Projet PERSEE
« Schémas Perceptuels et Codage vidéo 2D et 3D »
n° ANR-09-BLAN-0170

Livrable **D5.2** *30/03/2011*

2D/3D Codec architecture

Patrick	LE CALLET	IRCCyN
Vincent	RICORDEL	IRCCyN
Junle	WANG	IRCCyN
Josselin	GAUTIER	IRISA
Christine	GUILLEMOT	IRISA
Laurent	GUILLO	IRISA
Olivier	LE MEUR	IRISA
Emilie	BOSC	INSA
Luce	MORIN	INSA
Marco	CAGNAZZO	LTCI
Béatrice	PESQUET-POPESCU	LTCI

Contents

1 Introduction 3

2 Data format for multiview and 2D/3D compatibility 5
2.1 Multiple views plus depth 5
2.2 Layered depth image (LDI) representation 6

3 3D-specific conding tools 7
3.1 Disparity field estimation 7
3.2 Depth-map coding techniques 9
3.2.1 Bit-rate allocation between depth and texture in the context of MVD data compression 10
3.2.2 Reliability of an objective metric for the evaluation of views synthesized from decompressed depth maps 12
3.2.3 Conclusion 13
3.3 RD segmentation of dense disparity fields 13
3.3.1 Proposed scheme for MVD coding 14
3.3.2 Experimental results 15
3.4 LDI coding 16
3.5 Inpaiting-based virtual view synthesis 17
3.5.1 Problem introduction 17
3.5.2 Algorithm 18
3.5.3 Implementation 20
3.5.4 Results 20
3.5.5 Conclusion and perspectives 21
3.6 A way to increase the Quality of Experience of 3DTV 21

4 Performance evaluation system 23
4.1 Purpose of Mpeg-3DV CFP 23
4.2 Test Material, Coding Classes and Anchors 24
4.2.1 Test Material 24
4.2.2 Data format 25
4.2.3 Anchors 25
4.3 Test conditions 25
4.3.1 Input data 25
4.3.2 Test scenarios and coding conditions 25
4.3.3 Rendering Conditions (View Synthesis) 26
4.4 Subjective tests 28
4.5 Perceptual test methodology 29
4.6 Quality of experience for an autostereoscopic display 31

References 34

1 Introduction

In this report we describe contributions for the 2D/3D codec architecture developed for the task 5 within the PERSEE project.

The target of the task 5 is to provide all partners with a new video codec integrating their contributions. These contributions are core technologies identified as relevant in the other tasks (i.e tasks 1 to 4). These contributions will be integrated in a common software platform. This platform is described at the figure 1. This software has been initiated with a state-of-the-art 2D codec and its architecture will evolved in order to take into account contributions. The most promising resulting video codecs will be used to encode/decode selected videos that will be used for perceptual and subjective assessments during the task 6.

The programme of this task provided a preliminary description of the common test conditions (i.e the corpus of selected 2D and 3D videos, their description and their content classification, and the coding conditions) which has been performed during the first 6 months.

In this new deliverable, a description of relevant contributions for the codec architecture is done. The document is organized as follows:

- Section 2 is about the retained formats for the 3D videos:
  - Section 2.1 describes the multiple-views-plus-depth format;
  - Section 2.2 describes the Layered Depth Image format.
- Section 3 presents the specific contributions for the 3D codec. In particular, we have:
  - Section 3.1 presents contributions dense disparity field estimation;
  - Sections 3.2 and 3.3 present contributions for the coding of the depth maps;
  - Section 3.5 presents contributions for the synthesis of the virtual views;
  - Section 3.6 presents a way to increase the QoE of 3DTV.
- Finally, Section 4 presents the tests protocol:
  - Sections 4.1 presents the purpose of the MPEG normalization context;
  - Section 4.2 describes the video sequences to be used to perform the tests;
  - Section 4.3 deals with the test conditions;
  - Section 4.4 depicts the subjective tets;
  - Section 4.5 presents the performance tests of synthesis view systems;
  - Section 4.6 deals with the problem of the quality of experience for an autostereoscopic display.

The intended codec structure is shown in Fig. 1. The input to this encoder is a multiview video, made up by N views, acquired by a camera array. A further input can be the ensemble of depth maps associated to these view. This is represented as

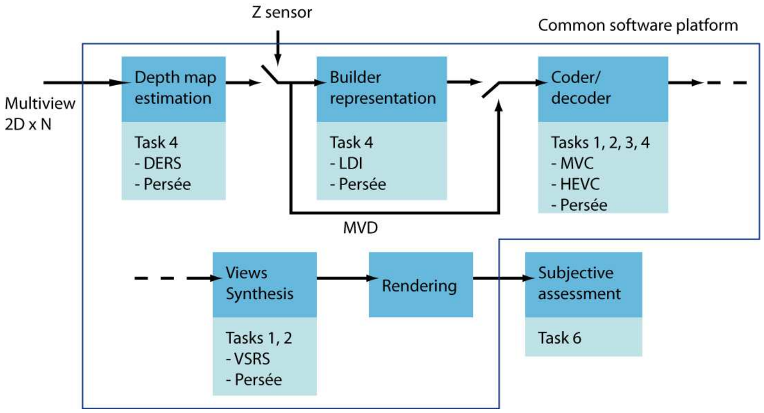


Figure 1: Common software platform of the project.

“Z sensor” input in the figure, since the depth information is often indicated as “Z data”. As an alternative, the depth information can be computed from the multiple views: this problem is one of the subjects of task 4. The proposed solutions for this problem are described in Section 3.1. Once we have the views and the depths, we have to decide which data representation to use. Two main candidates have been retained: multiple views plus depth (MVD) described in section 2.1 and layered depth image (LDI), described in section 2.2. In this second case, a further joint processing of views and depths is needed, and it is called “builder representation”’ in Fig. 1.

Then we have the actual codec block. The problems related to this block are addressed in taks 1 to 4 of the project. Standard based solutions such as MVC and Simulcast (e.g. with HEVC) can be considered as a reference (cf. Fig.2.1). We aim to propose a backward-compatible encoder. This means that it should be easy to extract from the encoded stream a substream describing a single view and hence easily decodable by a standard 2D decoder. This can be obtained by encoding one view with a standard (or nearly standard) method, and then encoding the other information (other views, depth) with novel methods. The proposed methods here should take into account the visual quality of the encoded video.

At the decoder side, we have the problem of view synthesis. This means that we want create an artificial point of view, which did not exist in the orginal data set. This method is necessary to implement such services as free viewpoint TV. Methods for view synthesis include the standard VSRS software and the solutions proposed in this project (task 4). While the rendering part is not a subject of research in this project, it is necessary to proceed to the last part of the scheme, that is the subjective quality assessment of the encoded video, which is the target of task 6 of the project.

2 Data format for multiview and 2D/3D compatibility

2.1 Multiple views plus depth

One of the most popular representations of multiview video and arguably the most adapted to free-viewpoint television [29] is the so-called multiple-views-plus-depth (MVD) format [21, 24]. When MVD is used, for each view we dispose of a texture video sequence and of a depth map sequence, representing, for each temporal instant, the distance of the current pixel from the point of observation. An example of MVD video is shown in Fig. 2.

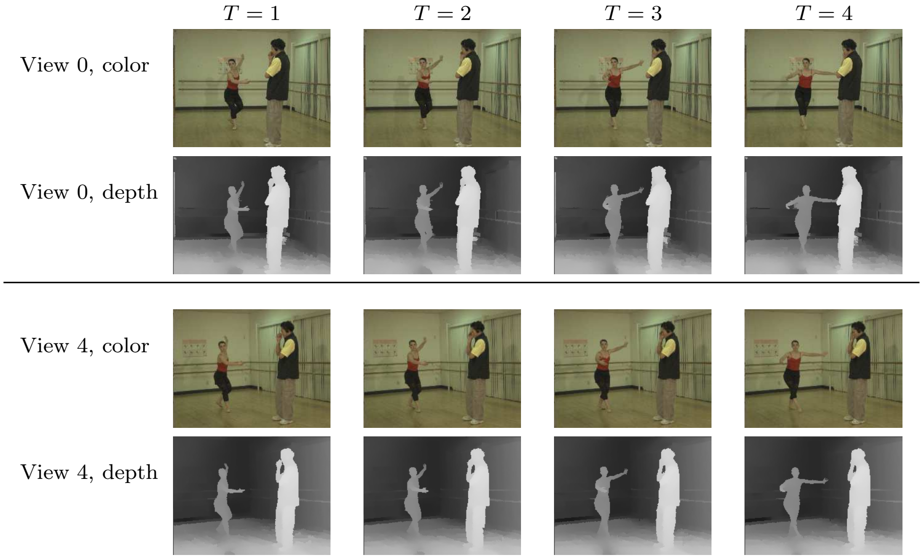


Figure 2: Example of multiple-view-plus-depth video.

MVD is extremely demanding in terms of storage space and transmission bandwidth, therefore compression is mandatory in order to manage this representation. Several approaches exist for MVD compression [34]. A simple, first one, is to independently compress each texture and depth sequence from each view. This approach is commonly referred to as Simulcast, see Fig. 2.1(a). Simulcast has the advantage of being simply implementable, backward compatible, and of allowing to decode immediately a single view for 2D screens. It has been chosen as reference in the Call for Proposal issued by the MPEG committee for the standardization of MVD [2]. Of course, one expects that more sophisticated schemes, taking into account the redundancy between views and between texture and depth, would achieve far better compression performance than the multicast scheme (this is actually the rationale behind the CfP). For example, as shown in Fig. 2.1(b), one can apply H.264/MVC [11] over texture sequences and (separately) over depth maps. Since depth and texture have very different content, no coding gain is expected by jointly coding texture and depth with H.264/MVC. Nevertheless, some

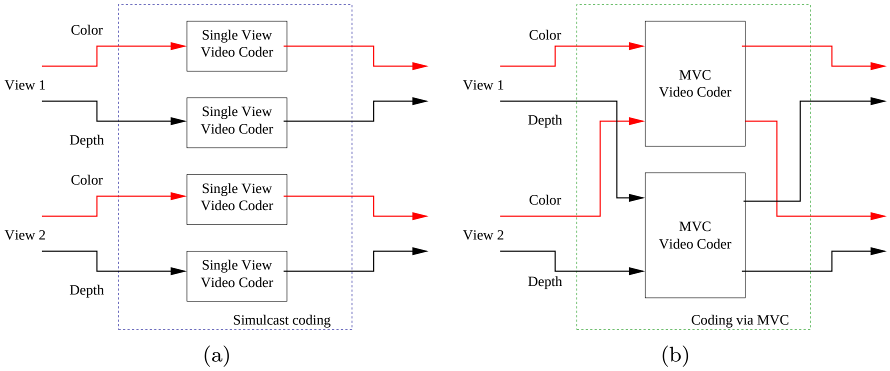


Figure 3: Reference MVD coding methods: (a) Simulcast; (b) MVC of texture and depths.

redundancy between texture and depth does exist, and this scheme does not exploit it. For example they partially share movement and disparity information, and above all, rate allocation between them should be jointly performed. However the latter is a quite difficult issue, and one of the key problems to be solved in order to achieve efficient coding [17].

The schemes shown in Fig. 2.1 allow a certain backward compatibility with respect to the 2D video. For example, in the Simulcast case, a user interested to the 2D video can acces only one texture stream and decode it with a single view video decoder such as AVC. In the case where MVC is used for coding all the views, it is often possible to extract a single view from the encoded stream without a complete transcoding. When more complex coding algorithms are used, a particular attention must be paied to assure backward compatibility. In the proposed architecture this will be obtained by allowing to extract a single view from the encoded stream with a very simple data processing, and without transcoding.

