Performance Study of View Synthesis with Small Baseline for Free Navigation

P. Nikitin[1 2] J. Jung[1] M. Cagnazzo[2] B. Pesquet[2]

[1] Orange Labs ,[2] LTCI, Télécom ParisTech, Université Paris-Saclay

{pavel.nikitin, joel.jung}@orange.com, {marco.cognazzo, beatrice.pesquet}@telecom-paristech.fr

Abstract

In a typical Free Navigation service, view synthesis is expected to provide virtual views between the real captured views, in order to improve the smoothness of the navigation. Practical constraints prevent from capturing views with a very small baseline, so view synthesis is required. One way of synthesizing views is to use texture and depth information. It is of interest to understand how much current view synthesis technology is able to provide acceptable quality for synthesized views, in the framework of Free Navigation.

A new super multi-view content has been recently provided by the University of Brussels. This high-density content has the characteristic to have a very small baseline of 1mm and is particularly adapted for this study.

In this study, some experiments of view synthesis with small baseline were performed. Experimental results are reported to understand how far view synthesis can be used, both from an objective and from a subjective point of view. It was shown that according to subjective point of view, more views can be synthesized while maintaining acceptable quality.

Key words

Free Navigation, View synthesis, Super multiview

1 Introduction

In a future video services user should have a possibility to freely navigate within the scene. In order to achieve this a huge number of views should be captured, which is not possible, because of the physical constraints. In addition to physical limitations, it is currently not possible to handle thousands of cameras and distribute such content. As a consequence, view synthesis is required.

In the most ambitious scenario, a user will be able to move freely and stop on any view, no matter if it is a synthesized view or a captured view. In a shorter term scenario, we might consider that a user is able to move freely, but to stop only on captured views. In this case, view synthesis can be seen as a means to make the navigation smoother.

One way of synthesizing views is to use texture and depth information. For instance, the MPEG reference software VSRS 4.1 [1] is well known for its ability to synthesize views out of a pair of views and corresponding depths. Similarly, during the 3D-HEVC standardisation process, VSRS1d-fast [2] has been extensively used.

Recent FTV Call for Evidence [3] has been too optimistic, trying to synthesize too many views in between two captured views (seen differently, considering a distance between two cameras too huge), so that even the anchor had unacceptable visual quality. Today, it is of interest to understand how much current view synthesis technology is able to provide acceptable quality for synthesized views.

A new super multi-view content has been recently provided by the University of Brussels [4]. This content has the characteristic to have a very small baseline of 1mm. Although the content is static (single frame), it is a first step that allows to test very different configurations of view synthesis, while always having a reference that has been captured, to compare with.

In this study, we perform some tests of view synthesis with small baseline. Experimental results are reported to understand how far view synthesis can be used. It is very important to understand that any result that claims a given distance between two cameras is tight to the sequence itself, because the distance between the objects of the scene and the cameras also needs to be taken into account. Initial conclusions are mostly derived from visual quality inspection by a group of four experts. Section 2 briefly describes the new test set provided by the ULB. Section 3 gives some preliminary results on how many views can be synthesized from an objective point of view, while Section 4 gives some preliminary clues from a subjective point of view.

2 Description of the content

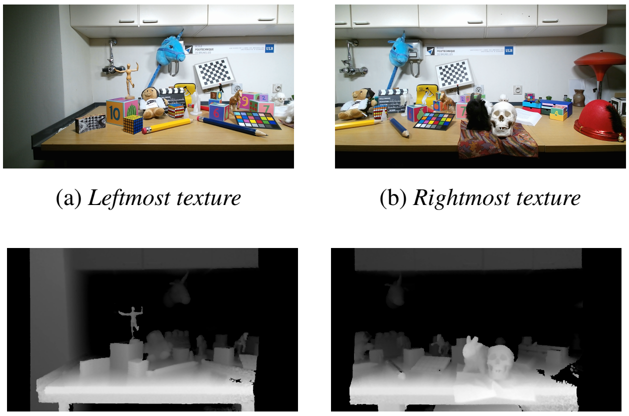
The ULB test set [4] is a high density LightField content captured with a 2D rail robotic system. The resulting scene is static and was captured using Kinect2 RGB sensor (1920x1080@24bits) and a Kinect2 depth sensor (512x424@16bits).

