Available online at www.sciencedirect.com

ScienceDirect

00 (2015) 1–23

Subjective evaluation of Super Multi-View compressed contents on high-end light-field 3D displays

Antoine Dricot[a,b], Joel Jung[a], Marco Cagnazzo[b], Béatrice Pesquet[b], Frédéric Dufaux[b], Péter Tamás Kovács[c,d], Vamsi Kiran Adhikarla[c,e]

[a] *Orange Labs*
[b] *Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI*
[c] *Holografika Kft.*
[d] *Department of Signal Processing, Tampere University of Technology*
[e] *Pazmany Peter Catholic University, Faculty of information Technology*

Abstract

Super Multi-View (SMV) video content is composed of tens or hundreds of views that provide a light-field representation of a scene. This representation allows a glass-free visualization and eliminates many causes of discomfort existing in current available 3D video technologies. Efficient video compression of SMV content is a key factor for enabling future 3D video services. This paper first compares several coding configurations for SMV content, and several inter-view prediction structures are also tested and compared. The experiments mainly suggest that large differences in coding efficiency can be observed from one configuration to another. Several ratios for the number of coded and synthesized views are compared, both objectively and subjectively. It is reported that view synthesis significantly affects the coding scheme. The amount of views to skip highly depends on the sequence and on the quality of the associated depth maps. Reported ranges of bitrates required to obtain a good quality for the tested SMV content are realistic and coherent with future 4K/8K needs. The reliability of the PSNR metric for SMV content is also studied. Objective and subjective results show that PSNR is able to reflect increase or decrease in subjective quality even in presence of synthesized views. However, depending on the ratio of coded and synthesized views, the order of magnitude of the effective quality variation is biased by PSNR. Results indicate that PSNR is less tolerant to view synthesis artifacts than human viewers. Finally, preliminary observations are initiated. First, the light-field conversion step does not seem to alter the objective results for compression. Secondly, the motion parallax does not seem to be impacted by specific compression artifacts. The perception of the motion parallax is only altered by variations of the typical compression artifacts along the viewing angle, in cases where the subjective image quality is already low. To the best of our knowledge, this paper is the first to carry out subjective experiments and to report results of SMV compression for light-field 3D displays. It provides first results showing that improvement of compression efficiency is required, as well as depth estimation and view synthesis algorithms improvement, but that the use of SMV appears realistic according to next generation compression technology requirements.

Keywords: 3D, super multi-view, SMV, light-field, video compression, video coding, subjective evaluation

1. Introduction

3D video provides an enhanced experience to the viewers, compared with usual 2D images. Current 3D technologies available on the consumer market present however several limitations [1]. The use of glasses in stereoscopic 3D introduces a lack of comfort, combined to annoying perception stimuli such as a vergence-accommodation conflict

(i.e. the eyes converge on a 3D image in front or behind the screen but focus on the screen plane), which can cause headaches and eyestrain. In current glass-free autostereoscopic display systems, the limited number of views cannot provide a smooth motion parallax (i.e. the visualization is not continuous when moving in front of the display) and have a restricted viewing zone with a limited number of sweet spots.
Super Multi-View video (SMV) is a glass-free 3D video technology that uses tens or hundreds of views of a scene to obtain a light-field representation of that scene [2, 3]. The light-field representation allows eliminating most of current 3D technologies drawbacks (e.g. the vergence-accommodation conflict) and can provide a smooth motion parallax, which is a key cue in the perception of depth. Several companies already show interest in this technology by working on light-field 3D display systems [4], as for example Holografika [5] which provides glass-free light-field display systems with smooth horizontal motion parallax on a large viewing angle.
SMV content can be acquired using a large number of cameras (or virtual cameras in the case of Computer Generated content), each of them providing a view of the scene from a different angle. The increasing number of views needed for SMV represent a large amount of data which is challenging to encode. H.264/AVC [6] and HEVC [7] standard encoders have multi-view extensions [8], respectively MVC and MV-HEVC, which provide additional high level syntax allowing inter-view prediction. Moreover, the Multi-View plus Depth (MVD) format [9] allows encoding only a subset of the views and their associated depth maps (a gray level image which represents the depth of each pixel). The views that are not encoded are then synthesized [10] at the decoder side. 3D-HEVC extension provides depth maps related tools and new tools at Coding Unit level (CUs in HEVC replace H.264/AVC macroblocks) for side views.
Efficient video compression of SMV content is a key factor for enabling future 3D video services. The in-depth understanding of the interactions between video compression and display is of prime interest. Evaluating the quality of 3D content is a challenging issue [11, 12]. In the context of SMV content, increased number of views and (depending on the configuration) increased number of synthesized views make it even more challenging. The main goal in this work is to assess the impact of compression on perceived quality for light-field 3D video content and displays. To the best of our knowledge, this paper is the first to carry out subjective experiments and to report results of this kind.
Assessing a range of bitrates required to provide an acceptable quality for compressed light-field content will give a cue on the feasibility of transmitting this kind of content on future networks. It is also needed to understand how much view synthesis disturb the general quality (both subjectively and objectively). Moreover, depth based rendering and synthesized views make the PSNR less relevant [13], but no other metric is currently accepted as more appropriate. One of the goals of this paper is to evaluate how much the use of the PSNR remains relevant, and if future codec developments can keep on relying on this basic indicator. Finally, as classical compression is well known to generate artifacts such as blocking, ringing, etc. One of the goals of this paper is to observe possible new compression artifacts that may affect the specific aspects of visualization of light-field content like the motion parallax, the perception of depth, etc. Our experiments provide first results showing that improvement of compression efficiency is required, as well as depth estimation and view synthesis algorithms improvement, but that the use of SMV appears realistic according to next generation compression technology requirements.
This paper is organized as follows. Section 2 describes the main configurations considered for Super Multi-View video coding. The principle of SMV display systems is described in Section 3 with a focus on Holografika's Holovizio system [5]. In Section 4, preliminary experiments are conducted in order to select the most relevant coding configurations for SMV content. Several configurations with varying ratios of coded/synthesized views are compared in Section 5 and objective results are shown. Subjective evaluation of the tested configurations is described in Section 6, and subjective results are presented and analyzed. Conclusions and perspectives are drawn in Sections 7.

2. Super Multi-View video coding

2.1. Multi-View plus Depth and synthesis

SMV content consists of tens or hundreds of views, with each view representing the scene from a different point of view. This corresponds to the number of input views for the display. It should be noted that this number varies depending on the model of the display system. In the first coding configuration considered, all the views are encoded. An example with an MV-HEVC based encoding is illustrated in Figure 1.
A second configuration is considered where only a subset of the views is encoded as well as the associated depth maps, as illustrated in Figure 2. For computer generated content, depth maps are generally perfectly known. For

natural video content, depth maps can be captured with dedicated cameras or estimated from the views with depth estimation algorithms (see Section 4.2). After decoding, the views that were skipped (not encoded) are synthesized (Figure 3).

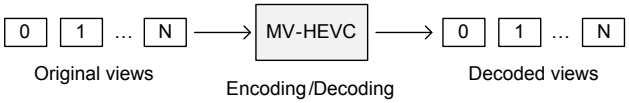


Figure 1. MV-HEVC encoding scheme (N views)

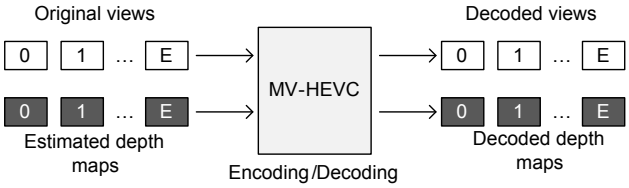


Figure 2. MV-HEVC encoding scheme (E views + E depth maps)

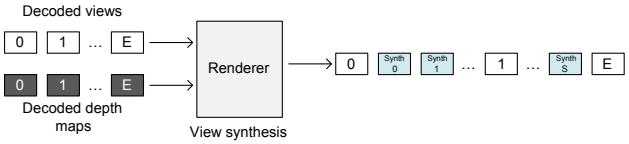


Figure 3. Rendering of S synthesized views from E views and associated depth maps

2.2. Inter-view prediction structures

Multi-view extensions of standard encoders (e.g. MV-HEVC) provide high level syntax which allows inter-view prediction. Inter-view prediction is based on the same principle as temporal prediction, i.e. intra frames I are coded independently, and predicted P (or bi-predicted B) frames are coded using other already coded frame(s) as reference. A combination of inter-view and temporal prediction structure is illustrated in Figure 4.