2.2 Layered depth image (LDI) representation

When dealing with the compression of multi-view content, two main approaches can be envisaged: either coding the MVD (Multi-view plus Depth) input data directly using a solution in the same vein as MVC (Multi-View Video Coding, see Del D4.1) or either first constructing an intermediate representation, such as the LDI (Layered Depth Image) Representation (see Del D4.1) which will first aim at getting rid of the redundancy between views. The LDI is thus formed of several layers, the first layer containing the color data of a reference view and the other layers are sparse and contain color data seen from the other cameras and not seen from the reference camera. The two approaches have been investigated in the project and tools compatible with these two approaches are presented below.

MVC coding is backward compatible with teh H.264 2D video codec, in the sense that the reference view is encoded with this solution. The other views are then encoded in a predictive manner taking the reference view to predict, with the help of disparity information, the other input views. Similarly, in the case one constructs an intermediate LDI-based representation, the different layers can be encoded either using standalone 2D video codecs or using the MVC codec, after padding the layers by propagating the pixel values from one layer to the adjacent one. This allows using classical 2D video codecs, the information propagated from one layer to the next being then removed by the inter-layer prediction modes of the MVC codec.

3 3D-specific conding tools

3.1 Disparity field estimation

In this section we describe a method for dense (*i.e.* one vector per pixel) disparity field estimation. Even though dense disparity fields cannot be directly used for compression because of their huge coding cost, we can use them to derive an R-D efficient representation of the disparity field, taking advantage of the global formulation of the disparity estimation problem. In other word we are able to produce a disparity field that standard approach would not take in consideration given their local and causal approach to the estimation problem. We have proved the effectiveness of this approach for multiple view video coding (without depth) in a previous work [15]. In the present paper we want to extend this concept to the MVD case and moreover to explore the critical issue of parameter tuning for the dense disparity estimation (DDE). At this end, it is necessary to recall the main ideas of DDE, which is the objective of this section.

Let I_t^n be the rectified frame taken by the n-th camera at time t. Therefore the disparity vectors can only have a the horizontal component, which we call d. Dense disparity estimation has the target of finding the disparity field $d(\mathbf{p}) = d(x,y)$ (that is the disparity vector for any pixel position $\mathbf{p} = (x,y)$) which best matches pixel $\mathbf{p}$ in current frame I_t^n in view n with pixel $\mathbf{p} + d(\mathbf{p})$ in the reference frame I_t^m in view m. This is a typical example of inverse problem, which needs suitable regularization to be solved. In the following, for the sake of simplicity, we will consider only the case $m = n - 1$.

At the basis of the estimation methods, there is the hypothesis that the image intensity is roughly constant once one has compensated for the disparity. As a consequence, a common method to estimate d is to minimize a cost function such as the sum of squared differences between the current image and the one compensated by disparity.

$$d^*(\cdot) = \operatorname*{argmin}_{d \in \Omega} \sum_{(x,y)\in\mathcal{P}} \left[I_t^n(x,y) - I_t^{n-1}(x + d(x,y), y)\right]^2 \tag{1}$$

where $\mathcal{P}$ is the picture support and Ω is the range of candidate disparity fields. However this criterion is hardly if ever useful, since any disparity field linking equally luminous pixels would make it equal to zero. In order to find a significant solution, we have to inject into the criterion other constraints, accounting for known characteristics of the solution (*a priori* information). This is the regularization needed to solve the problem.

However, before introducing regularization, we want to simplify the criterion. If we assume that an initial coarse estimate $\bar{d}$ of d is available (*e.g.* thanks to block-matching method), and that the difference between $\bar{d}$ and d is small, the warped image can be approximated as:

$$\begin{aligned} I_t^{n-1}(x+d,y) &\simeq \\ & I_t^{n-1}(x+\bar{d},y) + (d-\bar{d})\frac{\partial}{\partial x} I_t^{n-1}(x+\bar{d},y) \end{aligned} \tag{2}$$

This linearization allows to rewrite the criterion $J[d(\cdot]$ as a quadratic convex functional:

$$J[d(\cdot)] = \sum_{\mathbf{p}\in\mathcal{P}} [r(\mathbf{p}) - L(\mathbf{p})\ d(\mathbf{p})]^2 \tag{3}$$

where

$$\begin{aligned} L(\mathbf{p}) &= \tfrac{\partial}{\partial x} I_t^{n-1}(x+\bar{d}(\mathbf{p}),y) \\ r(\mathbf{p}) &= I_t^n(\mathbf{p}) - I_t^{n-1}(x+\bar{d}(\mathbf{p}),y) + \bar{d}(\mathbf{p})\ L(\mathbf{p}) \end{aligned}$$

As pointed out before, the minimization of J is an ill-posed problem, demanding for additional constraints, which reflect the *a priori* knowledge about the disparity.

This problem can be solved in the context of the set theory. We introduce M constraints. The m-th of them is represented by a closed convex set S_m in a Hilbert space $\mathcal{H}$. We call S the intersection of all the M sets S_m. Then, S is the set of candidate solutions [25], *i.e.* the set where we have to look for the field minimizing J:

$$d^*(\cdot) = \underset{d\in\bigcap_{m=1}^M S_m}{\operatorname{argmin}}\ J(d) \tag{4}$$

This formulation is useful, since the constraints can be described as level sets of suitable continuous convex real functions $\{f_m\}_{m\in\{1,\dots,M\}}$:

$$\forall m \in \{1,\dots,M\}, \qquad S_m = \{d \in \mathcal{H} \mid f_m(d) \le \delta_m\} \tag{5}$$

where $(\delta_m)_{1\le m\le M}$ are real-valued parameters such that $S = \bigcap_{m=1}^M S_m \neq \emptyset$.

Now we shall define the constraints. We consider two simple but effective constraints. The first one specifies the range of values of the disparity field $[d_{\min}, d_{\max}]$, and can be expressed by the constraint set S_1:

$$S_1 = \{d \in \mathcal{H} \mid d_{\min} \le d \le d_{\max}\} \tag{6}$$

The second imposes the regularity of the disparity field, limiting the amount of variability of d. This can be achieved by limiting the total variation of the disparity field. The total variation $\mathsf{tv}(d)$ is defined as the sum over $\mathcal{P}$ of the norm of the (discrete) spatial gradient of d [27]. As a conclusion, the total-variation based regularization constraint amounts to impose an upper bound τ on tv:

$$S_2 = \{d \in \mathcal{H} \mid \mathsf{tv}(d) \le \tau\} \tag{7}$$

The total variation limit τ depends on the characteristics of the scene and of the camera configuration: therefore finding an optimal value for it can be an hard task [13]. One of the contribution of this work is to explore the relationship between this parameter and the quantization parameters of the compressed. MVD sequence.

We introduce a last regularization term, which penalizes solutions too much different from the initial one. This is accounted for by a weight α. In conclusion, the criterion to minimize becomes:

$$J(d) = \sum_{\mathbf{p} \in \mathcal{P}} [r(\mathbf{p}) - L(\mathbf{p})\ d(\mathbf{p})]^2 + \alpha ||d - \bar{d}||^2 \tag{8}$$

In our implementation, described in [15], we used the efficient constrained quadratic minimization technique developed in [12, 25] which is adapted to problems with quadratic convex objective functions.

3.2 Depth-map coding techniques

Multivi-view data require efficient compression schemes, because of the large amount of data to process. Up to now, proposed methods were inspired from state-of-the-art 2D imaging compression methods. In other words, those 2D compression methods were meant to minimize visible artifacts in 2D and favour visual comfort. Proposed optimization tools are consequently based on visual perception criteria.

Depth maps are not real images. The depth data are gray-scales images and are considered as a monochromatic signal. Each pixel of a depth image, also called depth map, indicates the distance of the corresponding 3D-point from the camera.Depth maps consist of homogeneous regions, and sharp edges. The whiter the region, the closer it is to the camera.

As a solution for depth maps compression issues, suggestions consisted in applying state-of-the-art 2D compression methods to depth data. However, as highlighted previously, depth maps are not actual images.There is a chance 2D optimization choices are not suited for depth maps.

Second, a recurrent issue concerns the bit-rate allocation that optimize synthesized views visual quality, from decoded texture and depth data. Considering the monochromatic nature of depth map, suggestions consisted in applying severe compression to depth information. However, depth maps involve specific properties that are essential for intermediate view synthesis quality. Coarse compression may lead to artifacts in depth map, that may be destructive on the synthesized views.

As part of PERSEE project, studies were led to answer the problem of depth compression, by using new tools that preserve depth maps properties and that optimize visual quality of synthesized views. These studies led to publications. The following sections present at the end of these studies.

3.2.1 Bit-rate allocation between depth and texture in the context of MVD data compression

These studies led to the following publications:[5] and [6]. This work evaluates the impact of bit-rate allocation for texture and depth data relying on the quality of an intermediate synthesized view, in the context of multi-view-plus-depth (MVD) compression. The results show that depending on the acquisition configuration, the synthesized views require a different ratio between the depth and texture bit-rate: between 40% and 60% of the total bit-rate should be allocated to depth.

The experimental protocol, Figure 3.2.1, consists in varying the bit-rate ratio and the total bit-rate: the quantization parameter QP varies from 20 to 44 for both depth and texture coding. Multiview Video Coding (MVC) reference software, JMVM 8.0 (Joint Multiview Video Model) was used to encode three views(texture and depth separately), as a realistic simulation of a 3DTV use.Then, from the three decompressed views (left, central and right), the intermediate view between the central view and the right one was computed, by using the reference software: VSRS, version 3.5, provided by MPEG. Synthesized views quality was evaluated through PSNR scores, with acquired views as a reference.

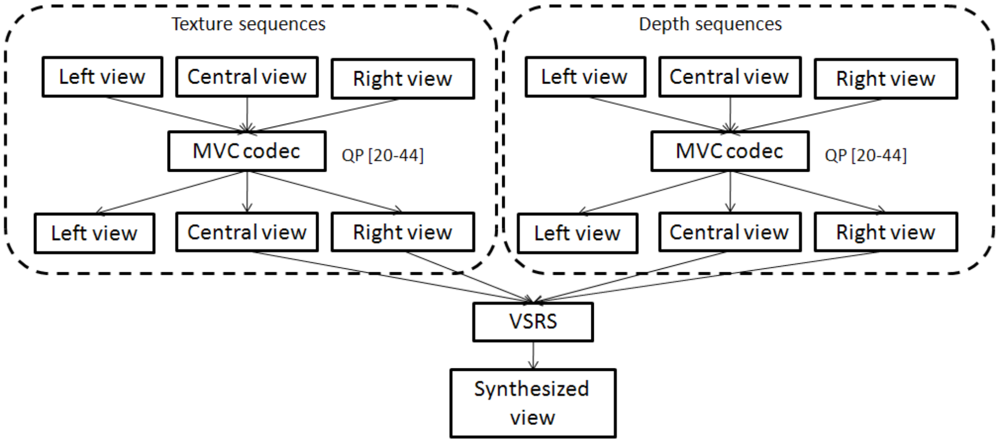
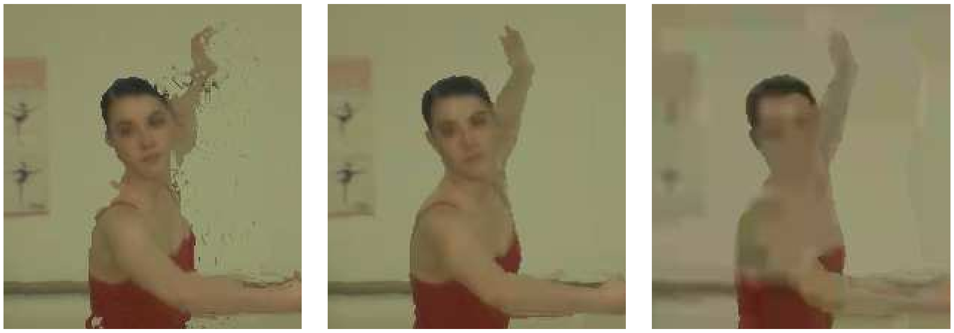


Figure 4: Experimental protocol

Bit-rate ratio between texture and depth determines final quality of synthesized views. For a given total bit-rate, 13 Mbit/s, Figure 5shows synthesized views from ("*Ballet*", Microsoft Research), but with different ratios between texture and depth. They have been computed through VSRS, from MVC-decompressed texture and depth data.

