from a viewpoint for which no direct information is actually available. Virtual view creation can be used for enabling 3D Television (3DTV) applications, in which cameras are often arranged with a relatively short baseline, and Free Viewpoint Video (FVV) applications, that can be based on a relatively sparse set of cameras that surround a scene [2].

MPEG included Free-viewpoint Television (FTV) in standardization efforts since 2001: from Multi view Video Coding, targeting to the efficient coding of multiple camera views, to 3D Video, for enabling viewing adaptation and display adaptation of multiview displays. Proposals on 3D Video Coding have been launched in March 2011 for designing a new coding framework for MVD data and several methods were presented as shown in [3]. Finally, FTV was launched in August 2013, targeting super multiview and free navigation applications. The goal is to design a new FTV framework for viewing of 3D scenes [4], [5]. In particular, the third phase of FTV depicts a scenario in which the input is the 3D scene, and the output is a single view with varying viewpoint, multiple views or super multiview. To cope with the low number of input camera views, forced by economical reasons, view synthesis algorithms are required. However,

as several tests show, artifacts may be introduced through the rendering process. Our aim, is to develop a metric for evaluating the quality of these methods.

The problem of generating views also arises in the multi-view data coding chain. In this case, for reducing the requested resources, alternative representations of original data have been proposed. Video-plus-depth representation [6] is based on a regular video stream, where each frame is enriched with a depth map providing a Z-value for each pixel. The final left and right views are reconstructed by means of Depth-Image-Based-Rendering (DIBR) [7] techniques. The MVD representation can be evolutionary built on the existing DVB infrastructure and is considered as the most broadcast-friendly representation.

We believe that the evaluation of synthesized views quality is of primary importance, due to its use in many 3D imaging applications. Moreover, despite the advancements in view modeling and synthesis, much less effort has been focused on developing algorithms for automatically assessing the visual quality of a synthesized view. In fact, it is well known that the most efficient way for assessing the quality of any content is to refer to the human evaluation. Unfortunately, the realization of subjective tests is expensive, time consuming, and the collected results may be influenced by many factors: i.e., loss of concentration of users is one of the factors that can affect the reliability of subjective tests. An objective quality assessment metric for synthesized views is thus of paramount importance. As observed in [8] most of existing objective metrics are not well adapted to assess the quality of the virtual views. A possible motivation, is that artifacts related to DIBR systems are mainly located around the disoccluded regions and they are not scattered in the entire image such as specific 2D video compression distortions. Consequently, commonly used 2D quality metrics, that were originally designed to address 2D video compression distortions, are not sufficient for assessing the visual quality of synthesized views. Those metrics allow to indicate the presence of errors but not the degree of visual annoyance.

To the purpose of defining ad hoc quality metrics it is possible to improve existing 2D metrics or to propose a new approach relying on the characteristics of the 3D content. Yasakethu *et al.* [9] propose a modified version of Video Quality Metric (VQM) for measuring 3D Video quality. It combines 2-D color information quality and depth information quality. Depth quality measurement is based on the analysis of the depth planes distortion. Results show higher correlation to subjective scores when it is compared to the basic VQM. Another approach based

on the improvement of existing 2D metrics is proposed by Ekmekcioglu *et al.* [10]. This depth-based perceptual tool 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 study in [10] shows that the proposed method enhances the correlation of Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity index (SSIM) with subjective scores. Recently, in [11], the joint use of Hausdorff distance with a block-matching based algorithm is proposed as quality tool for 3D synthesized images. The reported results the method capability to cope with ghost effect and object shifting.

In this contribution, we analyze the quality of synthesized frames, as a first step for assessing the quality of 3D contents. In particular, our attention is focused on the degradations perceived by the user as losses or modifications in the structural information of the image. It has been demonstrated in [12] that there is a strong correlation among neighboring pixels. This dependency conveys relevant information about the structure of the objects in the visual scene. In this work, we exploit the artifacts localization to objectively assess their impact on the perceived quality. In more details, the proposed method is based on two main assumptions. First, we believe that pixels, or regions, displacement can be introduced by the rendering process without affecting the visual quality of the synthesized images. Indeed, objects may be slightly shifted due to the projection process, nevertheless being the overall quality of the image still acceptable. This type of artifacts is penalized by pixel-by-pixel based quality metrics such as PSNR. Second, we believe that human beings are more sensitive to artifacts affecting regions containing human beings (i.e., faces or hands). This implies that modifications performed in such regions lead to severe subjective quality scores. To cope with this feature, we include a weighting function based on results of a skin detection procedure.

The rest of the paper is organized as follows. In Section II the virtual view synthesis through the DIBR mechanism is presented and in Section II-B an overview of video quality metrics is reported. In Section III the proposed metric, 3DSwIM, is detailed while in Section IV the results of the performed experiments are presented; finally, in Section VI the conclusions are drawn.

## II. DIBR TECHNIQUES AND VIEW SYNTHESIS: A NEW CHALLENGE FOR IMAGE QUALITY ASSESSMENT

As mentioned before, the process of virtual view synthesis can be based on the use of specific methods called DIBR. These techniques allow to generate a virtual view by means of an original image or video and the knowledge of depth information as shown in Figure 1 [13]. In brief, the DIBR procedure is based on three steps:

- 2D to 3D: back-projection mapping. The original view points are re-projected into the 3D world, by exploiting the camera intrinsics and the respective depth data;
- 3D to 2D: the 3D space points are projected into the image plane of a virtual view located at the required viewing position.
- Blending: after points from different perspective views are re-projected on the new virtual view, a fusion procedure is applied. In this step, it is possible that some occluded areas in the original views may become visible in the virtual one thus creating artifacts (holes) that are referred to as disocclusions. In order to fill the holes in the synthesized views, many different strategies can be applied: blend available pixels from two warped views with a linear weighting function, select one warped view as dominant one and use the pixels in the other view only to fill the holes in the dominant, or select the closest pixel based on the z-buffer method. A detailed review of these methods can be found in [14]. After the filling procedure, artifacts can still be present in the synthesized views. In this case inpainting techniques can be applied as in [15], [16].