2.1 Scene description

The whole scene is 2.3m wide and composed by conventional objects, semi-transparent ones and objects with fur. There is also a rotated checkerboard and color chart. The closest object of scene is 0.6m and the farthest is 1.6m from the sensor. The platform with camera is moved millimeter by millimeter, which provides a very high density of views. The 1mm precision for positioning the camera vertically and horizontally was obtained by several motors and a rail system composed of ball bearings.

2.2 Pre-processing of the content

The Kinect contains two distinct cameras for the color and depth images, consequently the extrinsic and intrinsic parameters for these images are different. Because of this, the depth images had to be reprojected onto the coordinate system of the color images. The resulting depth map needs to indicate for each pixel of the color image, an orthogonal distance to the color image's camera plane. The resulting most left and right views are shown in Figure 1 with their corresponding depth maps.



(a) *Leftmost texture* (b) *Rightmost texture*

(c) *Leftmost depth* (d) *Rightmost depth*

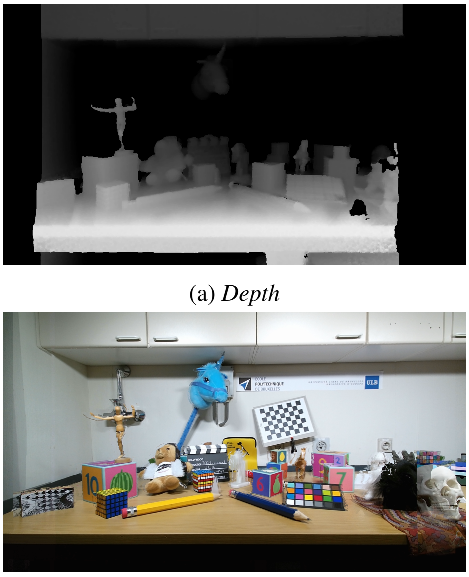
Figure 1: *Leftmost and rightmost views of the ULB test set, with corresponding depth maps.*

Due to the fact that RGB and depth sensors have different field of view, the depth information is missing on the left and right sides of the views, as shown in figure 1. When depths are used for synthesis, the resulting synthesized view exhibits severe artifacts on the borders too, as shown on figure 2. These artifacts prevent from drawing conclusions from the tests:

1.Objective results obtained from synthesis are largely biased by these left and right areas.

2.Subjective tests are biased (viewers assess the whole view).

3.Some bits are used to encode these useless areas of the depths, once compression is involved.



(a) *Depth*

(b) *Synthesized view*

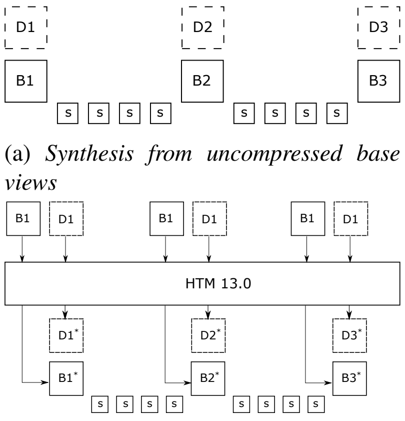
Figure 2: *Uncropped depth and resulting synthesized view.*

We consequently have cropped the original data. This cropping is a simple manual post-processing of the depth. The number of removed columns compared to the original is 192 on the left side, 320 on the right side, yielding to a resolution of 1408x1080. The texture input has been cropped similarly. In preliminary tests it was shown that quality of synthesis is better from cropped content, so only cropped version will be used for objective results and subjective quality evaluation.