Figure 6 presents the results. The average PSNR of the synthesized sequence are plotted over the bit-rate percentage assigned to depth. The different curves correspond to interpolation of the measured points for different ranges of bit-rate. We



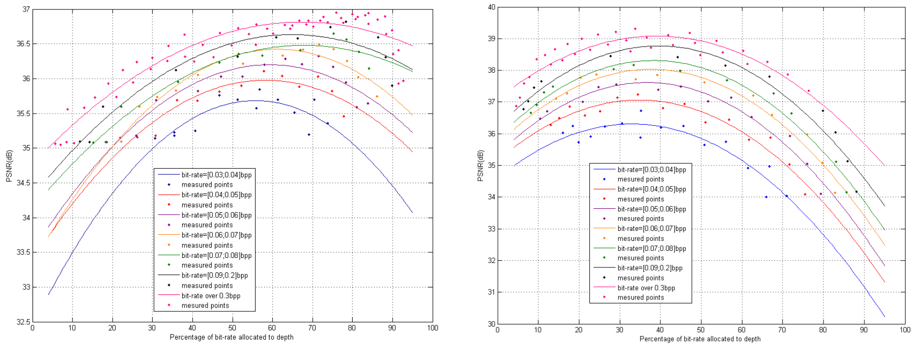
(a) PSNR = 30.0dB; Depth rate = 3%

(b) PSNR = 33.8dB; Depth rate = 60%

(c) PSNR = 30.8dB; Depth rate = 95%

Figure 5: Synthesized images from MVD data, with different bit-rate ratios between texture and depth.

observe that for a given sequence, no matter the bit-rate, the ratio that provides the best quality is the same: it seems to be around 60% for *Ballet*, and around 40% for *Book Arrival*. This suggests that the required depth information that enables a good reconstruction quality, in terms of PSNR depends on the content. Future work aims at determining the correct ratio between depth and texture for a given sequence, by analyzing automatically its properties.



(a) Ballet

(b) Book Arrival

Figure 6: Interpolated rate-distortion curves of synthesized views.

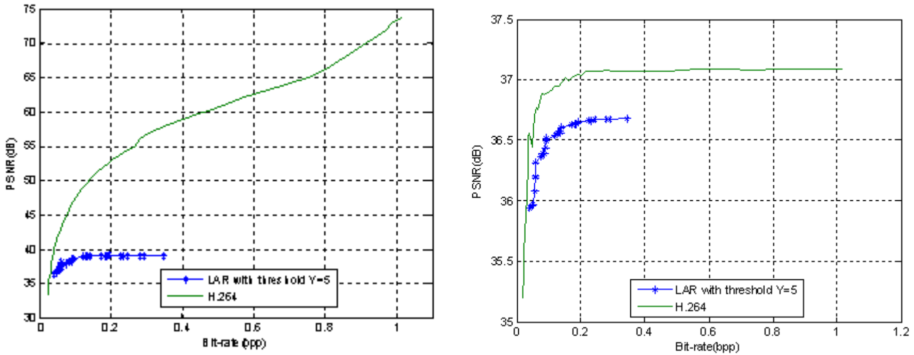
3.2.2 Reliability of an objective metric for the evaluation of views synthesized from decompressed depth maps

This study led to the following publications: [9] et[10]. The investigation concerns a very commonly used metric when assessing compression methods performances: PSNR (Peak-Signal-to-Noise-Ratio). This work questions its reliability in the case of 3D compression methods performance evaluation.

The experiment consists of encoding depth frames from two viewpoints, by either LAR compression[18], or H.264/MVC[22] intra mode. We have used the very first frame of views 2, and 4 from "*Breakdancers*" sequence (provided by Microsoft Research). Decoded depth map of views 2 and 4 were then used to synthesize view 3 and compute the PSNR score, with respect to the original color image from view 3. Only depth maps were encoded while texture data remain original. View 3 was synthesized through VSRS.

Figure 7 and Figure 8 show the results. In Figure 7, the PSNR values are plotted over the bit-rate. According to the PSNR scores, the H.264/AVC compression method outperforms the LAR compression method: the gap between H.264/AVC and LAR regarding the quality of depth map decoded frames is significant. However, scores differ from only 0.5dB concerning the quality evaluation of the synthesized view. This confirms our assumption that the LAR compression preserves essential depth information. Figure 8 shows a comparison of rendered views from encoded depth maps at different bit rates, with either H.264/AVC (intra-coding) or LAR. In Figure 8, the three presented areas show that although LAR provides lower PSNR than H.264/AVC intra-coding, the rendering quality evaluation suggests that LAR method achieves a rendering quality very close to H.264/AVC intra-coding or better, at the same bit-rates.

Then, the results show that for low PSNR scores, conclusions over 3D compression methods can hardly be argued without an additional visual analysis.



(a)

(b)

Figure 7: Rate-distortion curves for the "*Breakdancers*" (a) depth images, and (b) synthesized images, for LAR algorithm and H.264.

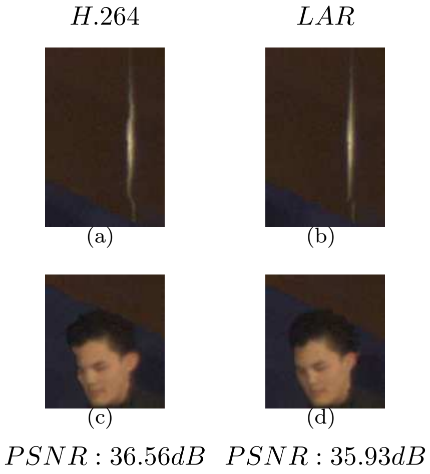


Figure 8: Synthesized images.

3.2.3 Conclusion

The last studies offer new options for depth maps compression.

They proved a correct compression method should preserve and consider sequences properties, that can be different depending on configuration parameters, content etc. The results suggested an optimal ratio between texture and depth, depending on the sequence.However, these observations concerned H.264/MVC codec, and they could differ with some other method.

Besides, studies also showed one of the most commonly used metric, namely PSNR, is not suited for assessing quality of views synthesized from decoded depth maps.

Future work aims at proposing a new joint texture-depth compression method. Future work should also investigate scenes complexity and its relationship with synthesized views quality.

3.3 RD segmentation of dense disparity fields

Another proposed coding scheme for MVD is based on a previous work [15], on which we build in order to take into account the depth maps.

The basic encoder allows to use a dense disparity field for efficiently encoding a multiple view video (without depth) with a standard encoder. For the sake of simplicity, the description will refer only to the stereo case (*i.e.*, two views). However the encoding schemes are promptly extended to the case of more than two views.

The reference encoder workflow is the following. First, a dense disparity field (DDF) is computed for the color sequence. This DDF is then segmented into 16×16 blocks, corresponding to the H.264 macroblocks (MBs). Then, for each MB we start from the 256 candidate vectors of the dense field, and we have to chose one, in order to

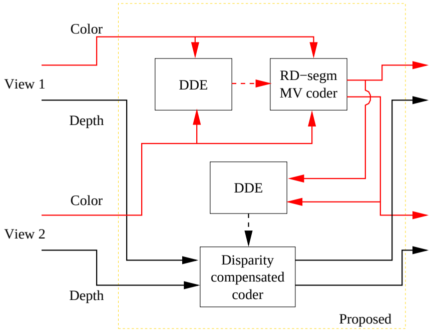


Figure 9: Proposed MVD coding methods.

represent the current block as it was an ordinary *INTER* block: we will encode the chosen vector and the corresponding motion-compensated residual. The representative vector is chosen with an RD criterion, from a set made up by the average vector of the 256 candidates, the median (in the sense of the norm) vector, and the 4 closest to the median vector.

This process is repeated for all the possible partition of an H.264 macroblock. This means that for smooth disparity regions the RD choice will tend to favor large partitions, while for "active" regions (*i.e.* those where the disparity varies significantly, like across object contours) small partitions will be more likely. Therefore we end up with a RD-driven segmentation of the disparity map, that allows to efficiently encode the stereo pair.

3.3.1 Proposed scheme for MVD coding

The proposed scheme takes advantage of the dense disparity estimation algorithm described in Section 3.1 and of the RD-driven segmentation-based multiview (RD-MV) coder described in Section 3.3. This coder is firstly used to encode the color sequences. In this case we set the values for the parameters (α, τ, $d_{\min}$ and $d_{\max}$ using the results of our previous work [15]. Then we use the same RD-MV coder to represent the depth maps. However, in order to save bit-rate taking advantage from the correlation between texture and depth, the depths disparity is computed using color images. In this way, we do not need to send the dense disparity map to the decoder, which instead can compute it from the compressed color sequence, and on the other hand, we take benefit from a dense disparity map, that is used to perform a disparity-compensated coding of the depth. Then, the compressed left depth map and the compressed residual of the disparity-compensated right depth map are sent to the output.

One of the key steps of the implementation of this algorithm is the choice of the DDE parameters for depths, in particular the amount of total variation τ. It is intuitive that they depend on the quality of the compressed color sequence. In particular, for high-quality color images, we expect that all small details are represented, and therefore

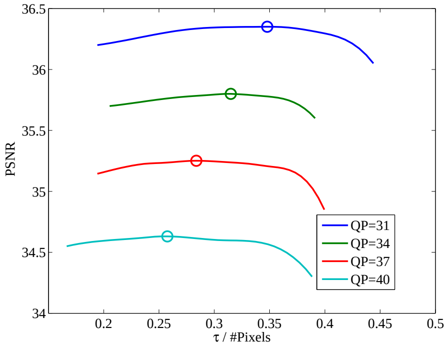


Figure 10: Effect of τ (normalized over the number of image pixels) on the disparity quality, for several values of the quantization step

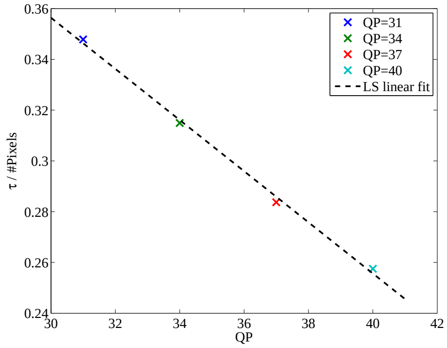


Figure 11: Best normalized τ values in function of QP. The best fitting first order polynomial is shown as well.

the corresponding disparity field can have a higher total variation. On the contrary, heavily quantized images are much smoother, and we should allow a smaller variation of the disparity field. This intuition is confirmed by the experiments, as shown in the next section.