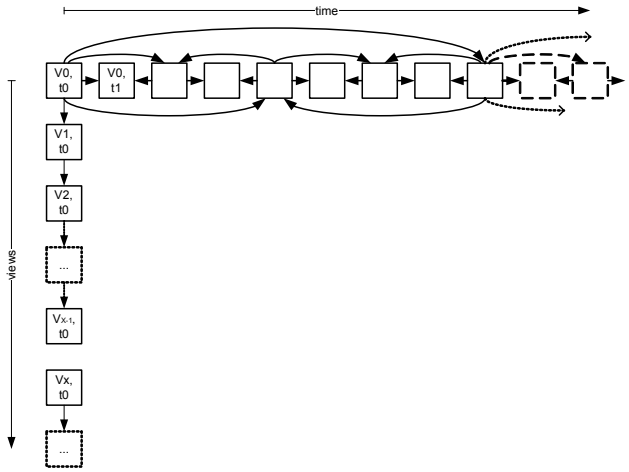


Figure 4. Group of X views with hierarchical temporal prediction structure and IPP inter-view prediction structure

3. Super Multi-View display system

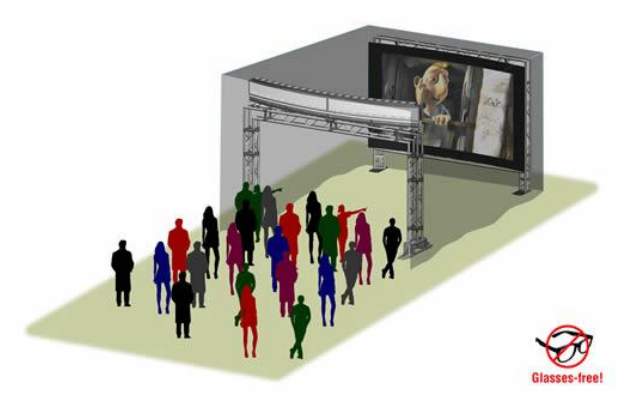


Figure 5. Holovizio C80 cinema system [5]

3.1. Example of light-field display system

SMV display systems, also called light-field displays, take tens to hundreds of views as input. Several display systems are based on a front or rear projection [2]. Each projection unit projects from a different angle onto a screen. The screen surface has anisotropic properties (which can be obtained optically or with a holographic diffuser for example), so that the light rays can only be seen from one direction which depends on the projection direction. Holovizio C80 display system, which has been used in our experiments, is illustrated in Figure 5 and consists of a large screen (3×1.8 meters) and of 80 projection units with a 1024×768 resolution, controlled by a rendering cluster. It offers a viewing angle of approximately 40°. Technical specifications and details are available at [5].

3.2. Light-field conversion

As illustrated in Figure 6, the input SMV content needs to be converted to be displayed on the Holovizio system. In the experiments described in the following of this paper, there are 80 views as input (N=80), captured by 80 cameras horizontally aligned in a linear arrangement. These views as well as the parameters of the camera rig (baseline, distance from the center of the scene, dimensions of the region of interest, etc.) are provided to the converter. Most of the common video and image formats (e.g. jpg, png, avi, etc.) are supported (as input and output) by the converter. It should be noted that the number of input views N is not fixed and can be more or less than 80. The converter outputs 80 light-field slices (P=80), which are provided to the player (software) at the display step. The whole image projected by a single projection unit cannot be seen from a single viewing position [14], therefore one projection unit represents a light-field slice, which is composed of many image fragments that will be perceived from different viewing positions. The number of light field slices P is fixed for a given display system as it corresponds to the number of projection units. Hence N should not necessarily be equal to P.

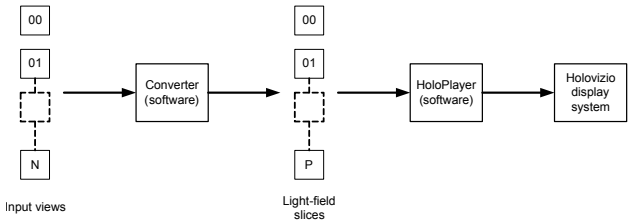


Figure 6. Conversion step of the input views before the display

4. Preliminary encoding configurations experiments

In the following, we report the results of preliminary tests performed in order to select the most relevant parameters, encoding configurations and encoding structures to encode the content included in the following subjective quality evaluation (Section 6).

4.1. Experimental content

The experiments in this paper include the SMV content described in Table 1. Dog, Pantomime, and Champagne Tower sequences [15] have been captured with the same camera system. Big Buck Bunny [16] and T-Rex [5] are sequences generated from 3D scenes (with Blender [17] and 3ds Max [18] respectively). There is a significant difference in the coding performance of content acquired with linear or arc camera arrangement [19]. As a consequence, only linear content is exploited in this work, in order to avoid that camera setup variations affect our conclusions. The comparison between the two kinds of contents will be studied in future work. As the coding efficiency and the quality provided by the light-field display system also depend on the number of input views, 80 views are used for each sequence.

Name	Fps	Duration (s)	Resolution	Views	Camera Setup	Type
ChampagneTower	30	6	1280x960	80	Linear	Real scenes
Pantomime	30	6	1280x960	80	Linear	
Dog	30	6	1280x960	80	Linear	
T-Rex	30	6	1920x1080	80	Linear	Computer generated
Bunny	24	5	1280x768	80	Linear	

Table 1. Description of the content used in our experiments

4.2. Depth estimation

As described in Section 2, depth maps can be captured, generated, or estimated. The depth maps used for experiments in this paper are estimated with DERS6.0 (Depth Estimation Reference Software [20]). Preliminary experiments are performed in order to compare several values for the following parameters of this software: Precision (1: Integer-Pel, 2: Half-Pel, or 4: Quarter-Pel), corresponding to the level of precision chosen to find correspondences, Search Level (1: Integer-Pel, 2: Half-Pel, or 4: Quarter-Pel), corresponding to the level of precision of candidate disparities, and Filter (0: Bi-linear, 1: Bi-Cubic, or 2: MPEG-4 AVC 6-tap), corresponding to the upsampling filter used to generate image signals at sub-pixel positions.
In these preliminary experiments, the depth maps for views 37 and 39 of Champagne sequence are estimated on 30 frames. The view 38 is then synthesized (with VSRS4.0 and with the HTM10.0 renderer - see next section), and the PSNR of this synthesized view is computed against the original view 38. The depth maps provided with the Champagne sequence [15] (which are estimated semi-automatically with DERS) are also tested for comparison.

Configuration			PSNR Y (dB)	
Precision	Search Level	Filter	VSRS4.0	HTM10.0 Renderer
1	1	1	33,8	34,0
4	4	2	32,0	32,2
Provided depth maps			33,8	34,3

Table 2. Preliminary results for DERS configuration

Table 2 shows that the lower values for the tested parameters provide a better PSNR for the synthesized view, and that this result is closer to the result obtained with the semi-automatically estimated depth maps provided with the sequence. Selecting higher values for the tested parameters implies the use of more advanced tools (e.g. with higher precision). These values provide depth maps with a smoother aspect and apparently less artifacts, however this involves a decrease of the PSNR of the synthesized view (i.e. more synthesis artifacts appear). As view synthesis is the main purpose of the depth map estimation here, the configuration with lower parameter values is used to estimate all the depth maps included in our experiment phase.

4.3. View synthesis

We have performed experiments to compare the 3D-HEVC Renderer [8] and VSRS4.0 (View Synthesis Reference Software [21]) with several configurations obtained by assigning different values for the following parameters: Precision (1: Integer-Pel, 2: Half-Pel, or 4: Quarter-Pel) and Filter (0: Bi-linear, 1: Bi-Cubic, or 2: MPEG-4 AVC), which are used for values at sub-pixel positions, Boundary Noise removal (0: disable, 1: enable), which process artifacts on edges, Mode (0: General, 1: 1D Parallel), corresponding to the type of camera arrangement, and Blend (0: disable, 1: enable), used to blend the right and left input views. View 38 of Pantomime sequence is synthesized on 30 frames and the PSNR is computed against the original view 38.

Configuration					PSNR Y (dB)	Time (s)
Precision	Filter	Boundary noise removal	Mode	Blend		
2	1	1	1	1	37,4	33
2	1	0	0	1	38,3	60
2	1	0	1	1	37,8	26
2	1	1	0	1	36,4	61
2	1	0	0	0	38,3	58
2	0	0	0	1	38,5	57
2	2	0	0	1	38,4	58
1	1	0	0	1	37,7	91
4	1	0	0	1	38,3	61
4	0	0	0	1	38,5	60
4	2	0	0	1	38,4	59
3D HEVC Renderer (HTM10.0)					38,6	17

Table 3. Preliminary results for view synthesis software configuration

Table 3 shows the PSNR results and the processing time for this preliminary experiment. HTM10.0 Renderer provides better results than VSRS4.0 in our experiments conditions and is also faster (approximately one third of the time of most VSRS configurations). Hence HTM10.0 Renderer (with default configuration) is used to synthesize all the intermediate views in our experiment phase.