A detailed description of the depth-frame-based view synthesis can be found in [17].

### A. DIBR-related distortions

DIBR synthesized images often contain artifacts generated by different factors: regions occluded in both input views and visible in the target view lead to fill-in errors in the rendered view, errors can also occur because pixel coordinates do not locate at an integer position and are usually either interpolated or rounded to the nearest integer position. In-painting methods as well as interpolation filters are developed in order to reduce these synthesis artifacts. However, using such processes may lead to the generation of new artifacts. Moreover, although powerful solutions are available for depth estimation, depth estimation errors remain. Such errors induce

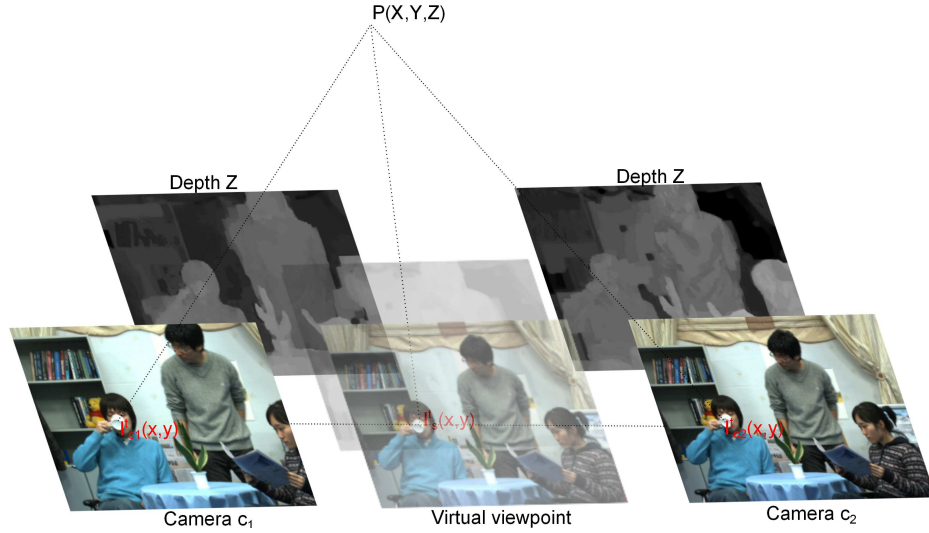


Fig. 1. Relationships between 3D real point and corresponding pixels in acquired and virtual views.

artifacts in the virtual synthesized view because of incorrect projections. Tech et al. in [18] pointed out the existence of such distortions in the synthesized views through the evaluation of several view synthesis methods in the framework of mobile 3DTV. The analysis is performed highlighting the constraints of computational power and display size and resolution given by the particular rendering device. The results of subjective tests allow the identification of the rendering artifacts and demonstrate the need for subpixel accuracy in the rendering process for avoiding errors in depth perception. In [19], different types of artifacts related to the use of DIBR systems have been defined. These artifacts can be summarized as follows.

**Object shifting:** it consists in a translation or change in the size of a region in the image. This can be due to depth pre-processings including low pass filtering, or to encoding methods smoothing object edges. Figure 2 depicts this type of distortion.

**Incorrect rendering of textured areas:** inpainting methods may fail to reconstruct complex textures. To overcome these limitations, hole filling techniques are proposed such as in [20] that is based on texture patches. In such approaches, candidate patches may be not perceptually close to the disoccluded region and consequently lead to the creation of artifacts. Furthermore, the use of rectangular patches can lead to blocky artifacts. Figure 3 depicts this type of distortion.

**Blurry regions:** this may be originated by the inpainting method used to fill in the disoccluded areas. It is more visible around the background/foreground transitions. These characteristics can



Fig. 2. Shifting/Resizing artifacts. The shape of the leaves, in this figure, is slightly modified (thinner or bigger). The vase is also moved.



Fig. 3. Incorrect rendering of textured areas. An example of texture stretching.

be observed in Figure 4 around the disoccluded areas. Behind the head and around the arms of the chair, thin blurry regions are perceptible.

**Flickering:** when errors occur randomly in depth data along the sequence, pixels are wrongly projected: some pixels suffer slight changes of depth, which appear as flickers in the resulting synthesized pixels. To avoid this effect, the use of a background sprite is proposed in [21]. Background image information is stored in the sprite and updated by using the information of the original and the synthesized images on previous frames for providing depth consistency. All along the frames, background texture is copied into the disoccluded region and then refined by patch-based texture synthesis. However, the flickering problem is still present in the hole boundary due to possible depth estimation inconsistency.

**Geometry distortion:** it includes depth estimation errors, depth quantization errors in the conversion from depth data to depth map, and inaccurate camera parameters.

**Depth coding induced distortions:** it refers to warping distortions due to quantization-related errors in decompressed depth data such as:

- Shifting effect, as previously described. Figure 5 illustrates this type of distortion.



Fig. 4. Blurry artifacts (Book Arrival).

- *Crumbling*: when artifacts occur in depth data around strong discontinuities, appearing like erosion. In this case objects' edges appear distorted in the synthesized view. This typically occurs when applying wavelet-based compression on depth data. Figure 6 depicts this artifact. It is perceptible around the arms of the chair.

**Depth distortions:** we consider, here, depth distortions as defined by Devernay et al. in [22]. In this paper the authors propose a postprocessing phase for detecting and reducing depth distortion artifacts by first estimating the pixel correctness, and then smoothing the 'error' area by exploiting an anisotropic diffusion.