3.3.2 Experimental results

In a first set of experiments, we looked for the optimal value of the total variation parameter τ. In particular, the intuition suggests that the best value for τ should depend on the quantization step of the color images used for the estimation. Therefore we ran the following experiment. We considered the multiview video sequence “break dance”and “ballet”, compressed with QPs in the range $\mathcal{Q} = \{31, 34, 37, 40\}$. For each QP value, we ran the DDE algorithm using the compressed images from several couples of views, using several values of τ. Then we used the resulting dense field to compute a disparity-compensated prediction of the depth map, and finally we compute the PSNR of this prediction with respect to the actual depth map. The results are shown in Fig. 10, where we report, for each QP, the PSNR as a function of the normalized value of τ (the normalization is computed with respect to the number of pixels). In this graph, we also point out the best value of normalized τ, *i.e.*, the one maximizing the PSNR. We refer to this values as τ^*. We remark that these values are strongly correlated to the QP. We compute the sample correlation coefficient obtaining:

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{QP_i - \overline{QP}}{\sigma_{QP}}\right)\left(\frac{\tau_i^* - \overline{\tau^*}}{\sigma_{\tau^*}}\right) \\ &= -0.9987 \end{aligned}$$

where, as usual, the bar represents the (sample) mean and σ represents the (sample) standard deviation. Finally we computed the least square linear fitting of QP and τ^*. We found the following regression equation:

$$\tau^* = -0.0101QP + 0.6587 \tag{9}$$

In Fig. 11 we show at the same time the experimental points and the least square linear fit. As expected, we obtained a very good match, and so Eq. (9) can be used in the proposed encoder in order to quickly find a good value for τ, at least in the considered range of QP values.

In a second set of experiments, we evaluated the performance of the disparity-compensated depth map coder, using the optimized values of total variation for the disparity field estimation. We encoded the first depth map as an ordinary video sequence, while for the second one, we considered the disparity-compensated residual. The disparity field rate was not take into account since this field is available at the decoder side as well. The resulting RD performances are compare to those of the reference schemes (Simulcast and MVC, see Fig. 2.1). We remark a non-negligible rate reduction (computed using the Bjontegaard metric [4]) with respect to the reference, estimated to 3% less than MVC and 17% less than Simulcast over the test sequences.

Global compression performance were misurated as well. Cumulating the gains obtained by the RD-segmentation driven encoder on the texture and those of the presented encoder for the depth maps, we register an average rate reduction of 11% with respect to an MVC-based scheme as the one shown in Fig. 2.1(b).

3.4 LDI coding

LDI representations can then be compressed using the Multi-view Video Codec (MVC) which is applied to both texture and depth information. The MVC codec, an amendment to H.264/MPEG-4 AVC video compression standard, is DCT-based and exploits temporal, spatial and inter-layer correlations. However, MVC does not deal with undefined regions on LDI layers. To produce complete layers, each layer is filled in with pixels from the other layer, at the same position, This duplicated information is detected by the MVC algorithm thanks to the disparity-based inter view coding tools so that this duplication does not have a significant impact on the bit rate.

3.5 Inpaiting-based virtual view synthesis

This section describes a depth-based inpainting algorithm which efficiently handles disocclusion occurring on virtual viewpoint rendering. A single reference view and a set of depth maps are used in the proposed approach. The method not only deals with small disocclusion filling related to small camera baseline, but also manages to fill in larger disocclusions in distant synthesized views. This relies on a coherent tensor-based color and geometry structure propagation. The depth is used to drive the filling order, while enforcing the structure diffusion from similar candidate-patches. By acting on patch prioritization, selection and combination, the completion of distant synthesized views allows a consistent and realistic rendering of virtual viewpoints.

3.5.1 Problem introduction

3DTV and FTV are promising technologies for the next generation of home and entertainment services. Depth Image Based Rendering (DIBR) are key-solutions for virtual view synthesis on multistereoscopic display from any subset of stereo or multiview plus depth (MVD) videos. Classical methods use depth image based representations (MVD, LDV [28]) to synthesize intermediate views by mutual projection of two views. Then, disoccluded areas due to the projection of the first view to the new one could be filled in with the remaining one.

However, in freeviewpoint video (FVV) applications, larger baseline (distance or angle between cameras) involves larger disoccluded areas. Traditional inpainting methods are not sufficient to complete these gaps. To face this issue the depth information can help to guide the completion process. The use of depth to aid the inpainting process has already been considered in the literature. Oh et al. [26] based their method on depth thresholds and boundary region inversion. The foreground boundaries are replaced by the background one located on the opposite side of the hole. Despite the use of two image projections, their algorithm relies on an assumption of connexity between disoccluded and foreground regions, which may not be verified for high camera baseline configurations. Indeed, upon a certain angle and depth, the foreground object does not border the disoccluded part anymore. Daribo et al. [16] proposed an extension to the Criminisi's [14] algorithm by including the depth in a regularization term for priority and patch distance calculation. A prior inpainting of the depth map was performed.

Our approach relies on the same idea. However, our contributions are threefold. The relevance of patch prioritization is improved by first using the depth as a coherence cue through a 3D tensor, and then by using a directional term preventing the propagation from the foreground. A combination of the K-nearest neighbor candidates is finally performed to fill in the target patch.

We present in Section 2 contributions to this priority calculation, based on tensor and then on depth, before describing depth-based patch matching. Section 3 describes the implementation of the method in a MVD context. Results are given in Section 4, as well as a comparison with existing approaches. Conclusions are drawn in Section 5.

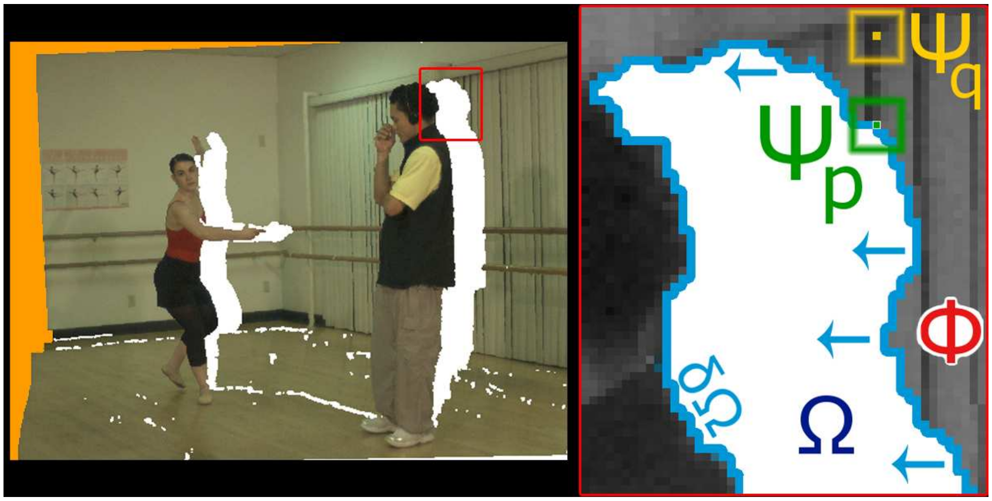


Figure 12: Illustration of principle. On (a) a warped view, (b) a zoom on the disoccluded area behind the person on the right, with the different elements overlaid.

3.5.2 Algorithm

The motivation to use a Criminisi-based algorithm resides in its capacity to organize the filling process in a deterministic way. As seen in fig.12, this technique propagates similar texture elements $\Psi_{\hat{q}}$ to complete patches Ψ_p along the structure directions, namely the isophotes. Their algorithm basically works in two steps. The first step defines the higher order patch priorities along the borders $\delta\Omega$. The idea is to start from where the structure is the strongest (in term of local intensity, with $D(p)$) and from patches containing the highest number of known pixels, $C(p)$. The priority is then expressed as $P(p) = D(p) \times C(p)$. The second step consists in searching for the best candidate in the remaining known image in decreasing priority order.

In the context of view synthesis, some constraints can be added to perform the inpainting and improve the natural aspect of the final rendering. The projection in one view will be along the horizontal direction. For a toward-right camera movement the disoccluded parts will appear on the right of their previously occluding foreground (Fig.12a), and oppositely for a toward-left camera movement.

Whatever camera's movement, these disoccluded areas should always be filled in with pixels from the background rather than the foreground. Based on this a priori knowledge, we propose a depth-based image completion method for view synthesis based on robust structure propagation. In the following, $D(p)$ is described.

Tensor-based priority First, the data term $D(p)$ of the inpainting method proposed by [14] involving the color structure gradient is replaced with a more robust structure tensor. This term is inspired by partial differential equation (PDE) regularization methods on multivalued images and provides a more coherent local vector orientation [30]. The Di Zenzo matrix [19] is given by:

$$J = \sum_{l=R,G,B} \nabla I_l \nabla I_l^T = \sum_{l=R,G,B} \begin{pmatrix} \frac{\partial I_l}{\partial x}^2 & \frac{\partial I_l}{\partial x}\frac{\partial I_l}{\partial y} \\ \frac{\partial I_l}{\partial x}\frac{\partial I_l}{\partial y} & \frac{\partial I_l}{\partial y}^2 \end{pmatrix}$$

with ∇I_l the local spatial gradient over a 3x3 window. This tensor can also be smoothed with a gaussian kernel G_σ to give robustness to outliers, without suffering from cancellation effects. We call it $J_\sigma = J * G_\sigma$. Finally, the local vector orientation is computed from the structure tensor J_σ. Its eigenvalues $\lambda_{1,2}$ reflect the amount of structure variation, while its eigenvectors $v_{1,2}$ define an oriented orthogonal basis. Of particular interest is v_2 the preferred local orientation and its “force” λ_2. Based on the coherence norm proposed in [32], the data term $D(p)$ is then defined as:

$$D(p) = \alpha + (1-\alpha) exp\left(\frac{-C}{(\lambda 1 - \lambda 2)^2}\right)$$

with C a constant positive value and $\alpha \in [0,1]$. Flat regions (when $\lambda_1 \approx \lambda_2$) do not favor any direction, it is isotropic, while with strong edges ($\lambda_1 >> \lambda_2$) the propagation begins along the isophote.

Depth-aided and direction-aided priority The priority computation has been further improved by exploiting the depth information, first by defining a 3D tensor product, secondly by constraining the side from where to start inpainting.

3D tensor

The 3D tensor allows the diffusion of structure not only along color but also along depth information. It is critical to jointly favor color structure as well as geometric structure. The depth-aided structure tensor is extended with the depth map taken as an additional image component Z:

$$J = \sum_{l=R,G,B,Z} \nabla I_l \nabla I_l^T$$

One side only priority

The second improvement calculates the traditional priority term along the contour in only one direction. Intuitively, for a camera moving to the right, the disocclusion holes will appear to the right of foreground objects, while out-of-field area will be on the left of the former left border (in orange in Fig.12a). We then want to prevent structure propagation from foreground by supporting the directional background propagation, as illustrated in Fig.12b with the blue arrows.

The patch priority is calculated along this border, the rest of the top, bottom and left patches being set to zero. Then for disocclused areas, the left border possibly connex to foreground will be filled at the very end of the process. For out-of-field areas, even if left borders are unknown, we will ensure to begin from the right border rather than possible top and bottom ones.

These two proposals have been included in the prioritization step.

Patch matching Once we precisely know from where to start in a given projected image, it is important to favor the best matching candidates in the background only.

