VIEW SYNTHESIS BASED ON TEMPORAL PREDICTION VIA WARPED MOTION VECTOR FIELDS

Andrei I. Purica[⋆†], *Marco Cagnazzo*[⋆], *Beatrice Pesquet-Popescu*[⋆], *Frederic Dufaux*[⋆], *Bogdan Ionescu*[†]

[⋆]Institut Mines-Telecom; Telecom ParisTech; CNRS LTCI
[†]University Politehnica of Bucharest,061071, Romania
Email:{purica, cagnazzo, pesquet, dufaux }@telecom-paristech.fr, bionescu@imag.pub.ro

ABSTRACT

The demand for 3D content has increased over the last years as 3D displays are now a common product. View synthesis methods, such as depth-image-based-rendering, provide an efficient tool in 3D content creation or transmission, and are integrated in future coding solutions for multiview video content such as 3D-HEVC. In this paper, we propose a view synthesis method that takes advantage of temporal and inter-view correlations in multiview video sequences. We use warped motion vector fields computed in reference views to obtain temporal predictions of a frame in a synthesized view and blend them with depth-image-based-rendering synthesis. Our method is shown to bring gains of up to 0.42dB in average when tested on several multiview sequences.

***Index Terms*—** view synthesis, multiview video, depth-image-based-rendering

1. INTRODUCTION

Technology advances in the past years have turned the display of 3D content into an expected functionality in high end television sets. Some of the common usage scenarios of 3D content involve Free view Point Television (FTV) [1], immersive teleconference systems, medical and entertainment applications [2].

Several formats for 3D video information exist. Some of the more commonly used ones include stereo video, MultiView Video (MVV), which is comprised of multiple video recordings of the same scene, acquired from different points of view, and Multiview-Video-plus-Depth (MVD) [3], where each texture sequence is accompanied by its corresponding depth information. This later format is of high interest because depth maps provide a less costly way of creating additional virtual sequences for new points of view. This process is known as view synthesis. In general view synthesis methods can be divided in three categories [4]. The first type of methods require implicit geometry information usually given in the form of depth maps, which are then used to compute pixel disparity between the original and synthesized views in order to perform a warping of the original view. These methods are known as Depth-Image-Based-Rendering (DIBR) [9]. The later two can either, require only implicit geometry, for example pixel correspondences that can be computed using optical flow [5] [6] between two views. Or, no geometry at all, interpolation and filtering is used to obtain the synthesis (e.g.: lumigraph [7], concentric mosaics [8]).

Because of MVD's format ability to support multiview applications, the Moving Picture Experts Group (MPEG) developed an experimental framework for this format during the standardization process of a 3D extension of High Effciency Video coding standard (HEVC) [10]. This framework also defined a View Synthesis Reference Software as part of the 3D-HEVC test model (3D-HTM) [11], which uses DIBR for the creation of virtual views.

When changing the point of view, some areas in the original video sequence, that were hidden, become visible and create holes or dissoclusions in the synthesis. This issue is usually resolved using inpainting algorithms [12] [13] [14]. Other methods use temporal correlations in a video sequence to extract information about the dissocluded areas [15] or extract the background of the scene [16] [17].

In our previous work [18], we used reversed motion vector fields warped in the synthesized view using an epipolar constraint [19], in order to extract information on disoccluded areas. This is combined with a sub pixel precision warping technique that uses an additional filtering step for background-foreground separation and pixel interpolation. In this paper, we use temporal correlations to obtain additional temporal predictions of the synthesized frame. We use forward motion vectors in the temporal sense, computed in the reference views and warped in the synthesized view to obtain up to four temporal predictions which are blended together with the DIBR predictions using either an average or adaptive approach.

The rest of this paper is organized as follows. In Section 2, we present the epipolar constraint. Sections 3 and 4 show how we obtained the predictions and the final synthesized frame. Experimental results are reported in Section 5 and Section 6 concludes the paper.

2. EPIPOLAR CONSTRAINT

The most widely spread view synthesis algorithms warp the texture of a given frame using the associated depth map (DIBR algorithms). However, when dealing with video sequences additional information can be utilized from the reference views such as motion information. The challenge is to appropriately utilize the correlations in the synthesized view. In our previous work [18], we used this information to reduce the size of the disoccluded areas in the synthesized frame. Here, our objective is to address the entire frame.