4.4. Group of views (GOV)

Encoding 80 dependent views is very demanding in terms of memory (RAM). To avoid memory limitations, a configuration with Groups Of Views (GOV) can be used, as illustrated in Figure 4. Table 4 compares the performance of encoding 80 views with groups of 16 views against groups of 9 views. For this experiment, MV-HEVC [8, 22] reference software version 10 is used (HTM v.10.0 with macro HEVC_EXT=1). 180 frames of Champagne, Dog and Pantomime sequences are encoded with QPs 20-25-30-35. IPP inter-view reference coding structure is used (see Sec. 4.5). The results are provided using the Bjøntegaard Delta rate (BD-rate) metric [23], which computes the average bitrate saving (in percentage) for a given quality between two rate-distortion curves, as described in [24].

GOV 16 vs. GOV 9 (mean PSNR on 80 views)	
ChampagneTower	-0,9%
Dog	-5,2%
Pantomime	-3,1%
Mean	-3,1%

Table 4. BD-rate performance of GOV size 16 against GOV size 9

Table 4 shows that using a larger group (from 9 to 16 views) provides an average BD-rate gain of 3.1%. The insertion of I-frames to create GOVs has a non-negligible impact on the BD-rate results. However, this limitation in the configuration is relevant for future use cases because the memory limitation is a practical reality. Moreover GOVs

allow parallel processing at both the encoder and decoder side, prevent from the loss of all the views when losing one view due to network errors for example, and provide some limits on error propagation into other views when losing one view.

4.5. Inter-view reference pictures structure

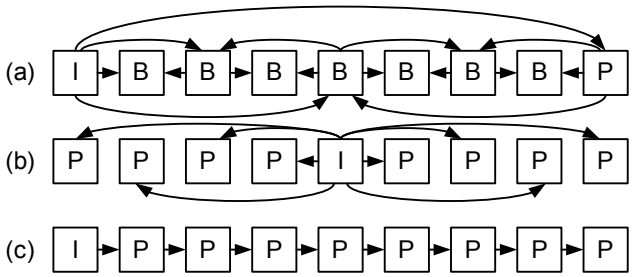


Figure 7. Inter-view reference structures within a GOV: (a) Hierarchical, (b) Central, (c) IPP

In this section we compare 3 inter-view reference structures inside the GOVs, illustrated in Figure 7 as follows: Hierarchical (a), Central (b), and IPP (c). Table 5 shows the BD-rate performance for these 3 structures with groups of 9 views. IPP is the most efficient inter-view reference structure for this experiment.

Ref: Central (b) (mean PSNR on 80 views)		
Sequence	IPP (c)	Hierarchical (a)
ChampagneTower	-7,5%	-0,9%
Dog	-6,1%	-5,5%
Pantomime	-2,9%	2,3%
Mean	-5,2%	-1,0%

Table 5. BD-rate performance depending on the interview reference structure within GOVs

The experiment is extended with some views skipped to simulate the encoding of a subset of the views as in configurations including view synthesis. The main goal is to confirm that IPP structure remains the most efficient in these configurations. 9 views are encoded with an increasing baseline from 1 view skipped (referred to as skip1) to 9 views skipped (referred to as skip9) between two coded views. Results are shown in Tables 6, 7, and 8. As expected when the baseline increases (more distance between the coded and the reference views), IPP remains the most efficient.

Baseline: skip1	Ref: Central (b)	
Sequence	IPP (c)	Hierarchical (a)
ChampagneTower	-8,1%	-1,3%
Dog	-2,9%	1,1%
Pantomime	-8,4%	2,0%
Mean	-6,5%	-1,3%

Table 6. BD-rate performance with different inter-view reference structures (with 1 view skipped)

5. Objective experimental results

Based on the preliminary results obtained in Sec. 4, the content is encoded with IPP inter-view reference structure and groups of 16 views (i.e. one intra frame every 16 views). For the configuration where all the views are encoded

Baseline: skip3	Ref: Central (b)	
Sequence	IPP (c)	Hierarchical (a)
ChampagneTower	-8,9%	-5,7%
Dog	-9,2%	-4,4%
Pantomime	-15,8%	-6,2%
Mean	-12,6%	-5,6%

Table 7. BD-rate performance with different inter-view reference structures (with 3 views skipped)

Baseline: skip9	Ref: Central (b)	
Sequence	IPP (c)	Hierarchical (a)
ChampagneTower	-7,4%	-4,9%
Dog	-8,2%	-1,3%
Pantomime	-11,3%	-3,9%
Mean	-9,6%	-3,4%

Table 8. BD-rate performance with different inter-view reference structures (with 9 views skipped)

(i.e. no skipped views), QPs 15, 17, 20 to 30, 32, 35, 37 and 40 are used in order to provide a large and dense range of bitrates. For the configurations with views skipped at the encoder, QPs 20, 25, 30, 35 are used. Resulting PSNR-bitrate curves are illustrated in Fig. 8, 9, 10, 11, and 12.

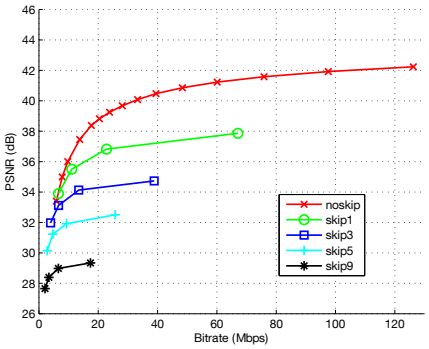


Figure 8. PSNR-bitrate curve (Dog)

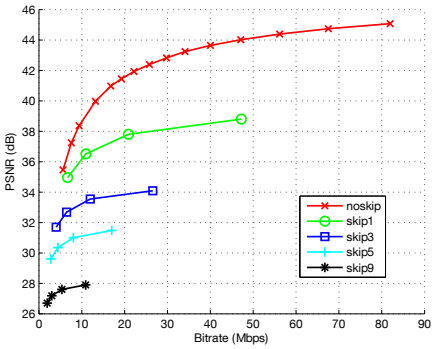


Figure 9. PSNR-bitrate curve (Champagne)

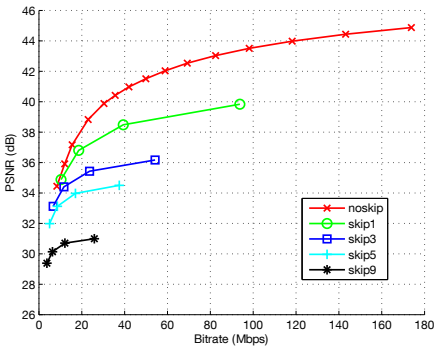


Figure 10. PSNR-bitrate curve (Pantomime)

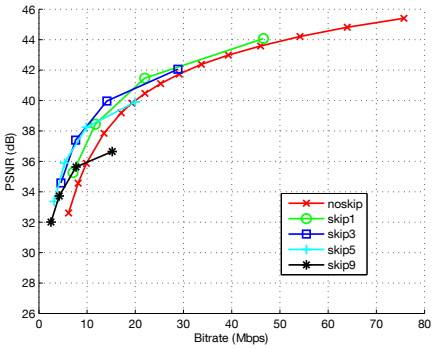


Figure 11. PSNR-bitrate curve (T-Rex)

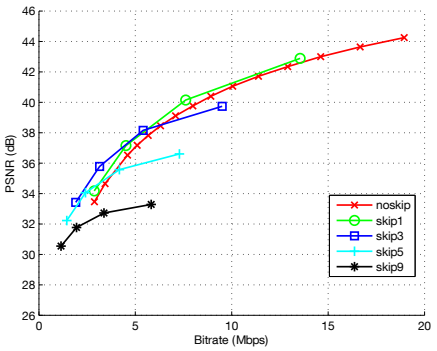


Figure 12. PSNR-bitrate curve (Bunny)

PSNR values between 30 and 40 dB are generally considered as corresponding to acceptable to good qualities for 2D images. Most of the curves in this experiment are above this 30 dB limit, except for some skip9 curves (Dog, Champagne).

For the content captured from real scenes (Dog, Champagne, Pantomime), the PSNR decreases as the number skipped/synthesized views increases. For example, with the Dog sequence (Figure 8), the gap between the different curves for a given bitrate is approximately from 2 to 3 dB. This is mainly due to the limitations of the PSNR, which is severe with synthesis artifacts (leading to a significant decrease of objective quality) that are hardly percep-

tible by human observers (see Section 6).
For the computer generated content (T-Rex, Bunny), the decrease in PSNR from one configuration to another is less severe. The skip1 and skip3 can even provide better results than the configuration without synthesis. One of the reasons can be the noise-less aspect of the content. Because of this characteristic, first the depth estimation (which is based on block-matching algorithms) is more accurate, and thus allows a better synthesis. Secondly, the PSNR metric is also less disturbed by hardly perceptible noisy variations. These results show that the weaknesses of the synthesis algorithms affect the coding scheme.