Nevertheless, starting from a non-foreground patch does not prevent it from choosing a candidate among the foreground, whatever the distance metric used. Thus, it is crucial to restrict the search to the same depth level in a local window: the background. We simply favor candidates in the same depth range by integrating the depth information in the commonly used similarity metric, the SSD (Square Sum of Differences):

$$\Psi_{\hat{q}} = \underset{\Psi_q \in \Phi}{\arg\min}\; d(\Psi_{\hat{p}}, \Psi_q) \;\; \text{with } d = \sum_{p,q \in \Psi_{p,q} \cap \Phi} \alpha_l \left\| \Psi_{\hat{p}} - \Psi_q \right\|^2$$

The depth channel is chosen to be as important as the color one ($l \in R, G, B, Z$ with $\alpha_{R,G,B} = 1$ and $\alpha_Z = 3$). Then it will not prevent the search in foreground patches, but will seriously penalize and unrank the ones having a depth difference above, i.e in front of the background target patch.

As proposed by [33], a combination of the best candidates to fill in the target patch shows more robustness than just duplicating one. We use a weighted combination of the K-best patches depending on their exponential SSD distances to the original patch. ($K = 5$ in our experiments).

3.5.3 Implementation

Experiments are performed on an unrectified Multiview Video-plus-Depth (MVD) sequence “Ballet” from Microsoft [34]. The depth maps are estimated through a color segmentation algorithm [34] and are supplied with their camera parameters. The choice of this sequence is motivated by the wide baseline unrectified camera configuration as well as its highly depth-and-color contrast resulting in distinct foreground-background. This makes the completion even more visible and the issue even more challenging.

First, the central view 5 is warped in different views. Standard cracks (unique vacant pixels) are filled in with an average filter. We then suppress certain ghosting effects present on the borders of disoccluded area in the background: the background ghosting. Indeed, as we start the filling process by searching from the border, it is of importance to delete ghostings containing inadequate foreground color values. A Canny edge detection on the original depth map, followed by a deletion of color pixels located behind that dilated border successfully removes this ghosting.

Finally, our inpainting method is applied on each warped image, using the depth of the final view. The depth inpainting issue is out of the scope of this paper, but encouraging methods are proposed in the literature [16]. In the context of MVD applications, it is realistic to consider a separate transmission of depth information through geometric representation (currently under investigation).

3.5.4 Results

Fig.13 illustrates the results obtained with the proposed method, comparatively with methods from the literature [14], [16], when rendering views located at varying distances from the reference viewpoint. The three versions take in input the same color and depth information, except for the approach in [14] using color only. Our method not only preserves the contour of foreground persons, but also successfully reconstructs the structure of missing elements of the disoccluded area (i.e. edges of the curtains and bars behind the person on the right, background wall behind the left one).

Thanks to our combination term, we can even extend the synthesis to very distant views, without suffering of aliasing effects. As illustrated, the view 5 is projected to view 2 ($V_{5 \to 2}$) and the out-of-field blank areas occupying one quarter width of the warped image are reconstructed. The counterpart of the patch combination is the smoothing effect appearing on the bottom part of this area. By taking different numbers of patches for combination, it is possible to limit this effect. We encourage people to refer to additional results available on our webpage[1] with videos illustrating the priority-based progressive inpainting principle. The results can indeed be essentially address visually, as argued by [23].

3.5.5 Conclusion and perspectives

A robust depth based completion method for view synthesis has been presented. We address the disocclusion issue by going beyond the limitations of scene warping. To start inpainting, coherent depth and color structures are favored along contour through a robust tensor-based isophote calculation while directional inhibition prevents to start from foreground borders. For target patch propagation, a combination of closest geometric and photoconsistent candidates manages effective natural filling. Future works will focus on completion of synthesized views extremely far from the reference view. The natural aspect of this filling in situation, i.e for video, will also be investigated.

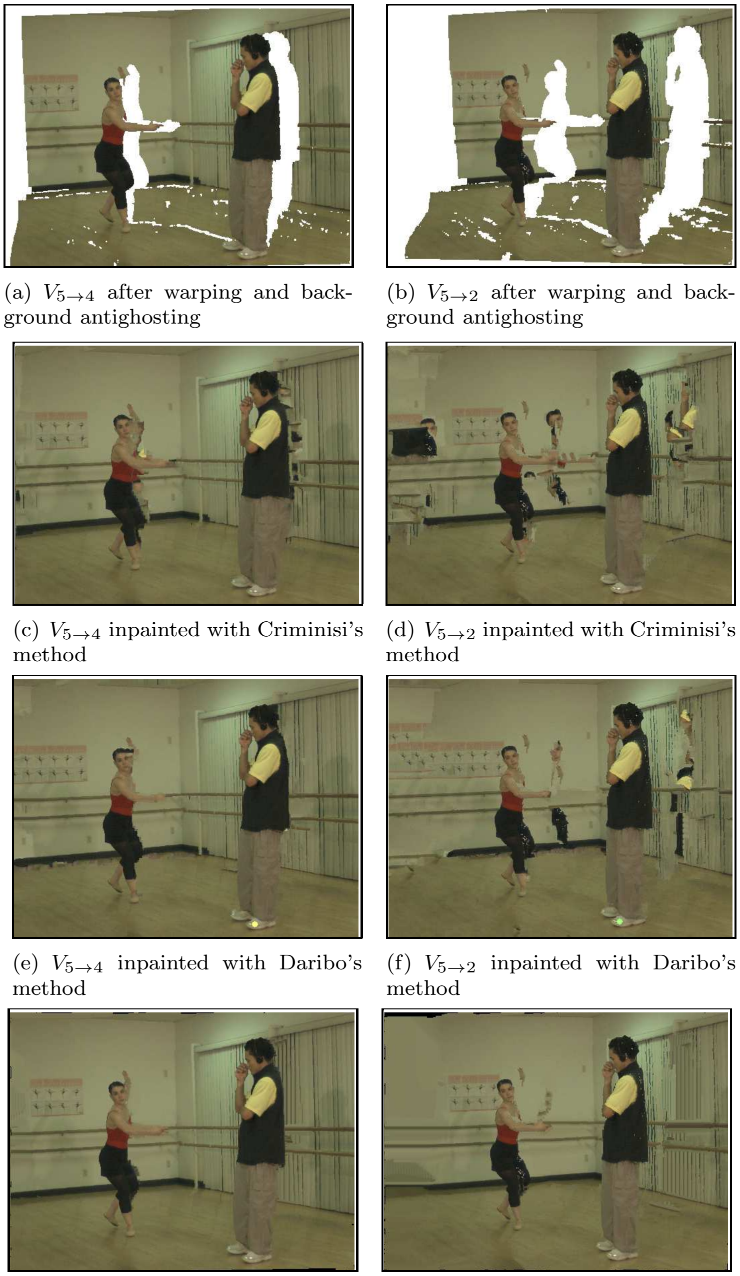
3.6 A way to increase the Quality of Experience of 3DTV

The influence of a monocular depth cue, blur, on the apparent depth of stereoscopic scenes was studied [31]. Recently, stereoscopic image and video production is gaining an increasing amount of attention. The displays nowadays used to show these 3D productions are usually planar so that binocular disparity on the display plane becomes a pre-eminent depth cue enabling viewers to perceive depth. As a binocular cue, disparity is stable, but on the current planar displays, it induces a conflict between accommodation and vergence of the eyes. This conflict is usually considered as one main reason for visual discomfort, especially when the disparity is large. If we decrease this binocular cue, disparity, to limit the visual discomfort, the apparent depth also decreases.

Several other depth cues besides binocular disparity affect also the apparent depth. We proposed decrease the (binocular) disparity of 3D presentations, and to reinforce (monocular) cues to compensate the loss of perceived depth and keep an unaltered apparent depth.

The limitation of depth-of-field of human eyes causes blur in the retinal image which is known as an important monocular depth cue. Some previous investigations have shown clear contributions of blur to depth perception, while others showed that blur has either no effect or only some qualitative effects on perceived depth ordering. In our study, we conducted the subjective experiment using a state-of-art stereoscopic

[1] http://www.irisa.fr/temics/staff/gautier/inpainting



(a) $V_{5\to4}$ after warping and background antighosting

(b) $V_{5\to2}$ after warping and background antighosting

(c) $V_{5\to4}$ inpainted with Criminisi's method

(d) $V_{5\to2}$ inpainted with Criminisi's method

(e) $V_{5\to4}$ inpainted with Daribo's method

(f) $V_{5\to2}$ inpainted with Daribo's method

(g) $V_{5\to4}$ inpainted with our method

(h) $V_{5\to2}$ inpainted with our method

Figure 13: Illustration of different methods of inpainting. Our approach relying on 3D tensor and directional prioritization shows efficient filling.

display system. The stimuli used in the experiment contained a background plane and a single object in the foreground, both of which were chosen closer to natural content compared to the stimuli used in previous publications. The image of a butterfly was used as the foreground object, since it is spatially complex enough, containing regions with both low and high frequency.

In our experimental observations, observers viewed stimuli in a two alternative forced choice (2AFC) task, being required to select the stimulus with largest depth interval between foreground and background. Two sources of perceived depth are used: disparity and blur. The perceived depth from disparity stems from the difference of disparity between the foreground object and the background. The perceived depth from blur stems from the amount of blur introduced to the background by convolution with a Gaussian kernel. Both the absolute position and relative distance between the foreground and background stay as a free parameter. This setup is able to evaluate how the combination of disparity and blur affects the perceived depth of objects located at different distance. By fitting the result to a psychometric function, we obtained points of subjective equality in terms of disparity.

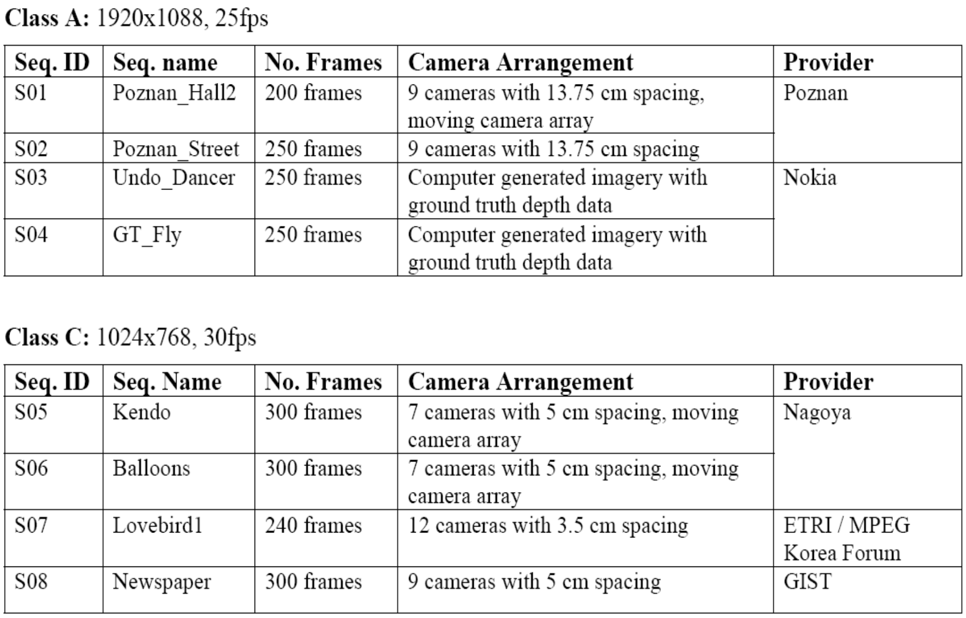
We found that when blur is added to the background of the image, the viewer can perceive larger depth comparing to the images without any blur in the background. The increase of perceived depth can be considered as a function of the relative distance between the foreground and background, while it is insensitive to the distance between the viewer and the depth plane at which the blur is added.The feasibility of enhancing the perceived depth by reinforcing a monocular cue, namely defocus blur, provides an interesting way to deal with the conflict between accommodation and vergence when 3D images are shown on a planar stereoscopic display.