As mentioned in the Introduction, only few studies targeting the assessment of DIBR-based synthesized views have been proposed. The following section addresses a review of the existing objective quality metrics targeting the quality assessment of synthesized views.

### B. *Quality assessment of DIBR-based synthesized views*

As already stated in the Introduction, many efforts have been devoted to the definition of Image Quality Assessment (IQA) methods. The overall goal is to mimic, as close as possible,



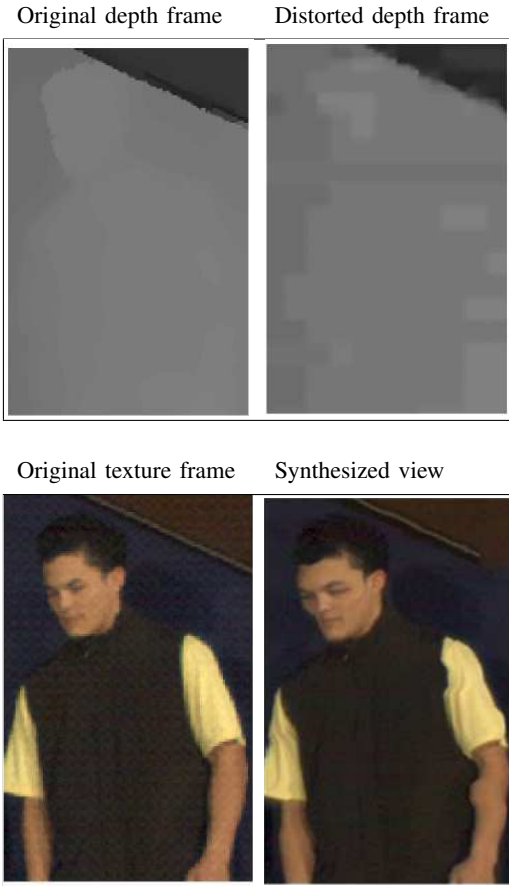


Fig. 5. Shifting effect from depth data compression results in distorted synthesized views (Breakdancers).

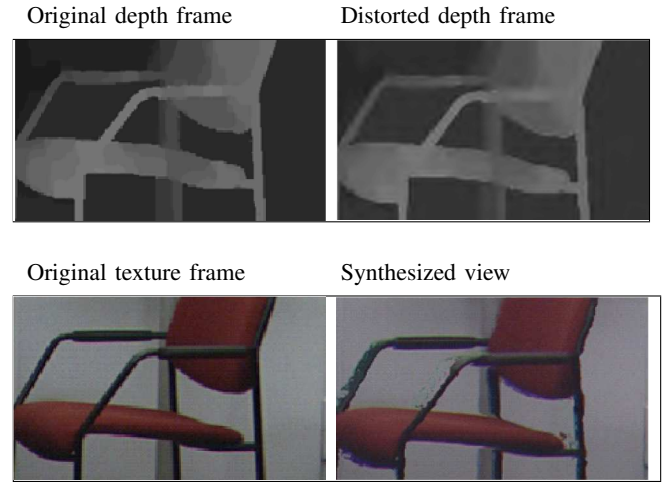


Fig. 6. Crumbling effect in depth data leads to distortions in the synthesized views (Book Arrival).

the average judgement of human subjects. This is a very challenging task, still to be completely solved, due to many reasons. In particular the application scenario and its related distortion may challenge the available quality metrics which are introducing new type of artifacts for which the existing quality metrics have not been designed for. DIBR techniques belong to this challenging case and recently few techniques have been proposed. In the following, we perform a brief literature review regarding the assessment of synthesized views with 2D-metrics based techniques and with depth-based methods.

1) *2D-like objective quality metrics*: in this section, we mention recent studies addressing the issue of objectively assessing DIBR-based synthesized views relying on 2D-like metrics.

Perceptual Quality Metric (PQM) [23] was proposed by Joveluro *et al.*. Although the authors target the quality assessment of decoded 3D data (2D+Z), the metric is applied on left and right views synthesized with a DIBR algorithm [7]. Thus, this method may also be applied for synthesized views. The quality score is a weighted function of the contrast distortion and the luminance difference between both reference and distorted color views. The method can thus be classified as HVS-based. The method is sensitive to slight changes in image degradation and error quantification. In [23] PQM method performances are validated by evaluating views synthesized from compressed data (both color and depth data are encoded at different bit-rates). Subjective scores are obtained by a SAMVIQ test, on a 3D 42-inch Philips multi-view auto-stereoscopic display. This experimental protocol assesses at the same time without distinction, the compression-related artifacts, the synthesis-related artifacts and factors inherent to the display. Zhao and Yu [24] proposed a Full Reference metric, Peak Signal to Perceptible Temporal Noise Ratio. This metric evaluates the quality of synthesized sequences by measuring the perceptible temporal noise within these impaired sequences.

Conze *et al.* [25] proposed a Full Reference objective quality metric dedicated to artifacts detection in synthesized view-points. The idea is to evaluate the distortion in areas where disparity estimation methods may fail: thin objects, object borders, transparency, variations of illumination or color differences between left and right views, periodic objects. The key feature of the proposed method is the use of three visibility maps which characterize complexity in terms of textures, diversity of gradient orientations and presence of high contrast. Moreover, the VSQA metric can be defined as extension to any existing 2D image quality assessment metrics.

2) *Depth-based objective quality metrics*: Ekmekcioglu *et al.* in [10] have proposed a depth-based perceptual quality metric. The method uses a weighting function based on depth data at the target viewpoint, and a temporal consistency function to take the motion activity into account. The final score includes a factor that considers non-moving background objects during view synthesis. Inputs of the method are the original depth map (uncompressed), the original color view (originally acquired, uncompressed), and the synthesized view. Validation of the performances is achieved by synthesizing different viewpoints from distorted data: color views suffer two levels of quantization distortion; depth data suffer four different types of distortion (quantization, low pass filtering, borders shifting, and artificial local spot errors in certain regions).