The main advantage of this new content provided by the ULB is that it gives the ability to synthesize large number of views between two views, while still having a reference anchor for objective tests, even if the distance between the two cameras is small. So far, this could only be obtained with computer generated content.

3 View synthesis – Objective results

For view synthesis, we have used VSRS1D-fast, which has been extensively used during the standardization process of 3D-HEVC. It was shown that VSRS1D-fast can provide better results compared to VSRS4.1 for 1D linear content. Comparison of these two synthesis tools is shown on the figure 4.



(a) *Synthesis from uncompressed base views*

(b) *Compression of base views before synthesis.*

Figure 3: *Overall scheme of study*

In the following experiments, intermediate views are synthesized between two captured (reference) views. The scheme of the study is shown on the figure 3. We consider two different scenarios. In the first one synthesis out of original uncompressed data is done, figure 3a. In the second case all the base views are compressed using HTM13.0 reference software, and the synthesis is done from decoded views. The synthesis out of compresses views is shown on the figure 3b, where views with 'star' denotes decoded texture or depth.

To present results in this paper the following notation is used: Synth-n,b means that n views are synthesized between two captured views distant from b millimeters.

3.1 Synthesis out of uncompressed views

The first part of study is related to the synthesis of intermediate views from original uncompressed data, as it is shown on the figure 3(a).

Experiment 1. The distance b between two reference views is progressively increased, and the corresponding PSNR loss is provided. Figure 4 shows the results for all configurations from Synth-b,b+1 for b=1...249. Among the 851 views available, only 751 views were used for PSNR calculation (view 1 to view 751). The figure 4 can be read this way: if for Synth-b,b+1 the PSNR is p dB, it means that when b views are synthesized between two cameras distant by b+1 millimeters, and the average PSNR of the n synthesized views is p dB. If we consider a Free Navigation application where the user can only watch real captured view, and move from one to the other, and synthesis is used to smooth the navigation effect, we believe this representation, computing the average of views, as represented in figure 4 is representative.

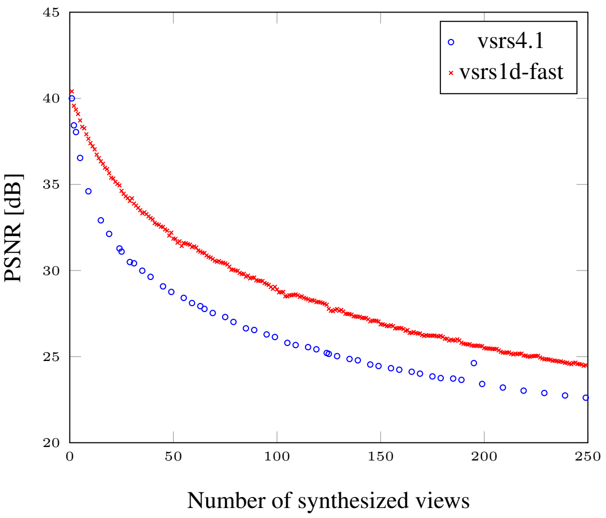


Figure 4: *Average PSNR for original data comparing VSRS4.1 and VSRS1D-fast.*

We can observe that the average PSNR drops from 40.4dB for Synth-1,2 configuration to 24.5dB for Synth-249,250. An average 35dB quality is achieved when synthesizing about 23 consecutive views, which correspond to 2.4cm distance between two cameras. An average 30 dB quality is achieved when synthesizing about 78 consecutive views, which corresponds to 7.9cm distance between two cameras.

Experiment 2. Between two reference views, the PSNR of the synthesized views is not constant: it depends on the distance with the closest reference view. This is depicted in figure 5 that show the PSNR variation for different Synth-b,b+1 configurations. We can observe that for all configurations, the PSNR decreases when the synthesized view is far from one of the two reference views. The worst quality is close to the middle point between the two base views. The PSNR is inconsistent from one view to another. This can be seen easily on configuration Synth-1,2 but the issue holds for other configurations. Some views are particularly degraded as for instance view 550. One of the possible sources of such errors can be erroneous camera parameters. In order to explain the issue, figure 6 represents the square error difference image between original and synthesized images for view 550 and for view 580. Blue color on this heat map represents the minimum error, and red the largest. While the difference for view 580 is small, all the contours are visible in the difference image of view 550, which tends to confirm that camera parameters for view 550 is erroneous or there is misalignment between depth and texture.