Generally, the foreground object popping out of the screen is the most important object in the scene, while the large disparity of the object may lead to visual discomfort when it is actually fixated by the observer. Our results show that it is feasible to decrease the disparity of this object without losing its pop-out effect by adding some blur on its background. This result suggests a possible way to increase the QoE (Quality of Experience) of 3DTV.

4 Performance evaluation system

4.1 Purpose of Mpeg-3DV CFP

In march 2011, the MPEG-3DV group has issued a call for proposal for compression of MVD (Multi-view plus Depth) data in the context of 3DTV. Candidate methods are planed to be evaluated in november and december 2011. The Call For Proposal (document ISO/IEC JTC1/SC29/WG11/MPEG2011/N12036) describes the evaluation protocol that will be used. In order to compare methods and algorithms developped in the PERSEE project to state-of-the-art methods such as the ones that will be proposed to standardization, the same protocol may be used. In this section we present an overview of test conditions described in the Mpeg-3DV Call For Proposal, both for objective and subjective tests while the next section 4.5 presents subjective



Class A: 1920x1088, 25fps

Seq. ID	Seq. name	No. Frames	Camera Arrangement	Provider
S01	Poznan_Hall2	200 frames	9 cameras with 13.75 cm spacing, moving camera array	Poznan
S02	Poznan_Street	250 frames	9 cameras with 13.75 cm spacing	
S03	Undo_Dancer	250 frames	Computer generated imagery with ground truth depth data	Nokia
S04	GT_Fly	250 frames	Computer generated imagery with ground truth depth data	

Class C: 1024x768, 30fps

Seq. ID	Seq. Name	No. Frames	Camera Arrangement	Provider
S05	Kendo	300 frames	7 cameras with 5 cm spacing, moving camera array	Nagoya
S06	Balloons	300 frames	7 cameras with 5 cm spacing, moving camera array	
S07	Lovebird1	240 frames	12 cameras with 3.5 cm spacing	ETRI / MPEG Korea Forum
S08	Newspaper	300 frames	9 cameras with 5 cm spacing	GIST

Figure 14: Test Material

tests that have already been performed in the PERSEE project.

The primary goal of the MPEG-3DV Call For Proposal is to define a data format and associated compression technology to enable the high-quality reconstruction of synthesized views for 3D displays. It is recognized that technology for depth estimation and view synthesis, as well as the data format itself, has a significant impact on the reconstruction capability and quality of reconstructed views. Therefore, contributions on such technology are also expected by Mpeg.

4.2 Test Material, Coding Classes and Anchors

4.2.1 Test Material

The data sets for 2 test classes, namely Class A and Class C, as displayed in Figure 4.2.1 will be used for evaluation. All data sets have a linear camera arrangement and the data sets are rectified . The depth data and camera parameters for view synthesis and rendering are provided. The depth maps have the same pixel resolution and bit-depth as the video data. Depth maps generated by other depth estimation algorithms are not permitted by Mpeg as additional input to the encoder.

4.2.2 Data format

In order to evaluate the benefits of a proposed data format and corresponding compression technology, video data and associated depth maps are provided as input. The proposed data format shall be reconstructed from a bitstream as an output of a decoding process. The reconstructed data format is then used by a view synthesis algorithm to generate the synthesized views. Proponents may use the VSRS reference software that is provided or a view synthesis algorithm that is optimized for the proposed data format.

The output of the view synthesis will be evaluated by objective and subjective evaluation. The objective quality of the reconstructed video of the decoder will be measured as an indication of overall compression efficiency according to the PSNR of each individual view relative to the input video. In the subjective tests, synthesized views that are provided by the proponents will be fed into 3D displays in order to assess the combined merits of the proposed data format, compression technology and view synthesis algorithm. The submissions will be evaluated on both stereoscopic as well as autostereoscopic displays.

4.2.3 Anchors

Anchors have been generated by encoding the above sequences using an MVC encoder (JMVC 8.3.1) and HEVC encoder with the high efficiency and random access encoder configuration (HM 2.0).

For anchor coding in the AVC-Compatible test category, MVC is used for video data and separately for depth data. For anchor coding in the HEVC-Compatible test category, HEVC is applied independently to the video data and depth data.

Anchor bit streams, decoders, the view synthesis tool (VSRS), all necessary configuration files and utilities are provided at the 3DV-CFP ftp-site ftp://ftp.merl.com/pub/avetro/3dv-cfp/ for all test classes and scenarios.

4.3 Test conditions

4.3.1 Input data

The input data for all submissions includes both video data and depth data as supplementary data for multiple views. The input views for both the 2-view and 3-view test scenarios are provided in Table 1.

4.3.2 Test scenarios and coding conditions

The following test scenarios and test categories are defined.

- Test Scenarios
  - 2-view: refers to the 2-view input configuration
  - 3-view: refers to the 3-view input configuration

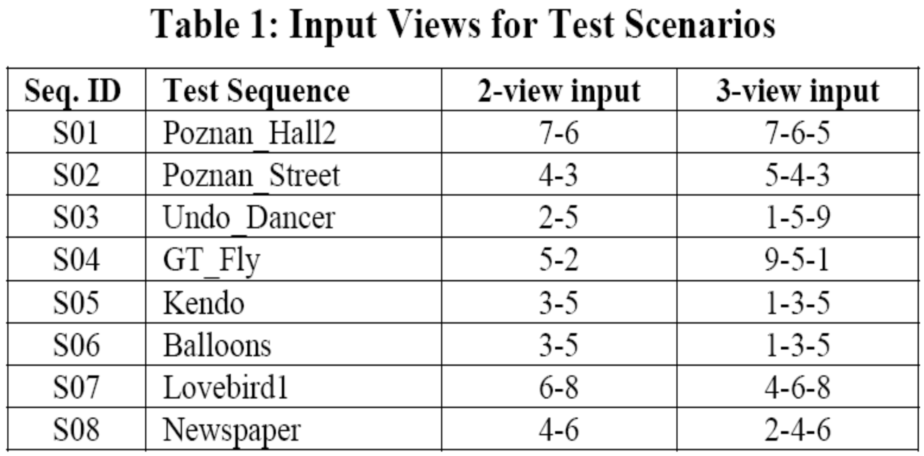


Table 1: Input Views for Test Scenarios

Seq. ID	Test Sequence	2-view input	3-view input
S01	Poznan_Hall2	7-6	7-6-5
S02	Poznan_Street	4-3	5-4-3
S03	Undo_Dancer	2-5	1-5-9
S04	GT_Fly	5-2	9-5-1
S05	Kendo	3-5	1-3-5
S06	Balloons	3-5	1-3-5
S07	Lovebird1	6-8	4-6-8
S08	Newspaper	4-6	2-4-6

- Test Categories
  - AVC-Compatible: refers to submissions in which the compressed data format satisfy the requirement on forward compatibility with AVC
  - HEVC-Compatible & Unconstrained: refers to submissions in which the compressed data formats satisfy the requirement on forward compatibility with HEVC, or submissions without any compatibility constraints

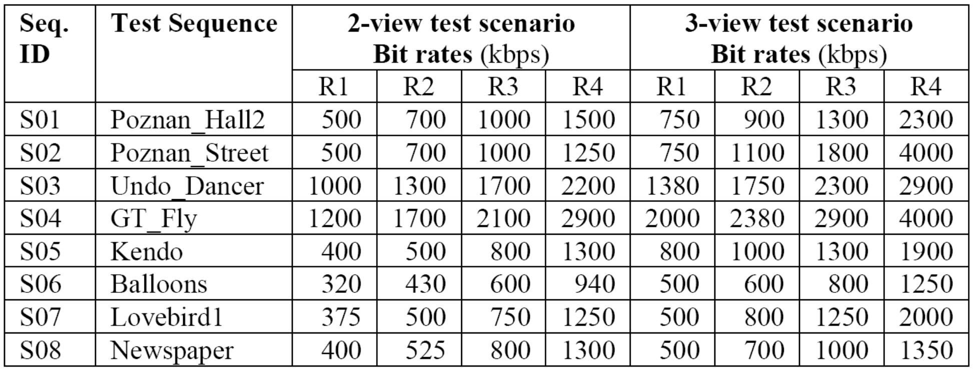
Rate points are listed in Table 2 for AVC-Compatible test category, and Table 3 for the HEVC-Compatible test category. It is not required by the CFP that the bit stream for the 2-view test scenario is a subset of the bit stream for the 3-view test scenario.

4.3.3 Rendering Conditions (View Synthesis)

Submissions shall produce synthesized views using a view synthesis algorithm, which may be either the VSRS software or their own method, for all sequences in all classes and all test scenarios, based on the decoded output. The views to be synthesized for both stereoscopic and autostereoscopic displays are specified in Table 4 (no missing or duplicate views).

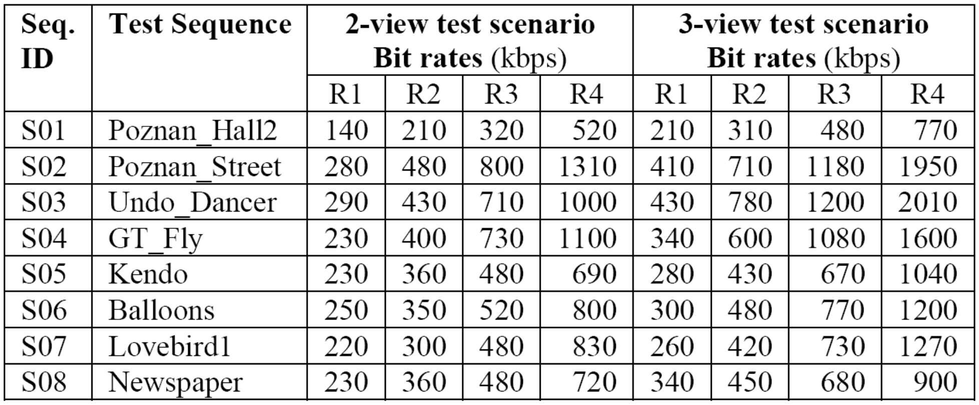
The VSRS configuration files that are used for the anchors for each sequence, as well as the corresponding camera parameter files are provided on the 3DV-CFP ftp site. The camera parameters are provided for each sequence along with the anchors with 5-digit floating point precision. The synthesized views shall represent the scene as if the cameras were positioned according to the views indicated in Table 4.