6. Subjective evaluation

In this section, we describe the subjective evaluation of SMV compressed content encoded in Section 5. Experimental conditions and evaluation methodology are first described, and subjective results are then presented and analyzed.

6.1. Experimental conditions

6.1.1. Raw video files constraint for the display step

As described in Section 3.2, the captured views are converted into light-field slices before the display step. Each light-field slice is associated to a projection unit of the Holovizio display system (the projection units are illustrated in Figure 13). The encodings performed in our experiments are done with raw video files (YUV4:2:0 raw data contained in .yuv files) as input of the encoder and as output of the decoder (and renderer), so that the video data suffers no degradation, apart from the compression effect which is aimed to be observed in the experiments. Additional software has been developed by Holografika in order to handle the YUV raw video format in the converter and the player. The use of a raw video format induces large file sizes. During our experiments, in order to have a smooth playback on the display, the light-field slices in raw video format had to be copied directly to the ramdisk (more specifically on a compressed ramdisk) of their associated nodes (e.g. light-field slices 72 to 79 are copied on Node #09's ramdisk, as illustrated in Figure 13).

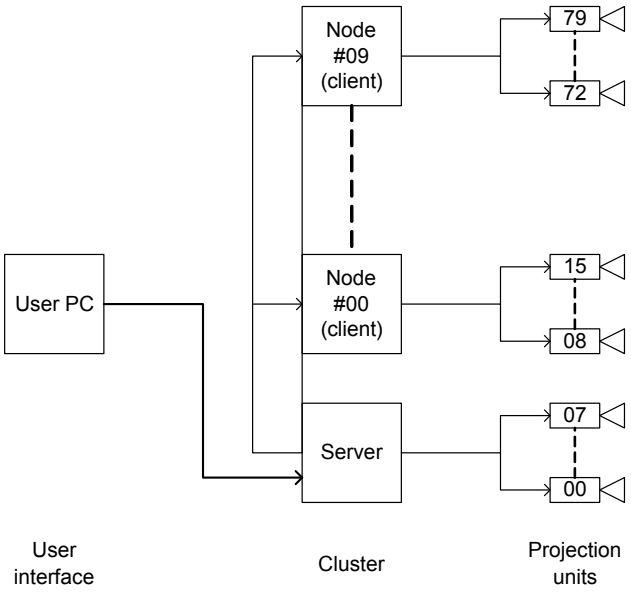


Figure 13. Display system structure

(a) Dog (b) Champagne (c) Bunny

Figure 14. Example frame for each sequence.

6.1.2. Content selection

Each Holovizio light-field display system has a field of depth in which the content of the scene must be included to be displayed correctly. The objects in the scene that are outside of these depth bounds (i.e. too far or too close) present ghost-like artifacts. In our experiments, it is the case for the objects in the background of the ChampagneTower sequence as well as for the background of the Dog sequence. A small part of Bunny is also too close in the foreground but in a slighter way which does not impact significantly the visualization.

Three sequences are included in the subjective evaluation: Dog, Champagne, and Bunny. In Dog, a woman plays with a dog in front of a colored curtain with dots. Champagne represents a woman serving champagne in glasses in front of a black curtain. Bunny is a computer generated scene with a rabbit coming out of a burrow with grass in the background. Example frames from the sequences are available in Figure 14.

The sequences encoded in the preparation phase as described in Section 5 have been evaluated in a preliminary subjective evaluation session in order to select relevant configurations to include in the limited time of one test session (see Section 6.1.3). Based on this preliminary visualization the following configurations are included in the evaluation: QPs 25-30-35 for the noskip configurations, QPs 20-30 for skip1 and skip3, and only QP 20 for skip5 and skip9.

6.1.3. Subjective evaluation methodology

The evaluation methodology used in our experiments is the double-stimulus impairment scale (DSIS) method (the EBU method) [25]. The double-stimulus method is cyclic. The assessor is first presented with an unimpaired reference, and then with the same picture impaired. In our experiments the first picture is the original (uncompressed) sequence, and the second picture is compressed and possibly synthesized. We followed the variant #2 of the DSIS method for which the pair is showed twice to the assessor. Following this, he is asked to vote on the second, keeping in mind the first unimpaired sequence. The rating scale is showed in Table 9.

Score	Impairments
5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Table 9. ITU-R impairment scale [25]

The choice of the DSIS method is motivated by the fact that the tested content and the display system already present flaws or artifacts that could prevent them from being rated as excellent. By using a comparative method like DSIS, we can ignore these aspects and only focus the evaluation on the compression/synthesis artifacts (which are the causes of the rated impairments).

6.1.4. Experimental set-up

Subjective experiments have been performed at Holografika's facilities (a surface of approximately 100m^2 with 3m height ceiling and a black curtain that halves the room) where the system is usually stored and used for demonstrations. The light-field display has been calibrated for geometry and intensity before the tests using proprietary Holografika tools, the same way calibration is performed for a new installation. Lighting conditions were the same as for demonstrations, i.e. a dark room with sparse sources of diffuse light (e.g. distant windows with curtains). No particular attention was paid to obtain a specific brightness measure (in cd/m^2) as no recommendation fits our experimental conditions on this point to our knowledge. Figure 15 illustrates the experimental set-up.

Figure 15. Experimental setup

For each viewing session there was between one and six subjects. Subjects were sitting at a viewing distance of approximately 6 meters from the screen (which has 3×1.8 meters dimensions), in the 40° angle recommended for the C80 display system. The experiments were performed with 16 subjects. The subjects are employees of Holografika and students from Budapest universities. Various categories of Holografika's employees participated in the evaluation (technical staff, engineers, programmers, etc.), which are not all working directly on light-field imaging, neither with the display systems, and there are no relevant differences between the results when separating the employees from Holografika and the others in two Expert/Non-expert groups. This can be explained by the fact that even if the panel contains light-field experts, the light-field related artifacts and characteristics in the sequences were present both in the original and compressed one. Thanks to the DSIS method, only compression artifacts are taken into account during the evaluation, therefore we have an homogenous panel of subjects who are non-experts in compression. Subjects have normal or corrected to normal vision.

Each session lasted roughly 30 minutes (as recommended in [25]). The duration required for 3DTV subjective evaluation is discussed in [26]. For 2D video, a 10 seconds duration is recommended in [25]. For 3D video, there are two conflicting arguments: i) as 3DTV is closer to the human natural viewing behavior, less time is needed to judge the quality, ii) more time is needed since more information is contained in the additional dimension. This discussion also applies for light-field video. In [27], the presentation time only had little effect on subjective evaluation results with durations of 5 and 10 seconds tested. Based on these considerations, and to limit the session duration as well as the encoding and conversion processing time, the duration of the tested content is limited to 5-6 seconds. From our observations, this 5-6 seconds duration did not appear to be too short. Pairs of sequences were displayed two times and artifacts were noticeable with one visualization.

Pencils and paper sheets with a scoring table to fill were provided. Before each evaluation session, written instructions describing the process in details as in Sec. 6.1.3 and [25] were provided and discussed with the subjects. One pair of sequences was first shown as a training phase, so that the subjects could experience the evaluation process and see examples of degradation types.

6.1.5. Light-field display specific aspects

In the evaluation of Super Multi-View compressed contents on light-field displays, different types of artifacts are observed at several steps in the process including capture, coding and display (transmission and/or storage are out of the scope here). The most significant ones are listed and discussed in the following. The very first artifacts in the content are created at acquisition (e.g. out-of-focus, low contrast, noise, etc.), however they should be considered as

characteristics of the content (because in some cases, distinguishing out-of-focus from artistic blur can be completely arbitrary for example) and should not impact the score in our case. Typical 2D compression artifacts (such as block artifacts) are introduced during the encoding. In configurations with synthesis, additional artifacts are present in the synthesized views. They are mainly block artifacts and synthesis errors flickering on objects edges. For multi-view (and SMV by extension) content, the differences between cameras (e.g. color calibration, brightness, etc.) can also be sources of artifacts. Additionally, the light-field conversion and the light-field display system itself are also potential sources for new artifacts that need to be listed and studied. Moreover, the impact of these artifacts (new ones and common ones listed above) on the perception of elements specific to light-field display and 3D imaging (e.g. the motion parallax, the depth, etc.) also needs to be studied.
The study in this paper is focused on the compression, therefore it does not take into account the variety of all the artifacts that occurs in the evaluated content. Thanks to the DSIS method the results of the evaluation are not impacted by those artifacts that are not related to compression or synthesis, because they are present in the original and the compressed content and therefore not scored. Our work provides preliminary hints and observations concerning the impact of light-field conversion and the perception of motion parallax in Sec. 6.6 and Sec. 6.7 respectively.