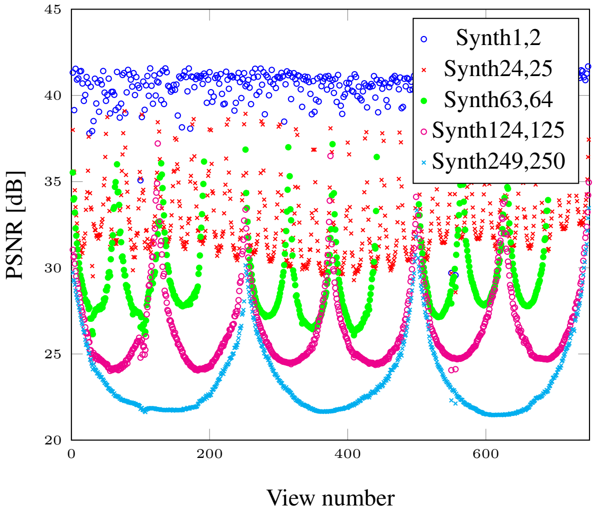
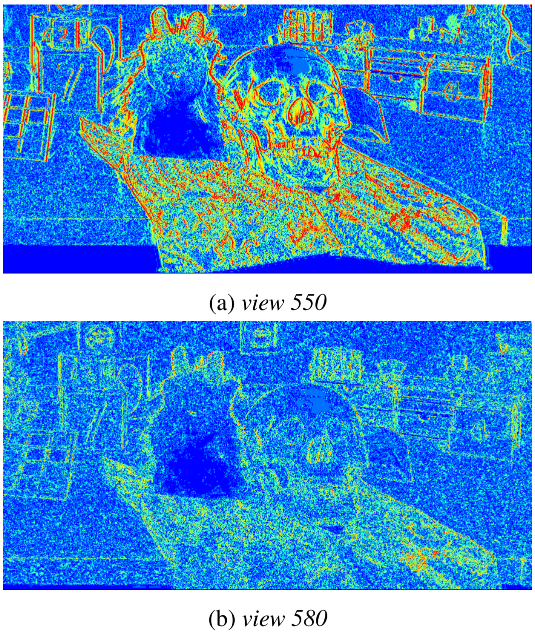


Figure 5: *PSNR evolution for different Synth-n,b configurations.*



(a) *view 550*

(b) *view 580*

Figure 6: *Square error heat map for Synth-1,2.*

Experiment 3. figure 7 represents the evolution of the quality for Synth-1,b configuration, with b varying from 2 to 848. This means that one single view is synthesized for two reference cameras distant by b millimeters. Said differently, we represent the PSNR of the worst point of figure 5. If we consider a Free Navigation application where the user can stop on any view (synthesized or captured), we believe the min value as reported in figure 7 is representative: it represents the worst case.

In this case, we observe that a quality of 35dB is achieved by skipping 15 views, this leads to a possible distance of 16mm, with the ULB content, while a quality of 30dB leads to a distance of 51mm.