Yasakethu *et al.* [9] proposed an adapted VQM for measuring the impact of packet loss on 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 measure 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.

Solh *et al.* in [26] introduced the 3D Video Quality Measure (3VQM) to predict the quality of views synthesized from DIBR algorithms. The method analyzes the quality of the depth map against an ideal depth map. Three different analysis lead to three distortion 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. Stereoscopic pairs include views synthesized from depth map and colored video compression, depth from stereo matching, depth from 2D to 3D conversion. Results show accurate and consistent scores compared to subjective assessments.

Even if the quality of the imaging system and the computational power are increasing, difficult situations still arise. Reduced depth-of-field, low-texture areas, depth discontinuities, repetitive patterns, transparencies, or specular reflections are still very challenging and cause local errors in most disparity computation methods, which result in 2D or 3D artifacts in synthesized views. Several studies have been conducted for understanding the difference in artifacts detectability in 2D and in 3D images. The results of these test campaigns are available in 3DTV and MPEG frameworks [27].

In this contribution, given the particular scenario of DIBR-based synthesized views we are aiming to compare the results achieved by using the most advanced state of the art metrics [23], [25], [26] which has not been done yet. In addition, we propose a new algorithm based on the evaluation of the artifacts introduced by DIBR-methods and on their impact of human judgment. The next section presents the proposed method.

### III. 3DSWIM: THE PROPOSED METRIC

This section presents 3DSwIM, the proposed method for the quality assessment of DIBR-synthesized views. It relies on a comparison of statistical features of wavelet subbands of two input images: the original image and the DIBR-based synthesized image. A registration step is included before the comparison step so that best matching blocks are always compared.

This ensures a "shifting-resilience" property: depending on the warping strategy, objects may be shifted in the synthesized frame while the whole image still presents a good visual quality. In addition, in the proposed approach, a skin detection step weights the final quality score so that distorted blocks containing "skin-pixels" are penalized. The block scheme of the proposed method is presented in Figure 7.

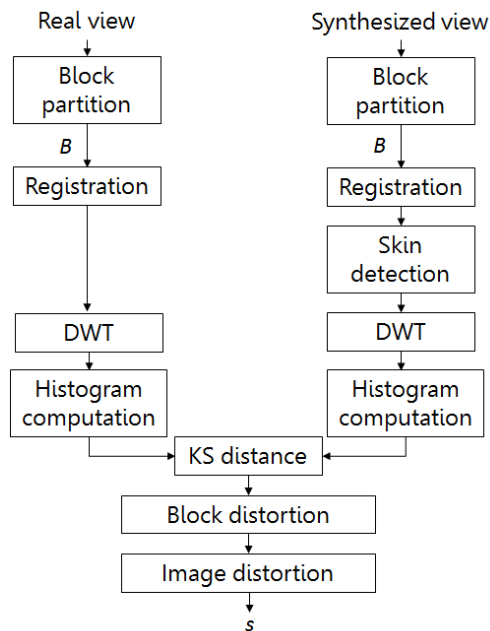


Fig. 7. Block diagram of the proposed method.

The details for each step of the proposed methods are reported in the following, while the values of the parameters used in our tests are reported in Section IV. Let us consider a frame  $F$  of size  $n \times m$  pixels. The quality assessment of  $F$  will be performed as follows:

- 1) *Block partition*: given both real and synthesized views of dimension  $n \times m$ , we set the frame partition into  $B = B_n \cdot B_m$  non overlapping blocks. Each block is of size  $b_n \times b_m$  pixels, being  $b_n = \frac{n}{B_n}$ ,  $b_m = \frac{m}{B_m}$ , with  $B_n$  and  $B_m$  set as metric parameters.
- 2) *Registration*: a registration procedure is performed for allowing the comparison of matching blocks between the original and the synthesized views. Since the presence of a displacement between different views is highly probable, a registration procedure is employed by using a block matching procedure before the analysis block. An Exhaustive Search (ES)-like

algorithm [28] is selected and a search window of size  $W$  pixels in horizontal direction only is used. This algorithm calculates the cost function at each possible location in the search window and the best matching candidate is chosen. The selection of the size of the search window is crucial: a larger windows corresponds to increased computational cost.

- 3) *Skin detection*: from the results obtained by the subjective tests in [29], [30], it has been noticed that the presence of human beings in the image under test increases the annoyance of the detected artifacts. Thus, in the proposed metric a skin detection procedure is performed to *perceptually* weight each block containing distorted faces, necks, etc. In more details, the adopted skin detector is based on the color segmentation performed on the H component of the HSV color space.
- 4) *Wavelet transform*: each block undergoes a first level Haar wavelet transform. In the proposed method, we measure the image degradation by analyzing the statistical variations in the wavelet sub-band related with the image horizontal details. This choice is motivated by visual observation of the synthesized views, as can be observed in Figure 8. As can be noticed, artifacts are present in both the images, especially close to the vertical edges. Errors in DIBR methods generally cause spatial outliers, temporal outliers, and temporal inconsistencies. In particular, the filled holes generated through the DIBR process are mainly characterized by high frequencies in the horizontal direction. If virtual views are located laterally in horizontal way as requested to obtain a stereo pair, the holes are mainly located close to vertical edges of the objects. These holes correspond to discovered areas that were not visible from the reference views. These holes are stretched in vertical way and correspond to horizontal details. Therefore, the image degradation can be measured by analyzing the statistical variations in the wavelet sub-band related with the image horizontal details. A detailed analysis of DIBR distortions can be found in [26].
- 5) *Histogram computation and block distortion computation*: the histogram computation of the original  $h_o$  and of the synthesized  $h_s$  blocks is computed. The Kolmogorov-Smirnov [31] distance between the two histograms is computed to quantify the distance between the distribution function of the real view  $F_{o_b}(x)$  and the distribution function of the synthesized view  $F_{s_b}(x)$ . The block distortion can be computed as follows:

$$d_b = \max (|F_{o_b}(x) - F_{s_b}(x)|). \quad (1)$$



Fig. 8. Left image: View 9 of *Book Arrival* after 3D warping and no hole-filling. Right image: View 4 of *Newspaper* after 3D warping and no hole-filling. Black areas correspond to the holes to be filled.