6.1.6. Statistical analysis methodology

According to chapter 2.3 in [28], it should be noted that since the panel size is relatively small (16 subjects), it is more relevant to compute the 95% confidence interval (CI) assuming that the scores follow a t-Student distribution, rather than a normal distribution as suggested in the ITU recommendation [25].
Two methods are used in our experiments to detect potential outliers. The first method is described in the recommendation [25] (for the DSIS evaluation). The principle of this method for screening the subjects is as follows. First, the β_2 test is used to determine if the distribution of scores for a given tested configuration t is normal or not. The kurtosis coefficient (β_2) of the function (i.e. the ratio between the fourth order moment m_4 and the square of the second order moment m_2) is calculated as in equations 1 and 2, with N the number of observers/scores. If β_2 is between 2 and 4, the distribution may be considered to be normal.

$$\beta_2 = \frac{m_4}{(m_2)^2} \tag{1}$$

$$\text{where } m_x = \frac{\sum_{i=1}^{N}(u_i - u_{mean})^x}{N} \tag{2}$$

Then for each tested configuration t, two values are processed as in equations 3 and 4: P_t corresponding to the mean value plus the associated standard deviation S_t times 2 (if normal) or times $\sqrt{20}$ (if non-normal), and Q_t corresponding to the mean value minus the associated standard deviation S_t times 2 or times $\sqrt{20}$.

$$\begin{aligned} P_t &= u_{mean} + 2 \times S_t \text{ (if normal)} \\ \text{or } P_t &= u_{mean} + \sqrt{20} \times S_t \text{ (if non-normal)} \end{aligned} \tag{3}$$

$$\begin{aligned} Q_t &= u_{mean} - 2 \times S_t \text{ (if normal)} \\ \text{or } Q_t &= u_{mean} - \sqrt{20} \times S_t \text{ (if non-normal)} \end{aligned} \tag{4}$$

Then for each observer i, every time a score is found above P_t a counter P_i associated with this observer is incremented. Similarly, every time a score is found below Q_t a counter Q_i associated with this observer is incremented. The following two ratios must be calculated: $P_i + Q_i$ divided by the total number of scores T for each observer, and $P_i - Q_i$ divided by $P_i + Q_i$ as an absolute value. If the first ratio is greater than 5% and the second ratio is less than 30%, then observer i must be eliminated as shown in equation 5.

$$\begin{aligned} &\text{if } \frac{(P_i + Q_i)}{T} > 0.05 \qquad \text{and} \qquad |\frac{(P_i - Q_i)}{(P_i + Q_i)}| < 0.3 \\ &\text{then reject observer } i \end{aligned} \tag{5}$$

The second method is described in chapter 2.3 of [28] as follows. For each tested configuration t, the interquartile range corresponds to the difference between the 25^{th} and the 75^{th} percentile. For a given tested configuration t, if the score $score_{i,t}$ of an observer i falls out of the interquartile range by more than 1.5 times, then this score is considered as an outlier score, as shown in equation 6. An observer is considered as outlier (i.e. is eliminated) if 20% of his scores are considered as outlier scores according to equation 6.

$$\begin{aligned} &\text{if } score_{i,t} < q_{t,25^{th}} - 1.5 \times (q_{t,75^{th}} - q_{t,25^{th}}) \\ &\text{or } score_{i,t} > q_{t,75^{th}} + 1.5 \times (q_{t,75^{th}} - q_{t,25^{th}}) \\ &\text{then } score_{i,t} \text{ is an outlier score} \end{aligned} \tag{6}$$

6.2. Subjective results

The raw data obtained after the subjective evaluation sessions is an array of 400 scores (25 sequences × 16 subjects). Table 10 shows the mean opinion score (MOS) for each tested configuration and its associated bitrate.

Sequence	configuration	#	QP	Bitrate (Mbps)	MOS
Dog	noskip	1	25	39,5	4,3
		2	30	17,7	4,1
		3	35	9,7	3,7
	skip1	4	20	67,1	4,6
		5	30	11,1	4,1
	skip3	6	20	38,9	4,4
		7	30	6,6	3,6
	skip5	8	20	25,7	4,1
	skip9	9	20	17,4	2,4
Champagne	noskip	10	25	34,1	4,7
		11	30	16,8	4,9
		12	35	9,4	4,3
	skip1	13	20	47,2	3,4
	skip3	14	20	26,6	3,1
	skip5	15	20	17,0	2,6
	skip9	16	20	10,9	1,9
Bunny	noskip	17	25	10,1	4,2
		18	30	5,7	3,9
		19	35	4,6	2,9
	skip1	20	20	13,5	4,6
		21	30	4,5	3,6
	skip3	22	20	9,5	4,8
		23	30	3,2	3,3
	skip5	24	20	7,3	4,8
	skip9	25	20	5,8	4,4

Table 10. Mean Opinion Scores (MOS) for each tested configuration and associated bitrate

Figure 16 (a), (b) and (c) show the results for Dog, Champagne and Bunny sequences respectively. The subjective scores (MOS) and associated confidence intervals (CI) are presented on the y-axis and the associated bitrates on the x-axis. The bitrate values are given in Mbps. Table 10 and MOS-bitrate curves show that the mean scores are globally coherent and increase/decrease as expected relatively to the QPs and configurations. The only obvious incoherence is the score (of 4.9) for Champagne sequence with the noskip configuration at approximately 16.8 Mbps (with QP 30) which is larger than the score (of 4.7) attributed to the same sequence also with noskip configuration at approximately 34.1 Mbps (with QP 25 which is less severe). However this could be explained by the fact that these two scores are very close to each other and to the highest score. Moreover, a statistical analysis of the distribution of the scores (e.g.

t-test [28]) would not define them as different scores because the CIs are superimposed. No outliers were detected in our panel using both methods. This, in addition to the reasonable size of the CIs, shows the reliability of the results of this evaluation.

For the Dog sequence, the curves are close to each other for noskip, skip1, skip3 and skip5 configurations. For example, the configurations skip3 and noskip (with QP 20 and 25 respectively) obtain approximately the same score (*Perceptible but not annoying*) at a bitrate of approximately 40 Mbps (rightest point on the dark blue curve, and rightest point on the red curve respectively). There is a tradeoff here because the reduction of bitrate due to reduced number of coded views allows a less severe compression, but induces synthesis artifacts. For this sequence with the skip9 configuration, impairments are rated between *Slightly annoying* and *Annoying* even with a bitrate of 17 Mbps for which the compression is not very severe (QP 20). This means that skip9 cannot be considered realistic here, because this low score is mainly due to the synthesis artifacts.
For the Champagne sequence, the noskip configuration is rated *Perceptible but not annoying* even at 10 Mbps (with QP 35). The curve shows that the limit with *Slightly annoying* score should be obtained at an even smaller bitrate value. The configurations with synthesis (only QP 20 on the curves) are not rated over *Slightly annoying* here, except for skip1 at 47 Mbps (QP 20), but this point is anyway associated to a bitrate already larger than for noskip rated *Imperceptible* at approximately 35 Mbps (QP 25). For this sequence, the configurations with view synthesis cannot be considered effective nor realistic (in our experimental conditions).
For the Bunny sequence, all the configurations with synthesis and QP 20 are rated between *Slightly annoying* and *Imperceptible*: skip1 at 13.5 Mbps, skip3 at 9.5 Mbps, and skip5 at 7 Mbps are very close to *Imperceptible*, noskip at 10 Mbps is closer to *Perceptible but not annoying* because of the compression artifacts (with QP 25), and skip9 at 6 Mbps also, because of the synthesis artifacts that appear. For this sequence, several configurations are close to *Slightly annoying* at a bitrate of approximately 4 Mbps. It should be noted that the curve for noskip configuration on Bunny is steeper than the other curves. This might be due to the fact that the Bunny sequence present fewer flaws than the Dog and Champagne sequences do, and so the compression distortions are more perceptible.