Table 2: Coding Conditions for 2-view and 3-view test scenarios for AVC-Compatible submissions



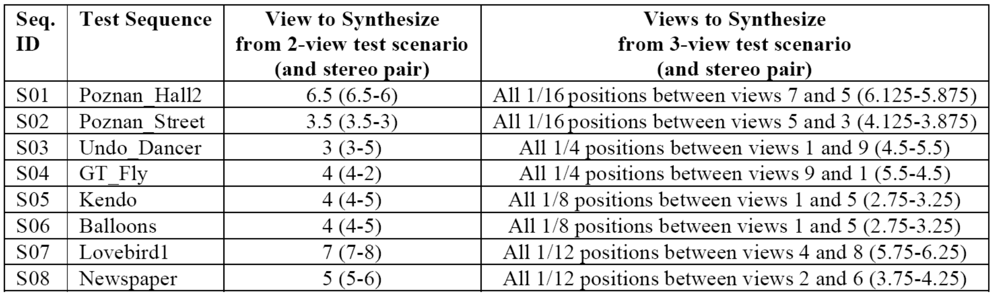
Seq. ID	Test Sequence	2-view test scenario Bit rates (kbps)				3-view test scenario Bit rates (kbps)			
		R1	R2	R3	R4	R1	R2	R3	R4
S01	Poznan_Hall2	500	700	1000	1500	750	900	1300	2300
S02	Poznan_Street	500	700	1000	1250	750	1100	1800	4000
S03	Undo_Dancer	1000	1300	1700	2200	1380	1750	2300	2900
S04	GT_Fly	1200	1700	2100	2900	2000	2380	2900	4000
S05	Kendo	400	500	800	1300	800	1000	1300	1900
S06	Balloons	320	430	600	940	500	600	800	1250
S07	Lovebird1	375	500	750	1250	500	800	1250	2000
S08	Newspaper	400	525	800	1300	500	700	1000	1350

Table 3: Coding Conditions for 2-view and 3-view test scenarios for HEVC-compatible and unconstrained submissions



Seq. ID	Test Sequence	2-view test scenario Bit rates (kbps)				3-view test scenario Bit rates (kbps)			
		R1	R2	R3	R4	R1	R2	R3	R4
S01	Poznan_Hall2	140	210	320	520	210	310	480	770
S02	Poznan_Street	280	480	800	1310	410	710	1180	1950
S03	Undo_Dancer	290	430	710	1000	430	780	1200	2010
S04	GT_Fly	230	400	730	1100	340	600	1080	1600
S05	Kendo	230	360	480	690	280	430	670	1040
S06	Balloons	250	350	520	800	300	480	770	1200
S07	Lovebird1	220	300	480	830	260	420	730	1270
S08	Newspaper	230	360	480	720	340	450	680	900

Table 4: Synthesized output views for stereoscopic and autostereoscopic displays



Seq. ID	Test Sequence	View to Synthesize from 2-view test scenario (and stereo pair)	Views to Synthesize from 3-view test scenario (and stereo pair)
S01	Poznan_Hall2	6.5 (6.5-6)	All 1/16 positions between views 7 and 5 (6.125-5.875)
S02	Poznan_Street	3.5 (3.5-3)	All 1/16 positions between views 5 and 3 (4.125-3.875)
S03	Undo_Dancer	3 (3-5)	All 1/4 positions between views 1 and 9 (4.5-5.5)
S04	GT_Fly	4 (4-2)	All 1/4 positions between views 9 and 1 (5.5-4.5)
S05	Kendo	4 (4-5)	All 1/8 positions between views 1 and 5 (2.75-3.25)
S06	Balloons	4 (4-5)	All 1/8 positions between views 1 and 5 (2.75-3.25)
S07	Lovebird1	7 (7-8)	All 1/12 positions between views 4 and 8 (5.75-6.25)
S08	Newspaper	5 (5-6)	All 1/12 positions between views 2 and 6 (3.75-4.25)

4.4 Subjective tests

For subjetive tests, the Double Stimulus Impairment Scale (DSIS), variant II, test method will be used with 11 quality levels, where 10 indicates the highest quality and 0 indicates the lowest quality. The tests will be carried out with naïve viewers.

The 2-view test scenario will be evaluated on a stereo display, while the 3-view test scenario will be evaluated on both an autostereoscopic display as well as a stereo display.

- Monitor for stereo display: Hyundai 46" stereo display (interlaced stereo IF) with passive glasses. (Model: S465D)
- Monitor for auto-stereo display: Dimenco 52" auto-stereoscopic display with no glasses (Model: BDL5231V3D)

In the 2-view test scenario, the stereo pair is formed with one of the decoded views among the input views and the synthesized view as specified in Table 4. In this case, there is always one decoded view and one synthesized view.

In the 3-view test scenario, a stereo pair is formed from two distinct viewpoints within the range of the input views specified in Table 1. The two viewpoints shall be selected so they are comfortable to view. The baseline distance for the stereo pairs to be viewed will be fixed for each sequence. Two stereo pairs will be selected: one that is centered around the center input view specified in Table 1, and another that is randomly with the same baseline distance between left and right views. In contrast to the 2-view test scenario, both views in this test scenario may be synthesized.

For autostereoscopic viewing tests, 28 views generated from each data set are selected. The 28 views are formed from the set of views including the decoded views and the synthesized views specified in Table 4. The 28 viewpoints are selected so that they are comfortable to view and provide a sufficient depth range, i.e., there is a sufficiently wide baseline distance between the 1st view and the 28th view.

It is expected that the results of the stereoscopic viewing produce the most meaningful results in terms of overall evaluation of proposals.

4.5 Perceptual test methodology

As part of PERSEE project, a study aimed at determining the reliability of 2D commonly used evaluation porotocols (subjective as objective) in the context of 3D video. This study led to submissions, currently under review: [7] et [8]. Experimental protocol consists in assessing seven different view synthesis algorithms through subjective and objective measurements. Three MVD sequences were used: (*Book Arrival*, *Lovebird* et *Newspaper*). For each sequence, four intermediate viewpoints are generated from three reference viewpoints.

Objective measures were computed through MeTriX MuX Visual Quality Assessment Package [1]. Subjective sessions were handled at IRCCyN and involved tow protocols: Absolute categorical rating (ACR) and Paired comparisons (PC) were used to collect perceived quality scores. ACR consists in presenting items to observers that score the test item according to a discrete category rating scale (from 1 to 5). Observers opinion scores are then averaged by computing a mean opinion score (MOS). PC consists in presenting the observers a pair of images. Then observers select the one that best satisfies the specified judgment criterion, i.e. the best image quality.

Results show that usual metrics are not sufficient for assessing 3D synthesized views, since they do not correctly express human judgment. Synthesized views contain specific artifacts located around the disoccluded areas, but usual metrics seem not to be able to express the ratio of perceived annoyance in the whole image.

In the following, the seven tested algortihms are noted $A1$ to $A7$.

The following tables show the results. Table 6 shows that the objective metrics are correlated to each other, in our experiment. Table 5 gives the algorithms rankings according tho the different measures. Subjective-based rankings point $A1$ as the best out of the seven algorithms. But objective-based rankings judge it as the worst. This point underlines the difference between subjective and objective assessments. Besides, table 7 shows the correlation bewteen objective measurements and subjective measurements: none of the tested metric is 50% close to human judgment when assessing synthesized views.

Tables 8 et9 show the results of the Student's t-test performed over the MOS scores, and over the PC scores for each algorithm. This is meant to provide a knowledge on the statistical equivalence of the algorithms. In the case of PC test, most of the algorithms can be statistically distinguish with less than 24 participants. However, in the case of ACR results, the final distinction seems stable when 32 observers participate. This suggests that less observers are needed for a PC based subjective assessment. These observeations are very important in order to set up 3D appropriate subjective tests protocols. Indeed, number of parrticipants and lenght of sessions are set by VQEG. These results suggest a reconsideration of the current protocols for the case of 3D video.

Statistical analyses show that less observers were required for Paired comparisons tests to establish the algorithms distinctions. However, this is a time-consuming method, often avoided by researchers. Concerning the objective metrics, the results show that usual objective assessments hardly correlate subjective assessments. Rankings of algorithms from objective metrics are not reliable, regarding the subjective

	A1	A2	A3	A4	A5	A6	A7
MOS	2.388	2.234	1.994	2.250	2.345	2.169	1.126
Classement	1	4	6	3	2	5	7
PC	1.4038	0.5081	0.2073	0.5311	0.9363	0.4540	-2.0547
Classement	1	4	6	3	2	5	7
PSNR	18.752	24.998	23.180	26.117	26.171	26.177	20.307
Classement	7	4	5	3	2	1	6
SSIM	0.638	0.843	0.786	0.859	0.859	0.858	0.821
Classement	7	4	6	1	1	3	5
MSSIM	0.648	0.932	0.826	0.950	0.949	0.949	0.883
Classement	7	4	6	1	2	2	5
VSNR	13.135	20.530	18.901	22.004	22.247	22.195	21.055
Classement	7	5	6	3	1	2	4
VIF	0.124	0.394	0.314	0.425	0.425	0.426	0.397
Classement	7	5	6	2	2	1	4
VIFP	0.147	0.416	0.344	0.448	0.448	0.448	0.420
Classement	7	5	6	1	1	1	4
UQI	0.237	0.556	0.474	0.577	0.576	0.577	0.558
Classement	7	5	6	1	3	1	4
IFC	0.757	2.420	1.959	2.587	2.586	2.591	2.423
Classement	7	5	6	2	3	1	4
NQM	8.713	16.334	13.645	17.074	17.198	17.201	10.291
Classement	7	4	5	3	2	1	6
WSNR	13.817	20.593	18.517	21.597	21.697	21.716	15.588
Classement	7	4	5	3	2	1	6
PSNR HSVM	13.772	19.959	18.362	21.428	21.458	21.491	15.714
Classement	7	4	5	3	2	1	6
PSNR HSV	13.530	19.512	17.953	20.938	20.958	20.987	15.407
Classement	7	4	5	3	2	1	6

Table 5: Rankings according to measurements.

	PSNR	SSIM	MSSIM	VSNR	VIF	VIFP	UQI	IFC	NQM	WSNR	PSNR$_{hsvm}$	PSNR$_{hsv}$
PSNR		83.9	79.6	87.3	77.0	70.6	53.6	71.6	95.2	98.2	99.2	99.0
SSIM	83.9		96.7	93.9	93.4	92.4	81.5	92.9	84.9	83.7	83.2	83.5
MSSIM	79.6	96.7		89.7	88.8	90.2	86.3	89.4	85.6	81.1	77.9	78.3
VSNR	87.3	93.9	89.7		87.9	83.3	71.9	84.0	85.3	85.5	86.1	85.8
VIF	77.0	93.4	88.8	87.9		97.5	75.2	98.7	74.4	78.1	79.4	80.2
VIFP	70.6	92.4	90.2	83.3	97.5		85.9	99.2	73.6	75.0	72.2	72.9
UQI	53.6	81.5	86.3	71.9	75.2	85.9		81.9	70.2	61.8	50.9	50.8
IFC	71.6	92.9	89.4	84.0	98.7	99.2	81.9		72.8	74.4	73.5	74.4
NQM	95.2	84.9	85.6	85.3	74.4	73.6	70.2	72.8		97.1	92.3	91.8
WSNR	98.2	83.7	81.1	85.5	78.1	75.0	61.8	74.4	97.1		97.4	97.1
PSNR$_{hsvm}$	99.2	83.2	77.9	86.1	79.4	72.2	50.9	73.5	92.3	97.4		99.9
PSNR$_{hsv}$	99.0	83.5	78.3	85.8	80.2	72.9	50.8	74.4	91.8	97.1	99.9	

Table 6: Correlation coefficients between objective metrics in percentage.

	PSNR	SSIM	MSSIM	VSNR	VIF	VIFP	UQI	IFC	NQM	WSNR	PSNR$_{HVSM}$	PSNR$_{HVS}$
CC$_{MOS}$	38.6	21.9	16.1	25.8	19.3	19.2	20.2	19.0	38.6	42.3	38.1	37.3
CC$_{PC}$	40.0	23.8	34.9	19.7	16.2	22.0	32.9	20.1	37.8	36.9	42.2	41.9

Table 7: Correlation coefficients between subjective and objective scores in percentage.

results.

New methods are required for assessing virtual synthesized views as pixel-based and perceptual-based metrics fail. Depth should be taken into account in such a metric as recently proposed in [20], because view synthesis produces geometric distorsions. Registration process according to the original view coupled with weighted critical areas could be investigated in future work to build a new metric.