- 6) *Overall image quality score*: the overall normalized image distortion can be computed as follows:

$$d = \frac{1}{D_0} \sum_{b=1}^B d_b \quad (2)$$

where  $D_0$  is a normalization constant. The image quality score is given by the following relation:

$$s = \frac{1}{1 + d}. \quad (3)$$

The score  $s$  ranges in the interval  $[0, 1]$  where a lower distortion corresponds to a higher score ( $d = 0$  and  $s = 1$ ) and a higher distortion corresponds to a lower score ( $d \rightarrow \infty$  and  $s = 0$ ).

Finally the presence of human subjects is taken into account through the weight  $w_{skin}$ , whose value is based on the skin detection procedure.

The overall image quality score is computed as follows:

$$s = \frac{1}{1 + \frac{1}{D_0} \sum_{b=1}^B w_{skin} \max(|F_{o_b}(x) - F_{s_b}(x)|)} \quad (4)$$

#### IV. VALIDATION PROTOCOL

This section describes the experimental protocol used for validating 3DSwIM.

Our goal is to propose a quality metric assessing views synthesized from DIBR in monoscopic

viewing conditions as a preliminary step. We consider the performances of the proposed metric when evaluating the quality of synthesized still images. In the following, we will first describe the material used for validating the performances of the proposed metric. Then, the results will be presented.

### A. Experimental Setup

1) *Stimuli*: A still image database was built from a video database. In particular, “key frames” extracted from the video sequences are considered. The video sequences were obtained from three different original MVD sequences, namely *Book Arrival*, *Newspaper*, *Lovebirds*. Table II summarizes the features of the test sequences. From each MVD sequence, four different intermediate viewpoints were generated using seven different DIBR algorithms. DIBR algorithms are labeled from *A1* to *A7*. In total, 84 synthesized views are considered. We thus obtain 84 still images from the still images database. The *IRCCyN/IVC DIBR Images* database [29], [30] is freely available.

The seven algorithms used for generating the databases are described in the following. The use of these algorithms is motivated by previous collaborative studies [19], [30]. The seven algorithms are:

- *A1*: based on [7], in which the depth map is pre-processed to filter out any insignificant depth discontinuities. Borders are cropped, and then an interpolation is processed to reach the original size. This can induce shifting artefacts.
- *A2*: based on depth map pre-processing as in [7] and the borders are inpainted as described in [32]. This can induce blurring around object discontinuities because the synthesized views are generated from low-pass filtered depth maps.
- *A3*: Tanimoto et al. [33], it is the recently adopted reference software for the experiments in the 3D Video group of MPEG. It can induce blurry regions in the reconstructed views.
- *A4*: Muller et al. [34], proposed a hole filling method aided by depth information.
- *A5*: Ndjiki-Nya et al. [20], the hole filling method is a patch-based texture synthesis.
- *A6*: Koppel et al. [21], uses depth temporal information to improve the synthesis in the disoccluded areas.
- *A7*: corresponds to the synthesized sequences when no inpainting techniques are applied.

2) *Subjective assessment method*: Forty three naive observers rated the quality of the still image database stimuli. Stimuli were rated following the Absolute Category Rating with Hidden Reference Removal [35] (ACR-HR) subjective assessment methodology. ACR-HR methodology involves observers to rate test objects (i.e. images or sequences) one at a time. Stimuli were rated based on a discrete quality scale as shown in Table I.

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very Annoying

TABLE I  
MEAN OPINION SCORE RATING SCHEME.

The reference version of each synthesized views was included in the test procedure and rated like any other stimulus. This explains the term mentioned as “hidden reference”. From the obtained scores, a differential score (DMOS for Differential Mean Opinion Score) was computed between the Mean Opinion Scores (MOS) of each test object and its associated hidden reference. The subjective evaluations were conducted in an ITU conforming test environment. The stimuli were displayed on a TVLogic LVM401W, and according to ITU-T BT.500 [36].

Sequence Name	Resolution (pxl)	Camera Arrangement
Book Arrival	1024 × 768	16 cameras with 6.5cm spacing
Newspaper	1024 × 768	9 cameras with 5 cm spacing
Lovebirds	1024 × 768	12 cameras with 3.5 cm spacing

TABLE II  
THREE MVD SEQUENCES USED IN THE EXPERIMENTS.



### B. Validation criterion

Any proposed image/video quality metric needs to be validated through correlation analysis with human judgment. A reliable objective quality metric should be highly correlated to human judgment. In this paper, the consistency between ACR-HR DMOS scores and objective quality metrics scores is assessed through the computation of the Pearson Linear Correlation Coefficients (PLCC) for the whole fitted measured points. Fitted measure points correspond to so called predicted DMOS, noted as  $DMOS_p$ . Video Quality Expert Group (VQEG) Phase I FR-TV [37] recommends the use of the following logistic function for the fitting step:

$$DMOS_p = a.score^3 + b.score^2 + c.score + d \quad (5)$$

where  $score$  is the obtained score from the objective metric and  $\{a, b, c, d\}$  are the parameters of the cubic function. They are obtained through the regression step to minimize the difference between  $DMOS$  and  $DMOS_p$ . The Pearson linear correlation coefficients are then computed through:

$$PLCC = \frac{\sum_{i=1}^N (DMOS_i - \overline{DMOS}) (DMOS_{p_i} - \overline{DMOS_p})}{\sqrt{\sum_{i=1}^N (DMOS_i - \overline{DMOS})^2} \sqrt{\sum_{i=1}^N (DMOS_{p_i} - \overline{DMOS_p})^2}} \quad (6)$$

where  $\overline{DMOS}$  and  $\overline{DMOS_p}$  are the average of  $DMOS$  and  $DMOS_p$  over the  $N$  stimuli.