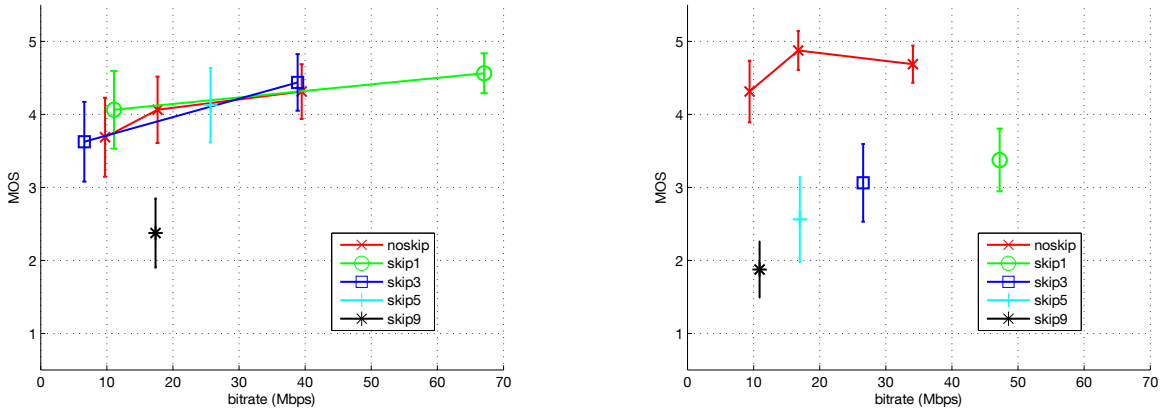
6.3. Impact of depth estimation and view synthesis

The experimental results in Section 6.2 first highlights the limitations of the configurations based on view synthesis. The results show that the efficiency and quality of the view synthesis greatly depend on the content of the sequence and on the quality of the associated depth maps. The 3 sequences used for the evaluation are quite representative of this dependency. For Champagne, the estimated depth maps present a lot of flickering from frame to frame, and do not show well shaped-objects (e.g. the lady is mingled with the background for most parts, as illustrated in Figure 17). The resulting synthesized 2D views present a lot of synthesis artifacts on objects edges. For Dog, the estimated depth maps are visually better, and the resulting synthesized views present fewer artifacts. The most severe artifacts generally appear only in skip5 and skip9 configurations. For Bunny, estimated depth maps and synthesized views are significantly better visually than for Dog or Champagne sequences. Even with the skip9 configuration, the synthesis artifacts are rare and hardly perceptible. The main reason might be that Bunny is a computer generated sequence (CG), so there are less misalignment issues (camera position, color calibration, capture noise, etc.) between the views than in a real world captured content, which allows the depth estimation and view synthesis algorithms to perform better (see Section 5).

The fact that the depth estimation and view synthesis algorithms performance is dependent on the content is an expected conclusion. However, in our experiments this inconsistency among sequences goes to an extent where for one sequence (Champagne) it is problematic even to synthesize only one view while for another (Bunny) it is possible to synthesize up to 9 views with a good quality. This shows that we cannot only rely on current depth estimation and view synthesis technologies for SMV video coding because they do not provide sufficient quality for some content.

6.4. Range of bitrate values for compressed light-field content

A second conclusion concerns the measured range of bitrate values and associated qualities. In our experiments the minimum bitrate values associated with *Slightly annoying* impairments are approximately 6.6 Mbps for Dog, 9.4 Mbps for Champagne, and 4.5 Mbps for Bunny (respectively with skip1 QP 30, noskip QP 35, and skip1 QP 30). Bitrate values associated to *Perceptible but not annoying* are about 11 Mbps for Dog (with skip1 configuration),



(a) Dog

(b) Champagne

(c) Bunny

(d) Bunny (close-up)

Figure 16. MOS scores and associated bitrates.

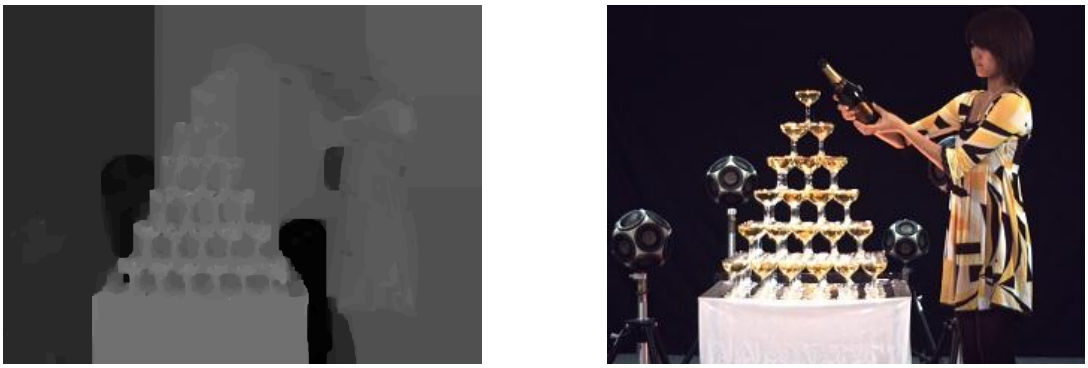


Figure 17. Estimated depth map and associated frame for Champagne (view 42)

less than 10 Mbps for Champagne (with noskip configuration), and about 5 Mbps for Bunny (with skip5 or skip9 configurations). The target bitrate values for encoding 4K content with HEVC are estimated at 10 to 15 Mbps, and 2 or 3 times more for 8K content. Moreover, encoding multiview content with 3D-HEVC can provide BD-rate gains from 20% to 25% over MV-HEVC in a configuration including 3 coded views (and associated depth maps) and 6 synthesized views. According to these values, the use of SMV content associated with MV-HEVC/3D-HEVC based encoding appears realistic for future 3D-services and broadcast. This conclusion cannot be considered as definitive because it is limited by the conditions of our experiments which largely depend on the characteristics of the display (like spatial and angular resolutions, field of depth, etc.) and of the tested content (resolution, camera arrangement, etc.). However these results provide a significant first hint on the feasibility of the 3D light-field video using SMV with current compression technologies.

6.5. Comparison between objective and subjective results

During the experiments we observed that some of the compression artifacts and synthesis artifacts are generally observable in the same way on the light-field display as they are when visualizing the 2D views separately, e.g. the typical compression block artifacts have the same recognizable aspect. Hence at this point, the experiments do not show any reason to prevent the measure of the quality for 3D light-field content by measuring the objective quality of the input views.

In Section 5, we provide objective results by comparing synthesized views against original views at the same position. The comparison of the performance for the different synthesis configurations (noskip, skip1, etc.) is not identical in the objective results and in the subjective results. In the objective results, the PSNR decreases as the number of synthesized views in the configuration increases for Dog and Champagne. The subjective results show the same decrease for Champagne, while for Dog, the noskip, skip3 and skip5 curves are very close and the skip3 curve is slightly better. For Bunny, the noskip, skip1 and skip3 curves are very close in the objective results and skip5 and skip9 are lower, while in the subjective results skip5 and skip9 are better. The PSNR is severe with synthesis artifacts that are not (or hardly) perceptible and do not impact the subjective quality. Figure 18, Figure 19, and Figure 20 show the residual images obtained by subtracting a view (#41) synthesized from the original uncompressed views (and depth maps) and the original view at the same position. Hence these captions only show the artifacts due to view synthesis. Table 11 shows the PSNR for views synthesized from uncompressed views computed against the original views at the same positions. These PSNR values (between 24.8 dB and 44.4 dB) are already impacted by the impairments due to the synthesis. PSNR is generally relevant for a given coding configuration, but not always with inter-configurations comparisons. Despite the limited number of points in our experiments, this aspect is coherent with the results provided.

Table 12 shows the Spearman and Pearson correlation coefficients [29] for the MOS and PSNR obtained on all configurations and sequences, with values of approximately 0.6 and 0.7 respectively, which suggest a correlation between the two variables (the correlation increases as the coefficients absolute value get closer to one). However, this correlation remains significantly smaller than in other mainstream video coding applications. Figure 21 plots the MOS relatively to the PSNR associated with each tested configuration. The curves are ascending functions, which

Sequence	Configuration	PSNR (dB)		
		Y	U	V
Dog	skip1	33,1	40,0	39,4
	skip3	31,8	39,9	39,0
	skip5	30,2	39,6	38,7
	skip9	27,4	39,4	38,4
Champagne	skip1	32,0	39,8	39,0
	skip3	29,4	39,1	38,3
	skip5	27,6	38,6	37,9
	skip9	24,8	37,2	36,9
Bunny	skip1	44,4	43,9	45,9
	skip3	38,7	42,8	45,9
	skip5	35,3	42,0	45,5
	skip9	31,9	40,3	44,1

Table 11. PSNR of uncompressed synthesized views

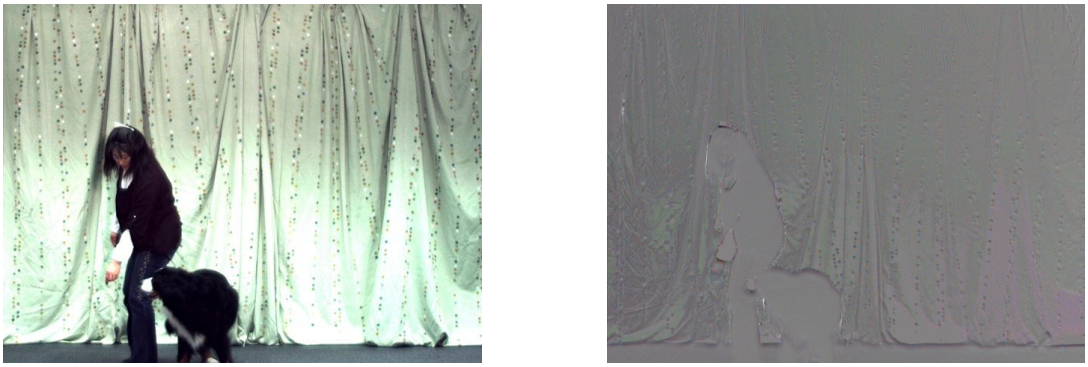


Figure 18. Synthesized view and residual synthesis artifacts (skip1, Dog)