A first step in achieving this goal, is to obtain usable Motion Vector Fields (MVFs) at the level of the synthesized view. Previously, we used reversed motion vector fields in order to obtain different dissoclusions when warping from multiple time instants. In this paper the interest is rather to have high accuracy. Indeed, the dissoclusions can be addressed after the process using any inpainting algorithm.

[†]Part of this work was supported under ESF InnoRESEARCH POSDRU/159/1.5/S/132395 (2014-2015).

[⋆]Part of this work was supported under CSOSG, ANR-13-SECU-0003, SUrveillance de Reseaux et d'InfrastruCtures par des systemes AeroporTes Endurants (SURICATE) (2013-2016).

We will compute forward MVFs in the reference view (r). More specifically, for a frame F at time instant t, we will compute two MVF using two reference frames at a future (F_{t_+}) and past (Ft_-) time instant.

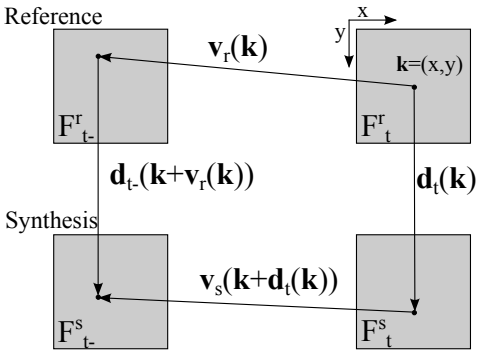


Fig. 1. Relation between projections of a real world point in two frames of the reference and synthesized views at t_- and t time instants.

The motion vectors in the reference view need to be adjusted to be usable in the synthesized view. In order to warp the MVFs at the level of the synthesized view, we impose a so-called epipolar constraint [19] on a point $\mathbf{k} = (x, y)$ in F_t. The ideea is that the projection of a real world point in different views at multiple time instants can be modeled using the disparity and motion information.

In Fig. 1, we show the relation between multiple projections of a real world point. Let us consider a reference view (r), a synthesized view (s) and two time instants t_- and t. $\mathbf{v}_r$ and $\mathbf{v}_s$ are the dense motion vector fields in the reference and synthesized views, respectively and $\mathbf{d}_{t\_}$, $\mathbf{d}_t$ are the disparity fields computed from depth information. If we consider the projection of a real world point $\mathbf{k}$ in frame F^r_t at position (x, y), we can determine the position of the projections in other frames relative to $\mathbf{k}$. In F^s_t and $F^r_{t\_}$ the projections can be found at $\mathbf{k} + \mathbf{d}_t(k)$ and $\mathbf{k} + \mathbf{v}_r(k)$, respectively. We can express the epipolar constraint in this case as:

$$\mathbf{d}_t(\mathbf{k}) + \mathbf{v}_s(\mathbf{k} + \mathbf{d}_t(\mathbf{k})) = \mathbf{v}_r(\mathbf{k}) + \mathbf{d}_{t\_}(\mathbf{k} + \mathbf{v}_r(\mathbf{k})) \quad (1)$$

Using Eq. 1 we can derive the dense motion vector field in the synthesized view as:

$$\mathbf{v}_s(\mathbf{k} + \mathbf{d}_t(\mathbf{k})) = \mathbf{v}_r(\mathbf{k}) + \mathbf{d}_{t\_}(\mathbf{k} + \mathbf{v}_r(\mathbf{k})) - \mathbf{d}_t(\mathbf{k}) \quad (2)$$

This can be interpreted as a motion compensation operation for $\mathbf{d}_{t\_}$ with $\mathbf{v}_r$ followed by an adjustment of the motion intensity with the difference between $\mathbf{d}_{t\_}$ and $\mathbf{d}_t$ and a warping of the motion vector field with $\mathbf{d}_t$.

Note that when using this formulation of the epipolar constraint we obtain the MVF for all positions $m \in \mathcal{M}$ where $\mathcal{M} = \{\mathbf{k} + \mathbf{d}_t(\mathbf{k}) \mid \mathbf{k} \in F^r_t\}$. As a consequence, our MVF in the synthesized view ($\mathbf{v}_s$) will have holes matching the disoccluded areas in a DIBR warping. Therefor, using this MVF to motion compensate will result in the same holes as the DIBR synthesis.