### C. Parameter setup

The metric parameters have been defined as follows.

- The first level of the decimated, discrete, Haar wavelet decomposition has been used. The length of the filter is two.
- Several frame partitions have been considered for properly tuning this parameter as shown in Table III. Experimental validation has shown that the the best performing results with respect to DMOS ranking are achieved by using a partition with  $B_n = B_m = 8$  blocks each of size 96x128 pixels.
- The search algorithm has been used in left and right directions with a search window of size  $W = +/- 10$  pixels in horizontal direction only. We noticed that a horizontal displacement of 20 pixels was greater than the maximum displacement generated by synthesis algorithms

Image partition dimension $B = B_n \cdot B_m$	Correlation	RMSE	Ranking	DMOS
1x1	40.38%	0.6392	5 4 6 1 2 3 7	1 5 4 6 2 3 7
2x2	57.63%	0.5824	1 3 2 5 6 4 7	1 5 4 6 2 3 7
4x4	67.99%	0.5370	1 4 5 6 2 3 7	1 5 4 6 2 3 7
8x8	76.17%	0.4269	1 5 6 4 2 3 7	1 5 4 6 2 3 7
16x16	70.67%	0.5136	1 6 5 4 2 3 7	1 5 4 6 2 3 7
32x32	62.79%	0.5652	6 5 4 2 1 3 7	1 5 4 6 2 3 7
64x64	55.64%	0.5954	5 6 4 2 3 1 7	1 5 4 6 2 3 7

TABLE III  
IMPACT OF BLOCK SIZE ON THE OVERALL METRIC PERFORMANCES

in the testing conditions. This value can be adjusted depending on the depending on the View synthesis algorithm and the distance between adjacent views to predict the new view.

- Skin region definition: based on the methods presented in [38] and after running preliminary tests, we define pixels belonging to skin regions to lie in the hue range values [0.064-0.085]. We use the morphological filters adopted in [38]. If skin is detected in a block, the weight  $w_{skin}$  is set to 15.

The next section presents the obtained results in the case of the image quality assessment and in the case of the video quality assessment.

## V. RESULTS

This section addresses the performances of 3DSwIM when assessing the quality of key synthesized frames, using the *IRCCyN/IVC DIBR Images* database. Concerning the computational complexity, the registration phase is the most time consuming. On a PC, CPU Xeon 3GHz, 8GB RAM, Windows 7, the non-optimized Matlab version of the metric requires 0.9s/image.

Table IV gives the obtained PLCC in percentage together with the RMSE values. This table shows that the proposed metric obtains the best correlation score (72.64% with DMOS). The 2D commonly used metric having the highest correlation score human judgment is MSSIM with 57.4%, that is lower than the proposed metric.

Since the agreement is different from the the correlation, as showed in [39], we also check for the agreement of the proposed metric with DMOS scores. For this purpose, Figure 9 gives

	PLCC (%)	RMSE
3DSwIM	<b>76.17</b>	<b>0.42</b>
Solh et. al [26]	47.7	0.61
VSQA [25]	53.78	0.58
PQM [23]	48.68	0.6
PSNR	47.27	0.61
SSIM	41.3	0.65
MSSIM	55.21	0.59
VSNR	36.25	0.65
VIF	31.3	0.66
VIFP	22.4	0.68
UQI	19.1	0.68
IFC	22.3	0.68
NQM	51.4	0.60
WSNR	47.7	0.61
SNR	40.85	0.64
PNSR-HVSM	42.53	0.63
PSNR-HVS	41.4	0.64

TABLE IV

PEARSON CORRELATION COEFFICIENTS (PLCC) BETWEEN DMOS AND OBJECTIVE SCORES IN PERCENTAGE AND RMSE.

the performances' ranking of the test DIBR algorithms, according to the DMOS scores, and according to the objective metrics. The darker the blue the better ranked by the metric. The lighter the blue the worse ranked by the metric. First line gives the ranking according to the DMOS scores. Next lines give the ranking with the different used objective metrics. While both test objective metrics are inconsistent with DMOS ranking (especially considering the ranking of *A1* and *A6*), the ranking obtained through the proposed metric's scores is very close to DMOS ranking, except for the ranking of *A4* and *A6* that are switched (as emphasized with red box in Figure 9). In particular, as explained in [30], *A1* is the best ranked algorithm according to DMOS scores. In other words, artifacts induced by *A1* may be the less annoying, increasing its perceived quality. However, both the objective metrics rank it as the worst, except for our proposed metric. This was explained in [30] by the fact that *A1* involves shifting artifacts that are costly when using signal-based or fidelity-like metrics. The proposed metric is not a fidelity measure since it

considers shifting blocks through the registration step. Moreover, the proposed metric integrates the fact that human are much more sensitive to artifacts occurring around representations of human beings.

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

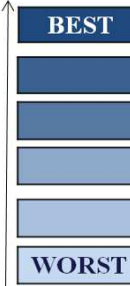


Fig. 9. Ranking of view synthesis algorithms according to obtained objective quality scores.