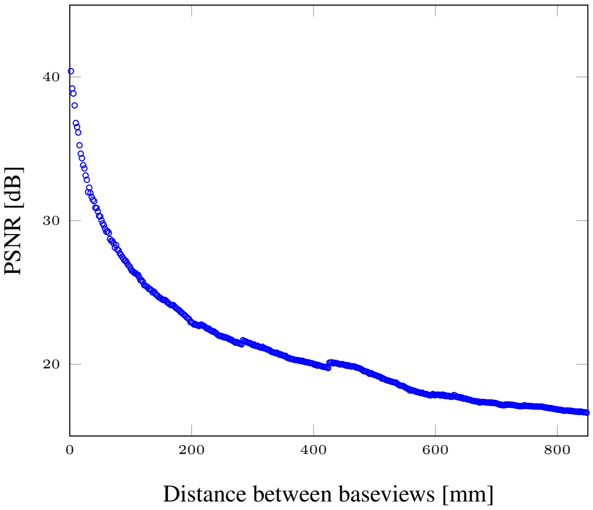


Figure 7: *Min PSNR for configuration from Synth-1,2 to Synth-1,848*

It has been demonstrated many times that the PSNR is very sensitive to view synthesis artifacts [5], such as shifts, while human quality evaluation makes abstraction of this parameter. So most of the time it reflects lower quality than the perceived one. In the next section, subjective quality is assessed.

3.2 Synthesis out of compressed views

For this study base views were compressed using HTM13.0 reference software, simulating the scenario where base views are available on decoder side and intermediate views can be synthesized to improve the smoothness of transition between base views, which allows the user to have better immersive experience.

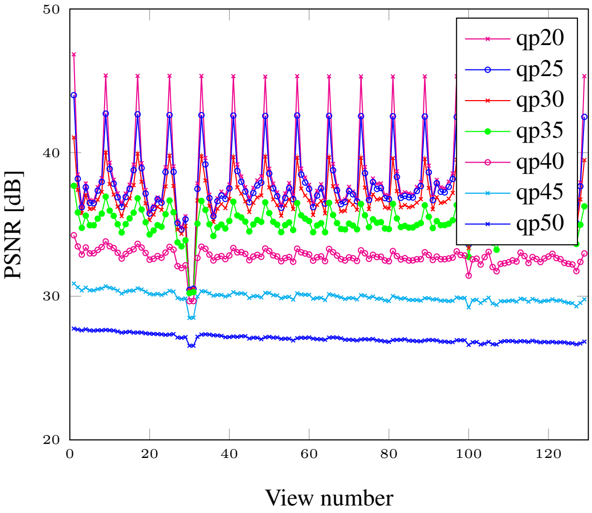


Figure 8: *Synth-7,8 with compressed baseviews*

On the figure 8 it is shown that for the synthesized views, the average PSNR, as well as the min value of the PSNR, are decreasing when the QP increases. The average PSNR of synthesized views for QPs equal to 20 and 50 respectively is 38.2dB and 26.5dB. This difference is huge, meaning that the synthesis is significantly impacted by the compression.

Increasing the distance between base views mainly influences on the synthesized views as it is shown on the figure 9. The coding does not suffer from increasing the distance to the inter-view prediction reference picture, as the content is very dense. The PSNR values of synthesized views become closer to each other. On the figure 10 all synthesized views for different QPs have practically the same quality in terms of PSNR. So if the synthesis is erroneous due to the large distance between base views the quality of base views does not play an important role.