3. TEMPORAL AND VIEW PREDICTION

In general, the view synthesis is performed from 2 reference sequences if possible, a left and a right one with respect to the synthesized view position. This is preferred over using a single reference view due to the reduction of dissocluded areas, especially the removal of border disocclusions [20]. As discussed in Section 2, we can use a past and a future reference frame for the MVF computation in the reference views. Thus, we can obtain up to four temporal predictions for each pixel, from a past or future reference frame using the left or the right available views.

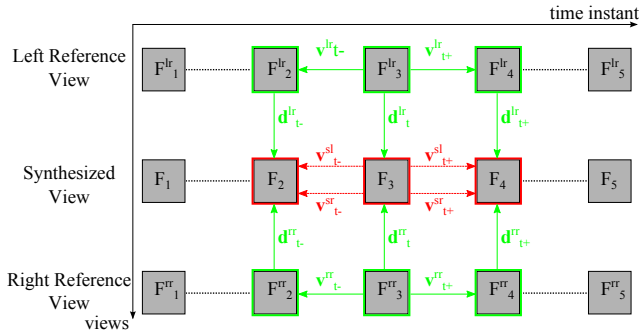


Fig. 2. Scheme for two reference views using past and future time instants (t_-, t_+). Green: MVF warping step; red: the motion compensation step.

In Fig. 2, we show the general scheme of the method. Considering a left (lr) and a right (rr) reference view, with their associated depth maps, we aim at synthesizing a middle view. With green we represent the MVF warping and the required inputs; the MVFs in the reference views for a past (t_-) and future (t_+) time instant ($\mathbf{v}^{lr}_{t\_}$, $\mathbf{v}^{lr}_{t+}$, $\mathbf{v}^{rr}_{t\_}$, $\mathbf{v}^{rr}_{t+}$) and the six disparity fields ($\mathbf{d}^{lr}_{t\_}$, $\mathbf{d}^{lr}_t$, $\mathbf{d}^{lr}_{t+}$, $\mathbf{d}^{rr}_{t\_}$, $\mathbf{d}^{rr}_t$, $\mathbf{d}^{rr}_{t+}$). The motion vector fields can be obtained using any motion estimation technique. The disparity fields are computed from the associated depth information. Assuming we are dealing with a 1D parallel camera setup, the disparity maps only have an x component, which is easily computed from the corresponding depth maps of each base view [2] as:

$$\mathbf{d}(\mathbf{k}) = f \cdot B \left[\frac{Z(\mathbf{k})}{255} \left(\frac{1}{Z_{min}} - \frac{1}{Z_{max}} \right) + \frac{1}{Z_{max}} \right] \quad (3)$$

where Z is the inversed depth, Z_{min} and Z_{max} are the minimum and maximum depth values respectively, f is the focal length of the camera, and B is the baseline, the distance between the base view and the synthesized one.

With red in Fig. 2, we show the motion compensation step in which four predictions of the current frame are obtained using the four MVFs. The red and green scheme can then be iterated through all the frames of the synthesized view. The temporal distance between the prediction and reference in the motion estimation process is 1 in Fig. 2.

The final steps in order to obtain the image are the blending of the temporal predictions and the inpainting of remaining holes. Note that during the blending step, the two DIBR predictions (from left and right) can also be taken into account. This aspect is discussed in the following section.

4. VIEW SYNTHESIS

As discussed in Section 3 we obtain four temporal and two DIBR predictions of a frame. A first solution to obtain the synthesized frame is to combine these predictions in a similar manner as VSRS-1DFast, by computing the average or median of the values for each pixel. This works well for sequences that have low intensity motion and provide good results as can be seen in Section 5. However,

Table 1. Average PSNR for all methods and sequences at each QP.