Figure 19. Synthesized view and residual synthesis artifacts (skip1, Champagne)

	Spearman	Pearson
Coeff	0.6356	0.7299

Table 12. Correlation coefficients between MOS and PSNR (on all sequences and tested configurations)

shows that the PSNR is able to reflect the increase in the effective quality (even in the presence of skipped views). However the curves also show the inconsistency of the relation between PSNR and MOS among configurations. For

Figure 20. Synthesized view and residual synthesis artifacts (skip1, Bunny)

Dog sequence, the MOS value 4 (associated to a score where the impairments are *Perceptible but not annoying*) approximately matches: 33dB with 5 skipped views, 34dB with 3 skipped views, 35.5dB with 1 skipped views, and 38dB without skipped views. This is mainly due to the inefficiency of the PSNR metric for synthesized views (i.e. the PSNR is too severe with synthesis artifacts). However for the noskip configuration, the curves of the 3 sequences cross the MOS value 4 at a PSNR of approximately 38 dB. Hence for this configuration, the PSNR values might be aligned to the MOS values. There is a consistency across sequences in the relation between PSNR and MOS for the configuration without synthesis only, and PSNR is able to reflect an increase of the effective quality. However the order of magnitude of the effective quality variation is biased by the PSNR and change for the different configurations.

6.6. Impact of the light-field conversion step

In this section, we compute the PSNR of the light-field slices (see Section 3.2 about light-field conversion) converted from compressed input views against the light-field slices converted from the original uncompressed views.

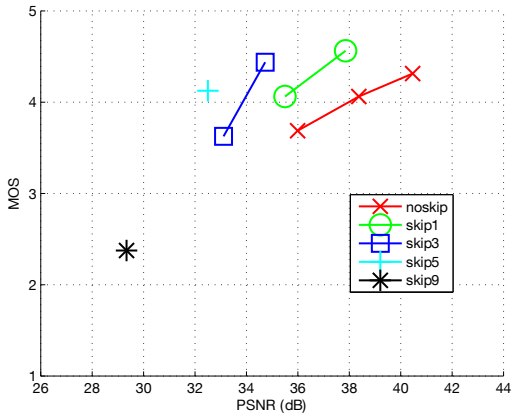
Figure 22 shows that the PSNR results on light-field slices for the Champagne sequence are consistent with the PSNR results obtained on the input views (see Figure 9). The range of PSNR values are different but relatively close, and the order of the configurations is very similar. The conversion step in our experiments conditions does not seem to have a large impact on the compression and synthesis artifacts, hence on the objective quality of the sequence.

6.7. Comments on motion parallax

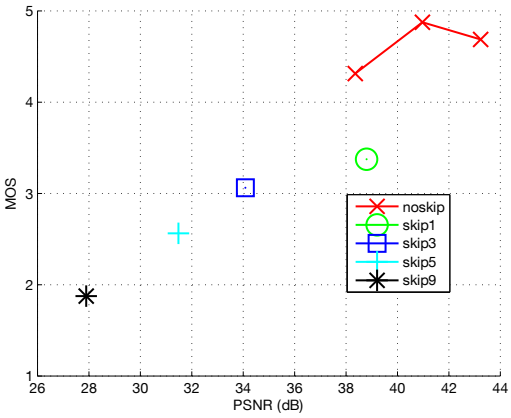
The effect of compression and synthesis on the motion parallax quality is discussed in this section. It should be noted that this is just based on a preliminary observation (based on one subject's comments). During one session, the subject watched the content while moving on a baseline of approximately 2 meters from left to right and right to left and commented the following aspect of the motion parallax. For compressed sequences which present many artifacts (i.e. with a low quality), a variation of the intensity of these artifacts (e.g. sizes of the blocks artifacts) has been observed when moving along the viewing angle of the display. For the sequences with only few artifacts (with impairments rated as *Imperceptible* or *Perceptible but not annoying*), the variations were not perceptible and did not disturb the perception of the motion parallax. As a first preliminary conclusion, it could be said that the perception of motion parallax is unsatisfying only when the image quality (in terms of compression artifacts and flickering synthesis artifacts) is already bad. More tests (by defining a scale rating the motion parallax from perfectly smooth to jerky for example) should be conducted to confirm these first hints.

7. Conclusion and future work

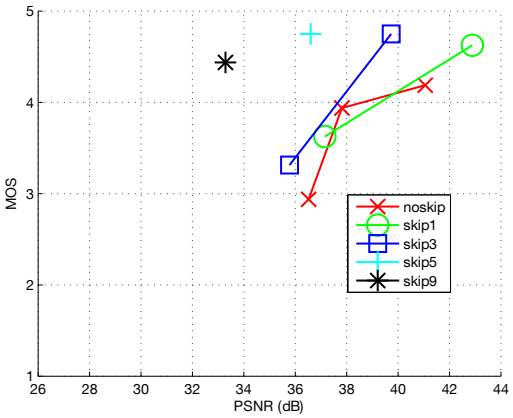
The study presented in this paper provides some initial conclusions on the feasibility of a video service that would require rendering about 80 views. We have observed that bitrates associated to impairments rated as not annoying are about 11 Mbps for the sequence Dog (with skip1 configuration), less than 10 Mbps for Champagne (with noskip configuration), and about 5 Mbps for Bunny (with skip5 or skip9 configurations). It is known that typical bitrates for encoding 4K content with HEVC are estimated to 15 Mbps and up to 80 Mbps for 8K content. We consequently conclude that bitrates required for rendering 80 views are realistic and coherent with future 4K/8K needs. In order to further improve the quality and avoid network overload, improved SMV video codec efficiency is mandatory. It



(a) Dog



(b) Champagne



(c) Bunny

Figure 21. MOS vs. PSNR

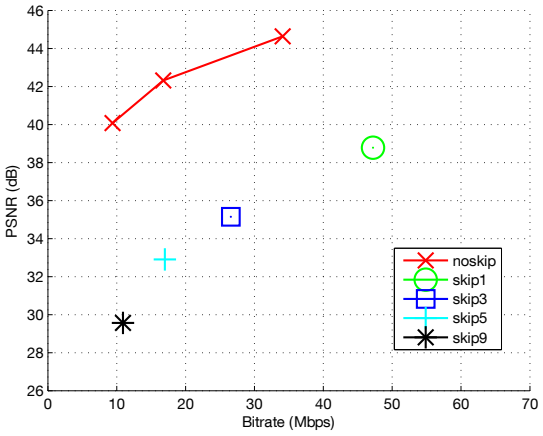


Figure 22. Bitrate variations with PSNR measured on light-field slices (Champagne sequence)

should also be noted that experiments results largely depend on the characteristics of the display (like spatial and angular resolutions, field of depth, etc.) and on the tested content (2 natural and 1 synthetic sequences) which we do consider as easy to encode contents because they contain still backgrounds, have small resolutions (1280x960 @30fps), and have a linear camera arrangement. This note does not change the feasibility conclusion, yet highlights the need for a better codec.

Preliminary experiments performed during this study lead to recommended coding configuration for SMV contents. In particular, IPP inter-view prediction structure with Groups of Views (GOVs) of size 16, with hierarchical temporal prediction structure with GOPs of size 8 is suggested. IPP inter-view prediction structure is more efficient than Central (with 5% BD-rate gains reported) and Hierarchical (3% BD-rate gains reported) structures. Results are similar when the coding scheme includes view synthesis. GOVs of size 16 bring about 3% coding improvement over size 9. GOVs enable compromise between memory limitations, coding efficiency and parallel processing.

Some conclusions are also drawn on the number of frames to skip at the encoder, and synthesize at the renderer after the decoding process. Several ratios for the number of coded and synthesized views are compared in our experiments. Subjective results suggests skipping 0, 1, 3 or 5 views for Dog, not skipping any view for Champagne, and skipping up to 5 or 9 views for Bunny. The amount of views to skip is highly sequence dependent, and varies from 0 to 9 (i.e. the minimum and maximum tested values). The ratio coded/synthesized depends on the quality of the synthesized views, hence is linked to the quality of the depth maps and the efficiency of the view synthesizer. It obviously also depends on the complexity of the scene that needs to be synthesized.