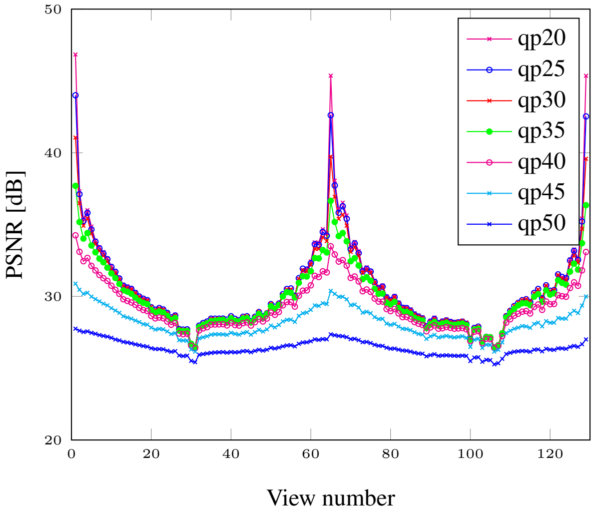


Figure 9: *Synth-63,64 with compressed baseviews*

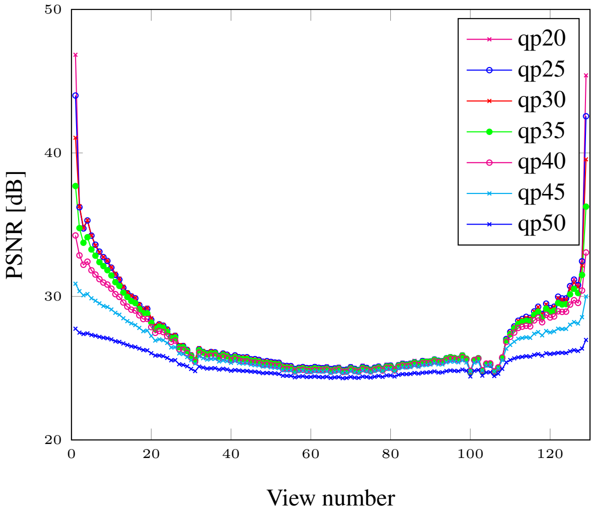


Figure 10: *Synth-127,128 with compressed baseviews*

4 View synthesis – Subjective results

According to the objective results, an acceptable quality is obtained only for a very small baseline. However, it has been shown that PSNR is more sensitive to artifacts from synthesis than the human eye is [5]. The goal of this section is to verify from a subjective point of view if it is possible to synthesize views with more distant cameras.

We have not performed subjective tests, as there is no standardized procedure to perform tests for this kind of scenario. The reminder should more be seen as a visual quality inspection, where four different expert viewers have shared their opinions.

4.1 Subjective results - Uncompressed data

For this quality inspection, viewers have basically expressed their opinion with the following words, to rate the quality of the transition between the two reference views separated by d millimeters:

1: Artifacts are not noticeable.

2: Artifacts are noticeable but not annoying.

3: Artifacts are noticeable and annoying.

4: Artifacts are too annoying (I would prefer to switch directly from one view to another without smoothness/synthesis.)

To simulate a scenario, when the user swaps the views on the tablet or smart phone, we have built the following sequence:

1.The viewer watches the same view (reference view) for 1.5 seconds

2.The viewer moves to the right, watches synthesized and reference views, and stops on a reference view located d millimeters away from the initial one.

3.The viewer watches this reference view for 1.5 seconds.

For this experiment the views in the range between 251 and 751 are used. The sequence is generated at 60fps, because in the preliminary tests this configuration provided for most of the Synth-n,b configurations better impression of the smooth transition between base views.

Several Synth-d,d+1 configurations have been presented to the viewers, starting from Synth-249,250. As reported in Table 1, viewers have generally agreed that the quality is unacceptable. As a consequence, d has been decreased progressively, etc.

From this table, we can conclude that:

- for d < 31, barely no artifact is observed.
- for 31< d < 124, some artifacts are noticeable but qualified as non-annoying (except for one view-er).
- for 124 < d < 249, artifacts are noticeable and qualified as annoying.
- for d > 249, the level of artifacts is too annoying, and synthesis it not accepted as a feature to make the transition smoother.

Configuration	PSNR [dB]	V1	V2	V3	V4	Avg
Synth-249,250	24.5	4	4	4	4	4.00
Synth-124,125	28.0	2	2	3	2	2.25
Synth-99,100	29.1	2	2	3	2	2.25
Synth-63,64	31.1	1	2	3	1	1.75
Synth-31,32	33.9	1	1	1	1	1.00
Orig, no synth	-	1	1	1	1	1.00

Table 1: *Average PSNR of each synthesized view for different Synth-d,d+1 configuration and the corresponding subjective scores. Tests performed on the cropped data set.*

Synth-n,n+1	QP35-QP25		QP50-QP35	
	psnr, dB	subjective	psnr, dB	subjective
31	-1.4	-0.3	-6.2	2.5
63	-0.8	0.5	-4.4	2.8
127	-0.4	0.3	-2.4	3
249	-0.2	0.4	-1.0	1.9

Table 2: *Comparison of deferences for different QPs for subjective scores and PSNR .*

4.2 Subjective results - Compressed data

For the synthesis out of compressed base views the following view inspection was done: the process of sweeping between two compressed base views is emulated.