Sequence	VSRS-1DFast PSNR (dB)				Wf PSNR (dB)				P+Bavg PSNR (dB)				P+Badapt PSNR (dB)			
QPs	25	30	35	40	25	30	35	40	25	30	35	40	25	30	35	40
Balloons	34.37	34.07	33.43	32.41	34.39	34.14	33.52	32.51	34.72	34.44	33.78	32.73	34.74	34.45	33.8	32.72
Kendo	34.98	34.51	33.77	32.75	35.37	34.9	34.15	33.08	34.86	34.42	33.75	32.79	35.37	34.87	34.13	33.06
NewspaperCC	29.2	29.05	28.78	28.31	29.81	29.69	29.39	28.83	29.91	29.8	29.5	28.95	29.85	29.74	29.44	28.9
PoznanHall2	36.24	35.87	35.36	34.55	36.35	36.02	35.51	34.77	36.44	36.11	35.7	34.88	36.49	36.2	35.7	34.86
Average	33.70	33.37	32.83	32	33.98	33.69	33.14	32.3	33.98	33.69	33.18	32.34	34.11	33.82	33.27	32.38
ΔPSNR	-	-	-	-	0.28	0.32	0.31	0.3	0.28	0.32	0.35	0.34	0.41	0.45	0.44	0.38

when dealing with high intensity motion, the temporal predictions can contain artifacts due to the motion estimation failing. In this situation, we should use only the DIBR predictions which are invariant with respect to the motion intensity, depending on the quality of the depth maps. The challenge here is to determine when to use DIBR prediction. Because there is no prior information about the texture, the accuracy of the six available predictions, four temporal and two DIBR, cannot be determined. However, we can reasonably assume that the motion estimation artifacts may vary at different time instants, which implies that four temporal predictions with matching values (very close values) probably indicate a good result of the motion estimation. The same reasoning can be applied for DIBR: two matching values predicted with DIBR indicate most likely a good prediction. Thus, we are interested in using the DIBR predictions when they have similar values and the temporal predictions are relatively different from each other and DIBR.

In order to better formulate this problem, let us consider four temporal ($\widehat{i}_1^t$, $\widehat{i}_2^t$, $\widehat{i}_3^t$, $\widehat{i}_4^t$) predictions, two DIBR ($\widehat{i}_1^D < \widehat{i}_2^D$) predictions and the vectors $\mathbf{p}_t = [\widehat{i}_1, \widehat{i}_2^t, \widehat{i}_3^t, \widehat{i}_4^t]$ and $\mathbf{p}_D = [\widehat{i}_1^D, \widehat{i}_2^D]$. When the temporal predictions are very close to the DIBR predictions or contained in the interval $[\widehat{i}_1^D, \widehat{i}_2^D]$ there are no reliable assumptions that can be made about the accuracy of each prediction type. Therefore, we should use the average or median of the six predictions ($[\mathbf{p}_t, \mathbf{p}_D]$). However, when some or all of the temporal predictions are outside of this interval and there is a relatively high difference between the two prediction types, we should use only the DIBR predictions. Based on this situations we can formulate the selection process as follows:

$$\widehat{i} = \begin{cases} \text{mean}(\mathbf{p}_D) & \text{if } \text{mean}(|\ [\mathbf{p_t}, \mathbf{p_D}] - \text{mean}([\mathbf{p_t}, \mathbf{p_D}])\ |) > \\ & \quad \text{mean}(|\ \mathbf{p_D} - \text{mean}(\mathbf{p_D})\ |) \\ \text{mean}([\mathbf{p_t}, \mathbf{p_D}]) & \text{otherwise} \end{cases} \tag{4}$$

For empirical reasons we decided to use the average over the median in the blending process as it provides slightly better results. The selection method described above is designed to replace the temporal predictions with DIBR when motion estimation errors occur. Because the reference frames used for motion compensation are also synthesized, there are situations when this selection process may fail and provide slightly worse results on some frames. However, as can be seen in the experiments section, this blending will correct motion estimation errors in sequences with high intensity motion, while maintaining similar results for the others.