By synthesizing intermediate views from original uncompressed views, a 25dB to 44dB PSNR is achieved (against the original uncompressed views). Apart from compression, view synthesis introduces severe distortions. View synthesis weaknesses affect the coding scheme and are tightly linked to the estimated depth maps quality. Improvement of view synthesis and depth estimation algorithms is mandatory. The curves representing the correspondence between PSNR and MOS are monotone increasing functions, which shows that the PSNR is able to reflect increase or decrease in subjective quality (even in the presence of skipped views). However, depending on the ratio of coded and synthesized views, we have observed that the order of magnitude of the effective quality variation is biased by the PSNR. PSNR is less tolerant to view synthesis artifacts than human viewers.

Finally, preliminary observations have been initiated. First, the light-field conversion step does not seem to alter the objective results for compression. Secondly, the motion parallax does not seem to be impacted by specific compression artifacts. The perception of the motion parallax is only altered by variations of the typical compression artifacts along the viewing angle, in cases where the subjective image quality is already low (i.e. cases with severe artifacts).

As mentioned above, the experiments depends on our test conditions and particularly on the tested content. As a consequence, future works should extend the evaluation towards additional content with different depth characteristics and encoding complexities. It should be noted that producing Super-Multi-View content is not a trivial task and is one

of the main issue for the research community working on this technology. For example, one of the main tasks of the FTV ad-hoc group in MPEG is to gather content [30].
The study should also be extended to content captured with different camera arrangements, like arc arrangement which is generally considered more appropriate for light field display and cannot be handled properly by current view synthesis and depth estimation algorithms and which provide less efficient coding performance with current multi-view standard encoder [19].
Further experiments could complete current results. Subjective evaluations with a denser range of bitrate values could allow refining the boundaries between the ranges of bitrate values associated with each quality level. Similarly, a lower range could allow determining the lowest bitrate value possible for an acceptable quality.
Using these denser ranges and limit values could allow finding a proper way to evaluate objectively the quality of compressed and synthesized SMV content by weighting efficiently the PSNR for synthesized views or by using a more convenient metric. This could allow associating ranges of subjective qualities with ranges of objective values.
The impact of the compression and synthesis artifacts on the perception of motion parallax should be further studied, as well as other specific aspects of light field content such as the perception of depth or the angle of view for example.

Acknowledgment

This work has been carried out in the context of a Short Term Scientific Mission (STSM) granted by the COST Action 3D-ConTourNet [31]. The authors want to thank the STSM coordinator, the Action Chair and the Management Committee of COST Action 3D-ConTourNet.
The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 32449, and from the DIVA Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 290227.
Dog, Pantomime, and Champagne Tower (Nagoya University's sequences) are provided by Fujii Laboratory at Nagoya University [15]. T-Rex sequence is provided by Holografika [5]. Big Buck Bunny is copyright Blender Foundation [32], and 3D camera setup and rendering is done by Holografika [16].

References

[1] F. Dufaux, B. Pesquet-Popescu, M. Cagnazzo, Emerging technologies for 3D video: content creation, coding, transmission and rendering, Wiley Eds, 2013.
[2] M. P. Tehrani, T. Senoh, M. Okui, K. Yamamoto, N. Inoue, T. Fujii, [m31103][FTV AHG] Introduction of super multiview video systems for requirement discussion, in: International Organisation For Standardisation, ISO/IEC JTC1/SC29/WG11, October 2013.
[3] M. Tanimoto, Overview of free viewpoint television, Signal Processing: Image Communication 21 (6) (2006) 454–461.
[4] M. P. Tehrani, T. Senoh, M. Okui, K. Yamamoto, N. Inoue, T. Fujii, [m31261][FTV AHG] Multiple aspects, in: International Organisation For Standardisation, ISO/IEC JTC1/SC29/WG11, October 2013.
[5] http://www.holografika.com/.
[6] D. Marpe, T. Wiegand, G. J. Sullivan, The H. 264/MPEG4 advanced video coding standard and its applications, Communications Magazine, IEEE 44 (8) (2006) 134–143.
[7] G. J. Sullivan, J. Ohm, W.-J. Han, T. Wiegand, Overview of the high efficiency video coding (HEVC) standard, Circuits and Systems for Video Technology, IEEE Transactions on 22 (12) (2012) 1649–1668.
[8] J.-R. Ohm, Overview of 3D video coding standardization, in: International Conference on 3D Systems and Applications, Osaka, Japan, June 2013.
[9] P. Merkle, A. Smolic, K. Muller, T. Wiegand, Multi-view video plus depth representation and coding, in: International Conference on Image Processing (ICIP), Vol. 1, IEEE, 2007, pp. 201–204.
[10] C. Fehn, Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv, in: Electronic Imaging 2004, International Society for Optics and Photonics, 2004, pp. 93–104.
[11] E. Bosc, P. Le Callet, L. Morin, M. Pressigout, Visual quality assessment of synthesized views in the context of 3d-tv, in: 3D-TV System with Depth-Image-Based Rendering, Springer, 2013, pp. 439–473.
[12] D. Howard, M. Green, R. Palaniappan, N. Jayant, Visibility of digital video artifacts in stereoscopic 3d and comparison to 2d, in: SMPTE Conferences, Vol. 2010, Society of Motion Picture and Television Engineers, 2010, pp. 1–15.
[13] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, L. Morin, Towards a new quality metric for 3-d synthesized view assessment, Selected Topics in Signal Processing, IEEE Journal of 5 (7) (2011) 1332–1343.
[14] P. T. Kovács, Z. Nagy, A. Barsi, V. K. Adhikarla, R. Bregovic, Overview of the applicability of H. 264/MVC for real-time light-field applications, in: 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), IEEE, 2014, pp. 1–4.
[15] http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/.

[16] P. T. Kovacs, A. Fekete, K. Lackner, V. K. Adhikarla, A. Zare, T. Balogh, [FTV AHG] Big Buck Bunny light-field test sequences, in: International Organisation For Standardisation, ISO/IEC, JTC1/SC29/WG11 M35721, Geneva, Switzerland, February 2015.
[17] http://www.blender.org/.
[18] http://www.autodesk.fr/products/3ds-max/overview.
[19] K. Wegner, O. Stankiewicz, K. Klimaszewski, M. DomaÅski, FTV EE3: Compression of FTV video with circular camera arrangement, in: International Organisation For Standardisation, ISO/IEC JTC1/SC29/WG11 MPEG2014/M33243, Valencia, Spain, April 2014.
[20] O. Stankiewicz, K. Wegner, M. Tanimoto, , M. Domanski, Enhanced Depth Estimation Reference Software (DERS), in: International Organisation For Standardisation, ISO/IEC JTC1/SC29/WG11 MPEG2013/M31518, Geneva, Switzerland, October 2013.
[21] M.Tanimoto, T. Fujii, K. Suzuki, View Synthesis Algorithm in View Synthesis Reference Software 2.0 (VSRS 2.0), in: International Organisation For Standardisation, ISO/IEC, JTC1/SC29/WG11 M16923, Lausanne, Switzerland, February 2008.
[22] G. Tech, K.Wegner, Y. Chen, M. Hannuksela, J. Boyce, MV-HEVC draft text 7, in: International Organisation For Standardisation, ISO/IEC JTC 1/SC 29/WG 11 M32661, San Diego, CA, USA, January 2014.
[23] G. Bjøntegaard, Calculation of average PSNR differences between RD-curves, in: VCEG Meeting, Austin, USA, April 2001.
[24] P. Hanhart, T. Ebrahimi, Calculation of average coding efficiency based on subjective quality scores, Journal of Visual Communication and Image Representation 25 (3) (2014) 555–564.
[25] Methodology for the subjective assessment of the quality of television pictures, Recommendation ITU-R BT.500-13, January 2012.
[26] W. Chen, J. Fournier, M. Barkowsky, P. Le Callet, New requirements of subjective video quality assessment methodologies for 3dtv, in: Video Processing and Quality Metrics (VPQM), 2010.
[27] W. A. IJsselsteijn, H. de Ridder, J. Vliegen, Subjective evaluation of stereoscopic images: effects of camera parameters and display duration, Transactions on Circuits and Systems for Video Technology 10 (2) (2000) 225–233.
[28] F. D. Simone, Selected contributions on multimedia quality evaluation, PhD thesis.
[29] M. G. Kendall, Rank correlation methods, London : Griffin, 1970.
[30] M. Tanimoto, K. Wegner, G. Lafruit, [m34604][AHG Report] AHG on FTV (Free-viewpoint Television), in: International Organisation For Standardisation, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, February 2015.
[31] http://www.3d-contournet.eu/stsms/.
[32] (c) copyright 2008, blender foundation / http://www.bigbuckbunny.org/.