The way, how the sequence is generated, is similar to the case with uncompressed data. But for this inspection viewers watch two sequences one after another. First is synthesis out of uncompressed data and second is synthesis from compressed base views. The viewers give theirs opinion on quality of the second sequence compared to the first using the continuous scale from -3 to +3, where +3 means second sequence is much better than the first one. Sequences appear randomly, so the viewer does not know which one is synthesized from the compressed data.

As it is shown in the Table 2, the difference between QP35 and QP25 subjectively was not visible by most of the expert viewers. For some experiments, for example, for Synth249,250 and QP 35 participants considered better quality for the configuration with compression rather than synthesis from uncompressed data.

5 Conclusion and future work

Several view synthesis experiments have been performed, using the new high density ULB content. From these experiments, some conclusion and suggestions can be made:

According to the PSNR, the acceptable distance between two cameras is very low.

The result of the visual quality inspection, performed by four expert viewers, confirms that from a subjective point of view, more views can be synthesized while maintaining acceptable quality. It could be suggested that objective (PSNR based) results can only be considered to rank two algorithms applied to the same configuration. PSNR gives an idea of the ranking (comparison), it evolves correctly (higher PSNR usually fits with higher subjective quality). But it does not reflect the overall quality. Even when considering subjective quality, the distance between two cameras that yields to acceptable quality for virtual views remains low.

The maximal distances, allowing acceptable quality for synthesized views, found by those experiments are quite small. However, the quality of synthesis depends not only on baseline between two cameras, but also the distance to the closest and farthest objects in the scene. With the ULB test set, the objects are very close to the scene, and this drastically reduces the allowed distance between two cameras.

In experiments with compression it was shown that the gap between compressed base views and synthesized intermediate views becomes smaller when the quality of base views becomes lower(QPs 45-50). So the transition between base views becomes smoother. Subjective inspection results showed a correlation with objective tests.

For the future perspectives improving the quality of the depths will drastically increase view synthesis quality and there is a room for further improvements of view synthesis algorithms in the near future. It is also important to mention, that depending on the application, constraints related to view synthesis will be different. The method described in [6] can be used in the future for comparing view synthesis algorithms.

References

[1] Wegner, Stankiewicz, Tanimoto, et Domanski. Enhanced view synthesis reference software (VSRS) for free-viewpoint television. Dans *ISO/IEC JTC1/SC29/WG11*. m31520, October 2013.

[2] Zhang, Tech, Wegner, et Yea. 3D-HEVC test model 5. Dans *ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*. JCT3V-E1005, July 2013.

[3] Lafruit, Wegner, et Tanimoto. Draft call for evidence on FTV. Dans *ISO/IEC JTC1/SC29/WG11*. m40293, February 2015.

[4] Bonatto, Lenertz, Li, Schenkel, et Lafruit. [MPEG-I-visual/apps] ULB high density 2D camera array data set, version 1. Dans *ISO/IEC JTC1/SC29/WG11*. m40293, April 2017.

[5] Dricot, Jung, Cagnazzo, Pesquet, Dufaux, Kovacs, et Kiran Adhikarla. Subjective evaluation of super multi-view compressed content on high end light field 3D display. *Elsevier Signal Processing: Image Communication*, 39:369–385, November 2015.

[6] Purica, Valenzise, Cagnazzo, Pesquet-Popescu, et Dufaux. Using region-of-interest for quality evaluation of DIBR-based view synthesis methods. September 2016.