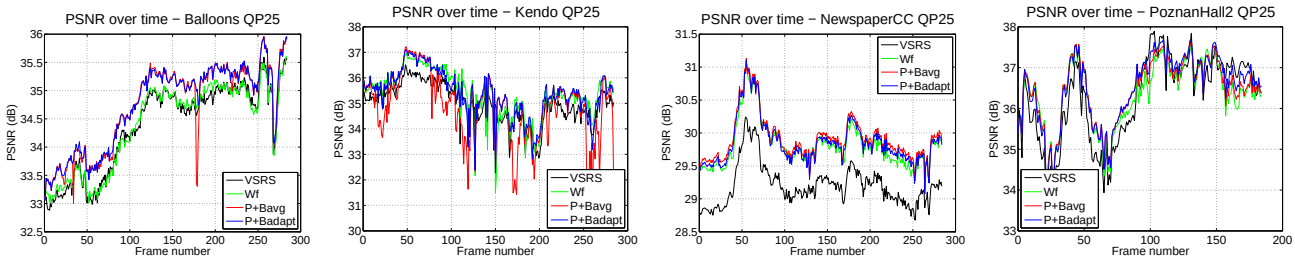
5. EXPERIMENTAL RESULTS

We test our method using the test model designed for 3D-HEVC (3D-HTM). Details can be found in the Common Test Conditions (CTCs) for conducting experiments with the reference software of 3D-HEVC [21]. The video sequences used in our tests are: Balloons, Kendo, NewspaperCC and PoznanHall2. The first three sequences have a resolution of 1024×768 with 30 frames per second while PoznanHall2 has a resolution of 1920×1088 with 25 frames per second, additional details can be found in [22]. We use the full sequences for our tests (300 frames for the first three sequences and 200 for the later). The left and right reference views used in the synthesis are $1\&5$ for the first two sequences, $2\&6$ and $5\&7$ for the NewspaperCC and PoznanHall2. The synthesized views are 3, 3, 4 and 6, respectively. We test the synthesis using different quality encoding for the reference texture and depth sequences. Each sequence is encoded using four QPs: 25 30 35 40, for the texture and corresponding QPs for the depth maps: 34 39 42 45, as indicated by the CTCs. The reference synthesis we compare against is performed with VSRS-1DFast [11], which is included in 3D-HTM.

For comparison purpose, we also include the results of the sub-pixel precision warping technique with adaptive filtering presented in [18]. For fairness of comparison, all methods use the same hole filling technique as that of VSRS-1DFast (a line-wise interpolation). We test our method using two different blending options, first an averaging of the predictions (avg) and second the blending described in Section 4. In our experiments, we use a temporal prediction distance of 2. As shown in Section 3, each frame in the synthesized view is motion compensated from a past and future reference frame. While the past reference frame is available, the future reference frame needs to be synthesized only from DIBR predictions before the motion compensation step. The DMVFs are computed using an optical flow implementation for motion estimation [23] and the motion compensation is performed with sub-pixel precision. The parameters used and additional details about the optical flow method can be found in [24].

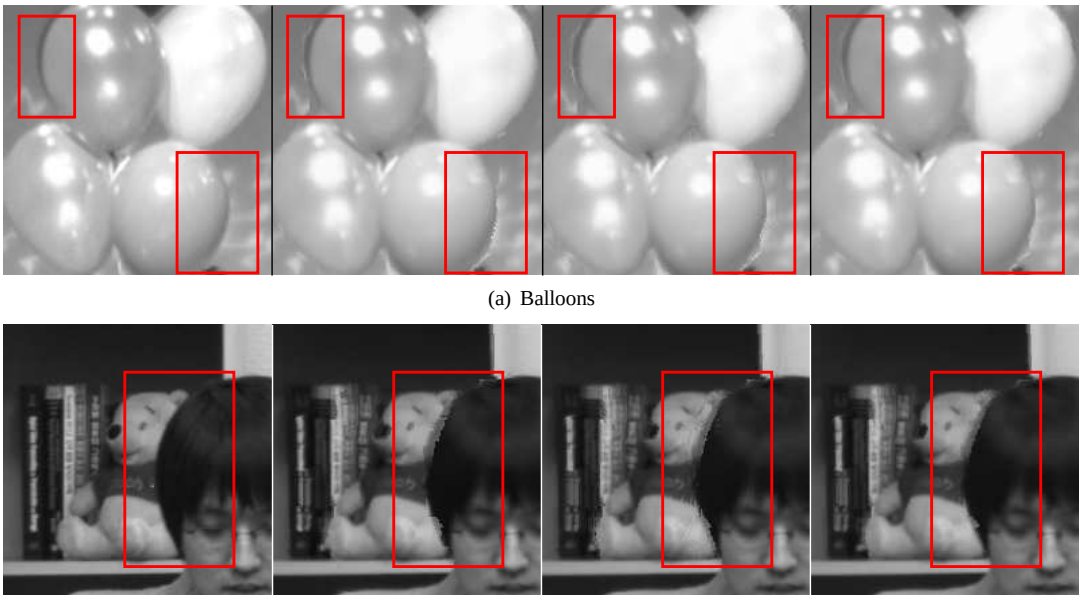
For each sequence and each QP, we synthesize the intermediate views using the reference software (VSRS-1DFast), the warping with adaptive filtering method (Wf) in [18], our proposed method with average blending (P+Bavg) and the adaptive blending (P+Badapt). We evaluate the PSNR of each synthesis using the original uncompressed sequences.