#### A. Validation of skin detection usefulness

The skin detection step included in the proposed metric is motivated by the assumption that humans are more sensitive to artifacts degrading the appearance of human beings in the reconstructed contents. For this reason, the proposed metric penalizes artifacts occurring in these areas. In this subsection, we validate the usefulness of the skin detection step. For this purpose, we analyze the performances of the proposed metric with and without the skin detection step, when assessing the *IRCCyN/IVC DIBR Images* database. Table V gives the Pearson correlation scores with the DMOS, obtained with Eq. 6, per content. This table shows that the skin detection improves the correlation of the objective quality score with the subjective scores. This is observed for two tested sequences (*Book Arrival* and *Lovebirds*, with an increase of 2.24 and 18.7 points respectively). Human beings appear in the three tested sequences. But, we assume that in *Book Arrival* and *Lovebirds*, the annoyance of artifacts occurring around human representations was more perceptible, which explains the improvement brought by the skin-detection-based weighting.

Tested sequence	3DSwIM (full)		3DSwIM without skin detection	
	PLCC	RMSE	PLCC	RMSE
<b>Book Arrival</b>	<b>96.91</b>	<b>0.39</b>	94.67	0.54
<b>Lovebirds</b>	<b>53.7</b>	0.6	35	<b>0.44</b>
<b>Newspaper</b>	67.6	0.49	<b>81.4</b>	<b>0.39</b>

TABLE V

PEARSON CORRELATION COEFFICIENTS (PLCC) BETWEEN DMOS AND OBJECTIVE SCORES IN PERCENTAGE AND RMSE.

We report the DMOS values over the proposed metric's fitted scores as an additional analysis aid in Figure 10. The confidence intervals plotted in this Figure describe the range of values which the MOS fall with a probability of 95%. The shorter the interval, the more reliable is the result.

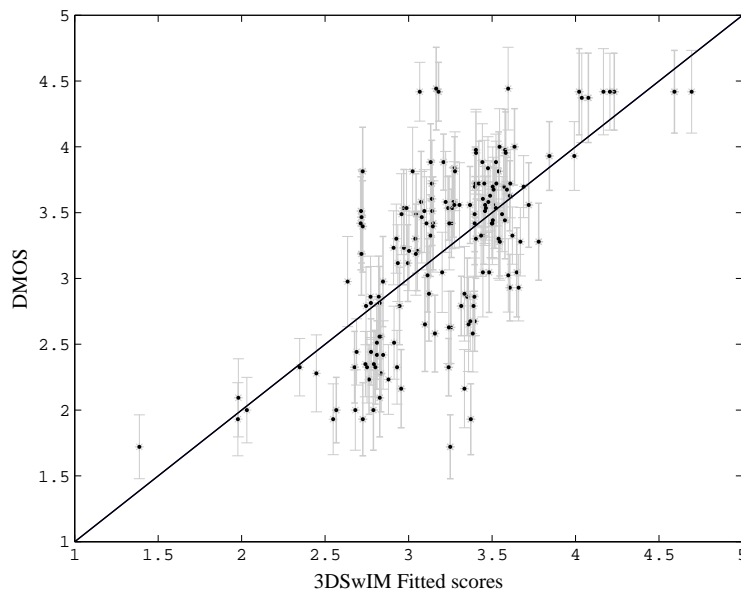


Fig. 10. DMOS over 3DSwIM fitted scores.

These analysis showed the promising performances of the proposed objective quality metric. The notable improvement compared to the objective metrics considered in this experiment comes from the use of the skin detection. However, the analysis also showed that the skin detection

increased the correlation with subjective scores only in specific contents.

### B. Frame partition impact

For sake of completeness in Table III the achieved results obtained by varying the number of blocks in which the image is partitioned are reported.

The frame partition  $B_n = B_m = 8$  has been selected since it results in better matching with DMOS ranking.

## VI. CONCLUSIONS

This paper presented a new full-reference objective quality assessment metric, 3DSwIM, addressing the image quality evaluation of DIBR-synthesized views. The proposed metric relies on a comparison of statistical features of wavelet subbands of two input images: the original image and the DIBR-based synthesized image. Based on the observation that DIBR can induce non visually annoying object shiftings, a registration step is included before the comparison step so that best matching blocks are compared, to ensure "shifting-resilience". In addition, a skin detection step weights the final quality score so that distorted blocks containing "skin-pixels" are penalized. The results of the validation process show that the proposed method outperforms the 2D conventional and DIBR-synthesized views dedicated quality metrics under test. The Matlab implementation of 3DSwIM is available for scientific purposes at <http://www.comlab.uniroma3.it/3DSwIM.html>. Extension to video quality assessment of synthesized views is under work and already shows promising results.

## REFERENCES

- [1] P. Merkle, A. Smolic, K. Mueller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proceedings of ICIP*, 2007, pp. 201–204.
- [2] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3D video and free viewpoint Video Technologies, applications and MPEG standards," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME06)*, 2006, pp. 2161–2164.
- [3] H. Schwarz, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, D. Marpe, P. Merkle, K. Muller, H. Rhee, G. Tech, M. Winken, and T. Wiegand, "3D video coding using advanced prediction, depth modeling, and encoder control methods," in *Picture Coding Symposium (PCS)*, 2012, May 2012, pp. 1–4.
- [4] M. Tehrani, S. Shimizu, G. Lafruit, T. Senoh, T. Fujii, A. Vetro, and M. Tanimoto, "Use Cases and Requirements on Free-viewpoint Television (FTV)," ISO/IEC JTC1/SC29/WG11 CODING OF MOVING PICTURES AND AUDIO, Geneva, Switzerland, 2013.