4.6 Quality of experience for an autostereoscopic display

In recent years, considerable effort has been made to propose 3D television as one of the next milestones in broadcast applications. For displaying the multiview content, an autostereoscopic multiview display is very well suited. It does not require the user to wear anaglyph, polarized, or shutter glasses and the motion parallax effect can be exercised by several observers simultaneously. Currently, most autostereoscopic displays are lenticular displays which use a tilted lens array in front of a TFT screen to project the different views. This leads to a significant reduction in image resolution and the displays also suffer from a considerable amount of crosstalk between the individual views. One of the most prominent displays in this category is the Philips 42-inch 3D display which is used for our experiments[3]. It features nine different views that are internally generated from the 2D texture and additional depth information at its input.

The measurement of subjective quality on a conventional 2D display has been stan-

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		↑**(32)**	↑(<24)	↑**(32)**	**o (>43)**	↑**(30)**	↑(<24)
$A2$	↓**(32)**		↑(<24)	**o (>43)**	**o (>43)**	**o (>43)**	↑(<24)
$A3$	↓(<24)	↓(<24)		↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A4$	↓**(32)**	**o(>43)**	↑(<24)		**o(>43)**	**o(>43)**	↑(<24)
$A5$	**o(>43)**	**o(>43)**	↑(<24)	**o(>43)**		↑(28)	↑(<24)
$A6$	↓**(30)**	**o(>43)**	↑(<24)	**o (>43)**	↓**(28)**		↑(<24)
$A7$	↓(<24)	↓(<24)	↓(<24)	↓ (<24)	↓(<24)	↓(<24)	

Table 8: Results of Student’s t-test with ACR results. Legend:↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "32" observers participate.

	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$
$A1$		↑(<24)	↑(<24)	↑(<24)	↑(<24)	↑(<24)	↑(<24)
$A2$	↓(<24)		↑**(28)**	o(<24)	↓(<24)	**o(>43)**	↑(<24)
$A3$	↓(<24)	↓**(28)**		↓(<24)	↓(<24)	↓(<24)	↑(<24)
$A4$	↓(<24)	**o(>43)**	↑(<24)		↓(<24)	↑**(43)**	↑(<24)
$A5$	↓(<24)	↑(<24)	↑(<24)	↑(<24)		↑(<24)	↑(<24)
$A6$	↓(<24)	**o(>43)**	↑(<24)	↓(<24)	↓(<24)		↑(<24)
$A7$	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	↓(<24)	

Table 9: Results of Student’s t-test with Paired comparisons results. Legend:↑: superior, ↓: inferior, o: statistically equivalent. Reading: Line"1" is statistically superior to column "2". Distinction is stable when "less than 24" observers participate.

dardized for many different scenarios, e.g. standard definition television or multimedia. For three dimensional presentations, the assessment is more difficult because most observers are only used to two dimensional presentations. The experience of having different views for each eye and the perception of depth is exciting for the viewer at first but the effect on the individual quality of experience is usually not normalized. When the subjects are asked to rate the video quality in a single stimulus test on an absolute scale they often only rate the 2D video quality.

In our work, a different approach is used. The observers compare two presentations side by side and they decide whether they prefer left or right in terms of quality of experience. Preliminary experiments demonstrated that the displayed 3D image appeared to be blurred compared to the texture information at its input, i.e., relative to the 2D presentation. Thus, even if there was no additional distortion in the texture or in the depth map, the 3D presentation is affected. The degradation is correlated with the values of the depth map at the corresponding spatial location. When the depth plane corresponded to the display plane, such that the content did not appear to be either in front or behind the display, the effect was minimized. The influence of the depth induced blur effect on the quality of experience has been analyzed in a subjective test. Several different protocols have been used and compared in order to learn about their suitability for the measurement of quality of experience as opposed to quality of 2D video presentation.

Exactly three different subjective test protocols were used. The first protocol was a “Pair Comparison” which can be expected to give the most precise results in terms of preference on the quality of experience scale. The second protocol can be termed “search for equal quality of experience”. The subjects are asked to select one sequence out of a set of sequences which were distorted along a certain axis of degradations. Two different axes were used, the reduction of the image resolution and the introduction of coding artifacts. Although the participants were asked to rate on the quality of experience scale, an indication was found that some participants may rate along the provided axis of degradations. In this study, the degradations were only introduced in the texture part and the added value of depth may not have been taken fully into account. The third protocol uses the “search for best quality of experience”. The participants chose the best sequence out of a set of sequences with increasing 3D effect. The split-screen setup provided a 2D presentation as a reference so they could base their decision on the comparison of preference over 2D.

The results of the experiments led to several conclusions in line with our goal to characterize the effect of the depth rendering technology. The first experiment demonstrated that the rendering of the depth for a sequence introduced distortions in a way that about 70 % observers preferred the 2D presentation without depth information on the same display. In the second experiment, the amount of this degradation was compared to a reduction of the resolution of the 2D presentation and it was shown that an equivalence is reached if the 2D sequence is downsampled by a factor of about 2.3 in both horizontal and vertical direction. In the third experiment, we evaluated the effect of the blur induced by the depth rendering on the perception of coding artifacts. The result is that a significantly higher bitrate may be necessary for the texture of 3D presentations in order to compensate for the depth rendering distortions.

These results differ between the various types of content included in the subjective test. Only one sequence out of the six sequences tested in experiment 1 provided a better quality when displayed in 3D. Starting from an analysis of this sequence it was concluded that a reduction of the depth map variation and moving the mean value of the depth map may lead to an improved 3D quality. In the fourth subjective experiment the viewers could choose the best 3D presentation from a set of four sequences. A setting for the depth range parameter was found for each sequence such that nearly all of the viewers preferred 3D to 2D presentation.

The next step would be to automatically adjust the depth range in order to maximize the quality of experience on the autostereoscopic display. Towards this goal, further investigations in the depth map rendering process and in the content dependency are necessary.

References

[1] MetriX MuXpage. http://foulard.ece.cornell.edu/gaubatz/metrix_mux/.

[2] Draft call for proposals on 3D video coding technology. Technical report, ISO/IEC JTC1/SC29/WG11, Daegu, Korea, January 2011. Doc. N11830.

[3] M. Barkowsky, R. Cousseau, and P. Le Callet. Influence of depth rendering on the quality of experience for an autostereoscopic display. In *Proc. International Workshop on Quality of Multimedia Experience QoMEX*. San Diego, CA, USA., 2009.

[4] G. Bjontegaard. Calculation of average PSNR differences between RD-curves. In *VCEG Meeting*, Austin, USA, April 2001.

[5] Emilie Bosc, Vincent Jantet, Luce Morin, M. Pressigout, and Christine Guillemot. Vidéo 3D : quel débit pour la profondeur ? In *Proc. of CORESA 2010*, Lyon, 2010.

[6] Emilie Bosc, Vincent Jantet, Muriel Pressigout, Luce Morin, and Christine Guillemot. Bit-rate allocation for multi-view video plus depth. In *to appear in 3DTV Conference 2011*, Turkey, 2011.

[7] Emilie Bosc, Martin Koppel, Romuald Pepion, Muriel Pressigout, Luce Morin, Patrick Ndjiki-Nya, and Patrick Le Callet. Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols? In *submitted to ICIP 2011*, Bruxels, 2011.

[8] Emilie Bosc, Romuald Pepion, Patrick Le Callet, Martin Koppel, Patrick Ndjiki-Nya, Muriel Pressigout, and Luce Morin. Towards a new quality metric for 3D synthesized views assessment. *submitted to IEEE Themes 2011*, 2011.

[9] Emilie Bosc, M. Pressigout, and Luce Morin. Focus on visual rendering quality through content-based depth map coding. In *Proceedings of Picture Coding Symposium (PCS)*, Nagoya, Japan, 2010.

[10] Emilie Bosc, Muriel Pressigout, and Luce Morin. Évaluation de la qualité des vues 3D synthétisées. In *to appear in ORASIS 2011*, Praz-sur-Arly, France, 2011.

[11] Y. Chen, Y. K. Wang, K. Ugur, M. M Hannuksela, J. Lainema, and M. Gabbouj. The emerging MVC standard for 3D video services. *EURASIP Journal on Advances in Signal Processing*, 2009, 2009.

[12] P.L. Combettes. A block-iterative surrogate constraint splitting method for quadratic signal recovery. *IEEE Trans. Signal Processing*, 51(7):1771–1782, 2003.

[13] P.L. Combettes and J.-C. Pesquet. Image restoration subject to a total variation constraint. *IEEE Trans. Image Processing*, 13(9):1213–1222, 2004.

[14] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[15] I. Daribo, M. Kaaniche, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu. Dense disparity estimation in multiview video coding. In *Proceed. of IEEE Worksh. Multim. Sign. Proc.*, Rio de Janeiro, Brazil, 2009.

[16] I. Daribo and B. Pesquet-Popescu. Depth-aided image inpainting for novel view synthesis. In *IEEE International Workshop on Multimedia Signal Processing*, 2010.

[17] Valentina Davidoiu, Thomas Maugey, Béatrice Pesquet-Popescu, and Pascal Frossard. Rate distortion analysis in a disparity compensated scheme. In *IEEE International Conference on Audio, Speech and Signal Processing*, ICASSP '11, Prague, Czech Republic, May 2011. To appear.

[18] O. Deforges, M. Babel, L. Bedat, and J. Ronsin. Color LAR codec: a color image representation and compression scheme based on local resolution adjustment and self-extracting region representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(8):974–987, 2007.

[19] S. Di Zenzo. A note on the gradient of a multi-image. *Computer Vision, Graphics, and Image Processing*, 33(1):116–125, 1986.

[20] 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. Palma de Mallorca, September 2010.

[21] C. Fehn, P. Kauff, M. Op De Beeck, F. Ernst, W. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton. An evolutionary and optimised approach on 3d-tv. In *In Proceedings of International Broadcast Conference*, pages 357–365, 2002.

[22] ITU-T Recommendation H.264. Advanced video coding for generic Audio-Visual services, 2009.

[23] N. Kawai, T. Sato, and N. Yokoya. Image inpainting considering brightness change and spatial locality of textures. In *Proc. Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 66–73, 2008.

[24] Philipp Merkle, Aljoscha Smolic, Karsten Müller, and Thomas Wiegand. Multiview video plus depth representation and coding. In *IEEE International Conference on Image Processing*, volume 1, pages 201–204, October 2007.

[25] W. Miled and J.C. Pesquet. Disparity map estimation using a total variation bound. In *3rd Canadian Conference on Computer and Robot Vision*, pages 48–48, 2006.

[26] K.J. Oh, S. Yea, and Y.S. Ho. Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video. In *PCS*, pages 1–4, 2009.

[27] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992.

[28] J. Shade, S. Gortler, L. He, and R. Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242. ACM New York, NY, USA, 1998.

[29] M. Tanimoto, M.P. Tehrani, T. Fujii, and T. Yendo. Free-viewpoint TV. *Signal Processing Magazine, IEEE*, 28(1):67 –76, January 2011.

[30] D. Tschumperlé. Fast anisotropic smoothing of multi-valued images using curvature-preserving pde's. *International Journal of Computer Vision*, 68(1):65–82, 2006.

[31] J. Wang, M. Barkowsky, V. Ricordel, and P. Le Callet. Quantifying how the combination of blur and disparity affects the perceived depth. In *Proc. SPIE Electronic Imaging*. San Jose, CA, USA., 2011.

[32] J. Weickert. Coherence-enhancing diffusion filtering. *International Journal of Computer Vision*, 31(2):111–127, 1999.

[33] Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video. *IEEE transactions on PAMI*, pages 463–476, 2007.

[34] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *ACM SIGGRAPH*, pages 600–608. ACM New York, NY, USA, 2004.