Tab. 1 shows the average PSNR for the reference methods and ours. We can see that our method provides a better synthesis. However, on Kendo sequence the averaging of predictions does not provide good results due to high intensity motion. This issue is resolved by the adaptive blending and the quality of the synthesis is highly increased compared to the averaging blend. The last rows show the average and ΔPSNR values of Wf, P+Bavg and P+Badapt over the VSRS-1DFast reference. We can see that all methods provide a gain over VSRS-1DFast (up to 0.42dB in average, with P-Badapt), while our proposed methods manage to outperform Wf (up to 0.1dB on Balloons sequence). It can be noticed that the gain depends on the



(a) Balloons (b) Kendo (c) NewspaperCC (d) Poznan Hall2

Fig. 3. PSNR variation of the synthesized view over time for the reference and proposed methods at QP 25 in Balloons 3(a), Kendo 3(b), NewspaperCC 3(c) and Poznan Hall2 3(d) sequences.



(a) Balloons

(b) NewspaperCC

Fig. 4. Details in Balloons 4(a) and NewspaperCC 4(b) sequences on frame 38. From left to right: original uncompressed, VSRS-1DFast, Wf, P+Badapt. Red squares show artifacts in the synthesis.

sequence. These aspects are further discussed at the end of this section.

In Fig.3, we show the PSNR variation over time for the four test sequences. For brevity reasons, we only show the QP 25 results as the behavior is similar across QPs. Black, green, red and blue colors indicate the methods VSRS-1DFast, Wf, P+Bavg and P+Badapt. We can see that our proposed methods outperform VSRS-1DFast by 0.65dB and Wf by 0.37dB at most, on NewspaperCC and Balloons sequences respectively. Fig. 4 shows some details of the synthesis with the tested methods. From left to right we show the original uncompressed, the VSRS-1DFast, Wf and P+Badapt. Red squares mark areas containing artifacts. In 4(a) we can see artifacts around the contours of the balloons and in 4(b) around the edge of the head. It is noticeable that in the images on the right these artifacts are diminished.

As can be seen in our experiments, the P+Bavg method provides a better quality synthesis when compared to VSRS-1DFast and Wf on most test sequences. However, this method uses temporal correlations in a video sequence and is dependent on the quality of the motion estimation technique used. While it is able to obtain a very good gain on Balloons sequence it falls behind on Kendo sequence due to motion estimation failure caused by high intensity motion. This problem is corrected using the adaptive blending presented in Section 4. In Fig. 3(a) and 3(b) we can see some drops in quality on some frames with P+Bavg which are corrected by P+Badapt.

6. CONCLUSION

In this paper we presented an improved synthesis method that uses the temporal correlations in the reference views in order to obtain temporal predictions of the synthesized frame from two different time instants. The temporal predictions are then blended with the DIBR predictions using either an averaging or an adaptive blending method designed to reduce the impact of motion estimation errors. The method was tested using the 3D-HEVC test model and compared with VSRS-1DFast synthesis and a sub pixel precision warping technique. Our method brings gains of up to 0.42dB and 0.1dB respectively.

7. REFERENCES

[1] Masayuki Tanimoto, Mehrdad Panahpour Tehrani, Toshiaki Fujii, and Tomohiro Yendo, "Free-Viewpoint TV," *IEEE Signal Processing Magazine*, vol. 28, pp. 67–76, 2011.

[2] Frederic Dufaux, Beatrice Pesquet-Popescu, and Marco Cagnazzo, Eds., *Emerging technologies for 3D video: content creation, coding, transmission and rendering*, Wiley, May 2013.

[3] Philipp Merkle, Aljoscha Smolic, Karsten Muller, and Thomas Wiegand, "Multi-view video plus depth representation and coding," *IEEE International Conference on Image Processing*, vol. 1, pp. 201–204, 2007.

[4] Harry Shum and Sing B. Kang, "Review of image-based rendering techniques," *SPIE Visual Communications and Image Processing*, vol. 4067, pp. 2–13, 2000.