- [5] M. Tanimoto, "Overview of Free Viewpoint Television," *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, July 2006.
- [6] C. Fehn, P. Kauff, M. O. de Beeck, F. Ernst, W. IJsselsteijn, M. Pollefeys, L. V. Gool, E. Ofek, and I. Sexton, "An evolutionary and optimised approach on 3D-TV," in *Proc. International Broadcast Conference*, 2002, pp. 357–365.
- [7] 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*, vol. 5291, 2004, pp. 93–104.
- [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," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 7, pp. 1332–1343, nov. 2011.
- [9] S. Yasakethu, S. Worrall, D. De Silva, W. A. C. Fernando, and 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, Jul. 2011.
- [10] E. Ekmekcioglu, S. T. Worrall, D. De Silva, W. A. C. Fernando, and A. M. Kondoz, "Depth based perceptual quality assessment for synthesized camera viewpoints," in *Proc. of Second International Conference on User Centric Media, UCMedia 2010*, Palma de Mallorca, 2010.
- [11] C.-T. Tsai and H.-M. Hang, "Quality assessment of 3d synthesized views with depth map distortion," in *Visual Communications and Image Processing (VCIP)*, 2013, Nov 2013, pp. 1–6.
- [12] W. Malpica and A. Bovik, "SSIM based range image quality assessment," in *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2009.
- [13] L. McMillan and Jr., "An image-based approach to three-dimensional computer graphics," Tech. Rep., 1997.
- [14] Z. Tauber, L. Ze-Nian, and M. Drew, "Review and preview: Disocclusion by inpainting for image-based rendering," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 4, pp. 527–540, 2007.
- [15] L. Azzari, F. Battisti, and A. Gotchev, "Comparative analysis of occlusion-filling techniques in depth image-based rendering for 3d videos," in *Proc. ACM Multimedia 2010 - MoViD: ACM Workshop on Mobile Video Delivery*, oct 2010.
- [16] L. Azzari, F. Battisti, A. Gotchev, M. Carli, and K. Egiazarian, "A modified non-local mean inpainting technique for occlusion filling in depth-image based rendering," in *Proc. SPIE International Conference on Electronic Imaging 2011, Stereoscopic Displays and Applications XXII*, jan 2011.
- [17] Y. Liu, Q. Huang, S. Ma, D. Zhao, and W. Gao, "Joint video/depth rate allocation for 3d video coding based on view synthesis distortion model," *Image Communication*, vol. 24, no. 8, pp. 666–681, Sep. 2009.
- [18] G. Tech, K. Muller, and T. Wiegand, "Evaluation of view synthesis algorithms for mobile 3dtv," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2011, may 2011, pp. 1–4.
- [19] E. Bosc, P. Callet, L. Morin, and M. Pressigout, "Visual quality assessment of synthesized views in the context of 3D-TV," in *3D-TV System with Depth-Image-Based Rendering*, C. Zhu, Y. Zhao, L. Yu, and M. Tanimoto, Eds. New York, NY: Springer New York, 2013, pp. 439–473.
- [20] P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand, "Depth image based rendering with advanced texture synthesis," in *Proc. IEEE International Conference on Multimedia & Expo (ICME)*, Singapore, July 2010.
- [21] M. Koppel, P. Ndjiki-Nya, D. Doshkov, H. Lakshman, P. Merkle, K. Muller, and T. Wiegand, "Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering," in *2010 17th IEEE International Conference on Image Processing (ICIP)*. IEEE, Sep. 2010, pp. 1809–1812.

- [22] F. Devernay and A. R. Peon, "Novel view synthesis for stereoscopic cinema: detecting and removing artifacts," in *Proceedings of the 1st international workshop on 3D video processing*, ser. 3DVP '10. New York, NY, USA: ACM, 2010, pp. 25–30. [Online]. Available: <http://doi.acm.org/10.1145/1877791.1877798>
- [23] P. Joveluro, H. Malekmohamadi, W. A. Fernando, and 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, pp. 1–4.
- [24] Y. Zhao and L. Yu, "A perceptual metric for evaluating quality of synthesized sequences in 3DV system," in *Proceedings of SPIE*, vol. 7744, 2010, p. 77440X.
- [25] P.-H. Conze, P. Robert, and L. Morin, "Objective view synthesis quality assessment," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 8288, Feb. 2012.
- [26] M. Solh, G. AlRegib, and J. Bauza, "3VQM: a vision-based quality measure for DIBR-based 3D videos," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, 2011, p. 16.
- [27] D. Howard, M. Green, R. Palaniappan, and N. Jayant, "Visibility of digital video artifacts in stereoscopic 3D and comparison to 2D," in *International Conference on Stereoscopic 3D for Media and Entertainment, SMPTE*, 2010.
- [28] F. Dufaux and F. Moscheni, "Motion estimation techniques for digital TV: a review and a new contribution," *Proceedings of the IEEE*, vol. 83, no. 6, pp. 858–876, 1995.
- [29] "IRCCyN/IVC DIBR Images," <http://www.irccyn.ec-nantes.fr/spip.php?article865>. [Online]. Available: <http://www.irccyn.ec-nantes.fr/spip.php?article865>
- [30] E. Bosc, M. Koppel, R. Pepion, M. Pressigout, L. Morin, P. Ndjiki-Nya, and P. Le Callet, "Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols?" in *ICIP 2011*, Brussels, 2011.
- [31] H. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, Jun. 1967. [Online]. Available: <http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1967.10482916>
- [32] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of Graphics, GPU, and Game Tools*, vol. 9, no. 1, pp. 23–34, 2004.
- [33] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3-D warping using depth information for FTV," *Elsevier Signal Processing: Image Communication*, vol. 24, pp. 65–72, 2009.
- [34] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3-D video systems," *EURASIP Journal on Image and Video Processing*, 2008, article ID 438148, 11 pages.
- [35] ITU-T, "Subjective video quality assessment methods for multimedia applications," Geneva, Tech. Rep. Rec. P910, 2008.
- [36] I. BT., *500, Methodology for the subjective assessment of the quality of television pictures*. November, 1993.
- [37] Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective models of video quality assessment," *VQEG*, Mar, 2000.
- [38] V. Oliveira and A. Conci, "Skin detection using HSV color space," in *SIBGRAPI*, 2009.
- [39] M. Haber and H. Barnhart, "Coefficients of agreement for fixed observers," *Statistical methods in medical research*, vol. 15, no. 3, p. 255, 2006.