[5] Frederic Dufaux, Marco Cagnazzo, and Beatrice Pesquet-Popescu, *Motion Estimation - a Video Coding Viewpoint*, vol. 5: Image and Video Compression and Multimedia of *Academic Press Library in Signal Processing*, Academic Press, 2014 (to be published).

[6] R. Krishnamurthy, P. Moulin, and J. Woods, "Optical flow techniques applied to video coding," in *IEEE International Conference on Image Processing (ICIP)*, 1995, vol. 1, pp. 570–573 vol.1.

[7] Chris Buehler, Michael Bosse, Leonard McMillan, and Steven Gortler, "Unstructured Lumigraph Rendering," in *Proc SIGGRAPH*, Los Angeles, California USA, August 2001, pp. 425–432.

[8] Heung-Yeung Shum and Li-Wei He, "Rendering with concentric mosaics," in *Proceedings SIGGRAPH*, Los Angeles, California USA, 1999, pp. 299–306.

[9] Christoph Fehn, "A 3D-TV approach using depth-image-based rendering," in *3rd IASTED Conference on Visualization, Imaging, and Image Processing*, Benalmadena, Spain, 8-10 September 2003, pp. 482–487.

[10] "High Efficiency Video Coding," ITU-T Recommendation H.265 and ISO/IEC 23008-2 HEVC, April 2013.

[11] Li Zhang, Gerhard Tech, Krzysztof Wegner, and Sehoon Yea, "3D-HEVC test model 5," ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11 JCT3V-E1005, July 2013.

[12] Christine Guillemot and Oliver Le Meur, "Image inpainting: Overview and recent advances," *IEEE Signal Processing Magazine*, vol. 31, pp. 127–144, 2014.

[13] Ismael Daribo and Beatrice Pesquet-Popescu, "Depth-aided image inpainting for novel view synthesis," in *IEEE MMSP*, Saint Malo, France, 4-6, October 2010.

[14] Antonio Criminisi, Patrick Perez, and Kentaro Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[15] Kuan-Yu Chen, Pei-Kuei Tsung, Pin-Chih Lin, Hsing-Jung Yang, and Liang-Gee Chen, "Hybrid motion/depth-oriented inpainting for virtual view synthesis in multiview applications," *3DTV-CON*, pp. 1–4, 7-9 June 2010.

[16] Wenxiu Sun, Oscar C. Au, Lingfeng Xu, Yujun Li, and Wei Hu, "Novel temporal domain hole filling based on background modeling for view synthesis," in *IEEE International on Image Processing (ICIP)*, Orlando, FL, 30 Sept. - 3 Oct. 2012, pp. 2721 – 2724.

[17] Katta Phani Kumar, Sumana Gupta, and K. S. Venkatesh, "Spatio-temporal multi-view synthesis for free viewpoint television," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, Aberdeen, 7-8 October 2013, pp. 1 – 4.

[18] Andrei Purica, Elie Gabriel Mora, Beatrice Pesquet-Popescu, Marco Cagnazzo, and Bogdan Ionescu, "Improved view synthesis by motion warping and temporal hole filling," in *ICASSP*, South Brisbane, 19-24 April 2014, pp. 1191–1195, IEEE.

[19] Ismael Daribo, Wided Milded, and Beatrice Pesquet-Popescu, "Joint Depth-Motion Dense Estimation for Multiview Video Coding," *Journal of Visual Communication and Image Representation*, vol. 21, pp. 487–497, 2010.

[20] Shafik Huq, Andreas Koschan, and Mongi Abidi, "Occlusion filling in stereo: theory and experiments," *Computer Vision and Image Understanding*, vol. 117, pp. 688–704, June 2013.

[21] Dmytro Rusanovsky, Karsten Muller, and Anthony Vetro, "Common Test Conditions of 3DV Core Experiments," ITU-T SG16 WP3 & ISO/IEC JTC1/SC29/WG11 JCT3V-D1100, April 2013.

[22] "Call for Proposals on 3D video coding technology," ISO/IEC JTC1/SC29/WG11 N12036, March 2011.

[23] C. Liu, "Optical flow Matlab/C++ code," http://people.csail.mit.edu/celiu/OpticalFlow/.

[24] C. Liu, *Beyond pixels: exploring new representations and applications for motion analysis*, Ph.D. thesis, Massachusetts Institute of Technology, May 2009.